|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство образования и науки Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ **Радиотехнический**

КАФЕДРА **СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ**

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ***

***НА ТЕМУ:***

**Модуль вычисления скорости реакции поисковой системы**

Студент \_\_РТ5-81Б\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_ Ф.А. Сафонов**\_\_\_

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель ВКР **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Ю.Е. Гапанюк**\_\_

(Подпись, дата) (И.О.Фамилия)

Нормоконтролер **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ Ю.Н. Кротов\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

# **РЕФЕРАТ**

Расчётно-пояснительная записка квалификационной работы бакалавра содержит 39 страниц. Вместе с приложениями объем работы составляет 51 страницы. Работа включает в себя 3 таблицы и 25 иллюстраций. В процессе выполнения было использовано 46 источников.

Объектом разработки является модуль вычисления скорости реакции поисковой системы.

Данный программный модуль позволяет измерять качество поисковой выдачи на запросах, связанных с новостными внезапными событиями.

Цель работы заключается в реализации метрики, измеряющей качество поисковой выдачи на запросах, связанных с внезапными новостными событиями, исследовании различных методов машинного и глубокого обучения, благодаря которым можно автоматизировать процесс подсчета метрики, а также непосредственно обучение моделей машинного и глубокого обучения.

В процессе выполнения квалификационной работы бакалавра подробно изучена предметная область, проведен анализ пользовательских запросов, обучены модели машинного и глубокого обучения. Также реализован итоговой процесс подсчета метрики качества поисковой выдачи на запросах, связанных с новостными внезапными событиями.

СОДЕРЖАНИЕ

[РЕФЕРАТ 2](#_Toc167895681)

[СОДЕРЖАНИЕ 3](#_Toc167895682)

[ВВЕДЕНИЕ 4](#_Toc167895683)

[1 Постановка задач разработки 6](#_Toc167895684)

[1.1 Постановка задач проектирования 6](#_Toc167895685)

[1.2 Описание предметной области 8](#_Toc167895686)

[2 Исследовательская часть 9](#_Toc167895687)

[2.1 Метрики ранжирования 9](#_Toc167895688)

[2.2 Mean reciprocal rank 10](#_Toc167895689)

[2.3 DCG & NDCG 11](#_Toc167895690)

[2.4 Рекурентные нейронные сети 12](#_Toc167895691)

[2.5 LSTM 16](#_Toc167895692)

[3 КОНСТРУККТОРСКО-ТЕХНОЛОГИЧЕСКАЯ ЧАСТЬ 25](#_Toc167895693)

[3.1 Описание поисковых запросов, на которых будет подсчитана метрика 25](#_Toc167895694)

[3.2 Запуск разметки на ультрасвежесть 28](#_Toc167895695)

[3.3 Ансамбль для классификации ультрасвежих запросов 29](#_Toc167895696)

[3.4 Обучение LSTM 31](#_Toc167895697)

[3.5 Обучение градиентного бустинга 35](#_Toc167895698)

[3.6 Поиск начала события 39](#_Toc167895699)

[3.7 Итоговый подсчет метрики 41](#_Toc167895700)

[ЗАКЛЮЧЕНИЕ 44](#_Toc167895701)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 46](#_Toc167895702)

[ПРИЛОЖЕНИЕ A ГРАФИЧЕСКИЕ МАТЕРИАЛЫ 47](#_Toc167895703)

[ПРИЛОЖЕНИЕ Б ТЕХНИЧЕСКОЕ ЗАДАНИЕ 48](#_Toc167895704)

[ПРИЛОЖЕНИЕ В ПРОГРАММА И МЕТОДИКА ИСПЫТАНИЙ 53](#_Toc167895705)

ВВЕДЕНИЕ

В мире существует множество поисковых систем: Bing, Google, Яндекс. Рынок поисковых систем очень конкурентный: любая поисковая система стремится привлечь к себе как можно больше пользователей, чтобы иметь большую прибыль с показа рекламы. Именно поэтому важными аспектами любой поисковой системы являются релевантность и качество поисковых страниц на выдаче. Как правило, учитываются первые 5-10 страниц, ведь именно их увидит пользователь. Яндекс каждый год вкладывает огромные средства на улучшения алгоритмов ранжирования. Запросы пользователя бывают разной тематики, например: медицина, спорт, новости. Такие тематики называются поисковыми срезами. Например, одна команда может улучшать выдачу на спортивном срезе, другая на медицинском. В этой работе речь пойдет о мониторинге качества поисковой выдачи на новостном срезе. Это задача крайне важна, т.к. доля новостных запросов очень велика и Яндексу крайне важно иметь качественную выдачу, которая сможет в полной мере удовлетворить интересы пользователей. Например, если человек идет по улице и видит, что на улице пожар, он может захотеть узнать, что случилось. Для этого он напишет запрос в поисковую систему: «пожар в Москве что случилось». На такой запрос хорошая поисковая система должна правильно обработать документы-кандидаты, которые пойдут пользователю на выдачу. В данном случае поиску важно показать новые, актуальные, «свежие», как мы говорим, документы.

В жизни часто происходят незапланированные внезапные новостные события: аварии, случаи смерти, стихийные бедствия. Как правило, такие внезапные новостные событие сильно интересуют людей в первые секунды их происшествия. В таком случае задача поисковой системы как можно раньше актуализировать выдачу по тому или иному новостному событию.

В данной работе представлена разработка метрики качества поисковой выдачи на новостных запросах, связанных с каким-либо событием. Чтобы иметь возможность построить такую метрику было решено множество задач внутри компании. Для предсказания новостных запросов о внезапных новостных событиях на которые Яндексу важно иметь актуальную выдачу с первых секунд были использованы алгоритмы машинного обучения, СatBoost (Categorical Boosting), различные нейросетевые признаки, а также нейросеть LSTM (Long short-term memory) для анализа временных рядов динамики запроса. На различных этапах подсчета метрики были применены нейронные сети, а именно BERT (Bidirectional Encoder Representations from Transformers), реализовано множество Map Reduce вычислений на языке запросов YQL (разработка Яндекса, аналог SQL). Не обошлось и без регулярной ручной, асессорской разметки. Стоит отметить, что затраты на разметку были снижены во много раз благодаря машинному обучению.

1. Постановка задач разработки
   1. Постановка задач проектирования

Задача поисковой системы — предоставлять пользователю возможность быстрого доступа к необходимой ему информации при помощи поиска в обширной коллекции доступных данных. Одна из самых распространенных реализаций поисковых систем — веб сервисы поиска текстовой или графической информации во Всемирной паутине. В данной работе речь пойдет об одной из самых популярных поисковых систем — Яндекс Поиск.

Яндекс Поиск не стоит на месте: тысячи разработчиков, аналитиков и менеджеров ежедневно работают над тем, чтобы алгоритмы Яндекс Поиска стали лучше. Чтобы увидеть улучшения необходимо иметь мониторинги и графики, которые бы отражали эффективность работы разных компонент Поиска. По таким мониторингам можно сделать вывод о проделанной работе и доказать, что проделанная работа, действительно, улучшила Поиск по тому или иному аспекту.

За любым графиком должен стоять какой-то смысл измеряемой величины, например график качественных и релевантных страниц на выдачи или график доли пользователей, успешно решивших свою задачу благодаря нашей выдачи.

При работе над построением таких мониторингов необходимо сформулировать критерия к измеряемой величине и разработать алгоритм ее измерения. Например, если мы хотим иметь график, показывающий среднюю долю качественных страниц на всех запросах пользователя за день, необходимо сформулировать критерии качества страницы и придумать, как этот критерий измерять, имея доступный ограниченный набор ресурсов.

Благодаря мониторингам разработчики могут систематически отслеживать тот или иной улучшаемый параметр, именно по графикам на мониторингах делается вывод о пользе новых алгоритмов и принимаются итоговые решения.

Яндекс Поиск — масштабный сервис, поэтому задача улучшения Поиска декомпозируется на множество других подзадач, так или иначе связанных между собой. Часто бывает, что поисковая система работает неидеально только на запросах определенной категории. Такие категории запросов называются срезами. Запросы из каждого среза имеют свою специфику: например, на медицинские запросы важно показывать проверенные врачами статьи, чтобы пользователь не прочитал ложную информацию и не причинил себе вред, а на новостные запросы важно показывать недавно созданные страницы, чтобы пользователь получал актуальную информацию. Из-за таких особенностей Яндекс Поиск применяет различные алгоритмы ранжирования в зависимости от среза запроса.

