Приложение Г

(обязательное)

Пояснительная записка к техническому проекту ИС МИР

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА К ТЕХНИЧЕСКОМУ ПРОЕКТУ

СОДЕРЖАНИЕ

ТЕРМИНЫ И СОКРАЩЕНИЯ

Термин	Определение
Алиас	Короткое, удобное для запоминания имя, использующееся вместо более длинного и сложного имени. В контексте данного документа – псевдоним для wiki-
Анализ аудитории	Оценка социально-демографических признаков вероятной пользовательской аудитории медиаматериала по имеющимся данным (лексика, источники)
Анализ медиаматериала	Методы исследования медиаматериала в целях установления его характеристик (достоверности, схожести, источников и др.), характеризующиеся обособлением и изучением отдельных частей объекта
Анализ охвата	Оценка объёмов вероятной пользовательской аудитории медиаматериала по имеющимся данным (лексика, источники)
Анализ эксплуатационных процессов	Сбор и анализ требований от заинтересованных лиц, отвечающих за эксплуатацию, с целью оценки текущих процессов
АРМ	Автоматизированное рабочее место
БД	База данных
Валидация результатов анализа	Алгоритмы по агрегации результатов выявления информационной цели в единую оценку, характеризующую качество выявления
Документ мониторинга	текстовый (если не указано иное) медиаматериал, подлежащий отслеживанию и анализу, включая содержание медиаматериала и/или основной набор выявляемых характерных признаков указанного

Термин	Определение
Зеркало, веб-зеркало, смысловая (контекстная) копия	Единица медиаматериала или его отдельный элемент, преднамеренно или непреднамеренно измененная, или обработанная в части ее отдельных аспектов (сентимент, авторство, последовательность событий и т.п.) таким образом, чтобы не вызывать противоречия у воспринимающего субъекта
Извлечение фактов	Технологии обнаружения в медиаматериале связей между сущностями, выражающими структуру фактов
ИИ (AI), искусственный интеллект (artificialintelligence)	Комплекс технологических решений, позволяющий имитировать когнитивные функции человека, включая самообучение, поиск решения без заранее заданного алгоритма и достижение инсайта, и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека
Интерпретируемость модели машинного обучения	Свойство модели машинного обучения, характеризующее возможность объяснения полученных с ее помощью результатов экспертом (человеком) и использования данных результатов как аргументации
Информационная среда, информационное пространство, информационное поле	Совокупность информации, объектов информатизации, информационных систем, сайтов в информационно-телекоммуникационной сети "Интернет", сетей связи, информационных технологий, субъектов, деятельность которых связана с формированием и обработкой информации, развитием и использованием названных технологий, обеспечением информационной безопасности, а также совокупность механизмов регулирования соответствующих общественных

Термин	Определение
ИС МИР	Информационная система мониторинга информационных ресурсов
ИС МИР-1	Первый этап создания информационной системы мониторинга информационных ресурсов
Классификатор сентимента	Обнаружение эмоциональной окраски, переданной автором в отношении упомянутой в медиаматериале именованной сущности
Краулинг	Технологии сбора данных, отвечающие за автоматизированное обнаружение новых источников медиаматериалов, просмотр источников, а также извлечение из структуры медиаматериалов характерных метаданных (к примеру, дата публикации, количество обращений, автор и др.)
Лексема	Распознанная группа символов, полученная в результате аналитического разбора входящей последовательности символов
Лемматизация	Процесс приведения словоформы к лемме (нормальной (словарной) форме) • для существительных — именительный падеж, единственное число; • для прилагательных — именительный падеж, единственное число, мужской род; • для глаголов, причастий, деепричастий — глагол в инфинитиве (неопределённой форме) несовершенного вида.
Методы оценки качества алгоритмов	Методология по формированию метрик качества вычислительных алгоритмов и моделей
Модальность	Способ отображения информации (текст, изображение, видео, аудио)

Термин	Определение
Мультимодальный медиаматериал (медиаобъект, медиаконтент)	Комплекс мультимедиа, рассматриваемый как одно новостное/информационное событие (к примеру, новость), содержащее несколько модальностей (текст, изображение, видео, аудио)
НИР	Научно-исследовательская работа
Токен	Объект, создающийся из лексемы в процессе лексического анализа
Расстояние Левенштейна	Метрика, измеряющая по модулю разность между двумя последовательностями символов. Она определяется как минимальное количество односимвольных операций (а именно: вставки, удаления, замены), необходимых для превращения одной последовательности символов в другую.
Чанк, chunk	Непрерывная последовательность токенов, образующая именованную сущность
Эвристика	Совокупность исследовательских методов, способствующих открытию ранее неизвестного
Эмбеддинг, embedding	Результат процесса преобразования языковой сущности (слова, предложения, параграфа или целого текста) в набор чисел (числовой вектор)
API	Программный интерфейс приложения
BERT	англ. Bidirectional Encoder Representations from Transformers — языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач
Dropout	Метод регуляризации искусственных нейронных сетей, предназначен для уменьшения переобучения сети за счет предотвращения сложных коадаптаций отдельных нейронов на тренировочных данных во время обучения

Термин	Определение
Faiss	англ. Facebook AI Research Similarity Search - библиотека алгоритмов поиска ближайших соседей в линейном пространстве, разработанная в Facebook AI
Fasttext	Библиотека для изучения встраивания слов и классификации текста, созданная исследовательской лабораторией Facebook AI Research
Inverted index	Структура данных, в которой для каждого слова коллекции документов в соответствующем списке перечислены все документы в коллекции, в которых оно
N-грамма, нграмма	Устойчивое словосочетание длиной n
Tf-idf	Статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов
Wikidata	Совместно редактируемая база знаний, созданная Фондом Викимедиа. Используется для обеспечения централизованного хранения данных, которые могут содержаться в статьях Википедии

1. ОБЩИЕ ПОЛОЖЕНИЯ

1.1. Назначение системы

ИС «МИР-1» предназначена для осуществления деятельности по автоматизированному мониторингу информационных ресурсов и выявления информации, распространение которой ограничено или запрещено на территории Российской Федерации, в соответствии с Указом Президента Российской Федерации № 646 «Об утверждении Доктрины информационной безопасности Российской Федерации» от 05.12.2016г., Федеральным законом № 126-ФЗ «О связи» от 07.07.2003г., Федеральный закон № 149-ФЗ «Об информации, информационных технологиях и о защите информации» от 27.07.2006г., Законом Российской Федерации № 2124-1 «О средствах массовой информации» от 27.12.1991г., Федеральным законом № 114-ФЗ «О противодействии экстремистской деятельности» от 25.07.2002г., Федеральным законом № 436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и (или) развитию» от 29.12.2010г., Федеральным законом № 152-ФЗ «О персональных данных» от 27.07.2006г., Постановлением Правительства РФ № 434 «О радиочастотной службе» от 14.05.2014г..

1.2. Область применения системы

Областью применения ИС «МИР-1» является автоматизация процессов мониторинга информационного пространства через глобальную сеть Интернет.

ИС «МИР-1» используется для:

- своевременного выявления признаков нарушений законодательства Российской Федерации в медиаматериалах (группах медиаматериалов), распространяющихся в информационном пространстве;
- поддержки принятия решений об отнесении медиаматериалов (групп медиаматериалов) к классам информации, распространение которой ограничено или запрещено в Российской Федерации;
- выявления зеркал (клонов), неточных (по смыслу) копий медиаматериалов, распространение которых ограничено или запрещено на территории Российской Федерации;
- выявлении наиболее вероятных путей распространения медиаматериалов (групп медиаматериалов) на зеркалах с признаками нарушений

законодательства Российской Федерации в зависимости от типа этих медиаматериалов;

- сбора, хранения, обобщения, систематизации и анализа данных по результатам мониторинга информационного пространства.

1.3. Перечень объектов автоматизации, на которых используется система

Объектом автоматизации ИС «МИР-1» является процесс мониторинга информационного пространства, который осуществляется в соответствии с Федеральным законом от 27.07.2006г. №149-ФЗ «Об информации, информационных технологиях и о защите информации» в целях выявления информации, распространение которой ограничено или запрещено на территории Российской Федерации.

2. ОПИСАНИЕ ПРОЦЕССА ДЕЯТЕЛЬНОСТИ

2.1. Краткая характеристика объекта управления

ИС «МИР-1» предназначена для обеспечения следующих процессов:

- Осуществлении краулинга русскоязычного медиапространства (СМИ, социальные сети, интернет-порталы, сервисы мгновенного обмена сообщениями) с возможностью масштабирования, автономного обнаружения и добавления новых ресурсов, а так же фиксации реального времени публикации медиаматериалов и их метаданных;
- Предоставлении сервисов обработки естественного языка, в том числе семантического анализа и онтологического моделирования (выявление именованных сущностей и их отношений), определение сентимента (тональности) по отношению к именованным сущностям;
- Предоставлении аналитических сервисов, в том числе анализа и поиска зеркал (клонов ресурсов) и неточных (смысловых) копий медиаматериалов, выявления трендов (тенденций изменений), авторства медиаматериалов в зависимости от контента (содержания) и источников.

2.2. Общие принципы функционирования

Для обеспечения выполнения, указанных в п. 2.1 данного документа процессов и функций, в составе ИС «МИР-1» реализованы следующие сервисы:

- Сервис анализа сентимента;
- Сервис извлечения именованных сущностей;
- Сервис идентификации именованных сущностей;
- Сервис классификации социально-демографического распределения;
- Сервис тематического моделирования;
- Сервис краулинга сети Интернет;
- Сервис выявления трендов;
- Сервис поиска документов-источников;
- Сервис обнаружения цепочки наследований документа;
- Сервис определения путей распространения медиаматериала;
- Сервис предсказания трафика домена;
- Сервис определения трафика медиаматериала;
- Сервис отображения интерфейсов пользователя.

Сервис анализа сентимента предназначен для осуществления анализа мнения автора в текстовом документе. Классификация сентимента, производимая сервисов, осуществляется по четырем классам: негативному, нейтральному, позитивному и не определенному.

Сервис извлечения именованных сущностей предназначен для обеспечения поиска, извлечения и классификацию отдельных именованных сущностей в текстовом медиаматериале. Классификация, производимая сервисом, производится по определенным заранее категориям, таким как: персоны, организации, локации.

Сервис идентификации именованных сущностей предназначен для осуществления идентификации именованных сущностей в базе знаний wikidata. Данный сервис производит семантическую редукцию (сокращение доли информации в языковой единице), в результате чего повышается полнота анализа информационного поля.

Сервис классификации социально-демографического распределения предназначен для предсказания вероятностного распределения социально-демографических категорий пользователей сети Интернет, которых заинтересует данный медиаматериал, по введенному тексту этого медаматериала. Классификация производится по категориям пол, возраст, образование и доход.

Сервис тематического моделирования предназначен для иерархического извлечения тематик из документа с разделением их по типовым рубрикам, подрубрикам, сюжетам и событиям (новостям). Рубрики и подрубрики классифицируются на основе рубрикатора, сюжеты и события обладают темпоральной (соотношение содержания текста с временной осью) природой. Так же данный сервис позволяет получить рейтинг сюжетов, связи сюжетов с документами и новостями. С помощью сервиса можно переносить новости между сюжетами, осуществлять слияние новостей и сюжетов.

Сервис краулинга сети Интернет предназначен для обнаружения и сбора поисковым роботом новых медиаматериалов в сети, который используется для дальнейшей обработки другими сервисами. Данный сервис обеспечивает мониторинг web-ресурсов, при помощи в том числе RSS-технологии сбора; мониторинг социальных сетей таких, как vk.ru, facebook.com. twitter, telegram. В сервисе реализована возможность извлечения структуры и обнаружения новых источников данных.

Сервис выявления трендов предназначен для обнаружения сигнатуры, характерной для последующего изменения метрики (упоминания, трафик, сентимент, социально-демографическое распределение) лексических и тематических размерностей и производить предсказание изменения метрики в заданном диапазоне.

Сервис поиска документов-источников предназначен для обнаружения документов-источников в части наличия точных или близких совпадения лексической формы (плагиат) и семантических заимствований (смысловая копия).

Сервис обнаружения цепочки наследований документа предназначен для обнаружения цепочки наследования исходного документа, что позволит трассировать исходный документ до его первоначального источника.

Сервис определения трафика медиаматериала предназначен для определения вероятного трафика (стационарного количества посетителей за день) для заданного медиаматериала при публикации его на заданном источнике.

Сервис определения путей распространения медиаматериала предназначен для моделирования распространения заданного текстового документа и его признаков. С помощью данного сервиса существует возможность обнаружить какие издания какие признаки наследуют (например: лексика, топики сентимент и т.д.)

Сервис предсказания трафика домена предназначен для предсказания вероятного значения трафика (количества посетителей за день) для заданного источника.

Сервис отображения интерфейсов пользователей предназначен для работы с информационной системой посредством пользовательского интерфейса. Сервис в том числе предоставляет возможности:

- конфигурирования пользовательских топиков;
- конфигурирования типов нарушений;
- управления размерностями и метриками в разделе создания правил универсального топика;
- рабочего места оператора по анализу выявленных материалов;
- интерфейса управления сервисом краулинга сети Интернет;
- составления отчетности эмоциональной окраски;
- составления отчетности трассирования (отслеживания) распространения сообщения;
- составления отчета анализа распространения информации;
- составления отчета по выявлению корреляций между объектами информационного поля и запрашиваемым оператором;
- составления отчета с отображением текущей картины по зарегистрированным нарушениям.

2.3. Состав процедур (операций)

Процедуры (операции), обеспечивающие достижение целей создания ИС «МИР-1», реализацию ее назначения (п.1.3 данного документа) и выполнение функций, подразделяются на:

- автоматические процедуры (операции), которые выполняются круглосуточно комплексом программно-технических средств системы;
- интерактивные процедуры (операции), которые выполняются по запросам пользователей или персонала в режиме диалога «человек-машина» по регламенту и в соответствии с должностными и технологическими инструкциями;

– неавтоматизированные процедуры (операции), которые выполняются персоналом системы без использования средств автоматизации по регламенту и в соответствии с должностными инструкциями.

В состав автоматических процедур (операций), реализованных с помощью программно-технических средств ИС «МИР-1» входят:

– прием, хранение, обработка и передача информации по алгоритмам программного обеспечения;
– управление комплексом программно-технических средств ИС «МИР-1» встроенными службами и специальными средствами администрирования на основе ПО с открытым исходным кодом;
– информационное взаимодействие со смежными ИС;
– обработка входящих запросов.

В состав интерактивных процедур (операций) входят:

– поиск, просмотр, аналитическая обработка информации;
– ввод новой информации, данных, настройка шаблонов;
– контроль ситуации и управление программно-техническими средствами ИС;
– административное управление ИС «МИР-1».

В состав неавтоматизированных процедур (операций), обеспечивающих функционирование ИС «МИР-1» и выполняемых персоналом (пользователями) системы без использования средств автоматизации, входят:

– принятие управленческих решений;
– техническое обслуживание (профилактика, ремонт, настройка, установка и т.д.)
– ведение эксплуатационной документации;
– управление персоналом и пользователями системы.

2.4. Организация работ

Автоматические процедуры (операции) выполняются круглосуточно. Допускается перерыв их выполнения на время технического обслуживания.

Все автоматические процедуры ИС «МИР-1» запускаются:

– событиями в базах данных (поступление новых данных/запросов);
– по таймеру (регулярный запуск краулера, запуск процесса ETL).

Автоматические функции получают и передают информацию в интерфейсы других программных модулей в формате JSON при помощи REST API, либо взаимодействуют с БД ИС «МИР-1» через SQL запросы. Скорость выполнения автоматических функций напрямую зависит от производительности оборудования, на котором они выполняются.

Интерактивные функции (операции) выполняются по запросам пользователей или обслуживающего персонала по регламенту и в соответствии с технологическими инструкциями. Допускается прекращение их выполнения на время технического обслуживания.

Интерактивные функции оформляются в виде web-приложений, исполняемых сервисом отображения интерфейсов и отображаемых на АРМ пользователей в веб-браузере. Взаимодействие между сервером приложений и рабочими станциями осуществляется в формате HTTP запросов. Взаимодействие сервиса отображения интерфейсов с БД ИС «МИР-1» и другими сервисами осуществляется в форме SQL запросов или запросов к конечным точкам сервисов при помощи JSON. Скорость выполнения интерактивных функций в некоторой степени зависит от производительности оборудования сервиса отображения интерфейсов, серверов БД и рабочих станций пользователей, но в основном определяется скоростью работы пользователя в интерфейсе приложения.

Выполнение неавтоматизированных операций (функций) производится персоналом ИС «МИР-1» в соответствии с действующими инструкциями, руководствами, регламентами и эксплуатационной документацией.

Скорость выполнения этих операций регулируется административными мерами.

3. ОСНОВНЫЕ ТЕХНИЧЕСКИЕ РЕШЕНИЯ

3.1. Структура системы

ИС «МИР-1» разработана с применением современных технологий, языков разработки и использовании дополнительного программного обеспечения с открытым исходным кодом, что позволяет обеспечить минимизацию трудозатрат на дальнейшее развитие системы.

ИС «МИР-1» построена с применением сервисной архитектуры с автономным исполнением. Использование данного подхода позволяет реализовать горизонтальное масштабирование как всех сервисов системы, так и только необходимых.

За счет сервисной архитектуры, в ИС «МИР-1» возможно в дальнейшем подключать дополнительные модули и компоненты, расширяя, таким образом, функциональное назначение системы.

В системе реализован механизм обработки потоков данных распределённо и параллельно.

Информационный обмен между сервисами ИС «МИР-1» осуществляется по протоколу HTTP посредством обращений к конечным точкам сервисов с использованием REST API.

Общая схема системы приведена в документе «Функциональная схема системы».

3.2. Взаимосвязь с внешними системами (описание потенциальных взаимосвязей)

Информационный обмен между сервисами ИС «МИР-1», а также с внешними системами осуществляется при помощи системы протоколов многоуровнего взаимодействия TCP/IP.

Для каждого сервиса существует набор конечных точек, которые позволяют передавать запрос в сервис.

3.3. Методы оценки качества моделей, используемых в сервисах

3.3.1. F1

F1-мера (f1 score) представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю.

$$F = 2\frac{Precision \times Recall}{Precision + Recall}$$

Точность (Precision) и полнота (Recall) являются метриками которые используются при оценке большей части алгоритмов извлечения информации. Точность системы в пределах класса – это доля документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому

классу. Полнота системы – это доля найденных сервисом документов, принадлежащих классу относительно всех документов этого класса в тестовой выборке.

Формулы расчета точности и полноты:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Обозначения в формуле:

- TP — истино-положительное решение;
- FP— ложно-положительное решение;
- FN — ложно-отрицательное решение.

3.3.2. MSE

Метрика MSE (Mean Squared Error, среднеквадратическая ошибка) измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных. Выполняется возведение во вторую степень, поэтому отрицательные значения не компенсируют положительными. А также в силу свойств этой метрики, усиливается влияние ошибок, по квадратуре от исходного значения. Это значит, что если в исходных измерениях была получена ошибка на 1, то метрика покажет 1, 2-4, 3-9 и так далее. Чем меньше MSE, тем точнее предсказание. Оптимум достигается в точке 0, то есть предсказание идеально.

По сравнению с средней абсолютной ошибкой, MSE имеет некоторые преимущества:

- подчеркивает большие ошибки на меньших ошибках;
- является дифференцируемым, что позволяет более эффективно использовать для поиска минимальных или максимальных значений с помощью математических методов.

Формула расчета MSE:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

где y_i фактический ожидаемый результат и $\hat{y}_i$ это прогноз модели.

3.3.3. MAE

Метрика MAE (Mean Absolute Error, средняя абсолютная ошибка) измеряет среднюю сумму абсолютной разницы между фактическим значением и прогнозируемым значением.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|, where \quad e_t = original_t - predict_t$$

3.3.4. Accuracy

Метрика Accuracy измеряет количество верно классифицированных объектов относительно общего количества всех объектов.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Обозначения в формуле:

- TP — истино-положительное решение;
- TN — истино-отрицательное решение;
- FP— ложно-положительное решение;
- FN — ложно-отрицательное решение.

3.4. Состав сервисов, используемых в системе

3.4.1. Сервис анализа сентимента

3.4.1.1. Общее описание сервиса

Сервис анализа сентимента предназначен для классификации эмоциональной окраски (сентимента) слов в предложении.

На вход в сервис подаются документы, разделенные на предложения и токены с их положением в предложении.

Ответом сервиса являются целевые токены, агрегированные по предложениям, которые, в свою очередь, агрегированы по документам.

Для каждого целевого токена возвращается значение класса его сентимента.

Классы сентиментов:

0 - негативный

1 - нейтральный

2 - позитивный

3 - не определен

Сентимент целевого токена может быть не определен, основная причина - целевой токен находится вне диапазона длины предложения с которой может работать модель.

Диапазон длины предложения от 0 до 128 токенов, если количество токенов в предложении больше 128, токены которые находятся по порядку за 128 токеном будут отброшены и значение сентимента для них будет определено как 3.

Для каждого предложения вложены индексы таргетов (сущностей, нграм) в тексте.

3.4.1.2.Конечные точки сервиса

Запрос сентиментов для сегментов

Конечная точка расположена по адресу http://сервер:порт/model.

Метод запроса на сервер - POST.

На вход подается JSON формата:

```
{
    "documents": {
        "2306": {
            "sentences": [
                {
                    "text": "Наночастицы для комплексной терапии рака разраб
отали в A&M.",
                    "items": []
                },
                {
                    "text": "rss Хоакин Соролья.",
                    "items": [
                        {
                            "item_type": "entity",
```

```
                            "item_pos": [
                                4,
                                18
                            ]
                        }
                    ]
                },
                {
                    "text": "Лаборатория.Лаборатория.Соролья.Хоакин Светочув
ствительный гидрогель для регулируемой терапии внутри организма разработали
исследователи из Техасского университета A&M, 3 июня сообщает журнал Advance
d Materials.",
                    "items": [
                        {
                            "item_type": "entity",
                            "item_pos": [
                                24,
                                31
                            ]
                        },
                        {
                            "item_type": "entity",
                            "item_pos": [
                                32,
                                58
                            ]
                        },
                        {
                            "item_type": "entity",
                            "item_pos": [
                                140,
                                163
                            ]
                        },
                        {
                            "item_type": "entity",
                            "item_pos": [
                                201,
                                210
                            ]
                        }
                    ]
                },
                {
                    "text": "Гидрогели обычно используются внутри организма,
 чтобы помочь в регенерации тканей и доставке лекарств.",
                    "items": []
                },
                {
                    "text": "Однако их трудно контролировать и оптимально ис
пользовать.Ученые из A&M разработали новый класс гидрогелей, которые могут в
заимодействовать со светом с различными результатами.Светочувствительные гид
рогели — это новый класс материалов, используемых для разработки неинвазивны
```

```
x, бесконтактных, точных и управляемых медицинских техник в широком спектре биомедицинских применений.",
                    "items": [
                        {
                            "item_type": "entity",
                            "item_pos": [
                                68,
                                69
                            ]
                        }
                    ]
                },
                {
                    "text": "К ним относятся фототермическая терапия, фотодинамическая терапия, доставка лекарств и регенеративная медицина.Поскольку видимый и ультрафиолетовый свет имеют слабую проникающую способность в тканях организма, ученые сосредоточились на ближней инфракрасной области излучения (NIR), которое имеет более высокую глубину проникновения.Новый класс двумерных наноматериалов, известных как дисульфид молибдена (MoS2), показал незначительную токсичность для клеток и превосходное поглощение NIR.",
                    "items": []
                },
                {
                    "text": "Эти наночастицы с высокой эффективностью фототермического преобразования могут поглощать и преобразовывать NIR-
свет в тепло, которое может активизировать термореактивные материалы.В экспериментах с помощью излучения NIR удавалось управлять передвижением массы гидрогеля в организме для точной доставки лекарств.",
                    "items": []
                },
                {
                    "text": "При лечении рака это дает возможность сконцентрировать большую часть лекарств в опухоли, что облегчит побочные эффекты химиотерапии.Кроме того, свет NIR может генерировать тепло внутри опухолей для удаления раковых клеток в процессе, называемом фототермическая терапия.Таким образом, разработанные наночастицы обладают синергетическим терапевтическим эффектом — с помощью них реализуется одновременно и фототермическая терапия и химиотерапия, что дает более высокую эффективность в уничтожении раковых клеток.",
                    "items": []
                }
            ]
        }
    },
    "return_sentences": true
}
```


Параметры:

documents - словарь, в который вложены все обрабатываемые документы;

2306 - id документа (уникальный), словарь с предложениями и сущностями в них;

sentences – массив с предложениями и сущностями в них

text - текст предложения;

items - массив целевых токенов;

item_type - тип токена (сущность или Нграмма);

item_pos - массив из двух элементов, где первый - индекс символа в предложении с которого начинается целевой токен, второй - индекс символа в предложении которым заканчивается целевой токен;

return_sentences – флаг возврата предложений в ответе (true/false).

Выходные данные – JSON формата:


```
{
    "success": true,
    "data": {
        "2306": {
            "sentences": [
                {
                    "text": "Наночастицы для комплексной терапии рака разраб
отали в A&M.",
                    "items": []
                },
                {
                    "text": "rss Хоакин Соролья.",
                    "items": [
                        {
                            "item_type": "entity",
                            "item_pos": [
                                4,
                                18
                            ],
                            "item": "Хоакин Соролья",
                            "item_sentiment": 1
                        }
                    ]
                },
                {
                    "text": "Лаборатория.Лаборатория.Соролья.Хоакин Светочув
ствительный гидрогель для регулируемой терапии внутри организма разработали
исследователи из Техасского университета A&M, 3 июня сообщает журнал Advance
d Materials.",
                    "items": [
                        {
                            "item_type": "entity",
                            "item_pos": [
                                24,
                                31
```

```
                        ],
                        "item": "Соролья",
                        "item_sentiment": 1
                    },
                    {
                        "item_type": "entity",
                        "item_pos": [
                            32,
                            58
                        ],
                        "item": "Хоакин Светочувствительный",
                        "item_sentiment": 2
                    },
                    {
                        "item_type": "entity",
                        "item_pos": [
                            140,
                            163
                        ],
                        "item": "Техасского университета",
                        "item_sentiment": 2
                    },
                    {
                        "item_type": "entity",
                        "item_pos": [
                            201,
                            210
                        ],
                        "item": "Materials",
                        "item_sentiment": 1
                    }
                ]
            },
            {
                "text": "Гидрогели обычно используются внутри организма, чтобы помочь в регенерации тканей и доставке лекарств.",
                "items": []
            },
            {
                "text": "Однако их трудно контролировать и оптимально использовать.Ученые из A&M разработали новый класс гидрогелей, которые могут взаимодействовать со светом с различными результатами.Светочувствительные гидрогели — это новый класс материалов, используемых для разработки неинвазивных, бесконтактных, точных и управляемых медицинских техник в широком спектре биомедицинских применений.",
                "items": [
                    {
                        "item_type": "entity",
                        "item_pos": [
                            68,
                            69
                        ],
                        "item": "A",
```

```
                                "item_sentiment": 2
                            }
                        ]
                    },
                    {
                        "text": "При лечении рака это дает возможность сконцентр
ировать большую часть лекарств в опухоли, что облегчит побочные эффекты хими
отерапии.Кроме того, свет NIR может генерировать тепло внутри опухолей для у
даления раковых клеток в процессе, называемом фототермическая терапия.Таким
образом, разработанные наночастицы обладают синергетическим терапевтическим
эффектом — с помощью них реализуется одновременно и фототермическая терапия
и химиотерапия, что дает более высокую эффективность в уничтожении раковых к
леток.",
                        "items": []
                    }
                ],
                "success": true
            }
        }
}
```


Параметры:

2306 - id документа, значение - это ответ для документа;

sentences – массив с предложениями и сущностями в них;

text - текст предложения (если на входе return_sentences = true);

items - список словарей с определенными сентиментами для токенов;

item - это строка в которой хранится извлеченный по полученным индексам из предложения целевой токен;

item_type - тип токена (значение берется из входящих данных);

item_sentiment - целочисленное значение от 0 до 3, это вычисленное сервисом значение сентимента;

item_pos - массив из двух элементов, где первый - индекс символа в предложении с которого начинается целевой токен, второй - индекс символа в предложении, которым заканчивается целевой токен;

success – флаг успешности обработки документа (true/false).

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Ответ содержит JSON формата:

```
{
    "success": true,
    "data": {
        "score": 0.9979138177625914,
        "metric": "f1"
    }
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

score – значение метрики

metric – наименование метрики

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер - POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:

```
{
    "success": true,
    "message": "Обучение инициировано"
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/last_train_metric.