В данной работе речь пойдет о создании мониторинга для отслеживания качества выдачи Яндекс Поиска на ультрасвежих запросах.

Запрос является **ультрасвежим**, если с запросом связано какое-то новостное событие и соблюдаются следующие пункты:

1. Событие по запросу незапланировано. Незапланированным считаем событие, о котором не было информации до дня запроса.
2. Для объекта запроса произошедшее событие значимо. Значимым считаем событие, заметно выходящее за рамки того, что обычно происходит с объектом запроса.

Примеры незапланированных и значимых событий:

* ЧП
* Катаклизм
* Неожиданные заявления
* Происшествия
* Неожиданная смерть

В данной работе ставится задача разработать программный модуль для определения качества поисковой системы на таких запросах. Поскольку Поиску крайне важно быстро реагировать на внезапные события, качество особенно важно в первое время с момента происшествия события.

Таким образом, в результате анализа целевой направленности разработки, можно сформулировать следующие задачи выполнения данной выпускной квалификационной работы бакалавра:

* Изучение и анализ предметной области метрик алгоритмов ранжирования
* Определение критериев к запросам, на которых будет подсчитана желаемая метрика
* Создания алгоритма извлечения необходимых запросов, среди общего множества поисковых запросов
* Создание алгоритмов определения начала события, связанного с запросом
* Создания алгоритма определения качества выдачи Поисковой системы Яндекс в первые часы с момента происшествия события по данному запросу
  1. Описание предметной области

Рассмотрим основные понятия и определения, касающиеся данной предметной области:

Ранжирование — это сортировка сайтов поисковыми системами и размещение их в результатах выдачи

Метрика — это качественный или количественный показатель, который отражает ту или иную характеристику и уровень успешности продукта.

Поисковый запрос — последовательность символов, которую пользователей вводит в поисковую строку, чтобы найти интересующую его информацию. Формат поискового запроса зависит как от устройства поисковой системы, так и от типа информации для поиска. Чаще всего, поисковый запрос задаётся в виде набора слов или фразы.

Поисковая выдача — это список документов, найденных и проиндексированных поисковой системой. Обычно он упорядочен по убыванию релевантности документов поисковому запросу согласно применяемым в поисковой системе алгоритмам ранжирования, но во многих системах предусмотрены также и другие виды сортировки, например, по дате документов.

Релевантность — степень смыслового соответствия результатов выдачи запросу пользователя.

Другие термины, придуманные командой, чтобы определять нужные нам понятие будут разъяснены по ходу изложения.

1. Исследовательская часть
   1. Метрики ранжирования

Ранжирование — это классическая задача машинного обучения, которая заключается в ранжировании списка элементов на основе их релевантности для конкретной задачи (например, ранжирование страниц в Google на основе их соответствия заданному запросу). Задача имеет широкий спектр применений в различных сервисах:

* Рекомендации фильмов и видео (Кинопоиск, YouTube)
* Ранжирование страниц в поиковых системах (Yandex, Google, Bing)
* Рекомендации товаров (Ozon, Amazon)
* Поиск жилья (Booking, Yandex travel)

Чтобы измерить качество системы ранжирования, необходимо знать истинно верный порядок элементов и выбрать соответствующую задаче метрику.

Вот основные метрики качества ранжирования:

* MRR (Mean reciprocal rank)
* Precision@ K
* DCG & NDCG
* MAP
* Kendall’s tau
* Spearman’s rho

Рассмотрим основные из них (первые 3). Практически все метрики реализованы в библиотеке Sklearn [1].

* 1. Mean reciprocal rank

Название этой метрики переводится как средний обратный ранг. Ранжирование работает тем лучше, чем ближе к началу выдачи релевантный для пользователя документ. Для каждого запроса найдём позицию первого релевантного документа, возьмём обратное от этого числа и усредним по всем запросам.

(1)

где: — множество запросов

— первый релевантный эдемент

Чтобы проиллюстрировать это, давайте рассмотрим приведенный ниже пример, в котором модель пытается предсказать форму множественного числа английских слов, делая 3 предположения. В каждом случае также дается правильный ответ.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Запрос*** | ***Предсказания модели*** | ***Правильный ответ*** | ***Ранг*** | ***Обратный ранг*** |
| Child | Childs, **Children**, Child | Children | 2 | 1/2 |
| Medium | Medium, Meds, **Media** | Media | 3 | 1/3 |
| Virus | **Viruses**, Viri, Virii | Viruses | 1 | 1 |

MRR в таком случае будет равен:

Одним из ограничений MRR является то, что он учитывает только позицию первого релевантного элемента и игнорирует другие. Это может быть неподходящей метрикой для случаев, когда мы хотим оценить весь список элементов.

* 1. DCG & NDCG

Normalized Discounted Cumulative Gain (NDCG), пожалуй, самая популярная метрика качества моделей ранжирования. В отличие от предыдущих метрик, NDCG учитывает порядок и относительную важность документов и поощряет размещение наиболее релевантных документов как можно выше в списке предсказания.

Прежде чем дать определение NDCG, давайте сначала определим два показателя: Cumulative Gain (CG) и Discounted Cumulative Gain (DCG).

Cumulative Gain (CG) набора элементов представляет собой сумму их оценок релевантности и определяется следующим образом:

где:

— множество элементов, предсказанное моделью.

Предполагается, что релевантность каждого элемента посчитана.

Discounted Cumulative Gain (DCG) — это, по сути, взвешенная версия CG, в которой используется логарифмический понижающий коэффициент, значение которого зависит от позиции элемента в предсказании модели. Это полезно, так как на практике ы хотим придать более высокий приоритет первым нескольким элементам при анализе качества ранжирования. DCG определяется как:

где:

— множество элементов, предсказанное моделью.

— релевантность i-го элемента.

Normalized Discounted Cumulative Gain (NDCG) нормализованная версия метрики DCG. Поскольку предсказанное множество элементов может отличаться по размеру в зависимости от запросов или моделей возникает необходимость нормализовать получившееся значение посчитанной метрики DCG (путем деления на DCG идеальной системы). Можно сказать, что при подсчете идеального IDCG считается DCG элементов, отсортированных по релевантности. Формула, по которой считается метрика:

Где:

— множество элементов, предсказанное моделью.

— DCG «идеального» ранжирования модели.

NDCG является одной из самых популярных метрик, но несмотря на это, имеет и недостатки. Основной ее недостаток в том, что она не штрафует за неприемлемые элементы в получившемся ранжировании.

* 1. Рекурентные нейронные сети

Каждая следующая мысль человека, как правило, связана с предыдущей. Читая какой-либо материал, человек понимает каждое слово, основываясь на своем понимании предыдущих слов. При чтении каждой главы статьи человек держит в голове знания, полученные в предыдущих главах, чтобы лучше понять текущую. Мысли отличаются постоянством.

Обычные нейронные сети не могут эффективно извлекать информацию из прошлого, и это кажется серьезным недостатком. Например, представим, что необходимо классифицировать, какие события происходят в каждом моменте фильма. Неясно, как обычная нейронная сеть могла бы использовать полученную информацию о предыдущих событиях в фильме для классификации последующих.

Рекуррентные нейронные [2] сети решают эту проблему. Это сети с циклами, которые умеют сохранять информацию о прошлом.

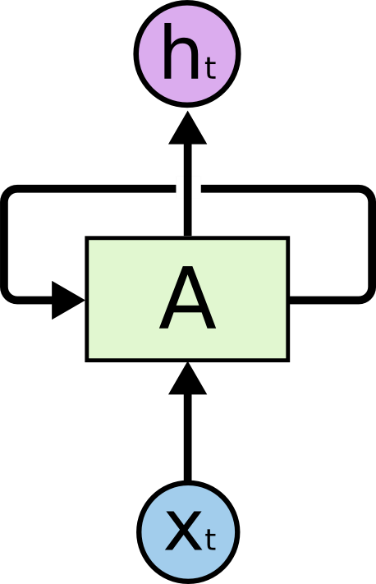


Рисунок 1 — фрагмент нейронной сети

На рисунке 1 представлен фрагмент нейронной сети. Блок A смотрит на входные данные и возвращает выходное состояние нейронной сети . Под блоком в данному случае подразумевается какой-то набор слоев нейронной сети. Цикл над блоком A позволяет нейронной сети получить информацию о предыдущих итерациях.