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{
    "success": true,
    "data": {
        "time": "2021-11-15 16:03:34.112002",
        "old_metric": "0",
        "new_metric": "0.9955168634146044",
        "update_model": "True"
    }
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

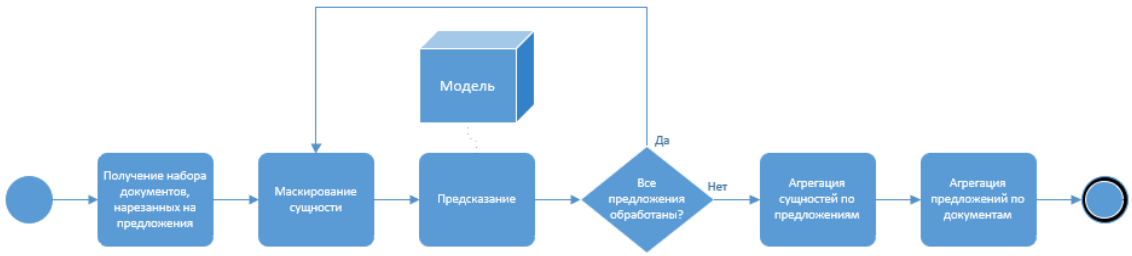
time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.1.3.Процесс работы сервиса



3.4.1.4.Модель

Модель сервиса анализа сентимента основана на семантической модели BERT и является ее модернизированной (Fine-Tuning) версией.

Данная модель была выбрана для задачи классификации сентимента по причине того, что учитывает контекст, в котором находится целевой токен.

Модернизация (Fine-Tuning) модели произведена с помощью «BertForTokenClassification» (обучение дополнительного слоя, предназначенного для классификации токена в тексте).

Задачей данной модели в сервисе является определение сентимента целевого токена.

Перед подачей на вход модели предложения, целевой токен маскируется с помощью тэга [MASK] непосредственно в тексте.

Пример маскирования целевого токена:

Исходный текст - "Цены на газ в Европе резко выросли в последние месяцы."

Маскированный текст - "Цены на газ в [MASK] резко выросли в последние месяцы."

В вышеприведенном примере целевым токеном является слово "Европе".

В каждом предложении маскируется только один целевой токен. Таким образом, если в предложении несколько целевых токенов, модели на вход будет подано столько замаскированных версий предложения, сколько в нем целевых токенов.

В ответе сервиса приводится расчет сентимента для каждого целевого токена, поданного на вход. Если на вход подаются одинаковые целевые сущности, но с разным расположением в тексте, то расчет будет произведен для каждой сущности в отдельности.

Такой алгоритм разработан для достижения наилучших результатов качества модели.

3.4.1.5.Метод оценки качества

Для оценки качества модели используется F1 мера (п.3.3.1 данного документа).

Качество модели составляет 0.7 f1.

3.4.2. Сервис извлечения именованных сущностей

3.4.2.1.Общее описание сервиса

Сервис извлечения именованных сущностей (англ. Named Entity Recognition, NER) реализует одноименную задачу, которая является одной из наиболее распространенных задач обработки естественного языка.

В большинстве случаев задача NER формулируется:

В последовательности токенов (слов и, возможно, знаков препинания) для каждого из них предоставляется тег из предопределенного набора.

Для задачи NER есть несколько распространенных типов токенов, используемых в качестве тегов:

- Персона
- Локации
- Организации

Дополнительно, для классификации смежных объектов с одним и тем же тегом, используется дополнительная схема тегов BIO («B» обозначает начало сущности, «I» означает «внутри» и используется для всех слов, составляющих сущность, кроме первого, а «O» означает отсутствие сущности).

В работе сервиса используется механизм разбиения текста на параграфы с сущностями. Количество сущностей, при которой производится разбивка текста задается в конфигурационном файле «~/папка с сервисом/Docker/ ner_rus_distilbert_torch.json», блок "chainer" - "pipe" - параметр "max_seq_length". По умолчанию при развертывании данный параметр равен 512.

При разбиении текста учитываются предложения в тексте. Таким образом, если в одно предложение заканчивается на 450 сущности, а следующее с 451 до 550, то разбивка произойдет после 450 сущности.

Сервис позволяет осуществлять работу как на GPU, так и на CPU ресурсах.

По умолчанию при развертывании сервис работает на GPU. Для переключения на CPU необходимо:

- остановить сервис, если он запущен
- открыть конфигурационный файл по адресу «~/папка с сервисом/Docker/ ner_rus_distilbert_torch.json»
- вставить на 52 строку файла строку "device": "cpu"
- пересобрать и запустить сервис

3.4.2.2.Конечные точки сервиса

Извлечение именованных сущностей

Конечная точка расположена по адресу http://сервер:порт/model.

Метод запроса на сервер - POST.

На вход подается JSON формата:


```
{"x": ["Российским властям известны люди, которых Великобритания подозрева
ет в отравлении экс-
офицера ГРУ Сергея Скрипаля и его дочери Юлии. Об этом заявил президент Ро
ссии Владимир Путин во время выступления на пленарной сессии Восточного эк
ономического форума, отвечая на вопрос модератора заседания — ведущего тел
еканала «Россия 24» Сергея Брилева. «В принципе, мы, конечно, посмотрели,
что это за люди, мы знаем, кто они такие, мы их нашли. Надеюсь, что они са
ми появятся и сами о себе расскажут. Это будет лучше для всех. Ничего там
особенного и криминального нет», — подчеркнул Путин. По его словам, речь и
дет о гражданских лицах. «Я хочу к ним обратиться, чтобы они нас услышали
сегодня. Пускай они куда-
нибудь придут — вот к вам в средства массовой информации», — отметил прези
дент, обращаясь к Брилеву. Скрипали были обнаружены без сознания в британс
ком городе Солсбери в марте. Спустя неделю британский премьер-
министр Тереза Мэй заявила, что они были отравлены нервно-
паралитическим веществом «Новичок», разработанным в России, и обвинила Мос
кву в причастности к происшедшему. В Москве все обвинения отвергали. Велик
обритания, США, а также 21 европейская страна выслали российских дипломато
в. Власти России ответили зеркально. В августе США ввели против России нов
ые санкции, придя к выводу о том, что Москва причастна к отравлению Скрипа
лей."]}
```


Параметры:

"x" - ключ, по которому сервис определяет массив текстов

Тексты в квадратных скобках, разделяются кавычками (") и запятой, например

{"x": ["ТЕКСТ1", "ТЕКСТ2"]}

Выходные данные – JSON формата:


```
{
    "entity_substr": [
        [
            "москва",
            "россии"
        ]
    ],
    "entity_lemm_substr": [
        [
            "москва",
            "россия"
        ]
    ],
    "entity_offsets": [
        [
            [
                0,
```

```
                6
            ],
            [
                17,
                23
            ]
        ]
    ],
    "entity_init_offsets": [
        [
            [
                0,
                6
            ],
            [
                17,
                23
            ]
        ]
    ],
    "tags": [
        [
            "LOC",
            "LOC"
        ]
    ],
    "sentences_offsets": [
        [
            [
                0,
                24
            ]
        ]
    ],
    "sentences": [
        [
            "Москва - столица России."
        ]
    ],
    "probas": [
        [
            0.9792,
            0.9893
        ]
    ],
    "status": [
          "ok"
    ]
}
```


Параметры:

entity_substr – найденные сущности

entity_lemm_substr – лемматизированные найденные сущности

entity_offsets – индексы найденных сущностей в предложениях

entity_init_offsets – индексы найденных сущностей в оригинальном тексте

tags – тип найденной сущности

sentences_offsets – индексы начала и окончания предложения в тексте

sentences – текст, разделенный на предложения

probas – точность

status – статус обработки

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Выходные данные – JSON формата:

```
{
    "metrics": 93.302
}
```

Параметры:

metrics – значение текущей метрики

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер – POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:

```
{
    "success": true,
    "message": "Обучение инициировано"
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/last_train_metric.

Метод запроса на сервер – GET.

На вход подается:

Входные данные не требуются.

Выходные данные – JSON формата:

```
{
    "success": true,
    "data": {
        "time": "2021-11-26 11:09:02.968223",
        "old_metric": 93.302,
        "new_metric": 93.302,
        "update_model": false
    }
}
```

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

Статус текущего процесса обновления

Конечная точка расположена по адресу http://сервер:порт/status.

Метод запроса – GET.

Входные данные не требуются.

Выходные данные – JSON формата:

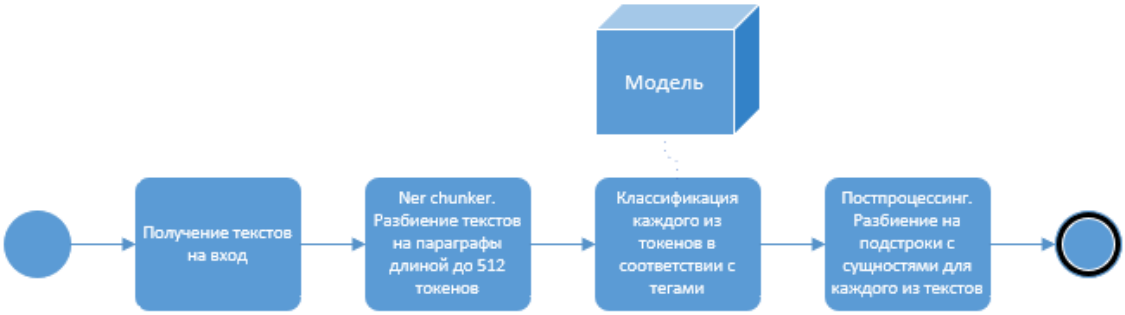
```
{
    "success": true,
    "message": "finished sucessfully"
}
```

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса, принимает 3 варианта:

- finished successfully – сообщает что предыдущий процесс отработал успешно;
- failed – предыдущий процесс завершился с ошибкой;
- running – предыдущий процесс в состоянии выполнения.

3.4.2.3.Процесс работы сервиса



3.4.2.4.Модель

В сервисе используется модель BERT с дополнительным классификатором BIO, которая классифицирует каждый из токенов текста на 7 тегов:

- B-PER – начало сущности «персона»
- I-PER – составляющая сущности «персона»
- B-LOC – начало сущности «локация»
- I-LOC – составляющая сущности «локация»
- B-ORG – начало сущности «организация»
- I-ORG – составляющая сущности «организация»
- O – отсутствие сущности

3.4.2.5.Метод оценки качества

Для оценки качества модели используется мера F1 (п.3.3.1 данного документа).

Пример оценки качества на данных с BIO-разметкой:

Слово	**Классификация**

А	B-PER
Силуанов	I-PER
назначен	O
управляющим	O
от	O
России	B-LOC
в	O
МВФ	B-ORG

Для каждого из тегов LOC, PER, ORG для токенов размеченного текста вычисляются индексы начала и конца искомых токенов (chunks) во всем массиве токенов.

Для примера выше (А.Силуанов назначен управляющим от России в МВФ):

для тега PER список chunks_per = [(0, 2)]

(соответствует токенам [“А”, “.”, “Силуанов”])

для тега LOC список chunks_loc = [(6, 6)]

(соответствует токену “России”)

для тега ORG список chunks_org [(8, 8)]

(соответствует токену “МВФ”)

Вычисляются списки chunks для размеченных эталонных тегов (gold-тегов) и тегов, предсказанных разработанным сервисом.

Далее для каждого из тегов вычисляются значения true positive (tp), false negative (fn) и false positive (fp):

```
tp = len(set(pred_chunks).intersection(set(true_chunks)))
fn = len(true_chunks) - tp
fp = len(pred_chunks) – tp
```

Функции:

Len() – длина в символах

Set().intersection() – пересечение множеств

pred_chunks – предсказанные значения chunks

true_chunks – эталонные значения chunks

Вычисляется общие параметры точности и полноты по формулам (состоят из вычисленных параметров точности и полноты каждого тега):

total_precision = precision_per * len(pred_chunks_per) / len(pred_chunks) +
precision_loc * len(pred_chunks_loc) / len(pred_chunks) +
precision_org * len(pred_chunks_org) / len(pred_chunks)

total_recall = recall_per * len(true_chunks_per) / len(true_chunks) +
recall_loc * len(true_chunks_loc) / len(true_chunks) +
recall_org * len(true_chunks_org) / len(true_chunks) +

В результате вычисляется значение меры F1 по формуле:
F1 = 2 * total_precision * total_recall / (total_precision + total_recall)

3.4.3. Сервис идентификации именованных сущностей

3.4.3.1. Общее описание сервиса

Сервис идентификации именованных сущностей (англ. Entity Linking, EL) реализует связывание подстрок в тексте, соответствующим именованным сущностям, с базой знаний Wikidata.

Связывание подстроки с сущностью Wikidata происходит следующим образом:

- нахождение сущностей в Wikidata, названия которых наиболее близки по расстоянию Левенштейна к подстроке;
- ранжирование сущностей по контексту.

3.4.3.2. Конечные точки сервиса

Идентификация именованных сущностей

Конечная точка расположена по адресу http://сервер:порт/model.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "entity_substr": [
        [
            "москва",
            "россия"
```

```
        ]
    ],
    "entity_offsets": [
        [
            [
                0,
                6
            ],
            [
                17,
                23
            ]
        ]
    ],
    "tags": [
        [
            "LOC",
            "LOC"
        ]
    ],
    "sentences_offsets": [
        [
            [
                0,
                24
            ]
        ]
    ],
    "sentences": [
        [
            "Москва - столица России."
        ]
    ],
    "probas": [
        [
            0.9792,
            0.9893
        ]
    ]
}
```


Параметры:

entity_substr – найденные сущности;

entity_offsets – индексы найденных сущностей в предложениях;

tags – тип найденной сущности;

sentences_offsets – индексы начала и окончания предложения в тексте;

sentences – текст, разделенный на предложения;

probas – точность.

Выходные данные – JSON формата:


```
{
    "entity_substr": [
        [
            "москва",
            "россия"
        ]
    ],
    "conf": [
        [
            [
                [
                    1,
                    134,
                    1
                ],
                [
                    1,
                    134,
                    0.9900000095367432
                ],
                [
                    1,
                    28,
                    0.7799999713897705
                ],
                [
                    1,
                    7,
                    0.20999999344348907
                ],
                [
                    1,
                    17,
                    0.10000000149011612
                ]
            ],
            [
                [
                    1,
                    203,
                    1
                ],
                [
                    1,
                    203,
                    1
                ]
            ]
        ]
    ],
```

```
    "entity_offsets": [
        [
            [
                0,
                6
            ],
            [
                17,
                23
            ]
        ]
    ],
    "entity_ids": [
        [
            [
                "Q649",
                "Q175117",
                "Q4303710",
                "Q1350689",
                "Q20643155"
            ],
            [
                "Q159"
            ]
        ]
    ],
    "entity_tags": [
        [
            "LOC",
            "LOC"
        ]
    ],
    "entity_labels": [
        [
            [
                "Москва",
                "Москва",
                "Москва",
                "Москва",
                "Москва"
            ],
            [
                "Россия"
            ]
        ]
    ],
    "status": [
        "ok"
    ]
}
```


Параметры:

entity_substr – список именованных сущностей

conf – массив точности (точность для каждого из вероятных wiki-сущностей из entity_ids) из параметров:

- title_confidence – насколько именованная сущность совпадает с wiki-сущностью
- graph_confidence – связность в графе сущности
- context_confidence -насколько контекст документа подходит под контекст wikidata

entity_offsets – индексы начала и окончания сущности в тексте

entity_ids – id вероятных wiki-сущностей для запрашиваемых сущностей (до 5 шт.)

entity_tags – теги сущностей (получены из NER)

entity_labels – имена вероятных wiki-сущностей из списка entity_ids

status – успешность обработки документа

Получение списка алиасов для сущности

Конечная точка расположена по адресу http://сервер:порт/aliases/get/.

В конце адреса конечной точки добавляется наименование сущности, для которой нужно получить алиасы. Например, нужно получить все алиасы сущности «Москва». Адрес конечной будет выглядеть так: http://сервер:порт/aliases/get/москва .

Метод запроса на сервер - GET.

Входные данные не требуются.

Выходные данные – JSON формата:

```
"['Q7747', 'Q4359688', 'Q7747', 'Q7747', 'Q7747']"
```

Параметры:

Массив wiki-алиасов запрашиваемой сущности, перечисленных через запятую.

Добавление Алиаса

Конечная точка расположена по адресу http://сервер:порт/aliases/add.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "label": "путин1",
```

```
    "entity_ids": [
        "Q7747"
    ]
}
```

Параметры:

label – сущность, к которой нужно добавить алиас

entity_ids - массив wiki-алиасов которые нужно добавить к сущности

Выходные данные – JSON формата:

```
null
```

Удаление алиасов для сущности

При запуске этого метода будут удалены все связанные алиасы у сущности.

Конечная точка расположена по адресу http://сервер:порт/aliases/delete/.

В конце адреса конечной точки добавляется наименование сущности, для которой нужно удалить алиасы. Например, нужно удалить алиасы у сущности «Москва». Адрес конечной будет выглядеть так: http://сервер:порт/aliases/delete/москва .

Метод запроса на сервер - GET.

Входные данные не требуются.

Ответ содержит JSON формата:

```
null
```

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Ответ содержит JSON формата:

```
{
    "precision": 0.6427672955974842,
    "recall": 0.8085443037974683
}
```

Параметры:

precision – значение точности модели

recall – значение полноты модели

Обновление версии Wikidata

Конечная точка расположена по адресу http://сервер:порт/update/wikidata.

Метод запроса на сервер – GET.

Входные данные не требуются.

Ответ содержит JSON формата:

При успешном запуске:

```
{
    "success": true,
    "message": "Process successfully started."
}
```

При запущенном процессе:

```
{
    "success": false,
    "message": "Update is already running."
}
```

Параметры:

Success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Статус текущего процесса обновления

Конечная точка расположена по адресу http://сервер:порт/status.

Метод запроса – GET.

Входные данные не требуются.

Выходные данные – JSON формата:

```
{
    "success": true,
    "message": "finished successfully"
}
```

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса, принимает 3 варианта:

- finished successfully – сообщает что предыдущий процесс отработал успешно;
- failed – предыдущий процесс завершился с ошибкой;
- running – предыдущий процесс в состоянии выполнения.

Обновление сущностей

Конечная точка расположена по адресу http://сервер:порт/update/model.

Метод запроса – GET.

Требуется предварительно запустить конечную точку по адресу http://сервер:порт/update/wikidata (описание приведено выше)

Входные данные не требуются.

Выходные данные – JSON формата:

При успешном запуске:

```
{
    "success": true,
    "message": "Process successfully started."
}
```

При запущенном процессе:

```
{
    "success": false,
    "message": "Update is already running."
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/last_train_metric.

Данный эндпоинт возвращает данные из файла с логом обновления, располагается в директории ~/data пользователя, который разворачивал сервис.

Метод запроса на сервер – GET.

Входные данные не требуются.

Ответ содержит JSON формата:


```
{
    "success": true,
    "data": {
        "time": "2021-11-26 11:06:22.175426",
        "old_precision": 0.643,
        "new_precision": 0.643,
        "old_recall": 0.809,
        "new_recall": 0.809,
        "update_model": false
    }
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_precision – старая метрика точности

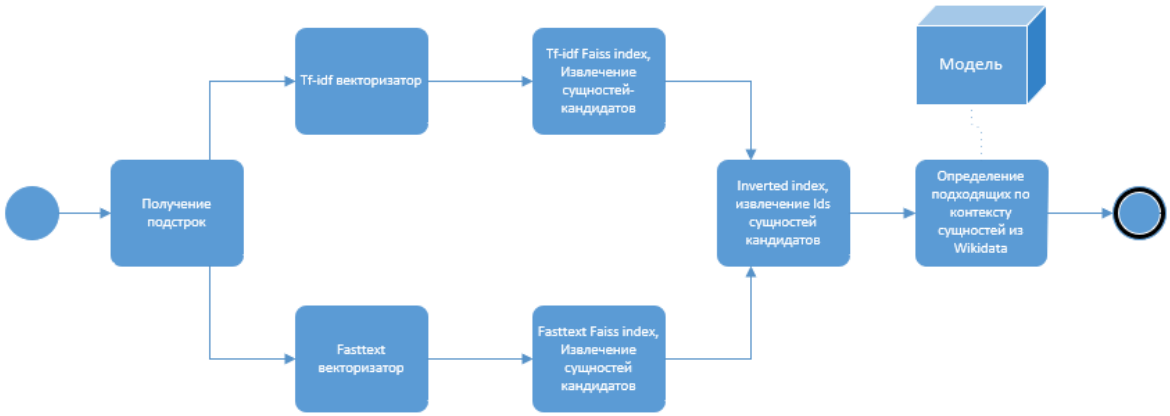
new_precision – новая метрика точности

old_recall – старая метрика полноты

new_recall – новая метрика полноты

update_model – параметр, который говорит о факте перехода на новую модель (true/false)

3.4.3.3.Процесс работы сервиса



3.4.3.4.Модель

Модель сервиса состоит из следующих компонентов:

- Faiss-индекс с tf-idf векторами из char n-грамм названий сущностей в Wikidata;
- Faiss-индекс с fasttext-векторами названий сущностей в Wikidata;
- Inverted index (соответствие названий и ids сущностей);
- BERT для ранжирования сущностей по контексту.

Связывание сущностей происходит в несколько этапов:

- Строка, которая подается на вход модели, векторизуется с помощью tf-idf и fasttext;
- Из tf-idf и fasttext-индексов извлекаются названия сущностей-кандидатов;
- С помощью inverted index извлекаются ids сущностей-кандидатов;
- Модель на основе BERT принимает на вход описание в Wikidata сущности-кандидата, параграф, в котором находится сущность и выдает оценку, насколько описание сущности подходит под контекст. На выходе модели top 5 сущностей, описания которых наиболее подходят под контекст.

3.4.3.5.Метод оценки качества

Для оценки качества модели используется F1 мера (п.3.3.1 данного документа).

Вычисление полноты производится по формуле:

Recall = ent_true / ent_wiki

ent_total - общее число сущностей в тестовых данных

ent_wiki - число сущностей в тестовых данных, для которых есть статья в Википедии

ent_true - количество сущностей, для которых сервис определил правильную ссылку на статью в Википедии

Вычисление точности производится по формуле:

Precision = ent_true / ent_found

ent_found - количество сущностей, для которых система нашла ссылку в Википедии (часть ссылок может быть определена правильно, часть - не правильно).

В результате вычисляется значение меры F1 по формуле:

F1 = 2 * Precision * Recall / (Precision + Recall)

3.4.4. Сервис классификации социально-демографического распределения

3.4.4.1. Общее описание сервиса

Сервис социально-демографического распределения предназначен для предсказания ожидаемых социально-демографических характеристики аудитории опубликованного документа, а именно - распределение аудитории по полу, возрасту, образованию и уровню дохода.

В качестве факторов для предсказания в сервисе используются:

- домен на котором размещено новостное сообщение;
- вектор топиков, полученный в результате обработки документа сервисом тематического моделирования.

Таким образом, для одного документа размещенного на разных ресурсах, будут предсказаны разные распределения аудитории по возрасту, полу, образованию и уровню дохода. Например, новость, размещённая на ресурсе iz.ru получит иной по составу траффик, чем новость, размещенная на безвестном новостном сайте местной газеты маленького города.

При работе сервиса домен векторизуется при помощи распределения по всем топикам у этого домена.

3.4.4.2. Конечные точки сервиса

Предсказание социально-демографического распределения

Конечная точка расположена по адресу http://сервер:порт/predict.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "id1": {
        "topic_vector": [
            0.01073046, 0.01182441, 0.00845626, 0.04557515, 0.00139581,
            0.00267226, 0.00286171, 0.00293985, 0.02671937, 0.00232662,
            0.00157005, 0.00597136, 0.00756816, 0.0038899, 0.01924294,
            0.0040252, 0.01886575, 0.01707102, 0.01273649, 0.0191225,
            0.01410219, 0.00571123, 0.05603516, 0.02741823, 0.01179593,
            0.00863104, 0.00251749, 0.00855498, 0.02367301, 0.00963152,
            0.00180897, 0.08079945, 0.00206469, 0.02671686, 0.00499152,
            0.02260217, 0.01405474, 0.01574102, 0.0085331, 0.02829218,
```

```
                0.01842694, 0.00955512, 0.01052718, 0.00339115, 0.00301299,
                0.01109661, 0.0133631, 0.00601077, 0.0092613, 0.0081462,
                0.01271417, 0.0017906, 0.00294103, 0.00283382, 0.00347364,
                0.00327252, 0.04924642, 0.00164319, 0.00370478, 0.00297635,
                0.13564222, 0.00780629, 0.02027697, 0.01721959, 0.00801877,
                0.01205254, 0.0158551, 0.0064999
            ],
            "link": "https://zen.yandex.ru/longreads/interview/_article/johnny-
        depp-quotes"
        },
        "id2": {
            "topic_vector": [
                0.01073046, 0.01182441, 0.00845626, 0.04557515, 0.00139581,
                0.00267226, 0.00286171, 0.00293985, 0.02671937, 0.00232662,
                0.00157005, 0.00597136, 0.00756816, 0.0038899, 0.01924294,
                0.0040252, 0.01886575, 0.01707102, 0.01273649, 0.0191225,
                0.01410219, 0.00571123, 0.05603516, 0.02741823, 0.01179593,
                0.00863104, 0.00251749, 0.00855498, 0.02367301, 0.00963152,
                0.00180897, 0.08079945, 0.00206469, 0.02671686, 0.00499152,
                0.02260217, 0.01405474, 0.01574102, 0.0085331, 0.02829218,
                0.01842694, 0.00955512, 0.01052718, 0.00339115, 0.00301299,
                0.01109661, 0.0133631, 0.00601077, 0.0092613, 0.0081462,
                0.01271417, 0.0017906, 0.00294103, 0.00283382, 0.00347364,
                0.00327252, 0.04924642, 0.00164319, 0.00370478, 0.00297635,
                0.13564222, 0.00780629, 0.02027697, 0.01721959, 0.00801877,
                0.01205254, 0.0158551, 0.0064999
            ],
            "link": "https://zen.yandex.ru/longreads/interview/_article/johnny-
        depp-quotes"
        }
    }
```


Параметры:

id1, id2 – id документов;

topic_vector - вектор топиков, созданный по результатам работы сервиса тематического моделирования;

link - ссылка на статью документа.

Выходные данные – JSON формата:


```
{
    "data": {
        "id1": {
            "data": {
                "age_0-11": 0.06291880458593369,
                "age_12-16": 0.06232666224241257,
                "age_17-21": 0.06618690490722656,
                "age_22-24": 0.06182613968849182,
                "age_25-29": 0.09451879560947418,
```

```
            "age_30-34": 0.09895917028188705,
            "age_35-39": 0.08977294713258743,
            "age_40-44": 0.09941617399454117,
            "age_45-49": 0.08308707177639008,
            "age_50-54": 0.07179877161979675,
            "age_55-59": 0.08174914121627808,
            "age_60-64": 0.06519130617380142,
            "age_65+": 0.06224812939763069,
            "education_doctor": 0.06309323012828827,
            "education_higher": 0.7083240747451782,
            "education_phd": 0.06611397862434387,
            "education_school": 0.06368046253919601,
            "education_special": 0.0987882912158966,
            "income_0-10": 0.05439368635416031,
            "income_10-20": 0.04798053950071335,
            "income_100-110": 0.048382628709077835,
            "income_110-120": 0.04687364771962166,
            "income_120-130": 0.04562076926231384,
            "income_130-140": 0.04069849103689194,
            "income_140-150": 0.039981987327337265,
            "income_150-160": 0.046754445880651474,
            "income_160-170": 0.05323905870318413,
            "income_170-180": 0.05445409193634987,
            "income_180-190": 0.04595192149281502,
            "income_190-200": 0.05253877490758896,
            "income_20-30": 0.049417540431022644,
            "income_30-40": 0.05723261833190918,
            "income_40-50": 0.05051540210843086,
            "income_50-60": 0.05669160187244415,
            "income_60-70": 0.050794441252946854,
            "income_70-80": 0.051240064203739166,
            "income_80-90": 0.054372839629650116,
            "income_90-100": 0.05286538600921631,
            "sex_female": 0.8404744267463684,
            "sex_male": 0.1595255732536316
        },
        "success": true
    },
    "id2": {
        "data": {
            "age_0-11": 0.06291880458593369,
            "age_12-16": 0.06232666224241257,
            "age_17-21": 0.06618690490722656,
            "age_22-24": 0.06182613968849182,
            "age_25-29": 0.09451879560947418,
            "age_30-34": 0.09895917028188705,
            "age_35-39": 0.08977294713258743,
            "age_40-44": 0.09941617399454117,
            "age_45-49": 0.08308707177639008,
            "age_50-54": 0.07179877161979675,
            "age_55-59": 0.08174914121627808,
            "age_60-64": 0.06519130617380142,
            "age_65+": 0.06224812939763069,
```

```
                    "education_doctor": 0.06309323012828827,
                    "education_higher": 0.7083240747451782,
                    "education_phd": 0.06611397862434387,
                    "education_school": 0.06368046253919601,
                    "education_special": 0.0987882912158966,
                    "income_0-10": 0.05439368635416031,
                    "income_10-20": 0.04798053950071335,
                    "income_100-110": 0.048382628709077835,
                    "income_110-120": 0.04687364771962166,
                    "income_120-130": 0.04562076926231384,
                    "income_130-140": 0.04069849103689194,
                    "income_140-150": 0.039981987327337265,
                    "income_150-160": 0.046754445880651474,
                    "income_160-170": 0.05323905870318413,
                    "income_170-180": 0.05445409193634987,
                    "income_180-190": 0.04595192149281502,
                    "income_190-200": 0.05253877490758896,
                    "income_20-30": 0.049417540431022644,
                    "income_30-40": 0.05723261833190918,
                    "income_40-50": 0.05051540210843086,
                    "income_50-60": 0.05669160187244415,
                    "income_60-70": 0.050794441252946854,
                    "income_70-80": 0.051240064203739166,
                    "income_80-90": 0.054372839629650116,
                    "income_90-100": 0.05286538600921631,
                    "sex_female": 0.8404744267463684,
                    "sex_male": 0.1595255732536316
                },
                "success": true
            }
        },
        "success": true
    }
```


Параметры:

data - словарь, в котором:

ключ - id документа;

значение - словарь, в котором содержатся пары “характеристика аудитории” - “значение характеристики”. Сумма характеристик по группам составляет 1, т.е. 100%

success – флаг успешности предсказания (true/false)

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются.

Ответ содержит JSON формата:


```
{
    "data": {
        "metric": "MSE",
        "Возраст": 0.010019822977483273,
        "Доход": 0.006741400342434645,
        "Образование": 0.00980359222739935,
        "Пол": 0.12879186868667603
    },
    "success": true
}
```


Параметры:

data - словарь с отчетом о выполнении

metric - наименование метрики

Возраст – значение отклонения от нуля метрики по классификации возраста

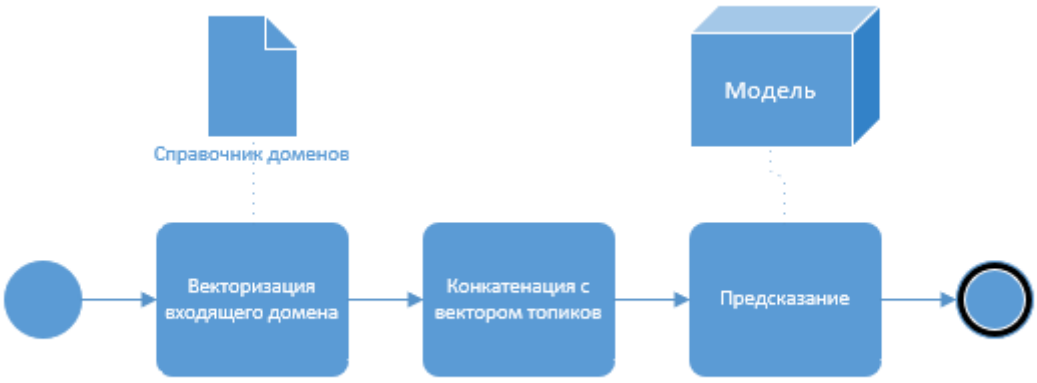
Доход - значение отклонения метрики по классификации дохода

Образование - значение отклонения метрики по классификации образования

Пол - значение отклонения метрики по классификации пола

success – параметр успеха выполнения запроса (true/false)

3.4.4.3.Процесс работы сервиса



3.4.4.4.Модель

В сервисе используется одна модель, которая решает задачу регрессии для всех признаков, а именно - имеет четыре выхода распределения: по полу, по возрасту, по образованию и по уровню дохода. Для каждого выхода, применяется слой SoftMax, чтобы полученный вектор был представлен вещественным числом в интервале [0,1] и

сумма координат равна единице. SoftMax это обобщение логистической функции для многомерного случая. Функция преобразует вектор z размерности K в вектор σ той же размерности, где каждая координата σ_i полученного вектора представлена вещественным числом в интервале [0,1] и сумма координат равна 1.

Координаты σ_i вычисляются следующим образом:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

Функция Softmax применяется в машинном обучении для задач классификации, когда количество возможных классов больше двух (для двух классов используется логистическая функция).

Структура модели (PyTorch):

ModelMultiOutput(

(layer_1): Linear(in_features=136, out_features=1024, bias=True)

(layer_2): Linear(in_features=1024, out_features=512, bias=True)

(layer_3): Linear(in_features=512, out_features=256, bias=True)

(layer_4): Linear(in_features=256, out_features=128, bias=True)

(layer_5): Linear(in_features=128, out_features=64, bias=True)

(layer_6): Linear(in_features=64, out_features=32, bias=True)

(layer_7): Linear(in_features=32, out_features=16, bias=True)

(fc_sex): Linear(in_features=16, out_features=2, bias=True)

(fc_age): Linear(in_features=64, out_features=13, bias=True)

(fc_education): Linear(in_features=64, out_features=5, bias=True)

(fc_income): Linear(in_features=64, out_features=20, bias=True)

(relu): ReLU()

(dropout): Dropout(p=0.2, inplace=False)

(batchnorm1):BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm2): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm3): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm4): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm5): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm6): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(batchnorm7): BatchNorm1d(16, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)

(softmax): Softmax(dim=None)

)

3.4.4.5.Метод оценки качества

Качество предсказания проверяется при помощи MSE (п.3.3.2 данного документа).

В результате измерений были получены следующие результаты:

label_sex = 0.165

label_age = 0.0515

label_education = 0.05

label_income = 0.0347

3.4.5. Сервис тематического моделирования

3.4.5.1.Общее описание сервиса

Сервис осуществляет иерархическое извлечение тематик из документа с разделением их по типовым рубрикам, подрубрикам, сюжетам и событиям (новостям). Рубрики и подрубрики классифицируются на основе рубрикатора.

Сервис тематического моделирования состоит из нескольких микросервисов:

1) Siamese embedder - сервис встраиватель (эмбеддер) текстов. Принимает на вход текст, возвращает векторные представления параграфов полученного текста.

2) Entities embedder - сервис встраиватель (эмбеддер) сущностей. Принимает на вход список сущностей, возвращает суммарное векторное представление всех сущностей.
3) Faiss Index - индекс векторных представлений новостей/сюжетов. Предназначен для поиска ближайших к документу новостей/ближайших к новости сюжетов.
4) PostgreSQL - база данных. Хранит информацию о документах/ новостях/ сюжетах. Из этой базы происходит восстановление Faiss Index.
5) Topic Modeling - центральный микросервис. Обеспечивает взаимодействие между остальными сервисами. В этом микросервисе происходит кластеризация документов в новости/кластеризация новостей в сюжеты. Дополнительно, в этом микросервисе находится backend рабочего места асессора.
6) Nginx – обратный прокси сервер с открытым исходным кодом. Предназначен для балансировки/репликации запросов к Faiss index, Siamese embedder и Entities embedder.

Описание микросервисов:

1) Siamese embedder

Микросервис представляется собой программный интерфейс приложения (API) который обеспечивает доступ к модели эмбеддера, основанной на BERT. Принимает на вход текст и идентификатор документа. На выходе возвращает векторное представление каждого параграфа текста, где каждый вектор — это одномерный массив float32 длинной 300.

2) Entities embedder

Микросервис представляется собой программный интерфейс приложения (API), который обеспечивает доступ к модели эмбеддера основанной на fast text (библиотека для изучения встраивания слов и классификации текста, созданная исследовательской лабораторией Facebook AI Research). Принимает на вход список сущностей и идентификатор документа. На выходе возвращает суммарное векторное представление сущностей - одномерный массив float32 длинной 100.

3) Faiss Index

Микросервис содержит индексы новостей и сюжетов, для построения индекса используется FlatIndexIP, пороговое значение для новостей 0.94, для сюжетов - 0.90. При получении запроса на поиск ближайшей новости/сюжета индекс проверяет дату пришедшего документа/новости, в случае если дата не совпадает дате сборки индекса - индекс пересобирается из базы данных PostgreSQL за дату пришедшего документа/ новости. В случае индекса новостей, индекс содержит новости за искомую дату минус 3 дня. В случае сюжетов - текущая дата минут 7 дней.

4) PostgreSQL

База данных сервиса. Хранит в себе информация о документах, новостях, сюжетах. Из нее происходит пересборка индексов новостей/сюжетов за требуемую дату.

5) Topic Modeling

Микросервис является связующим звеном для всех остальным микросервисов. Именно сюда приходит запрос на обработку документов, и отсюда происходят запросы ко всем остальным микросервисам. Здесь происходит получение тематического распределения документа, кластеризация в новости и сюжеты.

6) Nginx

Сервис предназначен для балансировки и репликации запросов при определенных обращениях к разным сервисам. При запросе на обновление индекса новостей/сюжетов микросервиса Faiss index, запрос реплицируется, то есть запрос отправляется одновременно на все реплики микросервиса. При запросе на поиск ближайших новостей/сюжетов запрос балансируется между всеми развернутыми репликами микросервиса Faiss index, то есть микросервис Nginx каждый такой запрос отправляет на одну из реплик микросервиса Faiss index. В случае с запросами к Siamese embedder и Entities embedder также происходит балансировка.

3.4.5.2.Конечные точки сервиса

Получение тематического распределения с обновлением кластеров

Конечная точка расположена по адресу http://сервер:порт/model/transform_update.

Конечная точка служит для заполнения баз данных индексов сервиса.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "3010": {
        "date": "2021-10-16T22:11:32",
        "text": "Спортивный директор «Эвертона» Марсель Брандс прокомментировал «СЭ» информацию об интересе английского клуба к российскому хавбеку «Монако» Александру Головину. Ранее, 10 ноября, Liverpool Echo сообщил, что моне гаски запросили у «Эвертона» 46 миллионов евро за 25-летнего футболиста. «Информация, что «Монако» запросил у нас за Головина 46 миллионов евро — это абсолютная чушь», — сказал Брандс «СЭ». В текущем сезоне Головин провел 16 матчей за «Монако», забил три гола и сделал три результативные передачи. Портал Transfermarkt оценивает полузащитника в 28 миллионов евро. Контракт игрока с французской командой действует до конца июня 2024-го.",
        "title": "Главный тренер \"Ахмата\" Рашид Рахимов",
        "entities": ["Рашид Рахимов"]
    },
    "3011": {
        "date": "2021-06-10 12:31:00",
        "text": "Премьера мини-сериала «Локи» в стрим-сервисе Disney+ состоялась 9 июня 2021 года. Кинокритики уже поделились первыми впечатлениями от пилотных эпизодов шоу. «Первые два эпизода «Локи» просто фантастические. Том Хиддлстон по-прежнему идеально подходит для роли Бога озорства», — написал в Twitter продюсер Стивен Вайнтрауб. \"«Локи» — это именно то, чего я ждала: забавный, странный и захватывающий сериал\", — заявила кинокритик Лиз Шеннон Миллер. Экспертами отмечена не только юмористическая составляющая, но и запутанность сюжета, которая создает интригу для зрителя. Актерская игра в сериале также не оставила никого равнодушным. Главные роли в сериале сыграли Том Хиддлстон, Оуэн Уилсон, Гугу Эмбата-Ро, Ричард Грант, Вунми Моссаку и др.",
        "title": "Сериал «Локи» от Disney+ вызвал восторг у критиков",
        "entities": [
            "Disney Channel",
            "кинокритик",
            "Том Хиддлстон",
            "стивен вайнтрауб",
            "лиз шеннон миллер",
            "Том Хиддлстон",
            "Оуэн Уилсон",
            "гугу эмбата-ро",
            "Ричард Э. Грант",
            "вунмить моссак"
        ]
    }
}
```

Параметры:

3010, 3011 – id документа;

date – дата публикации новости;

text – текст новости;

title - заголовок;

entities – сущности в новости из результата сервиса идентификации именованных сущностей.

Выходные данные – JSON формата:


```
{
    "data": {
        "3010": {
            "status": true,
            "topics": [
                {
                    "lvl1": {
                        "id": "lvl1_спорт",
                        "topic_label": "спорт"
                    },
                    "lvl2": {
                        "id": "lvl2_футбол",
                        "topic_label": "футбол"
                    },
                    "lvl3": {
                        "id": "2516e52b-7b1b-45f2-9b91-77f2de766615",
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "lvl4": {
                        "id": "67d990a1-28f9-4896-b949-07f20c84c768",
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "probability": "0.92479223"
                },
                {
                    "lvl1": {
                        "id": "lvl1_спорт",
                        "topic_label": "спорт"
                    },
                    "lvl2": {
                        "id": "lvl2_единоборства",
                        "topic_label": "единоборства"
                    },
                    "lvl3": {
                        "id": "2516e52b-7b1b-45f2-9b91-77f2de766615",
```

```
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "lvl4": {
                        "id": "67d990a1-28f9-4896-b949-07f20c84c768",
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "probability": "0.006960137"
                },
                {
                    "lvl1": {
                        "id": "lvl1_экономика_и_финансы",
                        "topic_label": "экономика и финансы"
                    },
                    "lvl2": {
                        "id": "lvl2_грузовые_перевозки",
                        "topic_label": "грузовые перевозки"
                    },
                    "lvl3": {
                        "id": "2516e52b-7b1b-45f2-9b91-77f2de766615",
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "lvl4": {
                        "id": "67d990a1-28f9-4896-b949-07f20c84c768",
                        "topic_label": "главный тренер \"ахмата\" рашид рахи
мов"
                    },
                    "probability": "0.0045508076"
                }
            ],
            "topics_vector": [
                0.0013448467943817377,
                0.0013581293169409037,
                0.0005186093039810658,
                0.0007193339988589287,
                0.000337220641085878,
                0.0017386515391990542,
                0.0010560646187514067,
                0.0008481445838697255,
                0.0003567746898625046,
                0.0007882007048465312,
                0.00038240724825300276,
                0.00048089242773130536,
                0.0011750489939004183,
                0.00033560575684532523,
                0.0020126833114773035,
                0.0002931936760433018,
                0.0003198668418917805,
                0.0006002318696118891,
                0.001061174669303,
                0.0007373489788733423,
```

```
            0.0005228542722761631,
            0.00044471881119534373,
            0.003107398748397827,
            0.0006531326216645539,
            0.0003705311974044889,
            0.0012666588881984353,
            0.0003507613728288561,
            0.0006442598532885313,
            0.00034335217787884176,
            0.0004562075773719698,
            0.0017711883410811424,
            0.0004450442793313414,
            0.000836075865663588,
            0.001023207907564938,
            0.0005409317091107368,
            0.001291695749387145,
            0.0005382995004765689,
            0.0004503126547206193,
            0.00045894228969700634,
            0.0004640522529371083,
            0.00039503295556642115,
            0.0011995802633464336,
            0.001460414729081094 3,
            0.001826175837777555,
            0.0014350279234349728,
            0.0036359464284032583,
            0.002906937850639224,
            0.0006133006536401808,
            0.0024886359460651875,
            0.0008326455135829747,
            0.0004455910238903016,
            0.00023580198467243463,
            0.0008355521713383496,
            0.001452188822440803,
            0.0008932484197430313,
            0.006960136815905571,
            0.0018588085658848286,
            0.9247922301292419,
            0.002180109964683652,
            0.0005125403986312449,
            0.004550807643681765,
            0.0006282291724346578,
            0.0012902762973681092,
            0.0004961909144185483,
            0.00038191265775822103,
            0.0006012347876094282,
            0.0011479539098218083,
            0.0014994239900261164
        ]
    },
    "3011": {
        "status": true,
        "topics": [
```

```
                {
                    "lvl1": {
                        "id": "lvl1_культура",
                        "topic_label": "культура"
                    },
                    "lvl2": {
                        "id": "lvl2_кино",
                        "topic_label": "кино"
                    },
                    "lvl3": {
                        "id": "3d42e1ea-c531-4427-b1ef-18b93fc4a204",
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "lvl4": {
                        "id": "23fa3698-b552-455a-8215-9e46351940f7",
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "probability": "0.97795004"
                },
                {
                    "lvl1": {
                        "id": "lvl1_общество",
                        "topic_label": "общество"
                    },
                    "lvl2": {
                        "id": "lvl2_пенсии",
                        "topic_label": "пенсии"
                    },
                    "lvl3": {
                        "id": "3d42e1ea-c531-4427-b1ef-18b93fc4a204",
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "lvl4": {
                        "id": "23fa3698-b552-455a-8215-9e46351940f7",
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "probability": "0.0031672996"
                },
                {
                    "lvl1": {
                        "id": "lvl1_наука_и_техника",
                        "topic_label": "наука и техника"
                    },
                    "lvl2": {
                        "id": "lvl2_интернет",
                        "topic_label": "интернет"
                    },
                    "lvl3": {
                        "id": "3d42e1ea-c531-4427-b1ef-18b93fc4a204",
```

```
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "lvl4": {
                        "id": "23fa3698-b552-455a-8215-9e46351940f7",
                        "topic_label": "сериал «локи» от disney+ вызвал вост
орг у критиков"
                    },
                    "probability": "0.0017366753"
                }
            ],
            "topics_vector": [
                0.0001423590147169307,
                0.00013699621194973588,
                0.0005357770714908838,
                0.000185827913810499,
                0.0003011987137142569,
                0.001641130424104631,
                0.9779500365257263,
                0.00046509309322573245,
                0.00012753982446156442,
                0.0002370615693507716,
                0.00041706705815158784,
                0.0002682309423107654,
                0.00016228537424467504,
                0.00011726392403943464,
                0.000532665231730789,
                0.00031048135133460164,
                0.00034153732121922076,
                0.0008097506361082196,
                0.0017366752726957202,
                0.00046043278416618705,
                0.00011925627768505365,
                0.0001184455250040628,
                0.0016652949852868915,
                8.759759657550603e-05,
                0.00019714758673217148,
                0.00021040219871792942,
                0.0031672995537519455,
                0.0002672033151611686,
                0.000339273625286296,
                0.00023173571389634162,
                0.00014093401841819286,
                0.000123604157124646,
                9.805648005567491e-05,
                0.0001427553070243448,
                0.00033925636671483517,
                0.00017121115524787456,
                7.754280522931367e-05,
                0.0001331569073954597,
                0.0001297164271818474,
                0.0003154654987156391,
                0.00010954472963931039,
```

```
                    0.00012735922064166516,
                    0.0001305426994804293,
                    0.0002897520025726408,
                    5.591001900029369e-05,
                    0.00027567948563955724,
                    0.00014479001401923597,
                    0.0010370537638664246,
                    8.139787678373978e-05,
                    0.00010222324635833502,
                    9.996777953347191e-05,
                    5.318289186106995e-05,
                    0.000114343965833541,
                    0.00015631693531759083,
                    7.064663077471778e-05,
                    0.00029855084721930325,
                    0.00010731125803431496,
                    0.00014679045125376433,
                    0.00046802754513919353,
                    8.1613652582746e-05,
                    0.0002557874540798366,
                    0.00012016709661111236,
                    0.0001289502251893282,
                    0.00010736889817053452,
                    6.0917125665582716e-05,
                    0.00015584017091896385,
                    0.00013621742255054414,
                    0.00012901899754069746
                ]
            }
        },
        "success": true
    }
```


Параметры:

data - словарь результатов

3010, 3011 – id документа

status - параметр успеха обработки документа (true/false)

topics – массив из 3 словарей с результатами предсказания для каждого документа в порядке убывания вероятности

lvl1, lvl2, lvl3, lvl4 – уровни «Рубрика», «Подрубрика», «Сюжет», «Новость»

id – идентификатор для уровней

topic_label – название уровня

probability – вероятность предсказания

topics_vector – тематический вектор документа

success - параметр успеха выполнения запроса (true/false)

Получение тематического распределения из документа

Конечная точка расположена по адресу http://сервер:порт/model/transform.

Метод запроса на сервер, параметры входа и выхода аналогичны конечной точке «Получение тематического распределения с обновлением кластеров»

“TM Assessor” – “Custom topic” - “Создание пользовательского топика”

Конечная точка расположена по адресу http://сервер:порт/assessor/fit_custom_topics.

Метод запроса на сервер - POST.

Входные данные:

Form-data с содержимым:

- Title – Тема первого уровня
- Subheading – тема второго уровня
- Documents – JSON файл формата:


```
{
    "196115839": {
        "date": "2021-06-10 12:30:59",
                                        "text": "Фото: пресс-
служба КХЛ Чемпион Высшей хоккейной лиги объявил о переходе защитника пете
рбургского \"Динамо\". Как сообщает пресс-служба ханты-
мансийской \"Югры\", состав обладателя Кубка Петрова пополнил Анатолий Ели
заров. В ВХЛ воспитанник питерской школы хоккея выступал за \"Торос\", \"Т
амбов\" и \"Динамо\". В 135 матчах игрок обороны записал на свой бомбардир
ский счёт 19 (6+13) очков. Что касается КХЛ, то на спортивном пути Елизаро
ва были \"Салават Юлаев\" и \"Сочи\". В составе уфимской команды с 2017-
го                              по                              2019-
й год он принял участие в 76 играх, отметившись 1 голевой передачей. Его п
оказатель полезности составил \"+1\", суммарное время штрафа - 10 минут.",
                                                    "title":    "Экс-
игрок «Салавата Юлаева» перешёл в чемпионскую команду",
        "entities": ["салават юлаева", "кхл", "высокий хоккейный лига", "Д
инамо",                                                    "Ханты-
Мансийский автономный округ — Югра", "Владимир Владимирович Петров", "анат
олий елизар", "Торос", "Тамбов", "Динамо", "кхл", "Владимир Николаевич Ели
заров", "салават юлай", "Сочи"]
        },
    "196086913": {
        "date": "2021-06-10 12:31:00",
        "text": "В рамках реализуемого Волгоградским региональным отделени
ем Российского общества \"Знание\" проекта \"Организация и проведение прос
```

ветительских мастер-классов по цифровой грамотности\" они успешно освоили ещё одну профессию - \"Консультант в области развития цифровой грамотности населения \"Цифровой куратор\" (квалификации 5-го уровня).\n\r\nБесплатное обучение проводилось в Сетевом университете \"Знание\". Участники познакомились с физиологическими и психологическими особенностями различных категорий населения, получили актуальные методические материалы.\n\r\nУспешно выдержанное итоговое тестирование позволило участникам группы получить новую дополнительную специальность. Список активных просветителей Волгоградского регионального отделения Российского общества \"Знание\" пополнится двенадцатью новыми именами, а жители города и областям узнают о том, что такое \"цифровой город\", что принесёт цифровизация обществу и стоит ли этого бояться, познакомятся с разновидностями мобильных устройств, их функциями и возможностями.\n\r\n\"Это увлекательный и полезный проект Российского общества \"Знание\", - отметила заведующая кафедрой Информационной безопасности Олеся Александровна Какорина. Он находится в русле задач, которые очертил Президент России Владимир Владимирович Путин в своём Послании Федеральному собранию - \"Знания должны вновь стать одной из важнейших ценностей общества, притягательной и доступной\"\n\r\nКакорина Олеся Александровна, заведующая кафедрой\n\r\nКафедра Информационной безопасности\n\r\nВолгоградский государственный университет",
"title": "Завершилось обучение в онлайн режиме группы преподавателей Института приоритетных технологий Волгоградского государственного университета",
"entities": ["Волгоградский государственный технический университет", "волгоградский", "российский общество \"знание\"", "Фонд SCP", "сетевой университет", "Знание", "ассоциация", "Знание", "ассоциация", "Бердник, Олесь Павлович", "александр какорин", "Россия", "Владимир Владимирович Путин", "федеральный собрание", "Какорин, Николай Иванович", "Бердник, Олесь Павлович"]
}
}

Параметры:

196115839, 196086913 – id документа;

date – дата публикации новости;

text – текст новости;

title - заголовок;

entities – сущности в новости из результата NER.

Выходные данные – JSON формата:

{

```
"data": {
    "date": "2021-11-28 08:31:47.158806",
    "heading": "Пользовательская рубрика",
    "id": "25acdde8-6bb4-44c7-88fa-8ef5a144e49c",
    "story": {
        "date": "2021-11-01 06:55:00",
        "id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
        "news": [
            {
                "date": "2021-11-01 06:49:00",
                "docs": [
                    "1e981973-8856-463e-b776-1b7eefeb9674"
                ],
                "id": "1f23c7d7-9abd-41ab-b524-d078ec610691",
                "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "title": "оксана гаврилина"
            },
            {
                "date": "2021-11-01 06:49:00",
                "docs": [
                    "a83ddf43-d9c1-48f4-8b03-7008a4d2bcc4"
                ],
                "id": "5a8f195f-b426-4194-9ca6-8a0397297f73",
                "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "title": "маг универ с верой ерохиной «секреты силы»"
            },
            {
                "date": "2021-11-01 06:51:00",
                "docs": [
                    "f9854bdb-085c-4b33-82f6-88d31816134d"
                ],
                "id": "da5c1744-1263-49ad-8826-9a15bf6b864b",
                "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "title": "наш хоккей [чемпионат мира 2021] кхл нхл"
            },
            {
                "date": "2021-11-01 06:52:00",
                "docs": [
                    "1e3f3888-5bba-4518-9a9d-14921a36de91"
                ],
                "id": "872cb684-2ade-4bd7-9ef7-a6335ce26d86",
                "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "title": "наша беседка в мичуринском"
            },
            {
                "date": "2021-11-01 06:53:00",
                "docs": [
                    "6e394e99-6ba2-4814-87e1-7926aa0a08a2"
                ],
                "id": "2c1fddc1-18f4-47c6-b25a-3260e85ef211",
                "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "title": "новости волгограда|volga-day.ru"
            },
```

```
                {
                    "date": "2021-11-01 06:54:00",
                    "docs": [
                        "af91fafc-bd00-4970-8e4b-5139f005e2fc"
                    ],
                    "id": "f56ac8d8-6fb0-4343-81ff-5f29b9739b4a",
                    "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                    "title": "люди неограниченных возможностей"
                },
                {
                    "date": "2021-11-01 06:55:00",
                    "docs": [
                        "e3f6d79a-2796-446b-af53-e541b8d37599"
                    ],
                    "id": "5478afd7-f21f-44d2-a617-9579b3a7ae67",
                    "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                    "title": "what is pennywise?"
                }
            ],
            "title": "Тестовая рубрика",
            "topics": [
                [
                    "общество/быт",
                    112.5702133178711
                ],
                [
                    "культура/кино",
                    103.51881408691406
                ],
                [
                    "наука_и_техника/интернет",
                    68.95503997802734
                ]
            ]
        },
        "subheading": "Тестовая подрубрика"
    },
    "success": true
}
```


Параметры:

data – словарь результатов выполнения запроса

date – дата создания пользовательского топика

heading – заголовок топика

subheading – подзаголовок топика

id – id созданного пользовательского топика

story – словарь сюжетов

date – дата сюжета

id – id сюжета

news – массив словарей новостей в сюжете

date – дата новости

docs – документы новости

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

title – заголовок сюжета

topics – тематическое распределение по рубрикам/подрубрикам

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Получение популярных сюжетов”

Конечная точка расположена по адресу http://сервер:порт/assessor/top_story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "moderated": false,
    "limit": 100,
    "page": 1,
    "context": "",
    "from_date": null,
    "to_date": null
}
```

Параметры:

moderated – признак внесения изменений в сюжет пользователем (true/false)

limit – количество сюжетов в выборке

page – страница с которой делать выборку (например, при значении 2 – сервис выберет 200 лучших и отдаст с 101 по 200-й)

context – подстрока для поиска

from_date – начало диапазона поиска

to_date – окончание диапазона поиска

Выходные данные – JSON формата:

```
{
    "data": [
```

```
            {
                "date": "2021-11-01 06:55:00",
                "docs_contained": 500,
                "docs_increase": 0,
                "id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
                "news_contained": 134,
                "title": "Тестовая рубрика"
            },
            {
                "date": "2021-06-10 12:58:57",
                "docs_contained": 106,
                "docs_increase": 0,
                "id": "35d1309a-ddc9-47ab-ad90-109c819c8110",
                "news_contained": 12,
                "title": "вогл - ввод сдвигается на 2022 год"
            },
            {
                "date": "2021-06-10 12:58:53",
                "docs_contained": 69,
                "docs_increase": 0,
                "id": "8196611d-321e-4228-bc85-8533dc476b05",
                "news_contained": 14,
                "title": "доску почета обновили в белогорске"
            },
            {
                "date": "2021-06-10 12:32:58",
                "docs_contained": 1,
                "docs_increase": 0,
                "id": "895e5eeb-ec19-4b06-88bc-fe3444e876bf",
                "news_contained": 1,
                "title": "марк цукерберг готовится к концу света, но пока лишь в
мемах. не стоило бизнесмену хвастать необычным хобби"
            }
        ],
        "number_of_pages": 1,
        "success": true
    }
```