Рекуррентную нейронную сеть можно рассматривать как множество копий одной и той же сети, каждая из которых передает сообщение следующей. Таким образом, если искусственно развернуть цикл, получится следующая интерпретация:

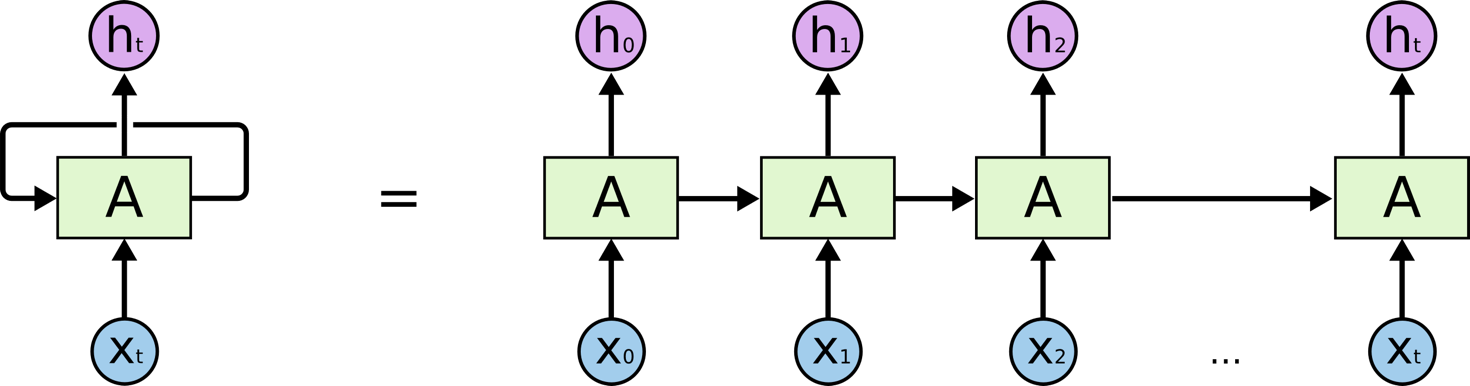


Рисунок 2 — рекуррентная нейронная сеть с развернутым циклом

Схема в виде цепи (Рисунок 2) показывает, что рекуррентные нейронные сети тесно связаны с обработкой последовательностей и списков. Это подходящая архитектура нейронной сети, используемая для обработки таких данных.

Одно из преимуществ рекуррентных нейронных сетей заключается в том, что они могут связать предыдущую информацию с текущей задачей, например, использование предыдущих видеокадров может помочь в понимании текущего кадра. Если бы рекуррентные нейронные сети могли это делать, они были бы чрезвычайно полезны. Но могут ли они? Зависит от контекста.

Иногда, нам необходимо посмотреть лишь на определенный фрагмент в прошлом. Например, рассмотрим языковую модель, которая пытается предсказать следующее слово на основе предыдущих. Если мы пытаемся угадать последнее слово в фразе “облака плывут по *небу*”, нам не нужен дополнительный контекст – совершенно очевидно, что следующим словом будет "небо". В таких случаях, когда разрыв между релевантной информацией и местом, где она необходима, невелик, рекуррентные нейронные сети могут эффективно использовать информацию о прошлом (Рисунок 3).

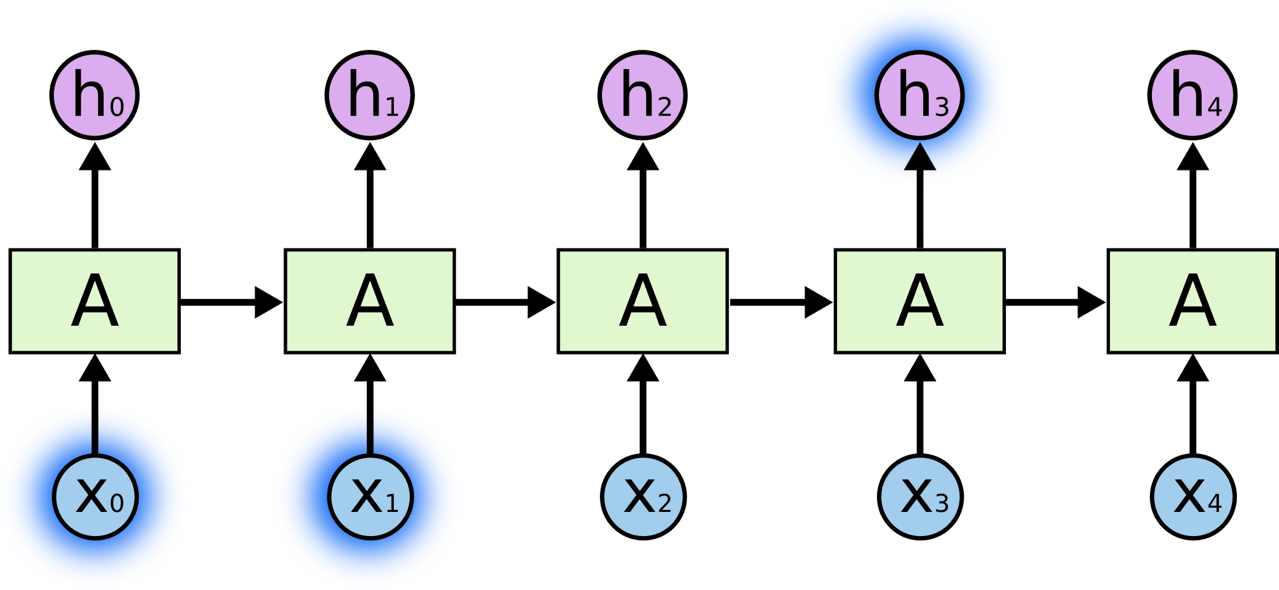


Рисунок 3 — рекуррентная нейронная сеть эффективно использует информацию о прошлом

Но бывают случаи, когда нам нужен более детализированный контекст. Попробуйте угадать последнее слово в тексте: “Я вырос во Франции… Я свободно говорю *по-французски*”. Последняя информация из контекста указывает на то, что следующее слово, вероятно, является названием языка, но если мы хотим уточнить, на каком именно языке, нам нужно не потерять информацию о Франции, которая встречалась в начале текста. Вполне возможно, что разрыв между релевантной информацией и местом, где она необходима, может оказаться слишком большим. К сожалению, по мере того, как этот разрыв растет, рекуррентные нейронные сети становятся неспособными связывать информацию воедино (Рисунок 4).

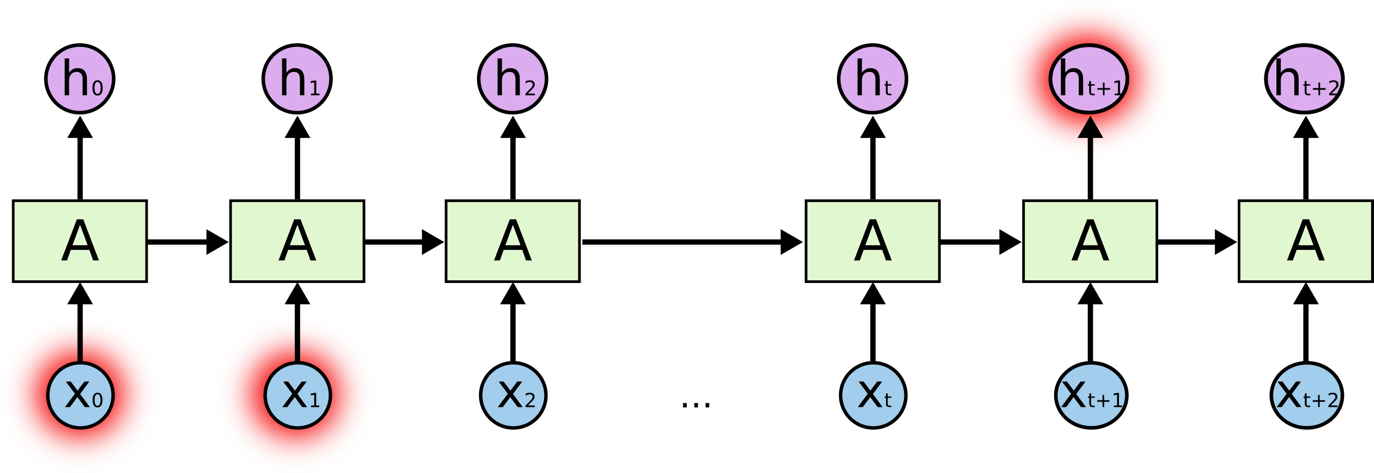


Рисунок 4 — рекуррентная нейронная сеть теряет информацию о прошлом из-за большого разрыва

Существуют другие архитектуры нейронных сетей, которые способны решить эту проблему.

* 1. LSTM

Рекуррентная нейронная сеть архитектуры Long Short Term Memory (LSTM) специально разработана для того, чтобы избежать проблемы отсутствия долговременной памяти. В отличие от обычных рекуррентных сетей, LSTM запоминает информацию, пришедшую спустя длительный период времени.

Все рекуррентные нейронные сети имеют форму цепочки повторяющихся ячеек нейронной сети. В стандартных RNNS эта повторяющаяся ячейка будет иметь очень простую структуру, например, один линейный слой с функцией активации tanh (Рисунок 5).

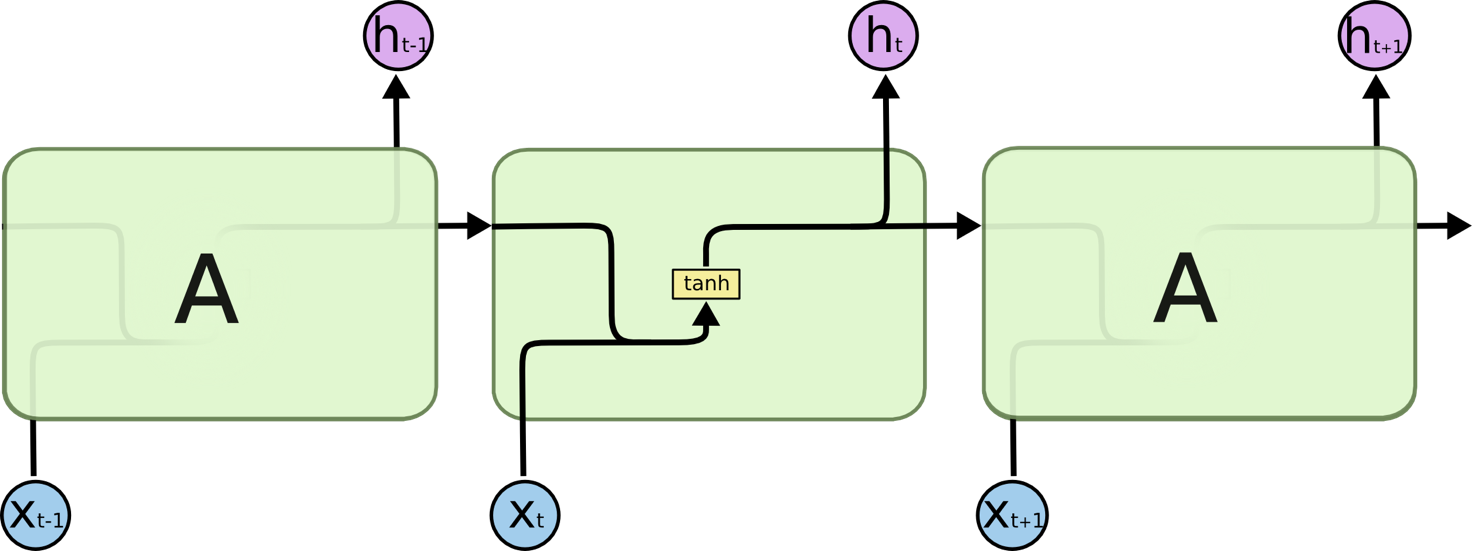


Рисунок 5 — повторяющиеся ячейки в обычной рекуррентной нейронной сети

LSTM также имеют цепочечную структуру, но повторяющаяся ячейка имеет другую структуру. Вместо одного слоя нейронной сети их четыре, которые взаимодействуют между собой особым образом (Рисунок 6).

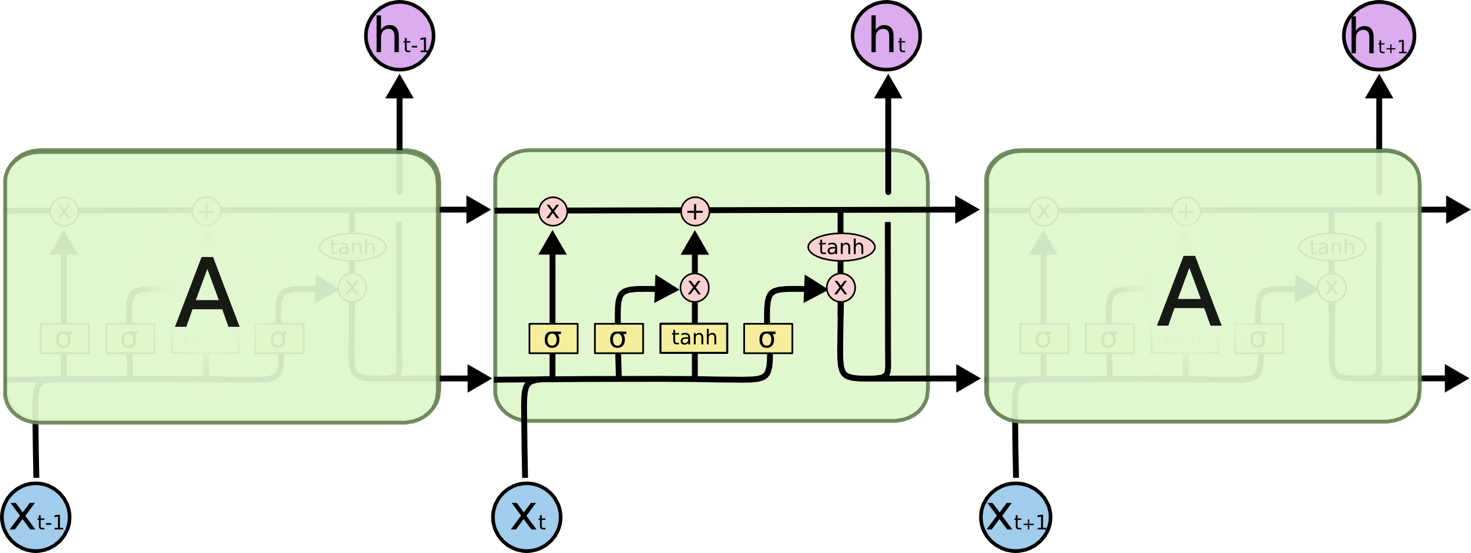


Рисунок 6 — Повторяющиеся ячейки в архитектуре LSTM

На приведенной выше схеме каждый элемент ячейки определяет определенное действие со входным вектором. Розовые кружки обозначают векторные операции, такие как сложение векторов, в то время как желтые прямоугольники обозначают обучаемые слои нейронной сети. Слияние линий означает объединение в единый вектор, в то время как разветвление строк означает, что их содержимое копируется и копии отправляются в разные места (Рисунок 7).



Рисунок 7 — Обозначение элементов ячейки LSTM

Ключевой особенностью LSTM является состояние ячейки (cell state), горизонтальная линия, проходящая через верхнюю часть диаграммы. Состояние ячейки похоже на ленту конвейера. Оно проходит прямо по всей ячейки, лишь с некоторыми незначительными линейными взаимодействиями (Рисунок 8).

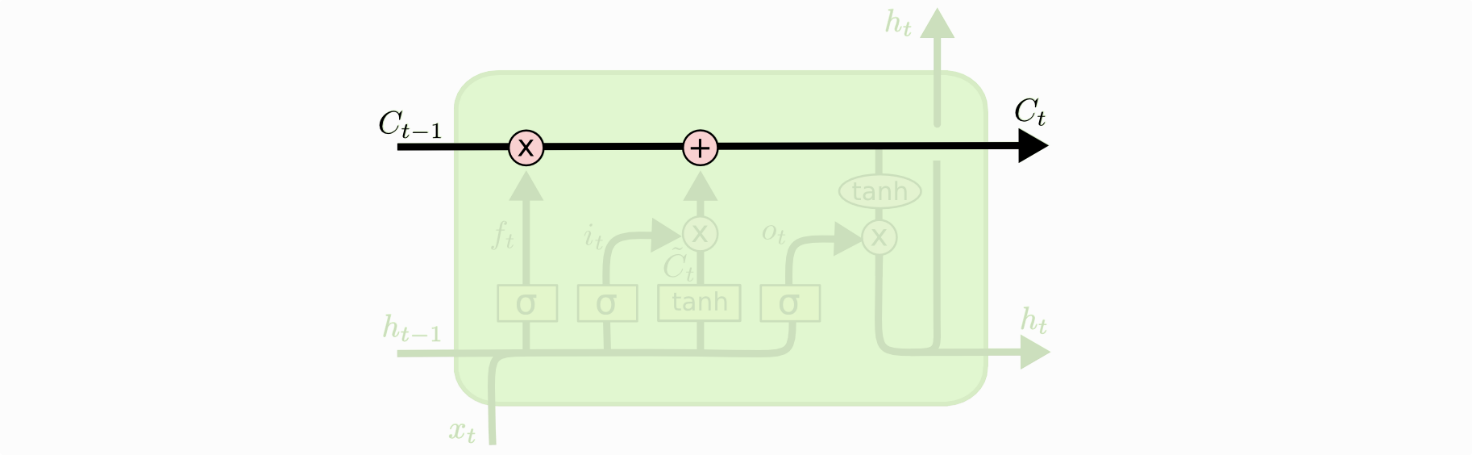


Рисунок 8 — Состояние ячейки LSTM

LSTM может добавлять или удалять определенную информацию из cell state с помощью специальных механизмов, которые называются **gates** (ворота или вентили в русскоязычной литературе).