Параметры:

data – массив результатов

date – дата сюжета

docs_contained – количество документов, содержащих сюжет

docs_increase – количество документов, содержащих сюжет за последние сутки

id – id сюжета

news_contained – количество новостей в сюжете

title – название сюжета

number_of_pages – количество страниц ответа

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Получение сюжета по id”

Конечная точка расположена по адресу http://сервер:порт/assessor/story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "story_id": "123123"
}
```


Параметры:

story_id – id сюжета

Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-06-10 12:58:33",
        "id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
        "news": [
            {
                "date": "2021-06-10 12:41:01",
                "docs": [
                    "196128927",
                    "196110827"
                ],
                "id": "504a2f09-8da3-46c6-9f44-395c1ac54611",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "в бельцах пройдёт день донора крови: где, когда и
кто может им стать?"
            },
            {
                "date": "2021-06-10 12:58:33",
                "docs": [
                    "196106377"
                ],
                "id": "86222bf0-c4f9-4fc4-a0ff-564350d52b77",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "китайские ученые разработали белок для защиты нейр
онов при инсульте"
            }
        ],
        "title": "в бельцах пройдёт день донора крови: где, когда и кто може
т им стать?",
        "topics": [
            [
                "наука_и_техника/другое",
```

```
                0.6886200904846191
            ],
            [
                "наука_и_техника/школа, университеты",
                0.48810485005378723
            ],
            [
                "общество/здравоохранение",
                0.41101008653640747
            ]
        ]
    },
    "success": true
}
```


Параметры:

data – массив результатов

date – дата сюжета

id – id сюжета

news – список новостей в сюжете

date – дата новости

docs – id документов, содержащие эту новость

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

topics – тематическое распределение по рубрике/подрубрике

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Получение новости по id”

Конечная точка расположена по адресу http://сервер:порт/assessor/news.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "news_id": "f3358dff-1fc5-4c00-b00b-1f01aa4c6de5"
}
```


Параметры:

news_id – id новости

Выходные данные – JSON формата:

```
{
    "data": {
        "date": "2021-11-01 06:54:00",
        "docs": [
            "e0aac67f-cdd4-489d-b5ea-1db8d34dd393",
            "aaec7e3c-ca73-4023-b3a7-02dcc1aa5700",
            "802cadbe-dfb1-4e02-9bd8-351b1b7412bd",
            "2310bf4b-6898-42e1-acc6-629a2e383abe",
            "80332217-2ede-4fe4-98eb-dbfaf12891c3",
            "fcfb5b34-5bbf-465c-8ca4-03f3030ecf87",
            "56852e42-9c05-4774-a1d9-f8ba2fd6bf6a",
            "31770e75-3c84-43b5-8724-0cfe28d2de15",
            "2aa31487-91d3-4799-ba39-b5d6dea93fc1",
            "76212885-9e50-439f-9ca6-201c39f1b8d8",
            "78c8d641-4081-43d5-b016-1577f39a8345",
            "87d8f54a-00ec-40ba-b6d8-96149ee9ecd9"
        ],
        "id": "181cbfb4-40f7-44f9-8578-fcb60d060eda",
        "story_id": "cc8da436-f99d-4347-bc39-e2cf9db3433e",
        "title": "овен"
    },
    "success": true
}
```


Параметры:

data – словарь результатов запроса

date – дата новости

docs – id документов, содержащие эту новость

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Получение документа по id”

Конечная точка расположена по адресу http://сервер:порт/assessor/document.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "document_id": "id документа"
}
```

Параметры:

document_id – id документа

Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-11-01 05:21:00",
        "id": "89225976-d7c7-4884-8fdf-518c2f844f6a",
        "news_id": "346fd4d0-40df-44fe-a7fe-d46f6b23e783",
        "text": ".HD Кино - Фильмы онлайн 2021сегодня в 8:00ПЕ|РЕРІОЖД|
ЕНИЕ(2019)Жанр: фантастика, фэнтези, боевик, триллер, драма, приключенияДействие ра
зворачивается в постапокалиптическом будущем и охватывает тысячелетний период. В ре
зультате секретного правительственного эксперимента на свободе оказываются кровожад
ные вампиры, зараженные опасным вирусом. Остановить распространение болезни, грозящ
ей уничтожить человеческий род, под силу лишь девочке-
сироте по имени Эми, которая обладает странной властью над вирусом.#кино #киношка #
киномания #кинопоиск #киноманы #кино2021 #киноман #киноночь #новинкикино #киноотзыв
 #кинопробы #кинофильм #киноха #кинолог #киноновинки #кинотеатр #кинозал7:28:57",
        "title": "Евгений Медведев"
    },
    "success": true
}
```


Параметры:

data – словарь результатов запроса

date – дата документа

id – id документа

news_id – id новостей в документе

text – текст документа

title – заголовок документа

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Перенос новости в другой сюжет”

Конечная точка расположена по адресу http://сервер:порт/assessor/transfer_news.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "news_id": "007c8478-0261-477d-ad9e-941f76ee34e8",
    "story_id": "e1ca6061-d9ff-4a83-b30e-7b9db839812d"
}
```


Параметры:

news_id – id новости

story_id – id сюжета

Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-11-01 06:54:00",
        "docs": [
            "e0aac67f-cdd4-489d-b5ea-1db8d34dd393",
            "aaec7e3c-ca73-4023-b3a7-02dcc1aa5700",
            "802cadbe-dfb1-4e02-9bd8-351b1b7412bd",
            "2310bf4b-6898-42e1-acc6-629a2e383abe",
            "abe99968-7680-47c7-bd87-8df647f8ed95",
            "50937df9-223f-4318-b714-0aaa6f97cd20",
            "a4315d7b-3f8c-4d31-a260-eafaa2ad3633",
            "2e659541-b20e-4da4-87e2-3f8795ef6481",
            "c3bcffed-bf6c-4001-8a8a-1394b3fa5265",
            "801340be-78af-44a2-810f-4a04e6158751",
            "17dbed29-ccf5-42b7-b1e6-33628906651a",
            "7798f726-7648-4b3b-84e8-f9af87e9f7b8",
            "80332217-2ede-4fe4-98eb-dbfaf12891c3",
            "fcfb5b34-5bbf-465c-8ca4-03f3030ecf87",
            "56852e42-9c05-4774-a1d9-f8ba2fd6bf6a",
            "31770e75-3c84-43b5-8724-0cfe28d2de15",
            "2aa31487-91d3-4799-ba39-b5d6dea93fc1",
            "76212885-9e50-439f-9ca6-201c39f1b8d8",
            "78c8d641-4081-43d5-b016-1577f39a8345",
            "87d8f54a-00ec-40ba-b6d8-96149ee9ecd9"
        ],
        "id": "007c8478-0261-477d-ad9e-941f76ee34e8",
        "story_id": "e1ca6061-d9ff-4a83-b30e-7b9db839812d",
        "title": "овен"
    },
    "message": "News successfully transferred",
    "success": true
}
```


Параметры:

data – словарь результатов

date – дата сюжета

docs – документы, перенесенные в другой сюжет

id – id новости

story_id – id сюжета

title – заголовок

message – сообщение о результате выполнения запроса

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Слияние сюжетов”

Конечная точка расположена по адресу http://сервер:порт/assessor/merge_story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "base_story_id": "5b5b5caa-2c22-462b-95d4-a1605ba6ddeb",
      "other_story_id": ["e1b0a301-8f30-4221-958b-f02e6a1abeff", "25afd122-fc94-4325-b35d-fde2442f3ceb"]
}
```


Параметры:

base_story_id – id сюжета к которому привязать

other_story_id – id сюжетов для привязки

Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-06-10 12:58:33",
        "id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
        "news": [
            {
                "date": "2021-06-10 12:33:57",
                "docs": [
                    "196125872"
                ],
                "id": "7344a25c-1696-4767-83b0-86d020f9bf4b",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "российского шпиона арестовали в польше, ему грозит 10 лет тюрьмы"
            },
            {
                "date": "2021-06-10 12:54:00",
                "docs": [
                    "196079774"
                ],
                "id": "b07811c1-4cf5-4cb8-af64-32c75a15907c",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "австрийский солдат получил срок за татуировку свастики между ног"
            },
            {
                "date": "2021-06-10 12:36:00",
                "docs": [
                    "196112369"
                ],
                "id": "fc92a9a0-7aab-4efe-8f04-83d8368ce9a0",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "джордан кларксон поможет азиатской семье, чей фудтрак в солт-лейк-сити исписали вандалы"
            },
```

```
                    {
                        "date": "2021-06-10 12:41:01",
                        "docs": [
                            "196128927",
                            "196110827"
                        ],
                        "id": "504a2f09-8da3-46c6-9f44-395c1ac54611",
                        "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                        "title": "в бельцах пройдёт день донора крови: где, когда и
кто может им стать?"
                    },
                    {
                        "date": "2021-06-10 12:58:33",
                        "docs": [
                            "196106377"
                        ],
                        "id": "86222bf0-c4f9-4fc4-a0ff-564350d52b77",
                        "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                        "title": "китайские ученые разработали белок для защиты нейр
онов при инсульте"
                    }
                ],
                "title": "в бельцах пройдёт день донора крови: где, когда и кто може
т им стать?",
                "topics": [
                    [
                        "культура/СМИ",
                        0.4758785367012024
                    ],
                    [
                        "наука_и_техника/другое",
                        0.47433939576148987
                    ],
                    [
                        "наука_и_техника/школа, университеты",
                        0.35049858689308167
                    ]
                ]
            },
            "message": "Stories successfully merged",
            "success": true
        }
```


Параметры:

data – словарь результатов

date – дата сюжета к которому привязывают

id – id сюжета к которому привязывают

news – массив словарей новостей в сюжете (с учетом слияния)

date – дата новости

docs – документы новости

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

title – заголовок сюжета

topics – тематическое распределение по рубрикам/подрубрикам

message – сообщение о результате выполнения запроса

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Слияние новостей”

Конечная точка расположена по адресу http://сервер:порт/assessor/merge_news.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "base_news_id": "6c94f6cd-cd10-489d-8078-05184a60e216",
    "other_news_id": ["bf28f38c-3902-4c83-bf1f-55351fc04e80"]
}
```


Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-06-10 12:58:25",
        "docs": [
            "196078358",
            "196078536",
            "196078588",
            "196078598",
            "196087472",
            "196088587",
            "196088655",
            "196088669",
            "196092557",
            "196092869",
            "196099832",
            "196099835",
            "196099841",
            "196097043",
            "196126727",
            "196079649",
            "196117901",
            "196086144",
            "196148221",
            "196098675",
            "196109261",
            "196075934",
```

```
            "196093692",
            "196086155",
            "196082399",
            "196097019",
            "196098581",
            "196085560",
            "196078957",
            "196076888",
            "196100369"
        ],
        "id": "fbab55c1-b6fb-4e1e-8ffc-0ee28a29ebcb",
        "story_id": "f02d5f88-a898-4462-8f74-ab4267094f54",
        "title": "в интервью юрия дудя обнаружили пропаганду наркотиков"
    },
    "message": "News successfully merged",
    "success": true
}
```


Параметры:

data – словарь результатов

date – дата новости

docs – документы новости

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

message – сообщение о результате выполнения запроса

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Создание сюжета из новости”

Конечная точка расположена по адресу http://сервер:порт/assessor/create_story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "title": "Новый сюжет",
    "news_id": "f17ba96e-5c67-4072-966e-686514a0f691"
}
```


Параметры:

title – название сюжета

news_id – id новости

Выходные данные – JSON формата:


```
{
```

```
    "data": {
        "date": "2021-11-01 06:54:00",
        "id": "1690b929-9dc6-4041-815f-2c762cc61f05",
        "news": [
            {
                "date": "2021-11-01 06:54:00",
                "docs": [
                    "52c4455a-754a-406c-84b6-16a03cf69d2f",
                    "ec1244e5-26c2-49fc-8cac-c30bbe766b79",
                    "b841986c-4ef6-427e-a647-01898bc0292c",
                    "4ff878d5-31b1-4594-88c3-76cb2dab8647",
                    "b6aeb9a4-347c-4931-9a33-cc222c295942",
                    "1f9cb2b9-5720-4a18-80a3-78126a950806",
                    "93009440-8fdf-49bd-92d7-5e5fde486704",
                    "9c1a44e2-a114-4d88-a5e3-d38dddc48eb9",
                    "290f134a-a205-4c39-a287-0f72c77c7c60",
                    "5a73bbd6-2c1d-4da4-8e83-2e3ee5cbafd7",
                    "2eaa3b6a-7dd3-471a-9eb9-3da0abf45c28"
                ],
                "id": "f17ba96e-5c67-4072-966e-686514a0f691",
                "story_id": "1690b929-9dc6-4041-815f-2c762cc61f05",
                "title": "питер :)"
            }
        ],
        "title": "Новый сюжет",
        "topics": [
            [
                "культура/музыка",
                0.6823224425315857
            ],
            [
                "культура/фестивали, выставки, мероприятия",
                0.6702058911323547
            ],
            [
                "культура/живопись",
                0.2804471552371979
            ]
        ]
    },
    "message": "Story successfully created",
    "success": true
}
```


Параметры:

data – словарь результатов

date – дата сюжета к которому привязывают

id – id сюжета к которому привязывают

news – массив словарей новостей в сюжете

date – дата новости

docs – документы новости

id – id новости

story_id – id сюжета уровнем выше

title – заголовок новости

title – заголовок сюжета

topics – тематическое распределение по рубрикам/подрубрикам

message – сообщение о результате выполнения запроса

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Переименовать сюжет”

Конечная точка расположена по адресу http://сервер:порт/assessor/update_story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
    "name": "Тестовое название"
}
```


Параметры:

story_id – id сюжета

name – новое название сюжета

Выходные данные – JSON формата:


```
{
    "data": {
        "date": "2021-06-10 12:58:33",
        "id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
        "news": [
            {
                "date": "2021-06-10 12:41:01",
                "docs": [
                    "196128927",
                    "196110827"
                ],
                "id": "504a2f09-8da3-46c6-9f44-395c1ac54611",
                "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                "title": "в бельцах пройдёт день донора крови: где, когда и
кто может им стать?"
            },
            {
                "date": "2021-06-10 12:58:33",
```

```
                        "docs": [
                            "196106377"
                        ],
                        "id": "86222bf0-c4f9-4fc4-a0ff-564350d52b77",
                        "story_id": "6c190a11-a899-4a55-a7ec-ef050578bed7",
                        "title": "китайские ученые разработали белок для защиты нейр
онов при инсульте"
                    }
                ],
                "title": "в бельцах пройдёт день донора крови: где, когда и кто може
т им стать?",
                "topics": [
                    [
                        "наука_и_техника/другое",
                        0.6886200904846191
                    ],
                    [
                        "наука_и_техника/школа, университеты",
                        0.48810485005378723
                    ],
                    [
                        "общество/здравоохранение",
                        0.41101008653640747
                    ]
                ]
            },
            "success": true
        }
```


Параметры:

data – словарь результатов

date – дата сюжета

id – id сюжета

news – массив словарей новостей в сюжете

- date – дата новости
- docs – документы новости
- id – id новости
- story_id – id сюжета уровнем выше
- title – заголовок новости

title – заголовок сюжета

topics – тематическое распределение по рубрикам/подрубрикам

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Автокомплит новости”

Данная конечная точка возвращает ближайшие новости к новости, поданной на вход с учетом контекста

Конечная точка расположена по адресу http://сервер:порт/assessor/autocomplete_news.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "context": "",
    "news_id": "38041b0c-ac2a-40dc-9fbd-6215102c3906",
    "day": 10000,
    "limit": 1
}
```

Параметры:

context – строка по которой искать совпадение (необязательное)

news_id – id новости

day – количество дней для поиска

limit - количество документов для ответа

Выходные данные – JSON формата:

```
{
    "data": [
        {
            "id": "b7c5eabd-89c5-4bd1-95c6-3a3bf958a516",
            "title": "евгений ощуев"
        }
    ],
    "success": true
}
```

Параметры:

data – массив результатов запроса

id – id новости

title – заголовок новости

success - параметр успеха выполнения запроса (true/false)

“TM Assessor” - “Автокомплит сюжета”

Данная конечная точка возвращает ближайшие сюжеты к сюжету, поданному на вход с учетом контекста.

Конечная точка расположена по адресу http://сервер:порт/assessor/autocomplete_story.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "context": "",
    "story_id": "38041b0c-ac2a-40dc-9fbd-6215102c3906",
    "day": 10000,
    "limit": 1
}
```

context – строка по которой искать совпадение (необязательное)

story_id - id сюжета

day - количество дней для поиска

limit – количество документов для ответа

Выходные данные – JSON формата:

```
{
    "data": [
        {
            "id": "35d1309a-ddc9-47ab-ad90-109c819c8110",
            "title": "вогл - ввод сдвигается на 2022 год"
        }
    ],
    "success": true
}
```

Параметры:

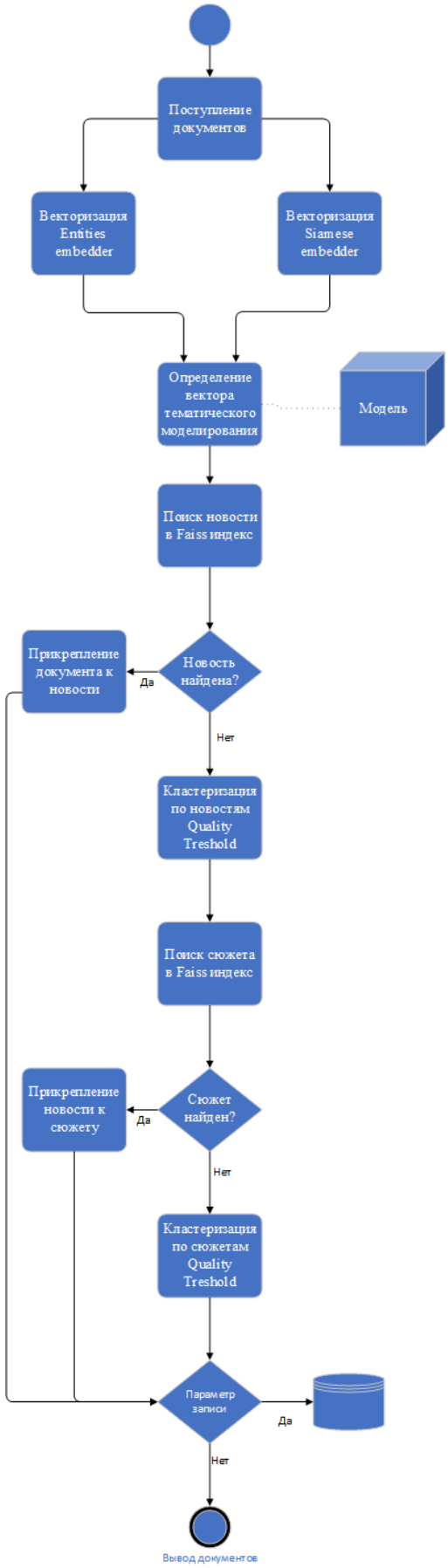
data – массив результатов запроса

id – id сюжета

title – заголовок сюжета

success - параметр успеха выполнения запроса (true/false)

3.4.5.1.Процесс работы сервиса


Поступление документов
Векторизация Entities embedder
Векторизация Siamese embedder
Определение вектора тематического моделирования
Модель
Поиск новости в Faiss индекс
Новость найдена?
Да
Прикрепление документа к новости
Нет
Кластеризация по новостям Quality Treshold
Поиск сюжета в Faiss индекс
Сюжет найден?
Да
Прикрепление новости к сюжету
Нет
Кластеризация по сюжетам Quality Treshold
Параметр записи
Да
Нет
Вывод документов

3.4.6. Сервис краулинга сети Интернет

3.4.6.1. Общее описание сервиса

Сервис краулинга сети Интернет позволяет производить обнаружение и сбор поисковым роботом новых медиаматериалов в сети для дальнейшей внутренней обработки другими сервисами, а именно:

1. Мониторинг Web-ресурсов:

Обеспечивает автоматизированный мониторинг различных внешних информационных ресурсов, в том числе RSS-лент, в том числе поддерживает:

- **Детекцию изменений верстки источника**.

Предложен и реализован метод и подход по своевременной реакции на обнаружение факта искажения верстки источника на основании метрики «Время обнаружения инцидента».

Определено значение метрики "Время обнаружения инцидента":

«Время обнаружения инцидента» соответствует значению текущего времени по TimeZone Msk, если время последней публикации (URL) этого источника, имеющейся в базе данных, превышает интервал обхода источника парсером.

Метод и подход по своевременной реакции на обнаружение факта искажения верстки источника:

В случае ошибки парсинга страницы из-за некорректной верстки время последней публикации в базе данных, полученной из данного источника рассчитывается по формуле:

Если Текущее время (ТВ) - ВП (время последней записи URL этого источника, имеющейся в базе данных) > ИП (интервала обхода источника парсером), то

Направить сообщение об инциденте во внешнюю треккер-систему.

Поля сообщения об обнаружении инцидента:

1) URL
2) Время обнаружения инцидента

3) Причина возникновения

- **RSS-технологию сбора**.
- **Сбор ссылок на источник информации**.

Парсер обладает возможностью обнаружения ссылки на источник информации в публикации.

В случае обнаружения новых источников в результате обработки публикации, происходит автоматическое создание задач на подключение этих источников.

- **Определение локации источника**.

Реализован функционал определения локации источника: ручной, GeoIP-заголовки страницы (TimeZone).

Реализована возможность автоматически определять тип источника (магазин, новостной ресурс и т.д.) при помощи обученного алгоритма классификации интернет-сайтов.

2. Мониторинг социальных сетей:

Сбор данных из социальных медиа в соответствии с требованиями технического задания на базе общего функционала краулинга:

1. Создание новых источников данных для парсинга
2. Удаление неиспользуемых источников
3. Активация и деактивирование источников
4. Настройка интервала обхода каждого из источников
5. Настройка спецификации для разметки целевой страницы источника на внутреннюю структуру контента системы (если применимо для данного типа источника) при помощи css и xpath -локаторов
6. Добавление разметки полей автора и просмотров для веб-страниц (если имеются)
7. Фильтрация полученного списка ссылок с источника при помощи регулярных выражений (regex)
8. Автоматическое формирование сообщений об ошибках парсинга источника
9. Автоматическое создание задач на подключение новых источников (в случае обнаружения)

10. Набор воркеров для парсинга источника на предмет обнаружения нового контента

11. Парсинг нового и обновленного контента и сохранение во внутреннем хранилище.

3. Микросервис регионов

Микросервис предназначен для предсказания вероятностного распределения по регионам для требуемого источника. Базовые данные для списка регионов и источников, по которым осуществляется предсказание, предоставляются открытым интернет ресурсом www.liveinternet.ru. При расчете вероятностного распределения для источников, отсутствующих в наборе базовых данных, микросервис отдает усредненное значение распределения по всем регионам в разрезе всех имеющихся источников.

4. Микросервис определения источников

Микросервис предназначен для выявления адресов источников информации публикаций на анализируемом ресурсе. Для работы данного микросервиса на его вход подается анализируемый вероятный источник и окружающие его слова. На выходе сервис отдает вероятность, на сколько поданный на вход источник может являться источником публикации на ресурсе. Микросервис использует модель Fasttext.

3.4.6.2.Конечные точки сервиса

Установка статуса краулера

Конечная точка расположена по адресу http://сервер:порт/api/v1/contents/status.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "isHealthy": true
}
```

Выходные данные:

201 Created

Проверка статуса краулера

Конечная точка расположена по адресу http://сервер:порт/api/v1/contents/status.

Метод запроса на сервер - GET.

Входные данные не требуются.

Выходные данные:


```
{
  "isHealthy": true,
  "metadata": {}
}
```


Получение списка всех источников

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/dataSources.

Возможно внести дополнительные параметры фильтрации добавив к адресу:

?sort={{fieldname}}

&page=1

&pageSize=10

&filters={

search={{datasourceitem.title search string}},

isAutoCreated=true

status=true,

types: ["messenges/telegram", "socialnetworks/facebook", "socialnetworks/vk", "web", "rss"]

}

Метод запроса на сервер - GET.

Выходные данные – JSON формата:


```
{
    "dataSourceItems": [
        {
            "uid": "6e1eef10-eaf6-41a5-a4b8-8d8f9f958dbb",
            "title": "",
            "description": "Рязань",
            "isActive": false,
```

```
            "updateInterval": 3600,
            "items": [
                {
                    "uid": "6f07e583-9102-4f33-a1bb-71c4d29c2da0",
                    "url": "https://vk.com/id254792090",
                    "type": "socialnetworks/vk",
                    "isActive": false,
                    "updateInterval": 3600
                }
            ]
        }
    ],
    "total": 18997
}
```


Получение источника по Id

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/dataSources/{{dataSourceUid}}.

Метод запроса на сервер - GET.