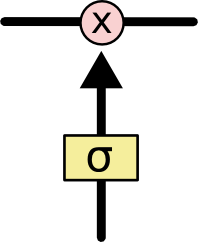


Рисунок 9 — Sigmoid слой гейта

На выходе после слоя с функцией активацией (ФА) Sigmoid получаются числа от нуля до единицы, описывающие, какое количество информации должно пройти к дальше по сети. Значения, близкие к нулю, означают “дальше информацию не пропускать ”, в то время как значения, близкие к единицам, означают “пропускать все!”.

Всего в LSTM 3 гейта, каждый из которых защищает и контролирует текущее состояние ячейки [3].

Первым шагом во время работы LSTM является принятие решения о том, какую информацию мы собираемся удалить из состояния ячейки (Рисунок 10). Это определяет выходная информация после слоя с ФА Sigmoid, который называется «forget gate layer». В этом гейте считывается информация и и вычисляются числа от 0 до 1 для каждого числа из сell state: 0 обозначает полностью забыть это число, 1 обозначает сохранения числа в памяти для дальнейшего использования.

Вернемся к нашему примеру с языковой моделью, которая предсказывает следующее слово на основе всех предыдущих. В такой задаче состояние ячейки может включать род существительного, чтобы можно было использовать правильные местоимения. Когда мы видим новый объект, мы хотим забыть пол старого объекта.



Рисунок 10 — Forget gate

Следующий шаг — решить, какую новую информацию мы будем сохранять в состоянии ячейки (Рисунок 11). Этот шаг состоит из двух частей. Сначала сигмоидальный слой, называемый «input gate layer», решает, какие значения мы будем обновлять. Затем слой с ФА tanh создает вектор новых возможных значений , которые могут быть добавлены в текущее состояние ячейки. На следующем шаге мы объединим эти два параметра, чтобы создать обновление состояния.

В примере с нашей языковой моделью мы хотели бы добавить пол нового субъекта в состояние ячейки, чтобы заменить старый, который мы забыли.

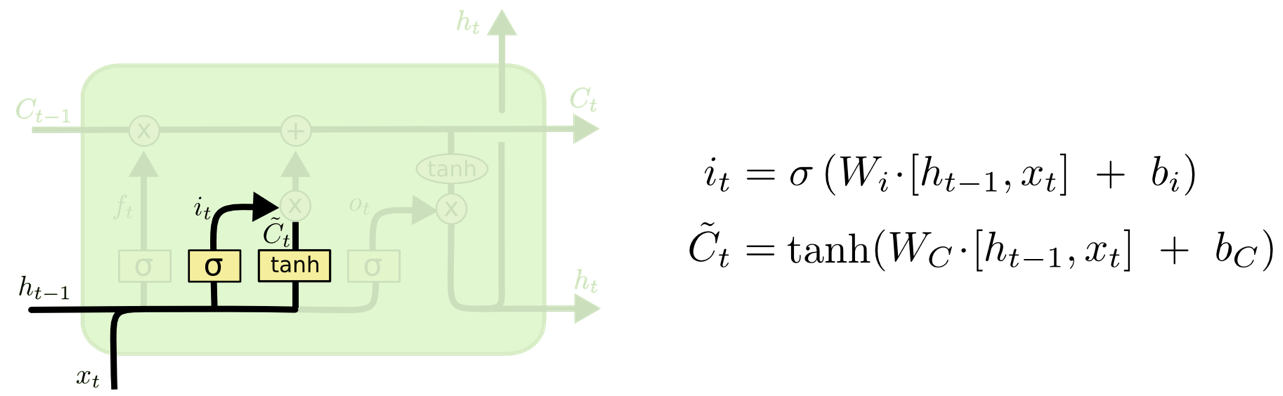


Рисунок 11 — Input gate

Теперь пришло время изменить старое состояние ячейки, , на новое состояние ячейки . На предыдущих шагах уже было посчитано, какую информацию удалить, а какую добавить, остается лишь сделать это.

Мы умножаем старое состояние на , удаляя ненужную информацию, которую решили удалить ранее. Затем мы добавляем *,* Это новые значения-кандидаты, умноженные на коэффициенты, определяющие силу обновления текущего числа скрытого состояния ячейки (Рисунок 12).

В случае с языковой моделью именно здесь мы фактически удалили информацию о поле старого субъекта и добавили новую информацию, посчитанную на предыдущих шагах.

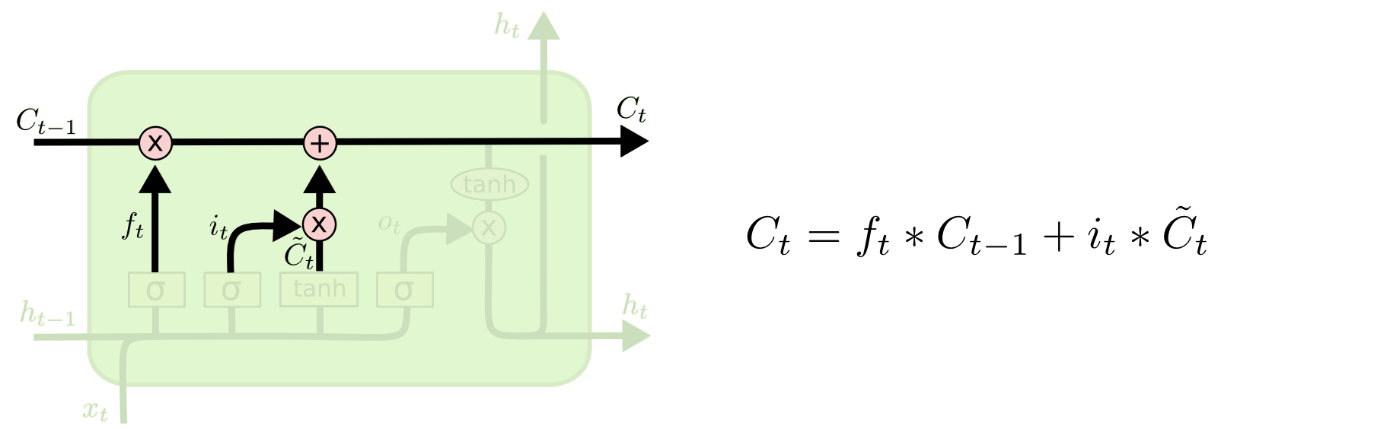


Рисунок 12 — Обновление текущего состояния ячейки LSTM

Наконец, нам нужно решить, что мы будем отправлять в качестве выхода нейронной сети. Этот вывод будет основан на состоянии нашей ячейки, но это будет отфильтрованная версия. Сначала получаем выход сигмоидального слоя, который определяет, какие части состояния ячейки мы будем отправлять на выход. Затем мы пропускаем состояние ячейки через ФА tanh (чтобы значения были между -1 и 1) и умножаем его на выходные данные слоя сигмойды, благодаря этому мы отправляем на выход только нужную часть скрытого состояния ячейки (Рисунок 13).

В примере с языковой моделью, поскольку он только что увидел подлежащее, он может захотеть вывести информацию, относящуюся к глаголу, на случай, если это то, что будет дальше. Например, он может выводить, является ли подлежащее единственным или множественным числом, чтобы мы знали, в какую форму следует спрягать глагол.