Выходные данные – JSON формата:


```
{
  "statusCode": 200,
  "description": "Success",
  "data": {
    "uid": "4dabc71e-83ca-49eb-ab51-f94842860593",
    "created": "2021-10-29T09:57:49.498585+00:00",
    "modified": "2021-12-01T00:28:11.028581+00:00",
    "title": "05TV ONLINE",
    "description": "",
    "updateInterval": 3600,
    "isActive": true,
    "isQuasy": false,
    "iconUri": "",
    "timezone": "Europe/Moscow",
    "dataSourceItems": [
      {
        "uid": "b7993f1c-b914-42d5-bdd0-a88c4629f29f",
        "created": "2021-10-29T09:57:49.804692+00:00",
        "modified": "2021-12-18T19:08:34.909769+00:00",
        "title": "05TV ONLINE",
        "description": "",
        "iconUri": "",
        "type": "socialnetworks/vk",
        "url": "https://vk.com/club206931290",
```

```
        "updateInterval": 300,
        "dataSource": {
          "uid": "4dabc71e-83ca-49eb-ab51-f94842860593"
        },
        "isActive": false,
        "isAutoCreated": true,
        "config": {},
        "account": {},
        "timezone": "Europe/Moscow",
        "updated": "2021-12-01T03:35:54.008906+00:00"
      }
    ]
  }
}
```


Создание нового источника

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/dataSources.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
  "dataSource": {
    "uid": "{{uuidDS}}",
    "title": "Банки.ру",
    "description": "Финансовые и банковские новости России и мира сегодня. Новости дня и онлайн-архив новостей прошлой недели",
    "iconUri": "https://www.banki.ru/static/favicons/favicon.ico",
    "updateInterval": 300,
    "isActive": true,
    "isQuasy":false,
    "timezone": "UTC"
  },
  "dataSourceItems": [
    {
      "uid": "{{uuidDSItem}}",
      "url": "https://www.facebook.com/dmitry.bocharov.967",
      "title": "DataSource Item Title",
      "account": {
        "username": "Account username",
        "password": "Account password"
      },
      "description": "DataSource item description",
      "modified": "2021-12-02T13:25:05Z",
      "config": {
        "mapping":{
```

```
            "title":{
              "area":[
                {
                  "method":"xpath",
                  "query":"//h1"
                }
              ],
              "exclude":[]
            },
            "content":{
              "area":[
                {
                  "method":"xpath",
                   "query":"//article[@class=\"article-text plain-text markup-
inside lenta\"]"
                }
              ],
              "exclude":[]
            },
            "publishedDateTime":{
              "area":[
                {
                  "method":"xpath",
                  "query":"//div[@class=\"news__info\"]/span/time"
                }
              ],
              "exclude":[]
            },
            "author":{
              "area":[
                {
                  "method":"xpath",
                       "query":"//div[@class=\"widget margin-top-small margin-
bottom-default\"]/a/text()"
                }
              ],
              "exclude":[]
            },
            "originSource":{
              "area":[
                {
                  "method":"xpath",
                  "query":"//div[@class=\"margin-top-x-small\"]/a"
                }
              ],
              "exclude":[]
            },
            "previewImage":{
```

```
            "area":[
              {
                "method":"xpath",
                "query":"//div[@class=\"preview\"]/img/@src"
              }
            ],
            "exclude":[]
          },
          "monitoring_area":{
            "area":[
              {
                "method":"xpath",
                     "query":"//div[@class=\"text-list text-list--date text-
list--date-inline\"]"
              }
            ],
            "exclude":[],
            "links_filter":[]
          }
        }
      },
      "dataSource": {"uid": "{{uuidDS}}"},
      "timezone": "Europe/Moscow",
      "iconUri": "http://datasource.com/favicon.ico",
      "type": "socialnetworks/facebook",
      "updateInterval": 300,
      "isActive": 1,
      "isAutoCreated": 0
    },
    ...
  ]
}
```

Выходные данные – JSON формата:

```
{
    "statusCode": 201,
    "description": "Created",
    "data": {
        "dataSource": {
            "uid": "02733573-f784-43d0-95fc-1ac440c9d4db",
            "created": "2021-12-18T13:46:41.673633+00:00",
            "modified": "2021-12-18T13:46:41.673634+00:00",
            "title": "Лента новостей — новости России и мира. Финансовый взг
ляд. | Банки.ру",
            "description": "Финансовые и банковские новости России и мира се
годня. Новости дня и онлайн-архив новостей прошлой недели.",
            "updateInterval": 300,
```

```
            "isActive": true,
            "isQuasy": false,
            "iconUri": "https://www.banki.ru/static/favicons/favicon.ico",
            "timezone": "UTC",
            "dataSourceItems": [
                {
                    "uid": "c7c0755b-0a4d-4586-b674-5394748aeb0a",
                    "created": "2021-12-18T13:46:41.675398+00:00",
                    "modified": "2021-12-18T13:46:41.675399+00:00",
                    "title": "Лента новостей — новости России и мира. Финанс
овый взгляд. | Банки.ру",
                    "description": "Финансовые и банковские новости России и
 мира сегодня. Новости дня и онлайн-архив новостей прошлой недели.",
                    "iconUri": "",
                    "type": "web",
                    "url": "https://www.banki.ru/news/lenta/",
                    "updateInterval": 300,
                    "dataSource": {
                        "uid": "02733573-f784-43d0-95fc-1ac440c9d4db"
                    },
                    "isActive": true,
                    "isAutoCreated": true,
                    "config": {
                        "monitoring_area": {
                            "area": [
                                {
                                    "query": "//div[@class=\"text-list text-
list--date text-list--date-inline\"]",
                                    "method": "xpath"
                                }
                            ],
                            "exclude": [],
                            "links_filter": []
                        },
                        "mapping": {
                            "title": {
                                "area": [
                                    {
                                        "query": "//h1",
                                        "method": "xpath"
                                    }
                                ],
                                "exclude": []
                            },
                            "content": {
                                "area": [
                                    {
```

```
                                                                "query": "//
        article[@class=\"article-text plain-text markup-inside lenta\"]",
                                                "method": "xpath"
                                            }
                                        ],
                                        "exclude": []
                                    },
                                    "publishedDateTime": {
                                        "area": [
                                            {
                                                                "query": "//
        div[@class=\"news__info\"]/span/time",
                                                "method": "xpath"
                                            }
                                        ],
                                        "exclude": []
                                    },
                                    "author": {
                                        "area": [
                                            {
                                                                "query": "//
        div[@class=\"widget margin-top-small margin-bottom-default\"]/a/text()",
                                                "method": "xpath"
                                            }
                                        ],
                                        "exclude": []
                                    }
                                }
                            },
                            "account": {},
                            "timezone": "Europe/Moscow",
                            "updated": "2021-12-18T13:46:41.675414+00:00"
                        }
                    ]
                }
            }
        }
```


Активация источника

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/dataSources/{{DSUid}}/dataSourceItems/{{DSItemUid}}

Пример: http://сервер:порт/api/v1/importer/dataSources/fdc4f344-5bb2-4fbc-ba03-d32e98c231fa/dataSourceItems/cead3fc2-6b53-4a45-83c9-eb16c1a6add4

Метод запроса на сервер - PATCH.

Входные данные – JSON формата:

```
{
"updateInterval":300,
"isActive":true
}
```

Выходные данные:

200 OK

Деактивация источника

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/dataSources/{{DSUid}}/dataSourceItems/{{DSItemUid}}

Пример: http://сервер:порт/api/v1/importer/dataSources/fdc4f344-5bb2-4fbc-ba03-d32e98c231fa/dataSourceItems/cead3fc2-6b53-4a45-83c9-eb16c1a6add4

Метод запроса на сервер - PATCH.

Входные данные – JSON формата:

```
{
"updateInterval":300,
"isActive":false
}
```

Выходные данные:

200 OK

Получение локации при помощи GeoIP

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/geoIP

Метод запроса на сервер – POST.

Входные данные – JSON формата:

```
{
  "urls": [
    "https://vk.com",
    "https://medium.com"
  ]
}
```

Выходные данные – JSON формата:

```
[
  {
    "ip": "87.240.190.72",
    "country_code2": "RU",
    "country_code3": "RUS",
    "country_name": "Russia",
```

```
    "state_prov": "Central Federal District",
    "district": "Kontakt SNT",
    "city": "Naro-Fominskiy Rayon",
    "zipcode": "143345",
    "latitude": "55.48913",
    "longitude": "37.01600",
    "timezone": "Europe/Moscow"
  },
  {
    "ip": "162.159.153.4",
    "country_code2": "US",
    "country_code3": "USA",
    "country_name": "United States",
    "state_prov": "California",
    "district": "China Basin",
    "city": "San Francisco",
    "zipcode": "94107-1907",
    "latitude": "37.78035",
    "longitude": "-122.39059",
    "timezone": "America/Los_Angeles"
  }
]
```


Инициация автосоздания источников

Конечная точка расположена по адресу http://сервер:порт/api/v1/ DataSources/ autocreate

Метод запроса на сервер – POST.

Входные данные – JSON формата:


```
{
    "links": [
        "https://vk.com/club123456",
        "https://fb.com/somelink",
        "https://fb.com/somel2ink"
    ]
}
```


Выходные данные:

202 Accepted

Получить классификацию по домену

Конечная точка расположена по адресу http://сервер:порт/api/v1/ importer/ domain2region?search={{адрес домена}}

Пример:

http://сервер:порт/api/v1/importer/domain2region?search=https://uslugi.mosreg.ru/services/6909

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{
  "success": "true",
  "data": {
    "Hi-Tech": 0.05246099052020236,
    "Авто": 0.026838668946752478,
    "Бизнес": 0.2531471557191488,
    "Дом": 0.07855449806603869,
    "Культура": 0.05363946103792374,
    "Общество": 0.263755537418605,
    "Отдых": 0.05305846776284193,
    "Развлечения": 0.025298705711442198,
    "СМИ": 0.04022389962338157,
    "Спорт": 0.025782679198328856,
    "Справки": 0.03663317533224138,
    "Учёба": 0.09060676066309288
  }
}
```


Получение мета-информации с источника

Конечная точка расположена по адресу http://сервер:порт/api/v1/importer/discover

Пример: http://сервер:порт/api/v1/importer/discover?uri=https://ya.ru

Метод запроса на сервер - GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{"id": "51ad33b2-e6f2-45fe-9e29-b8ef4c7e94dd",
"name": "web",
"metadata":
{"title": "\u042f\u043d\u0434\u0435\u043a\u0441",
"description": "Яндекс - поисковая система и интернет-
портал. Поиск по интернету и другие сервисы: карты и навигатор, транспорт и
такси, погода, новости, музыка, телепрограмма, переводчик, покупки в интерне
те. Бесплатная электронная почта и облачное хранилище. Найдется всё!",
```

```
"iconUri": "http://yastatic.net/s3/home-static/_/a6/
a6a296b741b51880ae8a9b04a67cfe3f.png",
"links": []}}
```


Получить классификацию по домену

Конечная точка расположена по адресу http://сервер:порт/regions/domain2region

Метод запроса на сервер – GET.

Входные данные – JSON формата:

```
[
    "lenta.ru",
    "rbc.ru"
]
```

Массив доменов в двойных кавычках, разделенные запятой.

Выходные данные – JSON формата:


```
{
    "success": "true",
    "data": [
        {
            "domain": "lenta.ru",
            "region": {
                "Москва": 0.53,
                "Башкортостан": 0.02,
                "Орловская область": 0.0,
                "Московская область": 0.0,
                "Кемеровская область": 0.0,
                "Иркутская область": 0.0,
                "Оренбургская область": 0.0,
                "Липецкая область": 0.0,
                "Калининградская область": 0.0,
                "Саратовская область": 0.0,
                "Ульяновская область": 0.0,
                "Санкт-Петербург": 0.11,
                "Томская область": 0.0,
                "Новгородская область": 0.0,
                "Тверская область": 0.0,
                "Псковская область": 0.0,
                "Вологодская область": 0.0,
                "Омская область": 0.0,
                "Тульская область": 0.0,
                "Рязанская область": 0.0,
                "Воронежская область": 0.02,
                "Ивановская область": 0.0,
                "Удмуртия": 0.0,
                "Самарская область": 0.04,
                "Тюменская область": 0.0,
                "Калмыкия": 0.0,
                "Татарстан": 0.02,
```

```
"Северная Осетия — Алания": 0.0,
"Ярославская область": 0.0,
"Сахалинская область": 0.0,
"Ленинградская область": 0.0,
"Нижегородская область": 0.03,
"Магаданская область": 0.0,
"Волгоградская область": 0.0,
"Кировская область": 0.0,
"Смоленская область": 0.0,
"Ростовская область": 0.02,
"Бурятия": 0.0,
"Свердловская область": 0.04,
"Амурская область": 0.0,
"Владимирская область": 0.0,
"Калужская область": 0.0,
"Алтайский край": 0.0,
"Пензенская область": 0.0,
"Челябинская область": 0.01,
"Белгородская область": 0.0,
"Новосибирская область": 0.08,
"Брянская область": 0.0,
"Якутия": 0.02,
"Костромская область": 0.0,
"Курганская область": 0.0,
"Курская область": 0.0,
"Карачаево-Черкесия": 0.0,
"Коми": 0.0,
"Красноярский край": 0.0,
"Краснодарский край": 0.05,
"Ямало-Ненецкий АО": 0.0,
"Архангельская область": 0.0,
"Тыва": 0.0,
"Хабаровский край": 0.0,
"Ставропольский край": 0.0,
"Дагестан": 0.0,
"Ханты-Мансийский АО — Югра": 0.0,
"Мурманская область": 0.0,
"Приморский край": 0.01,
"Чечня": 0.0,
"Чувашия": 0.0,
"Хакасия": 0.0,
"Тамбовская область": 0.0,
"Астраханская область": 0.0,
"Пермский край": 0.0,
"Чукотский АО": 0.0,
"Забайкальский край": 0.0,
"Камчатский край": 0.0,
"Мордовия": 0.0,
"Ингушетия": 0.0,
"Кабардино-Балкария": 0.0,
"Карелия": 0.0,
"Адыгея": 0.0,
"Марий Эл": 0.0,
```

```
            "Алтай": 0.0,
            "Ненецкий АО": 0.0,
            "Еврейская АО": 0.0
        }
    },
    {
        "domain": "rbc.ru",
        "region": {
            "Москва": 0.53,
            "Башкортостан": 0.0,
            "Орловская область": 0.0,
            "Московская область": 0.0,
            "Кемеровская область": 0.0,
            "Иркутская область": 0.0,
            "Оренбургская область": 0.0,
            "Липецкая область": 0.0,
            "Калининградская область": 0.0,
            "Саратовская область": 0.0,
            "Ульяновская область": 0.0,
            "Санкт-Петербург": 0.13,
            "Томская область": 0.0,
            "Новгородская область": 0.0,
            "Тверская область": 0.0,
            "Псковская область": 0.0,
            "Вологодская область": 0.0,
            "Омская область": 0.0,
            "Тульская область": 0.0,
            "Рязанская область": 0.0,
            "Воронежская область": 0.01,
            "Ивановская область": 0.0,
            "Удмуртия": 0.0,
            "Самарская область": 0.03,
            "Тюменская область": 0.0,
            "Калмыкия": 0.0,
            "Татарстан": 0.02,
            "Северная Осетия — Алания": 0.0,
            "Ярославская область": 0.0,
            "Сахалинская область": 0.0,
            "Ленинградская область": 0.0,
            "Нижегородская область": 0.03,
            "Магаданская область": 0.0,
            "Волгоградская область": 0.0,
            "Кировская область": 0.0,
            "Смоленская область": 0.0,
            "Ростовская область": 0.03,
            "Бурятия": 0.0,
            "Свердловская область": 0.04,
            "Амурская область": 0.0,
            "Владимирская область": 0.0,
            "Калужская область": 0.0,
            "Алтайский край": 0.0,
            "Пензенская область": 0.0,
            "Челябинская область": 0.03,
```

```
                "Белгородская область": 0.0,
                "Новосибирская область": 0.06,
                "Брянская область": 0.0,
                "Якутия": 0.03,
                "Костромская область": 0.0,
                "Курганская область": 0.0,
                "Курская область": 0.0,
                "Карачаево-Черкесия": 0.0,
                "Коми": 0.0,
                "Красноярский край": 0.0,
                "Краснодарский край": 0.04,
                "Ямало-Ненецкий АО": 0.0,
                "Архангельская область": 0.0,
                "Тыва": 0.0,
                "Хабаровский край": 0.0,
                "Ставропольский край": 0.0,
                "Дагестан": 0.0,
                "Ханты-Мансийский АО — Югра": 0.0,
                "Мурманская область": 0.0,
                "Приморский край": 0.01,
                "Чечня": 0.0,
                "Чувашия": 0.0,
                "Хакасия": 0.0,
                "Тамбовская область": 0.0,
                "Астраханская область": 0.0,
                "Пермский край": 0.0,
                "Чукотский АО": 0.0,
                "Забайкальский край": 0.0,
                "Камчатский край": 0.0,
                "Мордовия": 0.0,
                "Ингушетия": 0.0,
                "Кабардино-Балкария": 0.0,
                "Карелия": 0.0,
                "Адыгея": 0.0,
                "Марий Эл": 0.0,
                "Алтай": 0.0,
                "Ненецкий АО": 0.0,
                "Еврейская АО": 0.0
            }
        }
    ]
}
```


Параметры:

В ответе словарь, в котором:

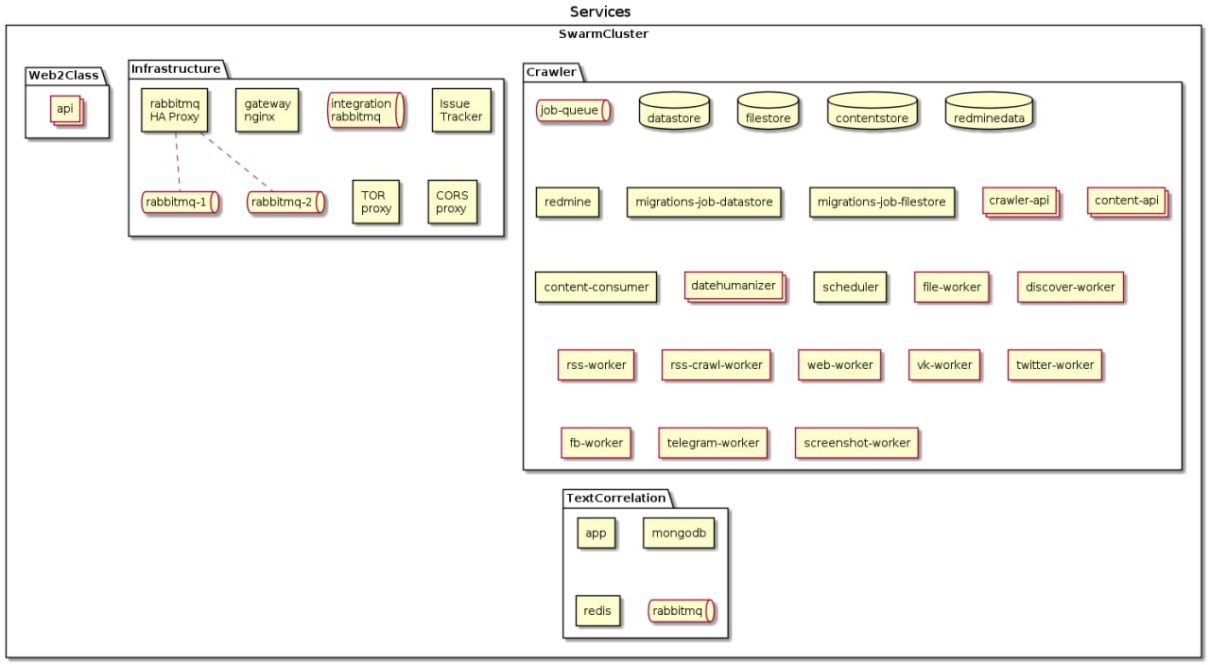
success – флаг успешности операции (true/false)

data – словарь с результатами обработки, в котором:

domain – домен, для которого производится определение распределения

region – словарь с распределением по регионам.

3.4.6.3.Архитектура сервиса



3.4.7. Сервис выявления трендов

3.4.7.1.Общее описание сервиса

Сервис выявления трендов предназначен для предсказания направления изменений временного ряда в будущем. В сервисе предусмотрена возможность настройки количества дня для предсказания. По умолчанию сервис предсказывает развитие на 7 дней вперед.

3.4.7.2.Конечные точки сервиса

Предсказать временной ряд

Конечная точка расположена по адресу http://сервер:порт/predict_time_series.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "data": [0.06492638, 0.06467812, 0.05656281, 0.05668245, 0.04613852,
       0.06164145, 0.05853276, 0.06216442, 0.05986168, 0.06095049,
       0.06279789, 0.07829134, 0.07745358, 0.08413915, 0.07881835,
       0.08337046, 0.08840743, 0.10130771, 0.1030707, 0.10304379,
       0.10009686]
}
```

Параметры:

data – временной ряд для предсказания

Выходные данные – JSON формата:


```
{
    "success": true,
    "data": [
        0.09964405745267868,
        0.10001574456691742,
        0.10711276531219482,
        0.11600495874881744,
        0.12166222929954529,
        0.12320959568023682,
        0.12031629681587219
    ]
}
```


success – флаг успешности операции (true/false)

data – предсказанный дальнейший временной ряд

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Ответ содержит JSON формата:


```
{
    "success": true,
    "data": {
        "score": 0.0014552619541063905,
        "metric": "mse_loss"
    }
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

score – значение метрики

metric – наименование метрики

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер – POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:

```
{
    "success": true,
    "message": "Обучение инициировано"
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/last_train_metric.

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:

```
{
    "success": true,
    "data": {
        "date": "2021-11-23 14:24:09.723804",
        "old_metric": "2.913808566518128e-05",
        "new_metric": "1.938150126079563e-05",
        "update_model": "True"
    }
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

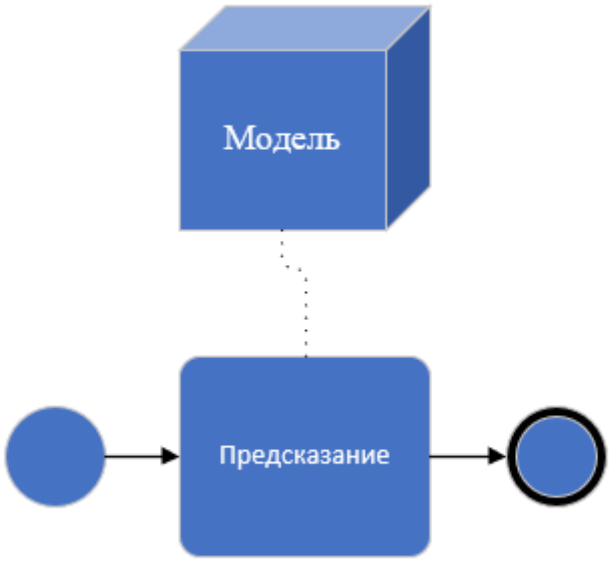
time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.7.3.Процесс работы сервиса



3.4.7.4.Модель

В качестве модели выбрана архитектура N-BEATS (архитектура прогнозирования временных рядов, разработанная Б.Орешкиным), которая использует простую, но мощную архитектуру ансамблевых сетей с прямой связью со сложенными остаточными блоками прогнозов и «ретроспективных прогнозов».

В качестве данных для предсказания используются значения отклонения трендов за последний 21 день.

Модель обучена на предсказание 7 дней вперед. При необходимости предсказания меньшего количества дней массив предсказанных значений обрезается под требования. При необходимости предсказания большего количества дней модель идет окном со смещением на 7 дней, используя при этом предсказанные значения на каждом шаге.

3.4.7.5.Метод оценки качества

Качество предсказание проверяется при помощи метрики MSE (п.3.3.2 данного документа).

3.4.8. Сервис поиска документов-источников

3.4.8.1.Общее описание сервиса

Сервис состоит из нескольких микросервисов:

1) Siamese embedder - сервис эмбеддер тектов, принимает на вход текст, возвращает векторные представления параграфов полученного текста.

2) Sparse index - спарсовая матрица hash значений нграмм/сущностей дкумента, предназначен для поиска локсографически близких к документу документов.

3) Faiss Index - индекс векторных представлений параграфов документов, предназначен для поиска семантически близких к документу документов.

4) Postgres - база данных сервиса. Хранит в себе информация о документах, векторных представлениях их параграфов, уникальные hash значения нграмм/ сущностей и их количественные значения. Из нее происходит пересборка индексов Faiss index и Sparse index.

5) Adopting Api - центральный микросервис, обеспечивающих взаимодействие между остальными сервисами.

6) Nginx - предназначен для балансировки/репликации запросов к Faiss index, Siamese embedder и Sparse index.

Описание микросервисов:

Siamese embedder

Микросервис представляется собой api который обеспечивает доступ к модели ембеддера основанной на BERT. Принимает на вход текст и идентификатор документа. На выходе возвращает векторное представление каждого параграфа текста, каждый вектор это одномерный массив float32 длинной 300.

Sparse index

Микросервис спарсовая матрица hash значений нграмм/сущностей дкумента. Поиск по индексу, обогащение и обносление происходит с помощью запросов из

Adopting Api. Возвращает список лексографически близких, к целевому документу, документов.

Faiss Index

Микросервис содержит индекс параграфов документов. Поиск по индексу, обогащение и обносление происходит с помощью запросов из Adopting Api. Возвращает список семантически близких, к целевому документу, документов.

Postgres

База данных сервиса. Хранит в себе информация о документах, векторных представлениях их параграфов, уникальные hash значения нграмм/сущностей и их количественные значения. Из нее происходит пересборка индексов Faiss index и Sparse index.

Adopting Api

Микросервис является связующим звеном для всех остальным микросервисов. Именно сюда приходит запрос на обработку документов, и отсюда же происходят запросы ко всем оставльным микросервисам. Запросы к макросервисам посылаются асинхронно и, благодаря микросервису Nginx достигается увеличение скорости работы алгоритма.

Nginx

Сервис предназначен для балансировки и репликации запросов при определенных обращениях к разным сервисам. При запросе на обновление индекса новостей/сюжетов микросервиса Faiss index/Sparse index запрос реплицируется, т.е. запрос отправляется одновременно на все реплики микросервиса. При запросе на поиск ближайших новостей/сюжетов запрос балансируется между всеми развернутыми репликами микросервиса Faiss index/Sparse index, т.е. микросервис Nginx каждый такой запрос отправляет на одну из реплик микросервиса Faiss index/Sparse index. В случае с запросами к Siamese embedder также происходит балансировка.

3.4.8.2.Конечные точки сервиса

Adopting model

Конечная точка расположена по адресу http://сервер:порт/api/model.

Метод запроса на сервер – POST.

Входные данные – JSON формата:

```
{
    "data": {
        "196106552": {
            "date": "2021-06-10 21:55:08",
            "link": "google.com",
            "description": "Актеру полушившему славу по роли Дина Винчестера
 в сериале \"Сверхъестественное\" достался образ супергероя Солдатика в трет
ьем сезоне проекта \"Пацаны\".
\n\r\n\n\r\nВ своем инстаграм Дженсон поделился первыми снимками в новом обр
азе:
\n\r\n\n\r\n\n\r\nПодписав фото: \"Мою идею сделать камуфляж Солдатика в вид
е банановой гамака отвергли. Но оно и к лучшему. Мне нравится мой костюм\".
\n\r\nОбраз Солдатика являет собой пародию на Капитана Америку. Фантастическ
ий сериал \"Пацаны\" -
 это новый взгляд на супергероев и их поведение, ведь оно же не всегда благо
родное?!",
            "entities": [
                "дженсон эклзти",
                "дин винчестер",
                "дженсон",
                "америка"
            ]
        },
        "196106554": {
            "date": "2021-06-10 21:55:08",
            "link": "google.com",
            "description": "Татьяна и Ольга Арнтгольц -
 русские сестры Олсен, одни из самых популярных сестер-близняшек в кино.
\n\r\n\n\r\nУспешные актрисы не очень то любят делиться личным, но, в честь
дня рождения великого поэта Александра Пушкина, Татьяна поделилась архивным
семейным фото и рассказала в честь кого им с сестрой дали имена.
\n\r\n\n\r\n\"Когда родители узнали, что родятся две девочки, то решили назв
ать нас в честь сестёр любимого ими романа А.С Пушкина \"Евгений Онегин\" -
 Татьяной и Ольгой. Нижайший поклон Вам, Александр Сергеевич, за каждую стро
чку и с Днем рождения!!!\" - написала актриса.",
            "entities": [
                "арнтгольц",
                "ольга арнтгольц",
                "олсный",
                "пушкин",
                "онегин",
                "татьяна",
                "ольга",
                "александр сергей"
            ]
        }
    }
}
```


Параметры:

data – словарь с данными для обработки

196106552, 196106554 – id документа (уникальный)

date – дата публикации

link – ссылка на публикацию

description – текст новости

entities – сущности из NER

Выходные данные – JSON формата:


```
{
    "data": {
        "196106552": {
            "adopted_from": [
                {
                    "delta": 39061.0,
                    "domain": "gameguru.ru",
                    "id": "01756f90-2c6b-4ac7-88f9-00f6a5c46ae2",
                    "ratio": 0.9001631736755371
                },
                {
                    "delta": 75968.0,
                    "domain": "vytegra.news",
                    "id": "196106554",
                    "ratio": 0.9104006290435791
                }
            ],
            "overall_ratio": 0.00023348120476301658,
            "plagiat_from": [
                {
                    "delta": 112258.0,
                    "domain": "pozitciya.com.ua",
                    "id": "f931ed2e-3ff4-4be9-a02a-701013a0c041",
                    "overall_ratio": 0.2701377539108102,
                    "ratio": 0.000864304235090752
                },
                {
                    "delta": 46785.0,
                    "domain": "shahta.org",
                    "id": "4879fe3a-d202-4090-9f9c-065683ce70cb",
                    "overall_ratio": 0.01634368433341116,
                    "ratio": 0.014285714285714285
                }
            ],
            "success": true
        },
        "196106554": {
            "adopted_from": [
                {
                    "delta": 103507.0,
                    "domain": "mirf.ru",
                    "id": "0a8fab2a-9307-44b3-bbf8-321f3057381a",
                    "ratio": 0.9014723300933838
```

```
                },
                {
                    "delta": 101058.0,
                    "domain": "eg.ru",
                    "id": "d8c93f1c-e065-4303-bed9-45e3ffbdda3f",
                    "ratio": 0.9063312411308289
                }
            ],
            "overall_ratio": 0.001722356183258698,
            "plagiat_from": [
                {
                    "delta": 104626.0,
                    "domain": "aif.ru",
                    "id": "5dca3525-011b-4bd2-acd8-f4c2894e599d",
                    "overall_ratio": 0.0923182914226662,
                    "ratio": 0.0037313432835820895
                },


                {
                    "delta": 59708.0,
                    "domain": "vg-news.ru",
                    "id": "40662947-590a-488c-9056-34c757b4737e",
                    "overall_ratio": 0.04822597313124354,
                    "ratio": 0.007142857142857143
                }
            ],
            "success": true
        }
    },
    "success": true
}
```


Параметры:

data – словарь с данными ответа сервиса

196106552, 196106554 – id документов (уникальный)

adopted_from – массив документов из которых произведено заимствование

delta - разница в секундах между публикацией

domain – домен, с которого произошло заимствование

id – id документа из которого произошло заимствование

ratio – процент заимствований/100

overall_ratio – процент плагиата для всех документов с текущим id

plagiat_from массив документов из которых произведен плагиат

delta – разница в секундах между публикацией

domain – домен, с которого произведен плагиат

id – id документа из которого произведен плагиат

overall_ratio – процент плагиата по всем документам для текущего id из п. выше

ratio – процент плагиата

success - параметр успеха выполнения запроса (true/false)

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – GET.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Выходные данные – JSON формата:

```
{
    "metrics": 0.9667
}
```

Параметры:

metrics – значение текущей метрики

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер – GET.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:

```
{
    "success": true,
    "message": "Обучение инициировано"
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/last_train_metric.

Файл расположен в корневой папке сервиса с именем «metrics_score_history.csv»

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:

```
{
    "success": true,
    "data": {
        "time": "2021-11-27 11:21:20.973078",
        "old_metric": 0.9667,
        "new_metric": 0.9667,
        "update_model": false
    }
}
```

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.8.3.Процесс работы сервиса

3.4.9. Сервис обнаружения цепочки наследований документа

3.4.9.1. Общее описание сервиса

Сервис обнаружения цепочки наследований документа выполняет задачу построения цепочек наследования для документов и выявления связей между документами.