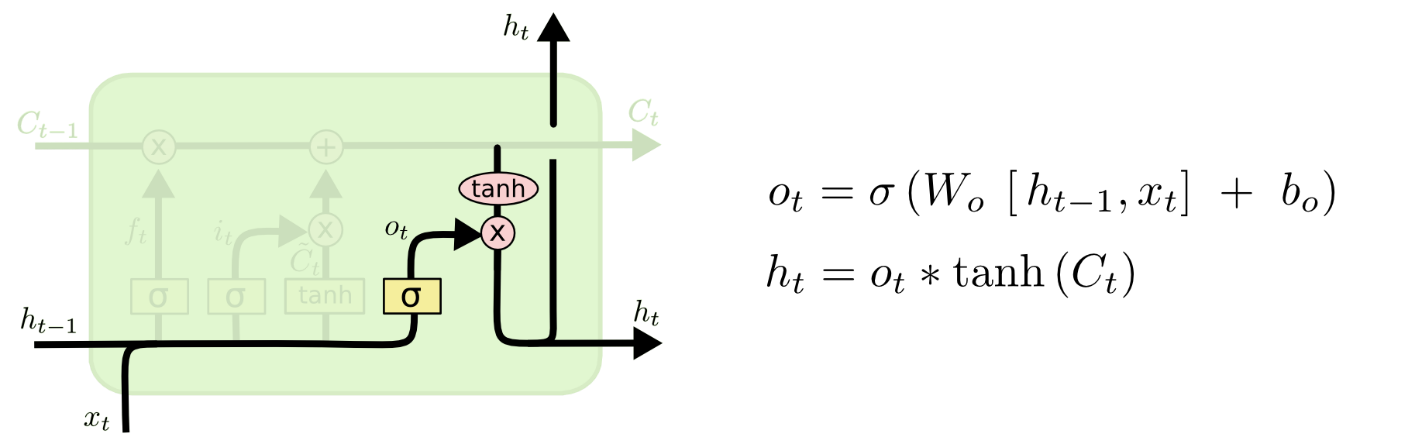


Рисунок 13 — Output gate

* 1. Грандиентные бустинг

В реализации процесса подсчета метрики скорости реакции одним из ключевых алгоритмов является градиентный бустинг. Расскажу подробнее суть данной модели машинного обучения.

Бустинг в машинном обучении воплощает идею последовательного построения линейной комбинации алгоритмов. Каждый следующий алгоритм стремится уменьшить ошибку текущего ансамбля моделей машинного обучения.

Градиентный бустинг — метод машинного обучения, который создает решающую модель прогнозирования в виде ансамбля слабых моделей прогнозирования, обычно деревьев решений. Он строит модель поэтапно, позволяя оптимизировать произвольную дифференцируемую функцию потерь. Этот алгоритм, чаще всего, показывает наивысшее метрики качества в задачах, где необходимо по табличным данным выполнить какую-либо задачу машинного обучения: чаще всего, классификацию или регрессию. Такой алгоритм способен эффективно находить нелинейные зависимости в данных различной природы.

Суть градиентного бустинга в последовательном обучении слабых самих по себе алгоритмов машинного обучения (неглубоких деревьев), каждый из которых уменьшает ошибку предыдущих. Достигается это засчет того, что каждая следующая модель обучается на антиградиент от функции ошибки текущего ансамбля, засчет чего с каждой итерацией обучения следующей модели оптимизируемая функция начинает приближаться с более высокой точностью. Визуализировать ансамбль ML моделей, в котором каждая следующая модель уменьшает ошибку предыдущих, можно следующим образом:

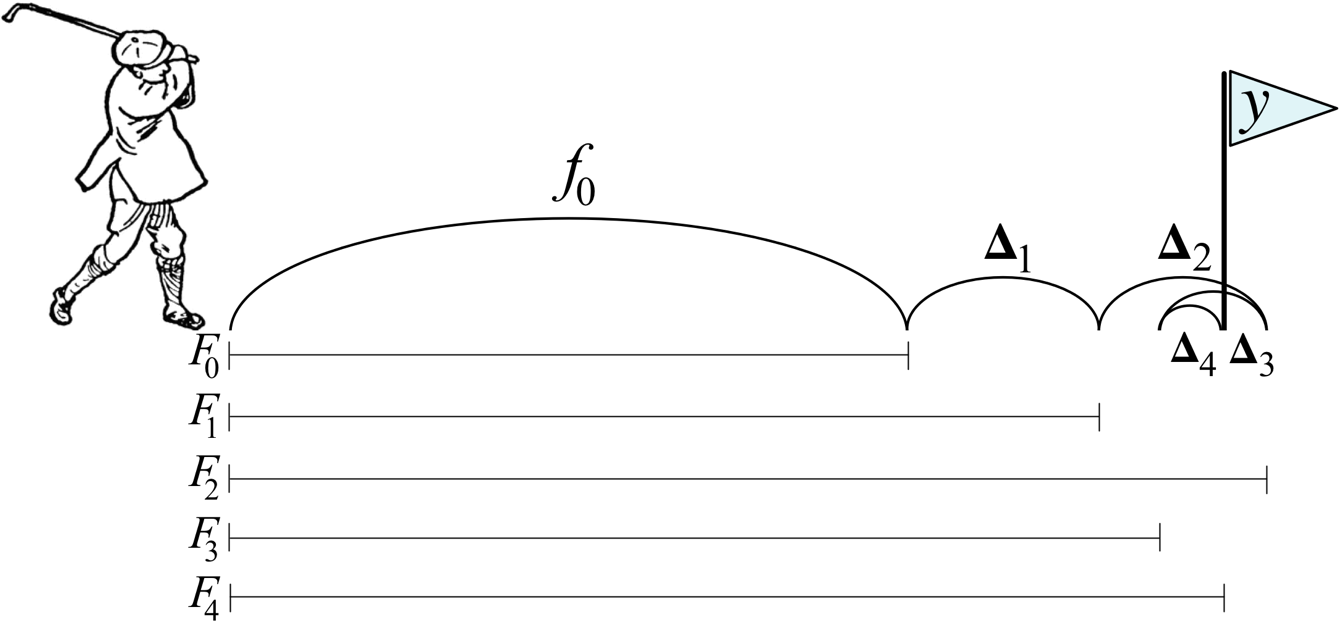


Рисунок 14 — обучении ансамбля, при котором каждая следующая модель уменьшает ошибку уже обученных моделей на предыдущих шагах

Градиентный бустинг играет важную роль в вычислении метрики скорости реакции, ведь именно благодаря нему определяются запросы, связанные с внезапными новостными событиями, на которых в дальнейшем и будет подсчитана целевая метрика.

* 1. Случайный лес

Случайный лес ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) random forest) — метод [машинного обучения](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), заключающийся в использовании ансамбля [решающих деревьев](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9). Основная идея заключается в использовании большого [ансамбля](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D1%81%D0%B0%D0%BC%D0%B1%D0%BB%D1%8C_(%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5)) [решающих деревьев](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9), каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