3.4.9.2. Конечные точки сервиса

Цепочка наследования

Конечная точка расположена по адресу http://сервер:порт/graph/inheritance.

Метод запроса на сервер – POST.

Входные данные – JSON формата:

```
{
        "pub_date": "2021-10-16T22:11:32",
        "text": "Спортивный директор «Эвертона» Марсель Брандс прокомментиро
вал «СЭ» информацию об интересе английского клуба к российскому хавбеку «Мон
ако» Александру Головину. Ранее, 10 ноября, Liverpool Echo сообщил, что моне
гаски запросили у «Эвертона» 46 миллионов евро за 25-
летнего футболиста. «Информация, что «Монако» запросил у нас за Головина 46
миллионов евро — это абсолютная чушь», — сказал Брандс «СЭ». В текущем сезон
е Головин провел 16 матчей за «Монако», забил три гола и сделал три результа
тивные передачи. Портал Transfermarkt оценивает полузащитника в 28 миллионов
 евро. Контракт игрока с французской командой действует до конца июня 2024-
го."
}
```

Параметры:

pub_date – дата публикации новости;

text – текст новости.

Выходные данные – JSON формата:

```
{
    "data": {
        "items": [
            {
                "date": "2021-11-02",
                "id": "00ae31f3-f157-43f6-871a-2a2983e120c7",
                "name": "Ольга Котова",
                "source": "https://vk.com/wall55658306_52550",
                "traffic": 1
            },
            {
                "date": "2021-11-02",
```

```
                "id": "22c069aa-950d-4399-a816-1fd3bb24c6d8",
                "name": "Владимир Судилов",
                "source": "https://vk.com/wall327587355_19005",
                "traffic": 1
            }
        ],
        "links": [
            {
                "force": 0.9004267,
                "source": "485443e3-e440-4198-ab4d-aee15288784c",
                "target": "0e698361-2a82-4ff5-9457-15c0e6a5d493"
            },
            {
                "force": 0.901823,
                "source": "20491a36-3971-4b27-83a4-8caa893f3cb0",
                "target": "f2271315-702e-47eb-8df1-5b02aa26b7e8"
            }
        ],
        "sources": [
            {
                "name": "https://vk.com/wall-48356765_23756",
                "topic": {
                    "name": "общество"
                }
            },
            {
                "name": "https://vk.com/wall31378028_10403",
                "topic": {
                    "name": "наука_и_техника"
                }
            }
        ]
    },
    "success": true
}
```


Параметры:

data – словарь с данными

items - массив узлов графа

id - идентификатор документа

date - дата публикации документа

name - имя документа (заголовок, title), один из источников в sources

source - источник документа (домен)

traffic - траффик документа

links - массив связей между документами (узлами графа) - ребра графа

force - сила связи.

source - начальная вершина ребра

target - конечная вершина ребра

sources - массив источников (доменов)

name - имя источника

topic - топик источника (1го уровня, самый распространенный)

name - название топика

success - параметр успеха выполнения запроса (true/false)

3.4.9.3.Процесс работы сервиса

Процесс работы сервиса реализует вывод через конечные точки REST API результатов запросов к одной или нескольким БД системы.

3.4.10.Сервис определения путей распространения медиаматериала

3.4.10.1.Общее описание сервиса

Сервис распространения решает задачу моделирования распространения текста новости и лексики по графу издателей. Для сервиса доступны подготовленные модели для русского и английского языков.

Сервис формирует отчет, который моделирует распространение текста новости другими издателями. Отчет состоит из всех доменов издателей и времени распространения. На вход сервис принимает два домена и текст распространения, на выходе - свойства связи между двумя источниками, отражающие спецификацию признаков (лексика, темы и т.д.), характерных для данной связи.

3.4.10.2.Конечные точки сервиса

Запуск путей распространения медиаматериала

Конечная точка расположена по адресу http://сервер:порт/predict_distribution

Метод запроса на сервер – POST.

Входные данные – JSON формата:

```
{
        "source": "politros.com",
        "doc": "Что такое искусство живописи? Живопись –
 это вид изобразительного искусства, когда художник передает на холсте разли
```

```
чные образы с помощью ярких, разноцветных красок. Иногда эти образы на карти
нах бывают абсолютно четкими и ясными –
 красивый пейзаж, который позволяет зрителю перенестись в пространстве, порт
рет красивой девушки, натюрморт из различных цветов или предметов.",
        "topic_vector": [
        0.02878, 0.00193, 0.05107, 0.04142, 0.04555, 0.07284, 0.05475, 0.030
08, 0.03594, 0.19542, 0.02752, 0.04107,
        0.10873, 0.08427, 0.17154, 0.01106, 0.0598, 0.0175, 0.02363, 0.05864
, 0.02715, 0.15638, 0.03024,
        0.01971, 0.26024, 0.06841, 0.04656, 0.10368, 0.09647, 0.01731, 0.068
34, 0.05883, 0.49138, 0.32978, 0.08802,
        0.00695, 0.05001, 0.12073, 0.15687, 0.13756, 0.45444, 0.19302, 0.052
29, 0.05128, 0.01512, 0.04615,
        0.0312, 0.03497, 0.0483, 0.04992, 0.02532, 0.10955, 0.18278, 0.06962
, 0.00972, 0.00738, 0.03814, 0.00885,
        0.0165, 0.04813, 0.01451, 0.12169, 0.00202, 0.01801, 0.02341, 0.0513
4, 0.03142, 0.06254
    ]
    }
```


Параметры:

source – домен источника

doc – текст документа

topic_vector – вектор топиков документа (результат обработки сервисом тематического моделирования)

Выходные данные – JSON формата:


```
{
    "uid": "c7472c94-a638-42fd-9f0e-dd1c4469a7ee",
    "completed_status": false,
    "result": []
}
```


Параметры:

uid – id запроса

completed_status – статус выполнения запроса

result – массив с результатом запроса

Получение пути распространения медиаматериала

Конечная точка расположена по адресу http://сервер:порт/predict_distribution/ *{{uid_запроса}}*.

Uid запроса – id запроса, полученный в ответе при запуске конечной точки «Запуск путей распространения медиаматериала»

Пример запроса – http://сервер:порт/predict_distribution/c7472c94-a638-42fd-9f0e-dd1c4469a7ee

Метод запроса на сервер – GET.

Выходные данные – JSON формата:


```
{
    "uid": "c7472c94-a638-42fd-9f0e-dd1c4469a7ee",
    "completed_status": true,
    "result": [
        [
            "kazanfirst.ru",
            {
                "proba": 0.3564540147781372,
                "time": 7.783333333333333
            }
        ],
        [
            "tvzvezda.ru",
            {
                "proba": 0.6482049226760864,
                "time": 1.1766666666666667
            }
        ],
    ]
}
```


Параметры:

uid – id запроса

completed_status – статус выполнения запроса

result – массив массивов с результатами предсказаний

массив результата состоит из:

- домен
- словарь с данными о вероятности и времени распространения

Моделирование распространения группы документов

Конечная точка расположена по адресу http://сервер:порт/predict_distribution_batch

Метод запроса на сервер – POST.

Входные данные – JSON формата:


```
{
    "data": {
        "key1": {
```

```
            "doc": "Президент Украины Владимир Зеленский после инаугурации в
 понедельник, 20 мая, провел первые переговоры в качестве главы государства,
 обсудив партнерство с президентом Грузии Саломе Зурабишвили. Об этом сообща
ет пресс-
служба Администрации президента. Зурабишвили поздравила Зеленского с победой
 и пожелала успехов на посту президента Украины. Она заявила о готовности да
льнейшего развития украинско-
грузинских отношений и углублении стратегического партнерства. Зурабишвили п
ланирует объединиться с Украиной для курса на евроинтеграцию и членство в НА
ТО. Всех нас ждет еще один вызов –
 выборы в Европарламент. Мы должны объединить позиции Украины и Грузии относ
ительно курса на евроинтеграцию и членство в НАТО, - сказала Зурабишвили.",
            "topic_vector": [
                0.02878, 0.00193, 0.05107, 0.04142, 0.04555, 0.07284, 0.0547
5, 0.03008, 0.03594, 0.19542, 0.02752,
                0.04107,
                0.10873, 0.08427, 0.17154, 0.01106, 0.0598, 0.0175, 0.02363,
 0.05864, 0.02715, 0.15638, 0.03024,
                0.01971, 0.26024, 0.06841, 0.04656, 0.10368, 0.09647, 0.0173
1, 0.06834, 0.05883, 0.49138, 0.32978,
                0.08802,
                0.00695, 0.05001, 0.12073, 0.15687, 0.13756, 0.45444, 0.1930
2, 0.05229, 0.05128, 0.01512, 0.04615,
                0.0312, 0.03497, 0.0483, 0.04992, 0.02532, 0.10955, 0.18278,
 0.06962, 0.00972, 0.00738, 0.03814,
                0.00885,
                0.0165, 0.04813, 0.01451, 0.12169, 0.00202, 0.01801, 0.02341
, 0.05134, 0.03142, 0.06254
            ],
            "entities": [
                "Энтони Блинкен", "МИД", "РФ", "Сергеем Лавровым", "поток",
"Госдепе", "США",
                "МИД", "Арктического совета", "Госдепартамента", "Данию", "Б
инкеном", "Лаврову",
                "Госдепа", "США", "Россией", "Лавровым", "Энтони Блинкен", "
МИД", "РФ",
                "Сергеем Лавровым", "Госдепе", "США", "МИД", "Арктического с
овета", "Данию",
                "Бинкеном", "Лаврову", "США", "Россией", "Лавровым"
            ],
            "nouns": [
                "Лавровым", "госсекретаря", "представитель", "Сергеем", "пот
ок",
                "Госдепартамента", "Россией", "Госдепе", "Комментарий", "США
",
                "Встреча", "Лаврову", "Энтони", "Госдепа", "дипломат", "Блин
кен",
                "Бинкеном", "РФ", "встрече", "совета", "конгресс", "Госсекре
тарь",
                "переговорах", "Данию", "отношение", "ранга", "МИД", "отноше
ниям",
                "вопросы", "потоку", "сессии", "Представитель", "полях", "Эн
тони Блинкен",
```

```
                "МИД", "РФ", "Сергеем Лавровым", "поток", "Госдепе", "США",
"МИД",
                "Арктического совета", "Госдепартамента", "Данию", "Бинкеном
", "Лаврову",
                "Госдепа", "США", "Россией", "Лавровым", "Энтони Блинкен", "
МИД", "РФ",
                "Сергеем Лавровым", "Госдепе", "США", "МИД", "Арктического с
овета", "Данию",
                "Бинкеном", "Лаврову", "США", "Россией", "Лавровым"
            ],
            "topics_with_probs": {
                "бизнес": 0.00334,
                "культура": 0.026449999999999998,
                "наука и техника": 0.024709999999999996,
                "общество": 0.02346,
                "политика": 0.83267,
                "проишествия": 0.011740000000000002,
                "силовые структуры": 0.01228,
                "спорт": 0.00868,
                "экономика и финансы": 0.056639999999999996
            },
            "domains": {
                "lenta.ru": 0,
                "avtoradio.ru": 20,
                "glas.ru": 160
            }
        },
        "key2":{
            "doc": "Госсекретарь Энтони Блинкен хочет обсудить с главой МИД
РФ Сергеем Лавровым \"Северный поток — 2\". Об этом сообщили в Госдепе США.
Встреча госсекретаря с главой российского МИД должна состояться на полях мин
истерской сессии Арктического совета. Комментарий дал представитель Госдепар
тамента высокого ранга, который поехал в Данию с Бинкеном. \"Я полагаю, что
это может возникнуть на двусторонних переговорах. Я думаю, что то же самое о
тносится и к Лаврову. Мы, как и конгресс, очень четко выразили свое отношени
е к \"Северному потоку — 2\", — отметил дипломат. Представитель Госдепа в оч
ередной раз напомнил, что США стремится к более предсказуемым и стабильным о
тношениям с Россией. На встрече они планируют обсудить с Лавровым и другие в
опросы. Госсекретарь Энтони Блинкен хочет обсудить с главой МИД РФ Сергеем Л
авровым \"Северный поток — 2\". Об этом сообщили в Госдепе США. Встреча госс
екретаря с главой российского МИД должна состояться на полях министерской се
ссии Арктического совета. Комментарий дал представитель Госдепартамента высо
кого ранга, который поехал в Данию с Бинкеном. \"Я полагаю, что это может во
зникнуть на двусторонних переговорах. Я думаю, что то же самое относится и к
 Лаврову. Мы, как и конгресс, очень четко выразили свое отношение к \"Северн
ому потоку — 2\", — отметил дипломат. Представитель Госдепа в очередной раз
напомнил, что США стремится к более предсказуемым и стабильным отношениям с
Россией. На встрече они планируют обсудить с Лавровым и другие вопросы",
            "topic_vector": [
        0.02878, 0.00193, 0.05107, 0.04142, 0.04555, 0.07284, 0.05475, 0.030
08, 0.03594, 0.19542, 0.02752, 0.04107,
        0.10873, 0.08427, 0.17154, 0.01106, 0.0598, 0.0175, 0.02363, 0.05864
, 0.02715, 0.15638, 0.03024,
```

```
        0.01971, 0.26024, 0.06841, 0.04656, 0.10368, 0.09647, 0.01731, 0.068
34, 0.05883, 0.49138, 0.32978, 0.08802,
        0.00695, 0.05001, 0.12073, 0.15687, 0.13756, 0.45444, 0.19302, 0.052
29, 0.05128, 0.01512, 0.04615,
        0.0312, 0.03497, 0.0483, 0.04992, 0.02532, 0.10955, 0.18278, 0.06962
, 0.00972, 0.00738, 0.03814, 0.00885,
        0.0165, 0.04813, 0.01451, 0.12169, 0.00202, 0.01801, 0.02341, 0.0513
4, 0.03142, 0.06254
    ],
            "entities": [
                "Энтони Блинкен", "МИД", "РФ", "Сергеем Лавровым", "поток",
"Госдепе", "США",
                "МИД", "Арктического совета", "Госдепартамента", "Данию", "Б
инкеном", "Лаврову",
                "Госдепа", "США", "Россией", "Лавровым", "Энтони Блинкен", "
МИД", "РФ",
                "Сергеем Лавровым", "Госдепе", "США", "МИД", "Арктического с
овета", "Данию",
                "Бинкеном", "Лаврову", "США", "Россией", "Лавровым"
            ],
            "nouns": [
                "Лавровым", "госсекретаря", "представитель", "Сергеем", "пот
ок",
                "Госдепартамента", "Россией", "Госдепе", "Комментарий", "США
",
                "Встреча", "Лаврову", "Энтони", "Госдепа", "дипломат", "Блин
кен",
                "Бинкеном", "РФ", "встрече", "совета", "конгресс", "Госсекре
тарь",
                "переговорах", "Данию", "отношение", "ранга", "МИД", "отноше
ниям",
                "вопросы", "потоку", "сессии", "Представитель", "полях", "Эн
тони Блинкен",
                "МИД", "РФ", "Сергеем Лавровым", "поток", "Госдепе", "США",
"МИД",
                "Арктического совета", "Госдепартамента", "Данию", "Бинкеном
", "Лаврову",
                "Госдепа", "США", "Россией", "Лавровым", "Энтони Блинкен", "
МИД", "РФ",
                "Сергеем Лавровым", "Госдепе", "США", "МИД", "Арктического с
овета", "Данию",
                "Бинкеном", "Лаврову", "США", "Россией", "Лавровым"
            ],
            "topics_with_probs": {
                "бизнес": 0.00334,
                "культура": 0.026449999999999998,
                "наука и техника": 0.024709999999999996,
                "общество": 0.02346,
                "политика": 0.83267,
                "проишествия": 0.011740000000000002,
                "силовые структуры": 0.01228,
                "спорт": 0.00868,
                "экономика и финансы": 0.056639999999999996
```

```
            },
            "domains": {
                "politros.com": 0,
                "pnp.ru": 20,
                "glas.ru": 160
            }

        }

    }
}
```


Параметры:

data – массив словарей с документами

"key1", "key2" – id документов

doc – текст документа

topic_vector – вектор топиков документа (результат обработки сервисом тематического моделирования)

entities – сущности документа (результат работы сервиса извлечения именованных сущностей)

nouns – существительные из сущностей

topics_with_probs – топики документов с вероятностями (результат обработки сервисом тематического моделирования)

domains – словарь с доменами и временем в часах, через которое была публикация документа относительно первого источника

Выходные данные – JSON формата:


```
{
    "result": {
        "key1": [
            [
                "sobesednik.ru",
                {
                    "proba": 0.8559091065336857,
                    "time": 36.977777777777774
                }
            ],
            [
                "tvzvezda.ru",
                {
                    "proba": 0.9927967542697079,
                    "time": 26.198611111111113
                }
            ]
```

```
            ],
            "key2": [
                [
                    "eg.ru",
                    {
                        "proba": 0.250666241197262,
                        "time": 31.92861111111111
                    }
                ],
                [
                    "vgoroden.ru",
                    {
                        "proba": 0.5209287405014038,
                        "time": 126.26944444444445
                    }
                ]
            ]
        }
    }
```


Параметры:

result – словарь с ответами

"key1", "key2" – ИД документов, массив массивов с результатами предсказаний

массив результата состоит из:

- домен
- словарь с данными о вероятности и времени распространения

Свойства связи между двумя источниками

Конечная точка расположена по адресу http://сервер:порт/predict_distribution_model

Метод запроса на сервер – POST.

Входные данные – JSON формата:


```
{
    "source_domain": "politros.com",
    "target_domain": "pnp.ru",
    "doc": "Госсекретарь Энтони Блинкен хочет обсудить с главой МИД РФ Сергеем Лавровым \"Северный поток — 2\". Об этом сообщили в Госдепе США. Встреча госсекретаря с главой российского МИД должна состояться на полях министерской сессии Арктического совета. Комментарий дал представитель Госдепартамента высокого ранга, который поехал в Данию с Бинкеном. \"Я полагаю, что это может возникнуть на двусторонних переговорах. Я думаю, что то же самое относится и к Лаврову. Мы, как и конгресс, очень четко выразили свое отношение к \"Северному потоку — 2\", — отметил дипломат. Представитель Госдепа в очередной раз напомнил, что США стремится к более предсказуемым и стабильным отношениям с Россией. На встрече они планируют обсудить с Лавровым и другие вопросы.
```

```
Госсекретарь Энтони Блинкен хочет обсудить с главой МИД РФ Сергеем Лавровым
\"Северный поток — 2\". Об этом сообщили в Госдепе США. Встреча госсекретаря
 с главой российского МИД должна состояться на полях министерской сессии Арк
тического совета. Комментарий дал представитель Госдепартамента высокого ран
га, который поехал в Данию с Бинкеном. \"Я полагаю, что это может возникнуть
 на двусторонних переговорах. Я думаю, что то же самое относится и к Лаврову
. Мы, как и конгресс, очень четко выразили свое отношение к \"Северному пото
ку — 2\", — отметил дипломат. Представитель Госдепа в очередной раз напомнил
, что США стремится к более предсказуемым и стабильным отношениям с Россией.
 На встрече они планируют обсудить с Лавровым и другие вопросы.",
    "topic_vector": [
        0.02878,
        0.00193,
        0.05107,
        0.04142,
        0.04555,
        0.07284,
        0.05475,
        0.03008,
        0.03594,
        0.19542,
        0.02752,
        0.04107,
        0.10873,
        0.08427,
        0.17154,
        0.01106,
        0.0598,
        0.0175,
        0.02363,
        0.05864,
        0.02715,
        0.15638,
        0.03024,
        0.01971,
        0.26024,
        0.06841,
        0.04656,
        0.10368,
        0.09647,
        0.01731,
        0.06834,
        0.05883,
        0.49138,
        0.32978,
        0.08802,
        0.00695,
        0.05001,
        0.12073,
        0.15687,
        0.13756,
        0.45444,
        0.19302,
```

```
            0.05229,
            0.05128,
            0.01512,
            0.04615,
            0.0312,
            0.03497,
            0.0483,
            0.04992,
            0.02532,
            0.10955,
            0.18278,
            0.06962,
            0.00972,
            0.00738,
            0.03814,
            0.00885,
            0.0165,
            0.04813,
            0.01451,
            0.12169,
            0.00202,
            0.01801,
            0.02341,
            0.05134,
            0.03142,
            0.06254
        ],
        "entities": [
            "Энтони Блинкен",
            "МИД",
            "РФ",
            "Сергеем Лавровым",
            "поток",
            "Госдепе",
            "США",
            "МИД",
            "Арктического совета",
            "Госдепартамента",
            "Данию",
            "Бинкеном",
            "Лаврову",
            "Госдепа",
            "США",
            "Россией",
            "Лавровым",
            "Энтони Блинкен",
            "МИД",
            "РФ",
            "Сергеем Лавровым",
            "Госдепе",
            "США",
            "МИД",
            "Арктического совета",
```

```
        "Данию",
        "Бинкеном",
        "Лаврову",
        "США",
        "Россией",
        "Лавровым"
    ],
    "nouns": [
        "Лавровым",
        "госсекретаря",
        "представитель",
        "Сергеем",
        "поток",
        "Госдепартамента",
        "Россией",
        "Госдепе",
        "Комментарий",
        "США",
        "Встреча",
        "Лаврову",
        "Энтони",
        "Госдепа",
        "дипломат",
        "Блинкен",
        "Бинкеном",
        "РФ",
        "встрече",
        "совета",
        "конгресс",
        "Госсекретарь",
        "переговорах",
        "Данию",
        "отношение",
        "ранга",
        "МИД",
        "отношениям",
        "вопросы",
        "потоку",
        "сессии",
        "Представитель",
        "полях",
        "Энтони Блинкен",
        "МИД",
        "РФ",
        "Сергеем Лавровым",
        "поток",
        "Госдепе",
        "США",
        "МИД",
        "Арктического совета",
        "Госдепартамента",
        "Данию",
        "Бинкеном",
```

```
            "Лаврову",
            "Госдепа",
            "США",
            "Россией",
            "Лавровым",
            "Энтони Блинкен",
            "МИД",
            "РФ",
            "Сергеем Лавровым",
            "Госдепе",
            "США",
            "МИД",
            "Арктического совета",
            "Данию",
            "Бинкеном",
            "Лаврову",
            "США",
            "Россией",
            "Лавровым"
        ],
        "topics_with_probs": {
            "бизнес": 0.00334,
            "культура": 0.026449999999999998,
            "наука и техника": 0.024709999999999996,
            "общество": 0.02346,
            "политика": 0.83267,
            "проишествия": 0.011740000000000002,
            "силовые структуры": 0.01228,
            "спорт": 0.00868,
            "экономика и финансы": 0.056639999999999996
        }
    }
```


Параметры:

source_domain – домен источник

target_domain – целевой домен

doc – текст документа

topic_vector – вектор топиков документа (результат обработки сервисом тематического моделирования)

entities – сущности документа (результат работы сервиса извлечения именованных сущностей)

nouns – существительные из сущностей

topics_with_probs – топики документов с вероятностями (результат обработки сервисом тематического моделирования)

Выходные данные – JSON формата:


```
{
    "success": true,
    "probability distribution": 0.5654661059379578,
    "topics": [
        "политика",
        "экономика и финансы",
        "культура"
    ],
    "tokens": [
        {
            "энтони": "neutral"
        },
        {
            "энтони": "neutral"
        },
        {
            "мид": "padding"
        },
        {
            "рф": "neutral"
        },
        {
            "рф": "neutral"
        },
        {
            "сергеем": "neutral"
        },
        {
            "сергеем": "neutral"
        },
        {
            "сергеем": "neutral"
        },
        {
            "лавровым": "neutral"
        },
        {
            "лавровым": "neutral"
        },
        {
            "поток": "padding"
        },
        {
            "госдепе": "neutral"
        },
        {
            "госдепе": "neutral"
        },
        {
            "сша": "neutral"
        },
        {
```

```
    "сша": "neutral"
},
{
    "мид": "padding"
},
{
    "госдепартамента": "neutral"
},
{
    "госдепартамента": "neutral"
},
{
    "данию": "neutral"
},
{
    "данию": "neutral"
},
{
    "бинкеном": "neutral"
},
{
    "бинкеном": "neutral"
},
{
    "бинкеном": "neutral"
},
{
    "лаврову": "neutral"
},
{
    "лаврову": "neutral"
},
{
    "лаврову": "neutral"
},
{
    "госдепа": "neutral"
},
{
    "госдепа": "neutral"
},
{
    "сша": "neutral"
},
{
    "сша": "neutral"
},
{
    "лавровым": "neutral"
},
{
    "лавровым": "neutral"
},
```

```
{
    "энтони": "neutral"
},
{
    "энтони": "neutral"
},
{
    "мид": "padding"
},
{
    "рф": "neutral"
},
{
    "рф": "neutral"
},
{
    "сергеем": "neutral"
},
{
    "сергеем": "neutral"
},
{
    "сергеем": "neutral"
},
{
    "лавровым": "neutral"
},
{
    "лавровым": "neutral"
},
{
    "поток": "padding"
},
{
    "госдепе": "neutral"
},
{
    "госдепе": "neutral"
},
{
    "сша": "neutral"
},
{
    "сша": "neutral"
},
{
    "мид": "padding"
},
{
    "госдепартамента": "neutral"
},
{
    "госдепартамента": "neutral"
```

```
            },
            {
                "данию": "neutral"
            },
            {
                "данию": "neutral"
            },
            {
                "бинкеном": "neutral"
            },
            {
                "бинкеном": "neutral"
            },
            {
                "бинкеном": "neutral"
            },
            {
                "лаврову": "neutral"
            },
            {
                "лаврову": "neutral"
            },
            {
                "лаврову": "neutral"
            },
            {
                "госдепа": "neutral"
            },
            {
                "госдепа": "neutral"
            },
            {
                "сша": "neutral"
            },
            {
                "сша": "neutral"
            },
            {
                "лавровым": "neutral"
            },
            {
                "лавровым": "neutral"
            }
        ]
}
```


Параметры:

success – флаг успешности обработки документа (true/false)

probability distribution – вероятность распространения

topics – список топиков

tokens – список сущностей с сентиментом

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/evaluate.

Метод запроса на сервер – POST.

Входные данные не требуются, опционально на вход можно подать файл формата JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Выходные данные – JSON формата:

```
{
"metrics": 0.9667
}
```

Параметры:

metrics – значение текущей метрики

Обучение

Конечная точка расположена по адресу http://сервер:порт /train.

Метод запроса на сервер – POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:

```
{
"success": true,
"message": "Обучение инициировано"
}
```

Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт /last_train_metric.

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{
"success": true,
"data": {
"time": "2021-11-27 11:21:20.973078",
"old_metric": 0.9667,
"new_metric": 0.9667,
"update_model": false
    }
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.10.3.Модель

В сервисе используется модель BERT

3.4.10.4.Метод оценки качества

Для оценки качества модели используется метрика F1 (п.3.3.1 данного документа).

Для топиков, сущностей и сентимента:

Ng - число топиков (сущностей/сентиментов), для которых есть данные, что они будут заимствованы.

Nm - число топиков (сущностей/сентиментов), для которых модель определила, что они будут заимствованы.

Nright - число топиков (сущностей/сентиментов), для которых модель правильно определила, что они будут заимствованы.

recall = Nright / Ng

precision = Nright / Nm

F1 = 2*precision*recall / (precision + recall)

3.4.11.Сервис предсказания трафика домена

3.4.11.1.Общее описание сервиса

Сервис прогнозирования трафика домена выполняет задачу оценки общего трафика домена за сутки.

3.4.11.2.Конечные точки сервиса

Предсказание трафика домена

Конечная точка расположена по адресу http://сервер:порт/api/domains_traffic/classify.

Метод запроса на сервер – POST.

Входные данные – JSON формата:


```
{
    "domains": [
        {
            "domain_name": "ua.tribuna.com",
            "domain_structure": [
                "tribuna.com",
                "vremyan.ru"
            ],
            "topics_document": [
                {
                    "lvl1": "спорт",
                    "lvl2": "хоккей",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось
, что он пропустит 3-4 месяца",
                    "probability": 0.25
                },
                {
                    "lvl1": "спорт",
                    "lvl2": "футбол",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось
, что он пропустит 3-4 месяца",
                    "probability": 0.24
                },
                {
                    "lvl1": "культура",
                    "lvl2": "знаменитости",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось
, что он пропустит 3-4 месяца",
                    "probability": 0.2
                }
```

```
            ],
            "link": "https://ua.tribuna.com/football/1101283480-brejtuejt-perenes-operacziyu-na-kolene-soobshhalos-chto-on-propustit-3.html",
            "date": "2021-09-16T17:02:08.000Z"
        },
        {
            "domain_name": "ua.tribuna.com",
            "domain_structure": [
                "tribuna.com",
                "vremyan.ru"
            ],
            "topics_document": [
                {
                    "lvl1": "спорт",
                    "lvl2": "хоккей",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось, что он пропустит 3-4 месяца",
                    "probability": 0.25
                },
                {
                    "lvl1": "спорт",
                    "lvl2": "футбол",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось, что он пропустит 3-4 месяца",
                    "probability": 0.24
                },
                {
                    "lvl1": "культура",
                    "lvl2": "знаменитости",
                    "lvl3": "футболисты «арсенала» вышли из отпуска",
                    "lvl4": "брэйтуэйт перенес операцию на колене.сообщалось, что он пропустит 3-4 месяца",
                    "probability": 0.2
                }
            ],
            "link": "https://ua.tribuna.com/football/1101283480-brejtuejt-perenes-operacziyu-na-kolene-soobshhalos-chto-on-propustit-3.html",
            "date": "2021-09-16T17:02:08.000Z"
        }
    ]
}
```


Параметры:

domains – массив доменов для запроса

domain_structure - embedding – структурный эмбеддинг домена (необяз.)

domain_name - название домена

topics_document - тематическое распределение документа (результат из сервиса тематического моделирования) (необяз.)

link - ссылка на документ (необяз.)

date - дата документа (необяз.)

Выходные данные – JSON формата:


```
{
    "amount_traffic": [
        {
            "success": true,
            "value": 461
        },
        {
            "success": true,
            "value": 109
        }
    ],
    "success": true
}
```


Параметры:

amount_traffic – массив с объемами трафика

success – статус обработки документа (true/false)

value – значение трафика

success – статус обработки всего запроса (true/false)

Эвалюация

Конечная точка расположена по адресу http://сервер:порт /api/domains_traffic/evaluate.

Метод запроса на сервер – POST.

Входные данные JSON с датасетом обучения модели (описан в документе «Методика валидации качества и дообучения моделей»).

Выходные данные – JSON формата:


```
{
    "mae": 0.6731796523480278,
    "success": true
}
```


Параметры:

mae – значение текущей метрики

success – статус обработки (true/false)

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер – POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:


```
{
    "message": "train initialize",
    "success": true
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/api/domains_traffic/last_train_metric.

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{
    "success": true,
    "data": {
        "time": "2021-11-27 11:21:20.973078",
        "old_metric": 0.6731796523480278,
        "new_metric": 0.6731796523480278,
        "update_model": false
    }
}
```


success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.11.3.Модель

В сервисе используется двухэтапное моделирование трафика.

На первом этапе данные обрабатываются классификаторами, определяющими категорию трафика:

- Менее 100 посещений
- От 100 до 1 000 посещений
- От 1 000 до 10 000 посещений
- От 10 000 до 100 000 посещений
- Более 100 000 посещений

После определения категории трафика работает регрессионная модель, уточняющая количество трафика в рамках данной категории - для каждой категории трафика обучена своя регрессионная модель.

Данные, на основе которых производится моделирование:

1) Тематический профиль домена

   Вычисляется как средний тематический профиль всех документов домена

2) Структурный эмбеддинг

   Эмбеддинг, построенный по связанным доменам с целевым доменом

Итоговый размер входного вектора в модель равен 1152.

В случае отсутствия данных реализованы следующие эвристики:

1) Если не передан тематический профиль домена, то он заменяется на средний тематический профиль документа, который был передан по этому домену за последние сутки
2) Если не передан структурный эмбеддинг, берется средний структурный эмбеддинг по известным доменам с идентичным тематическим профилем домена
3) Если передан только url домена, применяются эвристики 1-2

В итоге выбрана архитектура, показывающая наилучшее качество на тестовой выборке - полносвязная нейронная сеть с 2 слоями. Количество нейронов на первом слое - 128, на втором - 64 (число слоев и нейронов выбирается в зависимости от количества доменов в обучающей выборке), также применяется dropout к последнему слою с коэффициентом 0.5 для снижения эффекта переобучения модели.

3.4.11.4.Метод оценки качества

Для моделей классификации, используемых на первом этапе моделирования, используется метрика Accuracy (п.3.3.4 данного документа) на сбалансированной выборке доменов с целевой величиной трафика за сутки.

Этот показатель достигает 77% на отложенной тестовой выборке доменов, которые не были представлены в данных для обучения.

Для оценки моделей регрессии в рамках каждой категории трафика используется метрика MAE (п.3.3.3 данного документа). Прогнозируемое значение трафика от реального отличается в среднем на 10-30%.

3.4.12.Сервис определения трафика медиаматериала

3.4.12.1.Общее описание сервиса

Сервис позволяет определить вероятный трафик (стационарное количество посетителей за день) для заданного медиаматериала при публикации его на заданном источнике.

Сервис получает на вход текстовый медиаматериал, а также источник, на котором предполагается его публикация. и возвращает предсказательную величину трафика заданного медиаматериала.

3.4.12.2.Конечные точки сервиса

Предсказание трафика медиаматериала

Конечная точка расположена по адресу http://сервер:порт/api/documents_traffic / classify.

Метод запроса на сервер – POST.

Входные данные – JSON формата:

```
{
  "documents": [
    {
      "domain_traffic": 10000,
      "domain_name": "variant33.ru",
      "topics_document": [
        {
          "lvl1": "общество",
          "lvl2": "пенсии",
          "lvl3": "матвиенко: у россиян не будет накопительной пенсии",
          "lvl4": "13 жителей владимирской области скончались от коронавирус
а за сутки",
          "probability": 0.20266168733017192
        },
        {
          "lvl1": "политика",
          "lvl2": "региональная политика",
          "lvl3": "матвиенко: у россиян не будет накопительной пенсии",
          "lvl4": "13 жителей владимирской области скончались от коронавирус
а за сутки",
          "probability": 0.19261564214395935
        },
        {
          "lvl1": "общество",
          "lvl2": "коррупция",
          "lvl3": "матвиенко: у россиян не будет накопительной пенсии",
          "lvl4": "13 жителей владимирской области скончались от коронавирус
а за сутки",
          "probability": 0.17512797293151047
        }
      ],
      "document_metadata": {
        "name": "13 жителей Владимирской области скончались от коронавируса
за сутки",
        "text": "По данным на 14 сентября, за сутки в регионе ковид унёс жиз
ни 13-
и человек. В это же время 208 пациентов вылечились. Всего за время пандемии
с опасным недугом справились и выздоровели 47 310 наших земляков. Зафиксиров
ано 1 829 смертей.",
        "link": "https://variant33.ru/enews/trinadcat_umerli_kovid",
        "date": "2021-09-14T08:36:13.000Z"
      },
      "soc_dem_features": {
        "age_12-16": 1.67765137346793e-05,
        "age_17-21": 0.014889187512473648,
        "age_22-24": 0.0022982128152287374,
        "age_25-29": 9.499179061983807e-05,
        "age_30-34": 0.0006113037710650006,
        "age_35-39": 0.9430140010429796,
        "age_40-44": 0.027169572396751603,
        "age_45-49": 8.61126940657274e-06,
        "age_50-54": 0.0006567443758314541,
        "age_55-59": 0.00013551352087340035,
```

```
        "age_60-64": 0.0002547273116729852,
        "age_65+": 0.01085035767936249,
        "education_doctor": 4.9059040381127335e-05,
        "education_higher": 0.46200481488032,
        "education_phd": 0.0001005774045417216,
        "education_school": 6.862603240558453e-07,
        "education_special": 0.5378448624144331,
        "income_0-10": 0.0024398688746912656,
        "income_10-20": 5.109327828244061e-05,
        "income_100-110": 3.864007441591475e-05,
        "income_110-120": 0.0009457740790866287,
        "income_120-130": 6.366464841431894e-06,
        "income_130-140": 0.00018525740164164042,
        "income_140-150": 0.0002158656341225091,
        "income_150-160": 0.003239007927666754,
        "income_160-170": 0.0001279930114003643,
        "income_170-180": 0.0005877213877995525,
        "income_180-190": 0.0001684890392265287,
        "income_190-200": 0.0001332803556270509,
        "income_20-30": 0.030425532755297018,
        "income_30-40": 0.03795137350586401,
        "income_40-50": 0.0483296283419393,
        "income_50-60": 0.8750135399700815,
        "income_60-70": 1.6734647528048348e-05,
        "income_70-80": 5.570188168304757e-06,
        "income_80-90": 0.00011500421825439955,
        "income_90-100": 3.2588440653508513e-06,
        "sex_male": 0.5506456652263347,
        "sex_female": 0.44935433477366526
      }
    },
    {
      "domain_traffic": 1000000,
      "domain_name": "eaomedia.ru",
      "topics_document": [
        {
          "lvl1": "общество",
          "lvl2": "здравоохранение",
          "lvl3": "в иркутске проходит прививочная кампания против гриппа",
          "lvl4": "будет ли локдаун? инфекционист сообщил, какой может оказа
ться четвертая волна коронавируса",
          "probability": 0.7999869270700485
        },
        {
          "lvl1": "общество",
          "lvl2": "благоустройство территории",
          "lvl3": "в иркутске проходит прививочная кампания против гриппа",
          "lvl4": "будет ли локдаун? инфекционист сообщил, какой может оказа
ться четвертая волна коронавируса",
          "probability": 0.05247993729718599
        },
        {
          "lvl1": "общество",
```

```
          "lvl2": "пенсии",
          "lvl3": "в иркутске проходит прививочная кампания против гриппа",
          "lvl4": "будет ли локдаун? инфекционист сообщил, какой может оказа
ться четвертая волна коронавируса",
          "probability": 0.02532175133298206
        }
      ],
      "document_metadata": {
        "name": "Будет ли локдаун? Инфекционист сообщил, какой может оказать
ся четвертая волна коронавируса",
        "text": "Предыстория: \n \n\n Для привитых от COVID-19 россиян хотят
 ввести по три дополнительных выходных\n\n\n\n \n \nВрач-
инфекционист сообщил, какой может быть четвертая волна COVID-19 и ожидается
 ли локдаун. Новая волна окажется не меньше предыдущей, отметил эксперт. \nПо
 мнению главврача клинико-диагностической лаборатории \"Инвитро-
Сибирь\" Андрея Позднякова, четвертая волна будет не слишком отличаться от л
етней третьей. Вероятнее всего превалирующим останется тот же штамм \"дельта
\", однако на коронавирусную инфекцию могут наложиться другие — сезонные инф
екции. К тому же с началом учебного года больше начнут болеть дети, отметил
эксперт.
\n\n\n\n\n\n \nНе исключено сочетание коронавируса с другими инфекциями, поя
вление так называемых микст-
инфекций. При этом, каким будет поведение дельта-
штамма в этих условиях, пока неизвестно, добавил медик. По его словам, тяжес
ть будет не меньше летней, к тому же больше заболевших может быть среди дете
й. \nВакцинированные также будут болеть, но легче, чем непривитые, без госпи
тализаций, считает Поздняков. \n\"Новая волна пандемии, как и все предыдущие
, по нашим оценкам, продлится два-
три месяца\", — заключил эксперт в беседе с РИА Новости, предположив, что ло
кдаун введут едва ли, если население будет соблюдать уже ставшие привычными
меры защиты. \nАнна Кольцова Для привитых от COVID-19 россиян хотят ввести п
о три дополнительных выходных За эти дни сохранится зарплата",
        "link": "https://eaomedia.ru/news/1161158/",
        "date": "2021-09-14T08:36:14.000Z"
      },
      "soc_dem_features": {
        "age_12-16": 0.019059285145110533,
        "age_17-21": 1.4631420747516549e-06,
        "age_22-24": 0.006385705244699023,
        "age_25-29": 0.61177695875129,
        "age_30-34": 0.0012195493719351256,
        "age_35-39": 0.10715336572072377,
        "age_40-44": 0.06504798798104747,
        "age_45-49": 0.009956985790719021,
        "age_50-54": 0.1632919421531259,
        "age_55-59": 0.012991845020744339,
        "age_60-64": 6.756473444886137e-05,
        "age_65+": 0.0030473469440812956,
        "education_doctor": 0.0032886907628376684,
        "education_higher": 0.9350937905418224,
        "education_phd": 0.0023427161632222893,
        "education_school": 6.452160294731939e-05,
        "education_special": 0.059210280929170196,
```

```
            "income_0-10": 0.10220581356668579,
            "income_10-20": 0.06805069185970954,
            "income_100-110": 0.027669508868994157,
            "income_110-120": 0.01277832398531565,
            "income_120-130": 8.491874102274215e-05,
            "income_130-140": 0.01938789576641862,
            "income_140-150": 0.012030059633631853,
            "income_150-160": 0.060606343968670505,
            "income_160-170": 0.0018422540445362037,
            "income_170-180": 0.01523716457918301,
            "income_180-190": 0.0027189093315644545,
            "income_190-200": 0.00016714295117818532,
            "income_20-30": 0.031822895174834505,
            "income_30-40": 0.02210307904380822,
            "income_40-50": 0.020117642819584846,
            "income_50-60": 0.5374563786079872,
            "income_60-70": 0.0012478399041290081,
            "income_70-80": 5.6107768847445836e-05,
            "income_80-90": 0.0628432706452519,
            "income_90-100": 0.0015737587386462218,
            "sex_male": 0.4639566339368125,
            "sex_female": 0.5360433660631875
          }
        }
      ]
    }
```


Параметры:

documents – массив документов для запроса

domain_traffic – трафик домена

domain_name - название домена

topics_document - тематическое распределение документа (результат из сервиса тематического моделирования)

document_metadata – словарь метаданных документа:

- name – название документа
- text – содержание документа
- link - ссылка на документ
- date - дата документа

soc_dem_features – словарь с социально-демографическим распределением (результат работы сервиса классификации социально-демографического распределения)

Выходные данные – JSON формата:


```
{
    "success": true,
    "amount_traffics": [
```

```
            {
                "value": 5384.740829467773,
                "success": true
            },
            {
                "value": 544353.6639213562,
                "success": true
            }
        ]
    }
```


Параметры:

success – статус обработки пакета (true/false)

amount_traffics – массив с объемами трафика

success – статус обработки документа (true/false)

value – значение трафика

Эвалюация

Конечная точка расположена по адресу http://сервер:порт/api/documents_traffic / evaluate.

Метод запроса на сервер – POST.

Входные данные – JSON с данными обучения (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:


```
{
    "mse": 0.16102292714337701,
    "success": true
}
```


Параметры:

mse – значение метрики MSE

success – параметр успеха выполнения запроса (true/false)

Обучение:

Конечная точка расположена по адресу http://сервер:порт/train.

Метод запроса на сервер – POST.

Входные данные – JSON файл с датасетом (описан в документе «Методика валидации качества и дообучения моделей»)

Выходные данные – JSON формата:


```
{
    "message": "train initialize",
    "success": true
}
```


Параметры:

success – параметр успеха выполнения запроса (true/false)

message – сообщение о состоянии процесса

Получение последней метрики

Конечная точка расположена по адресу http://сервер:порт/api/documents_traffic /last_train_metric.

Метод запроса на сервер – GET.

Входные данные не требуются.

Выходные данные – JSON формата:


```
{
"success": true,
"data": {
"time": "2021-11-27 11:21:20.973078",
"old_metric": 0.16102292714337701,
"new_metric": 0.17304485617240305,
"update_model": false
    }
}
```


success – параметр успеха выполнения запроса (true/false)

data – словарь с отчетом о выполнении

time – время последнего обучения

old_metric – старая метрика

new_metric – новая метрика

update_model - параметр, который говорит о факте перехода на новую модель (true/false)

3.4.12.3.Модель

В данном сервисе используется двухэтапное моделирование трафика.

На первом этапе данные обрабатываются регрессионной моделью, которая оценивает нормированный трафик документа - величина от 0 до 1, показывающая долю трафика, которую получает документ в своем домене.

Для получения общего объема трафика эта величина умножается на предсказанную величину трафика домена.

Данные, на основе которых производится моделирование:

- Тематический профиль документа
- Тематический профиль домена
- Социально-демографические признаки аудитории
- Название документа
- Текст документа
- Именованные сущности документа

В случае отсутствия данных реализованы следующие эвристики:

- Если не передан тематический профиль домена, то он заменяется на тематический профиль документа
- Если не переданы именованные сущности документа, считается, что в тексте их нет, и вместо строкового представления списка именованных сущностей передается пустой текст
- Если не передано название или текст документа, то в модель передается пустая строка
- Если не переданы социально-демографические признаки аудитории, вектор аудитории сэмплируется случайным образом из распределения аудитории в обучающей выборке

В итоге выбрана архитектура, показывающая наилучшее качество на тестовой выборке - полносвязная нейронная сеть, кодирующая текстовые признаки документа BERT-эмбеддингом, а остальные признаки в виде числового вектора; количество слоев - 2; количество нейронов на первом слое - 64, на втором - 32 (число слоев и нейронов выбирается в зависимости от количества документов в обучающей выборке), также применяется dropout к последнему слою с коэффициентом 0.5 для снижения эффекта переобучения модели.

После прогнозирования нормированного трафика также используется оценка трафика домена для получения итогового трафика документа.

3.4.12.4.Метод оценки качества модели

Для оценки качества используется метрика MAE. Согласно этому показателю прогнозируемая величина трафика отличается от реального в среднем на 25-35% на отложенной тестовой выборке документов, которые не были представлены в данных для обучения.

3.4.13.Сервис отображения интерфейсов пользователя

3.4.13.1.Общее описание сервиса

Сервис отображения интерфейсов пользователя - сервис, выполняющий задачу отображения контента и функций, с которыми может взаимодействовать пользователь на различных устройствах.

3.4.13.2.Разделы интерфейса сервиса

Раздел «Проекты»

В разделе «Проекты» реализована следующая функциональность:

- Отображение созданных в Системе проектов и их состояний в виде доски;
- Изменение состояний проектов;
- Просмотр проектов;
- Отображение индикатора процесса цели проекта;
- Отображение индикатора срока проекта;
- Создание в Системе проектов с указанием описания, сроков, каналов распространения и автоматическим расчетом бюджета;
- Редактирование созданных проектов: описания, сроков, каналов распространения;
- Создание в Системе цели/угрозы;
- Настройка сегментации цели/угрозы;
- Настройка тематической картины и лексикона цели/угрозы: задание правил для универсального топика, добавление областей тематического

распределения с помощью диаграммы состояния информационного поля (пикер);

- Настройка требования соответствия для цели и требования дистанцирования для угрозы;
- Выбор созданной в Системе цели/угрозы для проекта;
- Редактирование созданной в Системе цели/угрозы.

Раздел «Реестр целей и угроз»

В разделе «Реестр целей и угроз» доступна следующая функциональность:

- Отображение в графическом виде текущего значения созданной в Системе цели/угрозы;
- Отображение на графике ретроспективы цели/угрозы до текущего значения, а также прогнозного тренда после текущего значения;
- Создание цели/угрозы:
  - Настройка сегментации цели/угрозы;
  - Настройка тематической картины и лексикона цели/угрозы: задание правил для универсального топика, добавление областей тематического распределения с помощью диаграммы состояния информационного поля (пикер);
  - Настройка требования соответствия для цели и требования дистанцирования для угрозы.

Раздел «Рабочее место оператора»

В разделе «Рабочее место оператора» реализована следующая функциональность:

- Просмотр списка документов, подходящих под цель/угрозу проекта;
- Поиск и фильтрация списка документов по статусу, периоду;
- Пагинация списка документов;
- Переход на источник документа;
- Просмотр текста документа;

- Просмотр доказательств, найденных Системой в тексте документа;
- Подтверждение или отклонение пользователем связи документа с целью/ угрозой проекта.

Раздел «Дискуссии»

В разделе «Дискуссии» реализована следующая функциональность:

- Переход в чат проекта, в котором уже производилась переписка.

Раздел «Аналитика»

В разделе «Аналитика» реализована следующая функциональность:

- Построение графика трендов и манипулирование содержимым:
  - Поиск и фильтрация для построения графика;
  - Отображение списка документов и их текстов, в которых упоминается выбранная на графике сущность в конкретную дату;
- Построение графика сентиментов и манипулирование содержимым:
  - Поиск и фильтрация для построения графика;
  - Отображение списка документов и их текстов, в которых упоминается выбранная на графике сущность в конкретную дату;
- Построение графика тримап и манипулирование содержимым:
  - Поиск и фильтрация для построения графика;
  - Построение графика по темам;
  - Построение графика по сущностям;
  - Построение графика по источникам;
- Построение графика новостей и манипулирование содержимым:
  - Поиск и фильтрация для построения графика;
  - Отображение даты новости, текста новости, источника и трафика.
- Построение графика связей и манипулирование содержимым:
  - Поиск и фильтрация для построения графика;
  - Отображение названия источника, трафика, часто публикуемых тем;
- Построение графика прогноза и манипулирование содержимым:

- Загрузка новости из файла и отображение предполагаемого распространения новости.

Раздел «Настройки», подраздел «Феномены»

В подразделе «Феномены» реализована следующая функциональность:

- Создание феномена с возможностью указания наименования феномена, его типа;
- Просмотр феномена;
- Редактирование феномена;
- Поиск феноменов в подразделе.

Раздел «Настройки», подраздел «Пользовательские топики»

В подразделе «Пользовательские топики» реализована следующая функциональность:

- Отображение списка пользовательских топиков с указанием наименования подрубрики, сюжета, количества документов, участвующих в обучающей выборке, динамики за день, динамики за месяц;
- Управление активностью пользовательских топиков (включение/ выключение);
- Удаление пользовательских топиков;
- Редактирование пользовательских топиков в разрезе изменения наименования сюжета;
- Создание пользовательских топиков с указанием подрубрики, наименования сюжета, описания и возможностью загрузки документов для него;
- Поиск и фильтрация пользовательских топиков по статусу (активен/ неактивен) и подрубрикам;
- Пагинация списка пользовательских топиков.

Раздел «Настройки», подраздел «Управление размерностями»

В подразделе «Управление размерностями» реализована следующая функциональность:

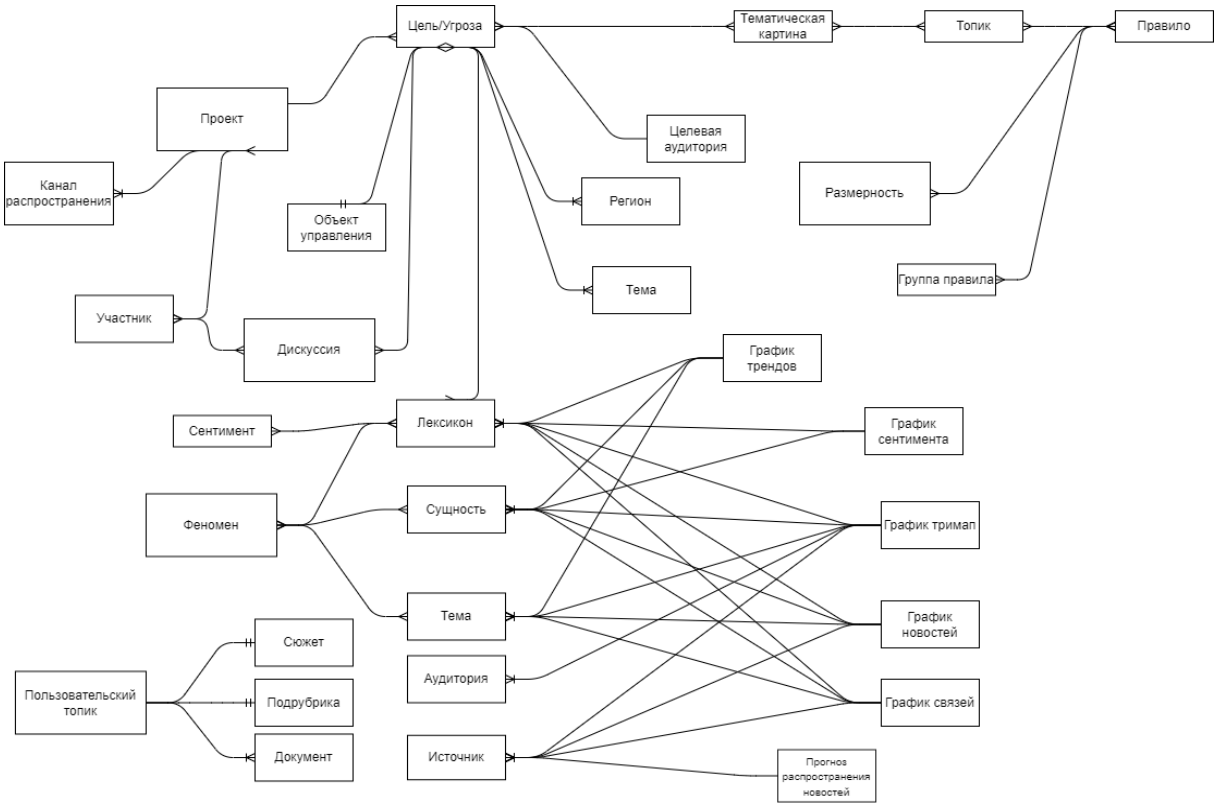
- Отображение списка всех размерностей целей и всех возможных правил в них с указанием статуса правила, условия, диапазона значений, допустимых типов ключей, описания и текста ошибки;
- Пагинация списка размерностей;
- Возможность настройки любого правила любой размерности: включение/ отключение правила размерности, указание диапазона, допустимых типов ключей, настройка описания правила, настройка текста ошибки;
- Поиск и фильтрация размерностей по объектам, свойствам, условиям.

Раздел «Настройки», подраздел «Настройки краулера»

В подразделе «Настройки краулера» реализована следующая функциональность:

- Отображение списка источников с указанием наименования источника, описания, url, типа страницы источника;
- Пагинация списка источников;
- Управление активностью парсинга источников;
- Удаление источников;
- Поиск и фильтрация списка источников по статусу и типу источника;
- Создание источников;
- Настройка парсинга источников;
- Редактирование источников;
- Загрузка списка источников VK.

3.4.13.3.Схема связей объектов сервиса



3.4.14.Сервис аналитики

3.4.14.1.Общее описание сервиса

Сервис аналитики выполняет задачу построения сводных отчетов о данных, полученных в результате обработки документов процессом ETL. Сервис включает в себя функции построения тренд-диаграмм, тримап-диаграмм, отчетов по сентименту и выявления связей между документами.

3.4.14.2.Конечные точки сервиса

Получение данных для графика трендов

Конечная точка расположена по адресу http://сервер:порт/analytics/trend_diagram.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "queries": [
        {
            "tag_value": [
```

```
                    {
                        "type": "entity",
                        "token": "путин",
                        "entity_param": "person"
                    },
                    {
                        "type": "entity",
                        "token": "путин",
                        "entity_param": "organization"
                    }
                ],
                "tag_param": "",
                "tag": "phenomena",
                "from": "2020-10-01",
                "filter": {"words" : []},
                "to": "2020-10-15",
                "smoothing" : true,
                "smoothing_values" : {
                        "rolling" : 1,
                        "seasonal" : 1
                    }

            }
        ]
    }
```


Параметры:

queries - список запросов

smoothing - применение сглаживания (true/false)

smoothing_values - значения сглаживания

from - дата с

to - дата по

filter - данные для фильтрации документов, по которым будет строиться график

tag - тип объекта, для которого будут искаться связанные объекты (topic, entity, source, ngram, phenomena)

tag_param - параметры объекта (обязательно при tag=entity. Значения: geo, person, organization)

tag_values - значение, для которого будут искаться связанные объекты (обрабатывается только при tag=topic)

id - Id фильтруемой сущности (обрабатывается только при tag=topic)

metric - метрика по которой происходит генерация отчета (traffic, count)

rolling – размер окна (по умолчанию - 1)

seasonal – размер сезона (по умолчанию - 1)

words – список слов

token – значение токена

type – тип токена (entity, ngram)

entity_param – тип сущности (geo, person, organization)

Выходные данные – JSON формата:


```
{
    "data": [
        {
            "color": "0,0,0",
            "id": "",
            "name": "",
            "tag": "phenomena",
            "tag_param": "",
            "tag_value": [
                {
                    "entity_param": "person",
                    "token": "путин",
                    "type": "entity"
                },
                {
                    "entity_param": "organization",
                    "token": "путин",
                    "type": "entity"
                }
            ],
            "values": [
                {
                    "date": "2020-10-01",
                    "value": 867,
                    "value_abs": 867
                },
                {
                    "date": "2020-10-02",
                    "value": 606,
                    "value_abs": 606
                },
                {
                    "date": "2020-10-03",
                    "value": 254,
                    "value_abs": 254
                },
```

```
{
    "date": "2020-10-04",
    "value": 253,
    "value_abs": 253
},
{
    "date": "2020-10-05",
    "value": 716,
    "value_abs": 716
},
{
    "date": "2020-10-06",
    "value": 1149,
    "value_abs": 1149
},
{
    "date": "2020-10-07",
    "value": 1448,
    "value_abs": 1448
},
{
    "date": "2020-10-08",
    "value": 631,
    "value_abs": 631
},
{
    "date": "2020-10-09",
    "value": 536,
    "value_abs": 536
},
{
    "date": "2020-10-10",
    "value": 352,
    "value_abs": 352
},
{
    "date": "2020-10-11",
    "value": 355,
    "value_abs": 355
},
{
    "date": "2020-10-12",
    "value": 557,
    "value_abs": 557
},
{
    "date": "2020-10-13",
    "value": 437,
```

```
                    "value_abs": 437
                },
                {
                    "date": "2020-10-14",
                    "value": 700,
                    "value_abs": 700
                },
                {
                    "date": "2020-10-15",
                    "value": 527,
                    "value_abs": 527
                }
            ]
        }
    ],
    "success": "true"
}
```


Параметры:

data – данные

color – цвет

id – id

name – название

tag – тип объекта

tag_param – параметры объекта (geo, person, organization)

tag_value – значение

values – найденные значения

date – дата

value – значение

value_abs – значение (абсолютное)

Получение данных для графика сентиментов

Конечная точка расположена по адресу http://сервер:порт/analytics/sentiment_diagram.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "queries": [
        {
            "from": "2019-06-04",
```

```
            "to": "2019-06-30",
            "metric" : "traffic",
            "object": {
                "tag_value": "путин",
                "tag_param": "person",
                "id": "",
                "tag": "entity"

            },
            "filter": {
                "words": ["фсб"]
            }
        },
        {
            "from": "2019-06-04",
            "to": "2019-06-30",
            "metric" : "count",
            "object": {
                "tag_value" : [
                    {"token" : "путин", "type" : "entity"},
                    {"token" : "президент", "type" : "unigram"},
                    {"token" : "россия", "type" : "entity"}
                    ],
                "tag_param": "",
                "id": "",
                "tag": "phenomena"

            },
            "filter": {
                "words": ["фсб"]
            }
        }
    ]
}
```

Параметры:

queries - список запросов

from - дата с

to - дата по

filter - данные для фильтрации документов, по которым будет строиться график

object – данные объекта

id – id фильтруемой сущности

metric - метрика по которой происходит генерация отчета (traffic, count)

words – список слов

tag - тип объекта, для которого будут искаться связанные объекты (topic, entity, source, ngram, phenomena)

tag_param - параметры объекта (обязательно при tag=entity. Значения: geo, person, organization)

tag_value - значение, для которого будут искаться связанные объекты (обрабатывается только при tag=topic)

token – значение токена

type – тип токена (entity, ngram, unigram)

entity_param – тип сущности (geo, person, organization)

Выходные данные – JSON формата:


```
{
    "data": [
        [
            {
                "date": "2019-06-04",
                "negative": 1553,
                "neutral": 2846,
                "positive": 1079
            },
            {
                "date": "2019-06-05",
                "negative": 4112,
                "neutral": 3378,
                "positive": 2589
            },
            {
                "date": "2019-06-06",
                "negative": 1064,
                "neutral": 4312,
                "positive": 382
            }
        ],
        [
            {
                "date": "2019-06-04",
                "negative": 126,
                "neutral": 139,
                "positive": 98
            },
            {
                "date": "2019-06-05",
                "negative": 137,
```

```
                "neutral": 118,
                "positive": 102
            },
            {
                "date": "2019-06-06",
                "negative": 161,
                "neutral": 128,
                "positive": 88
            }
        ]
    ],
    "success": "true"
}
```


Параметры:

data – данные

date – дата

negative – значение негатива

neutral – значение нейтральности

positive – значение позитива

Получение данных для графика тримап

Конечная точка расположена по адресу http://сервер:порт/analytics/treemap.

Метод запроса на сервер - POST.

Входные данные:

1. token – токен авторизации
2. query - JSON запрос:


```
{
    "date_to": "2018-12-01",
    "date_from": "2018-11-01",
    "tag": "entity",
    "tag_value": "фсб",
    "id": "",
    "tag_param": "organization",
    "search": "ngram",
    "filter": {
        "words": [
            "неуплата",
            "господдержка"
        ]
    },
    "with_sentiment": true
}
```

Параметры:

tag – тип объекта, для которого будут искаться связанные объекты (topic, entity, source, ngram, phenomena)

tag_param – параметры объекта (обязательно при tag=entity. Значения: geo, person, organization)

tag_value – значение, для которого будут искаться связанные объекты (не обрабатывается при tag=topic)

phenomena_values – массив сущностей и нграм для поиска по феномену

search – тип связанных объектов, по которым нужно построить тримап

date_from – дата с

date_to – дата по

filter – данные для фильтрации документов, по которым будет строиться график

id – id фильтруемой сущности (Обрабатываться только при tag=topic)

limit – кол-во возвращаемых данных (по умолчанию, 100)

with_sentiment – флаг, определяющий необходимость расчета сентимента

words – список слов

token – значение токена

type – тип токена (entity, ngram)

entity_param – тип сущности (geo, person, organization)

tag – тип объекта, для которого будут искаться связанные объекты (topic, entity, source, ngram, auditory)

tag_param - параметры объекта (обязательно при tag=entity. Значения: geo, person, organization)

Выходные данные – JSON формата:


```
{
    "data": [
        {
            "count": 873,
            "tag": "source",
            "tag_param": "",
            "tag_value": "google.com"
        },
        {
            "count": 506,
            "tag": "source",
```

```
            "tag_param": "",
            "tag_value": "hornews.ru"
        },
        {
            "count": 479,
            "tag": "source",
            "tag_param": "",
            "tag_value": "tass.ru"
        }
    ],
    "success": "true"
}
```


Параметры:

data – данные

count – количество

tag – тип объекта (topic, entity, source, ngram, auditory)

tag_param – Параметры объекта

tag_value – Значение

success – флаг успешности выполнения запроса (true/false)

Получение данных для графика распространения новостей

Конечная точка расположена по адресу http://сервер:порт/analytics/graph/spread.

Метод запроса на сервер - POST.

Входные данные – JSON формата:


```
{
    "token": "fae6bdfaa87d0ff84145e47c05bf68ba",
    "event_id": "da9ee3ab-06f7-4848-b5c4-a14caa0a40cb",
    "docs": 50,
    "sources": 50
}
```


Параметры:

token – токен авторизации

event_id – id новости

docs – количество документов

sources – количество источников

Выходные данные – JSON формата:


```
{
    "data": {
```

```
            "items": [
                {
                    "date": "2017-04-06",
                    "id": 75086881,
                    "name": "Узбекистан: В Ташкенте прекратил свою работу регион
альный офис НАТО",
                    "source": "fergananews.com",
                    "traffic": 44
                },
                {
                    "date": "2017-04-07",
                    "id": 75107816,
                     "name": "Глава МИД РФ примет участие в заседании СМИД СНГ в
Ташкенте",
                    "source": "regnum.ru",
                    "traffic": 3669
                }
            ],
            "links": [],
            "sources": [
                {
                    "name": "regnum.ru",
                    "topic": {
                        "color": "255,0,0",
                        "name": "политика"
                    }
                },
                {
                    "name": "fergananews.com",
                    "topic": {
                        "color": "255,0,0",
                        "name": "политика"
                    }
                }
            ]
        },
        "success": true
    }
```


Параметры:

items - массив узлов графа

id - идентификатор документа

date - дата публикации документа

name - имя документа (заголовок, title), один из источников в sources

source - источник документа (домен)

traffic - траффик документа

sources - массив источников (доменов)

name - имя источника

topic - топик источника (1го уровня, самый распространенный)

name - название топика

color - цвет топика

links - массив связей между документами (узлами графа) - ребра графа

source - начальная вершина ребра

target - конечная вершина ребра

force - сила связи

Получение данных для графика зависимости источников

Конечная точка расположена по адресу http://сервер:порт/analytics/graph/sources.

Метод запроса на сервер - POST.

Входные данные – JSON формата:

```
{
    "token": "qwdqwdwqdqwdwqdfgrgrgrgr",
    "count": 20,
    "date_from": "2019-10-01",
    "date_to": "2019-10-31",
    "params": [
        {
            "tag": "ngram",
            "tag_param": "",
            "tag_value": "голосовой помощник алиса"
        },
        {
            "tag": "entity",
            "tag_param": "person",
            "tag_value": "петров"
        }

            ]
}
```

Параметры:

token – токен авторизации

count – количество отдаваемых источников

date_from – дата с

date_to – дата по

params – список параметров для фильтрации

tag – тип объекта, для которого будут искаться связанные объекты (topic, entity, ngram)

tag_param – параметры объекта (обязательно при tag=entity. Значения: geo, person, organization)

tag_value – значение, для которого будут искаться связанные объекты (не обрабатывается при tag=topic)

id – идентификатор топика

Выходные данные – JSON формата:


```
{
    "data": {
        "items": [
            {
                "name": "tass.ru",
                "topics": [
                    {
                        "color": "0,0,0",
                        "name": "общество"
                    },
                    {
                        "color": "0,0,0",
                        "name": "культура"
                    },
                    {
                        "color": "0,0,0",
                        "name": "бизнес"
                    }
                ],
                "traffic": 1194
            },
            {
                "name": "rusfootball.info",
                "topics": [
                    {
                        "color": "0,0,0",
                        "name": "бизнес"
                    },
                    {
                        "color": "0,0,0",
                        "name": "силовые_структуры"
                    },
```

```
                    {
                        "color": "0,0,0",
                        "name": "наука_и_техника"
                    }
                ],
                "traffic": 113
            }
        ],
        "links": [
            {
                "force": 0.12954119447385892,
                "source": "tass.ru",
                "target": "life.ru"
            },
            {
                "force": 0.08654961266438477,
                "source": "tass.ru",
                "target": "znak.com"
            },
            {
                "force": 0.004705007188022137,
                "source": "rusfootball.info",
                "target": "aftershock.news"
            }
        ]
    },
    "success": true
}
```


Параметры:

data – словарь с результатами запроса

items – массив документов

links – массив связей между документами

name – имя документа

topics – источник документа

name – название топика

color – цвет топика

force – сила связи

source – начальная вершина ребра

target – конечная вершина ребра

3.4.14.3.Процесс работы сервиса

Процесс работы сервиса реализует вывод через конечные точки REST API результатов запросов к одной или нескольким БД системы.

3.4.15.Механизм обработки процесса ETL

3.4.15.1.Общее описание механизма

Механизм обработки процесса ETL отвечает за предварительную обработку данных из документов, полученных в результате работы сервиса краулинга сети Интернет. Обработка заключается как в последовательной, так и в параллельной обработке групп документов.

Механизм реализован на базе свободного ПО/

Балансировка при обработке данных механизмом между несколькими экземплярами развернутых сервисов реализована посредством свободно распространяемого обратного прокси сервиса.

Более подробное описание схематичной структуры, а также самого процесса ETL описано в документах «Функциональная схема системы» и «Схема потоков данных ETL».

3.4.16.Система мониторинга Системы

Система мониторинга представляет собой дашборд с набором панелей, на которых представлена индикация по статусам работы как Системы в целом, так и по сервисам. Также, для удобства администрирования, на панелях представлены элементы управления очередями/потоками документов, с возможностью ручного принудительного запуска сервисов для обработки документов с ошибками.

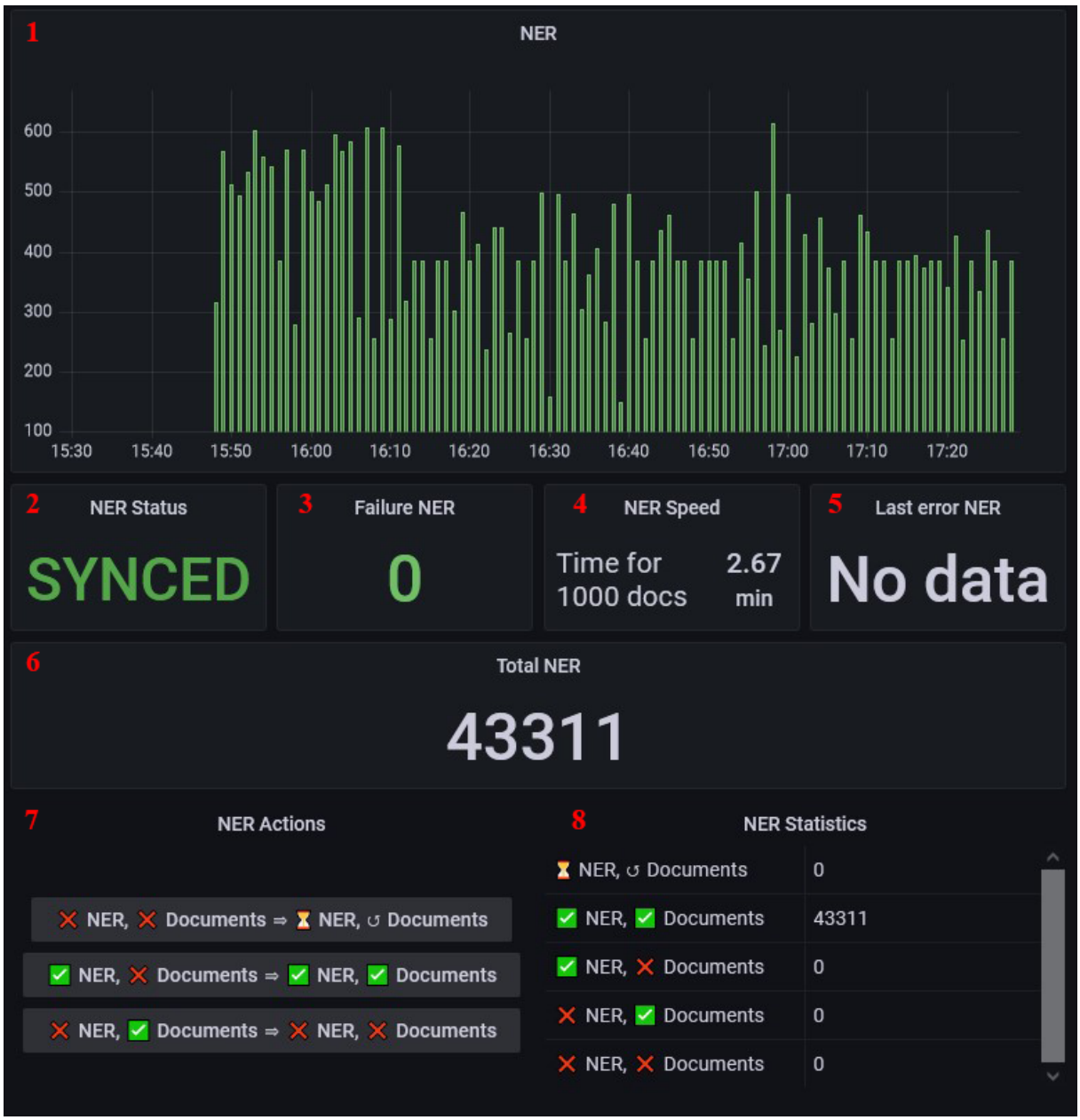
Панели со статусами работы в целом:

На панели цифрами обозначены:

1. Количество обработанных документов всего сервисами
2. График поступления очищенных документов во времени
3. Скорость поступления 1000 документов (в секундах)
4. Общее количество очищенных документов
5. График поступления документов из сервиса краулинга сети Интернет
6. Скорость поступления 1000 документов (в секундах)
7. Общее количество документов, поступивших из сервиса краулинга сети Интернет

Дополнительно для каждого сервиса, участвующего в обработке поступающих документов разработаны панели с элементами управления и статусами обработки данных.

Внешний вид панели на примере сервиса извлечения именованных сущностей:



На панели цифрами обозначены:

1. График количества обработанных документов во времени
2. Статус синхронизации сервиса
3. Количество ошибок при обработке сервисом
4. Скорость обработки сервисом 1000 документов
5. Информация о последней ошибке сервиса
6. Общее количество обработанных сервисом документов
7. Панель с интерактивными элементами управления сервисом. На панели сверху вниз:

- Помещение необработанных документов в очередь

- Установка признака обработки сервисом документов в общей таблице с документами
- Снятие признака обработки сервисом документов в общей таблице с документами при отсутствии данных по этим документам в таблице сервиса

8 Панель со статистикой работы сервиса. На панели сверху вниз:

- Количество документов в очереди
- Количество обработанных документов
- Количество обработанных документов с не проставленным признаком обработки сервисом в общей таблице с документами
- Количество документов у которых проставлен признак обработки сервисом в общей таблице с документами, но нет данных в таблице сервиса
- Количество необработанных документов (ошибки при обработке)

3.5. Режимы функционирования системы

ИС «МИР-1» функционирует в непрерывном круглосуточном режиме и находится в постоянной готовности к обслуживанию пользователей и взаимодействию с внешними системами.

Предусмотрены следующие режимы работы системы:

– Штатный режим – основной режим функционирования системы, когда выполняется полный набор требуемых функций с максимальной производительностью непрерывно и круглосуточно. Переход системы в данный режим происходит автоматически при первоначальном запуске системы в эксплуатацию или по завершению иных режимов;

– Режим проведения регламентных работ – предполагает частичное временное отключение части оборудования/сервисов Системы. При этом функционирование Системы в целом не предусматривается. При переходе сервисов Системы в режим проведения регламентных работ обеспечивается бесперебойная работа сервиса краулинга сети Интернет в целях обеспечения непрерывного накопления данных для последующей обработки. В перечень работ, производимых в данном режиме, входят: сжатие и переиндексация таблиц баз данных, обновление системного и прикладного программного

обеспечения, обновление системы мониторинга, операции с нейронными сетями и т.д. Выход из этого режима происходит автоматически после проведения запланированных операций;

– Аварийный режим работы – предполагает полное или частичное ограничение полнофункциональной доступности компонентов (сервисов) Системы, явившееся следствием сбоя в электроснабжении комплекса технических средств, на которых развернута Система, отказа аппаратно-программных средств, обеспечивающих функционирование Системы, а также в случае нештатного функционирования программного обеспечения. Переход в аварийный режим не может быть вызван некорректными действиями пользователей системы (исключая администраторов). Переход в аварийный режим происходит автоматически с уведомлением на адрес электронной почты ответственных лиц.

  Выход из режима осуществляется:

  - автоматически при устранении причин отказа системы/сервисов;

  - вручную администратором.

При проведении работ по реконфигурации и/или масштабированию допускается временная или частичная неработоспособность пользовательских функций системы.

Выявление неработоспособности технических и/или программных компонентов ИС «МИР-1» обеспечивается диагностированием системы. Диагностирование технических и программных компонентов осуществляется средствами операционной системы, СУБД, а также дополнительных систем мониторинга в реальном времени.

Система мониторинга ИС «МИР-1» обладает следующими средствами диагностики, служащими для предотвращения сбоев в работе и облегчающих поиск и идентификацию ошибок:

- средства проверки сервиса краулинга сети Интернет;
- средства мониторинга процесса ETL с возможностью интерактивного управления запуском сервисов для обработки документов с ошибками;
- средства логирования в БД запросов к сервису отображения интерфейсов.

При возникновении сбоев в функционировании сервиса отображения интерфейсов или совершении пользователем ошибочных действий выдаются системные сообщения на русском языке (за исключением внутренних ошибок сервиса), на основании которых пользователь может определить причину ошибки и способы ее устранения.

3.6. Обеспечение потребительских характеристик системы

Требования к надежности ИС «МИР-1» выражены количественным показателем доступности – 97%.

Показатель доступности Системы рассчитывается следующим образом:

Доступность = (Д — П) / Д × 100 %

Где:

- Д – Время плановой доступности
- П – Время простоя

Пользуясь данной формулой, время недоступности Системы при расчете на 1 сутки составляет 43 мин 12 сек.

Рекомендуется производить регламентные работы в Системе не реже 1 раза в квартал. В минимально производимые работы должен входить перезапуск контейнеров сервисов Системы.

Ориентировочное время перезапуска одного контейнера составляет около 5 минут.

При последовательной перезагрузке сервисов и их количестве равном 1, время недоступности составляет 1 час.

Для соблюдения 97% доступности в квартал Система должна безотказно работать в течение 88д 13ч 42м 41с и быть в состоянии недоступности до 2д 17ч 44м 37с.

При этом расчете использовано усредненное значение дней в квартале, равное 91,311 дней.

Данный показатель достигается за счет применения системного и базового программного обеспечения, соответствующих классу решаемых задач, а так же за счет организационных мероприятий, необходимых для обеспечения корректной работы Системы, таких как своевременное выполнение процессов администрирования и

соблюдение правил эксплуатации и технического обслуживания комплекса технических и программно-аппаратных средств Системы.

3.7. Техническое обеспечение

Комплекс технических средств требуемых для работоспособности ИС «МИР-1» должен обеспечивать:

- Работу всех сервисов и компонентов, перечисленных в п. 3.3 настоящего документа;
- Достижение показателей назначения, указанных в п. 4.1.10 ТЗ на разработку системы;
- Поддержку сервисной архитектуры системы.

Комплекс технических средств должен включать в себя:

- Серверные платформы, которые объединены средствами виртуализации и поддерживают функционирование виртуальных серверов;
- Стационарные и мобильные АРМ, которые призваны обеспечивать работу пользователей в системе на рабочих местах;
- Телекоммуникационное и сетевое оборудование, обеспечивающее обмен между сервисами системы и выход в сеть Интернет;
- Средства, обеспечивающие гарантированное бесперебойное питание.

Серверные платформы системы должны обеспечивать поддержку функционирования как минимум всех сервисов в единственном экземпляре.

Требования к комплексу технических средств приведены в документе «Спецификация перечня технических средств для развертывания ИС МИР-1».

Выбор состава и количества оборудования для реализации серверной платформы системы, а также выбор сетевого и телекоммуникационного оборудования, которое обеспечит информационный обмен между компонентами системы производится Заказчиком с учетом требований к системе.

Физическое серверное оборудование должно быть размещено в специализированных технологических и офисных помещениях зданий, оборудованных системами электроснабжения, связи, отопления, вентиляции и поддержки климатических условий, а также системами управления и контроля доступа, противопожарной и охранной сигнализацией.

Система поддерживает возможность развертывания в облачных системах. Выбор и настройка облака производится квалифицированным персоналом Заказчика.

В обоих случаях должен быть обеспечен доступ к системе с рабочих станций пользователей.

Телекоммуникационная инфраструктура обеспечивает взаимодействие сервисов ИС «МИР-1» и АРМ пользователей с гарантированной полосой пропускания не менее 100 Мбит/сек. При использовании системы с мобильных АРМ данный показатель определяется возможностями сетей 3G/4G LTE сетей.

Для обеспечения надежной работы системы в процессе эксплуатации рекомендуется проводить следующие мероприятия:

- Резервирование системных и хранимых данных;
- Защита по электропитанию посредством использования источников бесперебойного питания.

В качестве мер по обеспечению сохранности информации ИС «МИР-1» в эксплуатирующих подразделениях должны быть разработаны и утверждены регламенты по резервированию, хранению и восстановлению информации.

Помещения, в которых размещается оборудование ИС «МИР-1», должны быть оборудованы системой автоматического пожаротушения.

3.8. Информационное обеспечение

В состав информационного обеспечения системы входят нормативно-справочная информация, входные и выходные данные, структура организации баз данных, схема потоков преобразования данных.

Для хранения данных в ИС «МИР-1» используются:

1) реляционные базы данных, обеспечивающая реализацию механизмов построения индексов и контроля целостности данных. Основные возможности этой СУБД:
   - поддержка реляционной или объектно-реляционной модели базы данных;
   - поддержка многопроцессорной архитектуры;
   - наличие средств создания индексов и кластеров данных;

- автоматическое восстановление базы данных;
- возможность контроля доступа к данным;
- централизованное управление учетными записями пользователей;
- оптимизация запросов.

2) колоночная база данных, позволяющая обеспечить быстрые запросы к данным в том числе для целей аналитики.

Отличительные возможности этой СУБД:

- поддержка сжатия данных;
- поддержка возможности хранения данных на дисках;
- параллельная обработка запросов в многоядерных системах;
- параллельная обработка запросов на разных серверах;
- отсутствие блокировок при добавлении данных.

Деление внутримашинной БД описано в документе «Описание организации информационных баз ИС МИР-1»

3.9. Программное обеспечение

Программное обеспечение ИС «МИР-1» состоит из:

- Специального ПО (СПО) – программных средств, реализующих функциональное назначение ИС «МИР-1»
- Общего ПО (ОПО) – программных средств, создающих инфраструктурную платформу для функционирования ИС «МИР-1»

ОПО является открытым и обеспечивает включение в систему вновь разрабатываемых средств.

В состав ОПО ИС «МИР-1» входят:

- платформа управления рабочими процессами
- обратные прокси-сервера для балансировки нагрузки
- системы управления СУБД
- система мониторинга

В состав СПО ИС «МИР-1» входят сервисы системы.

Специальное программное обеспечение обеспечивает выполнение автоматизируемых функций, приведенных в настоящем документе.

ИС «МИР-1» реализована на основе свободно распространяемого программного обеспечения. Для работы пользовательского интерфейса достаточно функциональных возможностей последних версий наиболее популярных интернет-браузеров.

ИС «МИР-1» обеспечивает запись, чтение и обновление данных БД в части, обеспечивающей выполнение её функций, а также разграничение доступа пользователей к данным БД.

СПО ИС «МИР-1» устойчиво при функционировании и обеспечивает продолжение работы – восстановление после устранения отклонений, вызванных сбоями технических средств и ошибками во входных данных.

Программное обеспечение обеспечивает реализацию следующих функций управления доступом:

- идентификацию и проверку подлинности субъектов доступа при входе в систему по идентификатору и паролю условно-постоянного действия;
- идентификацию терминалов, вычислительных машин, узлов сети, каналов связи;
- идентификацию томов, каталогов, файлов, записей, полей записей по именам.

3.10. Обоснование выбора технических и программных средств

Выбор технических средств обусловлен:

- сервисной архитектурой ИС «МИР-1»
- функционированием СПО ИС «МИР-1» в круглосуточном режиме, которое требует обеспечение высокого уровня надежности.

Программные компоненты и библиотеки, используемые при разработке, являются свободно распространяемыми с открытым исходным кодом.

3.11. Перечень заданий на разработку специализированных технических средств

Разработка специализированных (новых) технических средств в рамках работ по разработке ИС «МИР-1» не предполагается.

3.12. Перечень заданий на разработку строительных, электротехнических, санитарно-технических и других разделов проекта, связанных с созданием системы

Разработка строительных, электротехнических, санитарно-технических и других разделов проекта в рамках работ по развитию ИС «МИР-1» не предполагается.

4. МЕРОПРИЯТИЯ ПО ПОДГОТОВКЕ ОБЪЕКТА АВТОМАТИЗАЦИИ К ВВОДУ СИСТЕМЫ В ДЕЙСТВИЕ

4.1. Приведение информации к виду, пригодному для обработки на ЭВМ

В ИС «МИР-1» для обеспечения информационного обмена как внутри системы, так и с другими взаимодействующими системами, используются стандартные протоколы взаимодействия:

- на информационном уровне протокол взаимодействия основан на использовании текстового формата обмена данными, основанном на JavaScript - JSON в рамках протокола обмена структурированными сообщениями REST API;
- на транспортном уровне взаимодействие обеспечивается транспортным интернет-протоколом TCP/IP;
- на физическом уровне передачи информации используются стандартные технологии передачи данных.

Для обеспечения автоматизированного информационного обмена ИС «МИР-1» с внешними ИС могут потребоваться некоторые работы по модернизации интерфейсов межсистемного взаимодействия, если внешние системы не поддерживают указанные стандарты.

Вся информация поступает в ИС «МИР-1» либо автоматически, либо в диалоговом режиме с рабочего места оператора, в связи с этим, каких-либо мероприятий по приведению информации к виду, пригодному для обработки на ЭВМ, не требуется.

4.2. Мероприятия по обучению и проверке квалификации персонала

Решение по организации обучения и проверке квалификации персонала ИС «МИР-1» принимается Заказчиком на основании разработанной в рамках данной работы эксплуатационной документации на систему.

Обучение персонала ИС «МИР-1» должно быть проведено до принятия системы в опытную эксплуатацию, его формы и методы также определяются Заказчиком системы.

Проверка квалификации персонала по владению навыками работы с ИС «МИР-1» может проводиться как в форме тестирования, так и в ходе проведения испытаний при вводе системы в опытную эксплуатацию, и организуется Заказчиком.

4.3. Мероприятия по созданию необходимых подразделений и рабочих мест

Для внедряемой ИС «МИР-1» предлагается организовать обслуживающие и эксплуатационные подразделения.

Заказчик:	**Подрядчик:**
_____ / С.В. Гуляев /	_____ / В.Д. Басков /
М.П.	М.П.