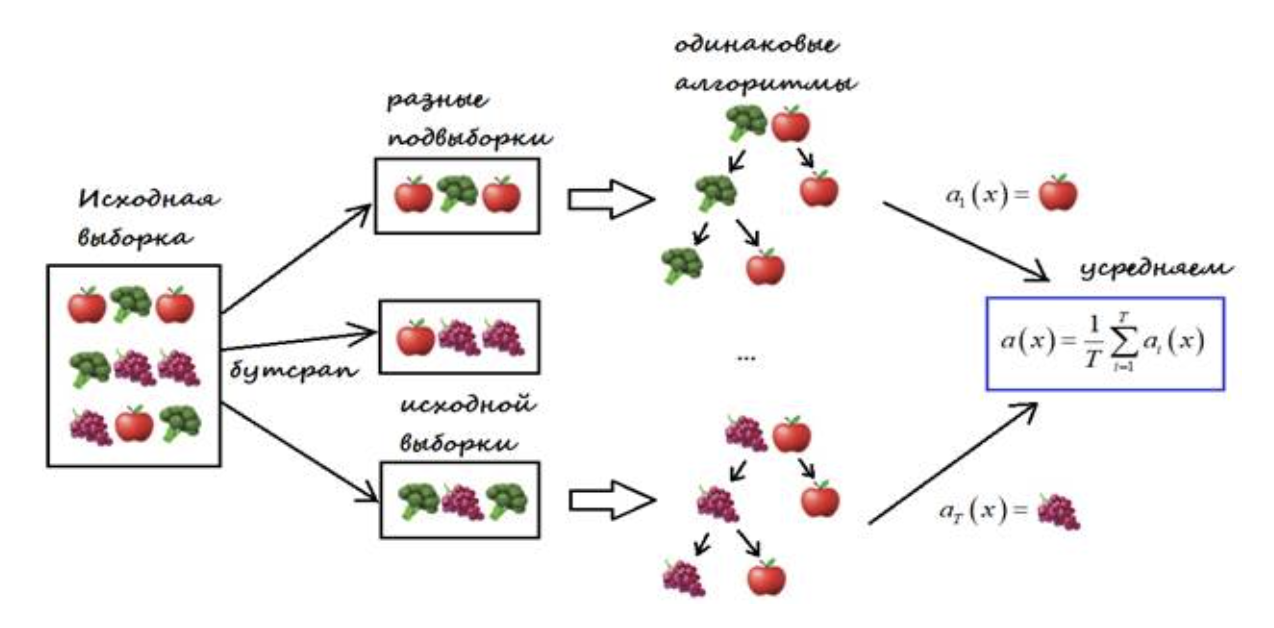


Рисунок 15 — простая схема случайного леса

При обучении случайного леса определяется количество деревьев из которых будет состоять ансамбль. Ошибка обобщения ответов машинного обучения уменьшается с уменьшением ошибки ответов отдельных деревьев и снижением корреляции между деревьями. Одним из способов снижения корреляции ответов между деревьями является обучение этих деревьев на разных данных. Но как из одного датасета сделать много различных датасетов? На помощь приходит статистический метод взятия выборки данных из генеральный совокупности — бутстрэп. Бутстрэп — это метод оценки различных статистик генеральной совокупности засчет усреднения полученных статистик из выборок, полученных с помощью случайного выбора элементов с повторением из этой генеральной совокупности.

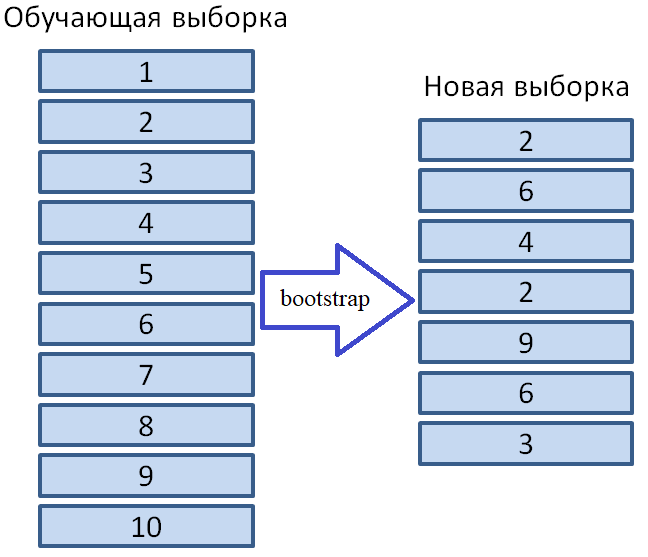


Рисунок 16 — схема бутстрэпа

Таким образом каждое дерево в ансамбле обучается на разной выборке данных. Чтобы таже понизить корреляцию предсказаний обученных деревьев применяют так называемый метод случайных подпространств. Суть его заключается в том, чтобы использовать случайную выборку признаков на каждом этапе выбора решающего правила для разбиения вершины случайного дерева. Таким образом деревья становятся еще более случайными и разнообразными по своей структуре, но засчет максимальной глубины они все еще остаются способными идеально выучить обучающую выборку. Устойчивость выходного значения ансамбля будет определяться усреднением независимых ответов от каждого дерева. засчет усреднения ответов каждой из модели.

1. КОНСТРУККТОРСКО-ТЕХНОЛОГИЧЕСКАЯ ЧАСТЬ
   1. Описание поисковых запросов, на которых будет подсчитана метрика

Задача Метрики скорости реакции — измерить качество поисковой

выдачи на запросах, который задавались сразу после наступления новостного события. Таким образом мы можем сформулировать следующие критерии для подходящего запроса (ниже будут даны соответствующие пояснения критериям):

1. С запросом должно быть связано какое-либо новостное событие. Новостное событие должно удовлетворять следующим критерием:
   1. Событие по запросу не является запланированным.
   2. Для объекта запроса произошедшее событие значимо.
2. За день до запроса не должно происходить значимых событий, подходящих под формулировку запроса.
3. Запрос должен быть взят из потока, т.е. являться настоящим существующим запросом, который пользователь задал в какое-либо время в Яндекс Поиск.

Запрос, отвечающий критериям выше будем называть «***Ультрасвежим***». Примеры незапланированных и значимых событий:

* ЧП
* Катаклизм
* Неожиданные заявления
* Происшествия
* Неожиданные смерти

Примеры запланированных событий (запросы, связанные с данными событиями не являются ультрасвежими):

* Телепередача/эфир/стрим
* Спортивный матч
* Премьера фильма / выход новой серии
* Конференции / заседания

Примеры незначимых событий (запросы, связанные с данными событиями не являются ультрасвежими):

* Частые непланируемые события
* Незаметные новости с фронта
* Регулярные заявления

Такие запросы на начальном этапе будут определяться асессорами. Асессоры — это специалисты, чья работа, как правило, связана с оценкой и анализом качества какого-либо контента, процедур или процессов. Одной из наиболее известных функций асессоров является оценка качества поисковых систем. В данном контексте асессоров часто называют "поисковыми асессорами". Их основная задача заключается в том, чтобы оценивать качество результатов поиска, предоставляемых пользователю, и улучшать алгоритмы сортировки контента. Когда вы вводите запрос в поисковую систему, вы хотите получить наиболее релевантные и полезные результаты. Однако, несмотря на сложные автоматические алгоритмы, используемые поисковыми системами, человеческий фактор остается важным для определения точности и полезности информации. Асессоры анализируют результаты и дают свою оценку, указывают на проблемы и предлагают возможные улучшения. Они работают по специальным гайдлайнам (руководствам), которые описывают, как оценивать различные типы контента и запросов.

В дальнейшим планировалось обучить модель, которая по временному ряду задавания запроса смогла бы классифицировать запрос на ультрасвежесть.

Для любых самых громких новостных событий характерно следующее поведение: сразу после наступления такого события, количество поисковых запросов, связанных с этим событием, резко возрастает. С какого-то момента ажиотаж вокруг события начинает затухать и динамика запросов по событию падает, ведь люди все меньше и меньше интересуются таким событием.

Проиллюстрируем динамику таких запросов:

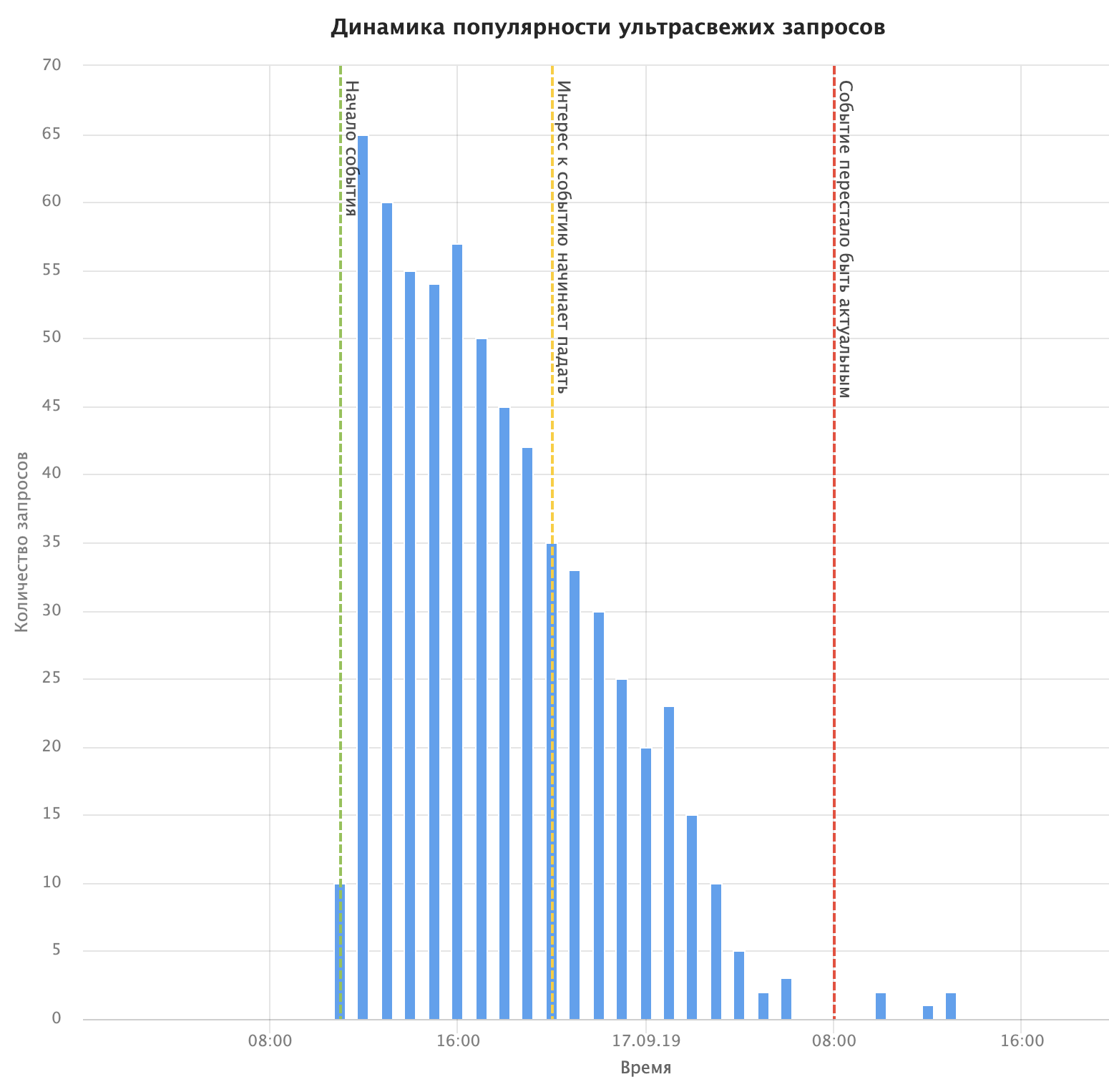


Рисунок 17 — Динамика популярности ультрасвежих запросов

По графику (Рисунок 14) видна отличительная особенность всех ультрасвежих запросов — всплеск популярности и плавный количества поисковых запросов по событию. Чем строже будет соблюдаться такой вид динамики запроса, тем легче в дальнейшем их будет классифицировать модели. Поэтому было решено не рассматривать запланированные события, ведь, поскольку о событии будет известно заранее, люди будут им интересоваться и до его наступления и выделить их среди потока с необходимой точностью не получится. Дополнительным условием для запроса является значимость события. Если событие незначимо, то оно не вызовет резонанс общественности, характерный всплеск не будет наблюдаться и моделе будет сложнее такой запрос верно классифицировать. Чтобы наблюдался резкий всплеск запроса также необходимо, чтобы за день до по данному запросу не было релевантных событий.

По нашим предположениям, если учтены нюансы, о которых сказано выше, то модель сможет с достаточной точностью классифицировать запросы на ультрасвежесть и тогда затраты на необходимый объем разметки асессорами будут укладываться в бюджет компании на соответствующее полугодие.

* 1. Запуск разметки на ультрасвежесть

После анализа природы необходимых запросов и формализованного описания термина ультрасвежести необходимо набрать необходимое количество асессоров и составить проект разметки в системе Yang.

Yang — краудсорсинговый проект для разметки большого количества данных, которые затем, как правило, используются для машинного обучения.

Асессоры должны были ознакомиться с инструкцией, пройти составленные мной обучающие и экзаменационные задания. Асессоры успешно справившееся с экзаменом допускаются до разметки реальных данных (Рисунок 15).

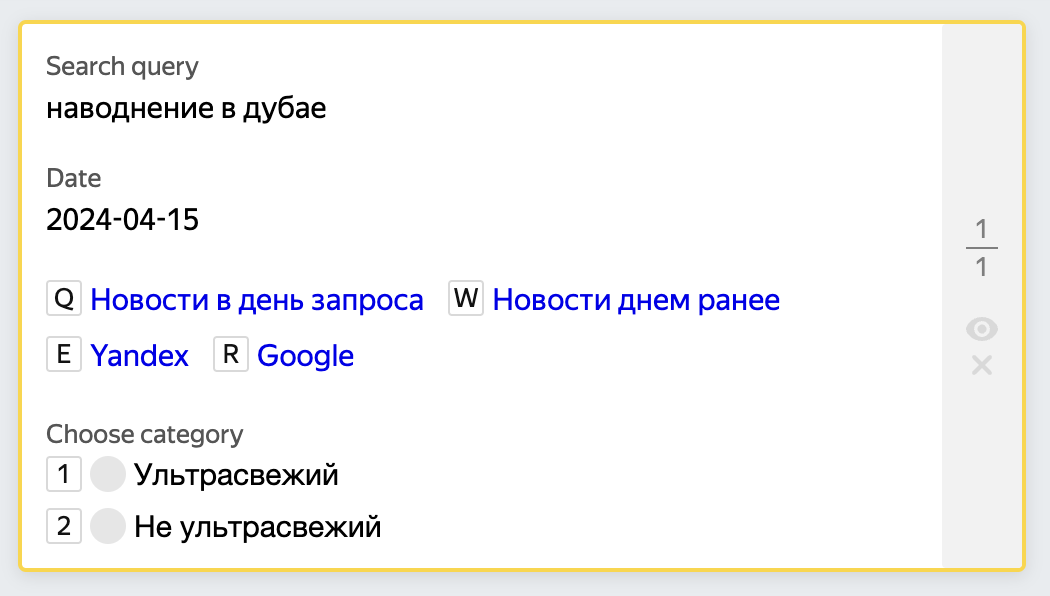


Рисунок 18 — Интерфейс асессорской разметки

* 1. Ансамбль для классификации ультрасвежих запросов

После получения необходимого количества разметки необходимо понять,

как на этом датасете обучить модель для классификации ультрасвежих запросов.

В нашем распоряжении были логи самого поиска и вектор признаков запроса, который автоматически считается Яндекс Поиском. Для классификации необходимо было обработать динамику запроса в соответствующий день (пример Рисунок 14), а также использовать уже имеющийся посчитанный вектор признаков, который описывает смысл запроса.

Было решено использовать LSTM для обработки временных рядов динамики запроса, т.к. эта модель хорошо себя показывает на работе с последовательностями, и ее архитектура не такая сложная, как, например, BERT, чтобы быстро переобучиться на нашем небольшом датасете (асессорский ресурс — небесплатный). Одна из ключевых причин, по которым LSTM хорошо работает для классификации временных рядов, заключается в ее способности запоминать важную информацию на длительные промежутки времени и забывать неважные данные.

Для финальной классификации поверх LSTM было решено использовать CatBoost т.к. эта модель, как правило, показывает самые высокие результаты на табличных данных. В нашем случае данными для классификации будут признаковое описание запроса и получившееся предсказания LSTM.

Одной из ключевых причин высокой эффективности градиентного бустинга на табличных данных является его способность естественно обрабатывать разнородные признаки (числовые, категориальные и т.д.) без необходимости масштабирования или нормализации, что снижает потребность в предобработке данных. Он также эффективно выявляет нелинейные зависимости и сложные взаимодействия между признаками, что трудно учесть линейными моделями. Градиентный бустинг устойчив к выбросам и шуму, поскольку деревья решений могут уменьшить влияние таких аномалий. Более того, он хорошо работает с большим количеством признаков, строя сильные предсказательные модели даже при наличии значительного числа фич. Этот подход масштабируется и стабильно работает на больших наборах данных, предоставляя высокую производительность и интеграцию с параллельными вычислениями, что делает его эффективным для промышленных приложений. Хотя итоговая модель градиентного бустинга может быть сложнее для интерпретации, методы объяснения, такие как "важность признаков", предоставляют понимание, какие признаки наиболее важны для предсказания, что критично в бизнесе. В итоге, способность обрабатывать разнородные признаки, выявлять сложные зависимости, быть устойчивым к шумам и выбросам, эффективно масштабироваться и обеспечивать интерпретацию результатов делает градиентный бустинг превосходным выбором для задач, связанных с табличными данными.

Итоговый пайплайн можно проиллюстрировать следующим образом (Рисунок 16). Данный подход позволяет учитывать данные анализа временного ряда (выход LSTM) и веткорное представления запроса на основе нейросетевых факторах Яндекс Поиска. Благодаря чему достигается наилучшая эффективность классификации запроса.

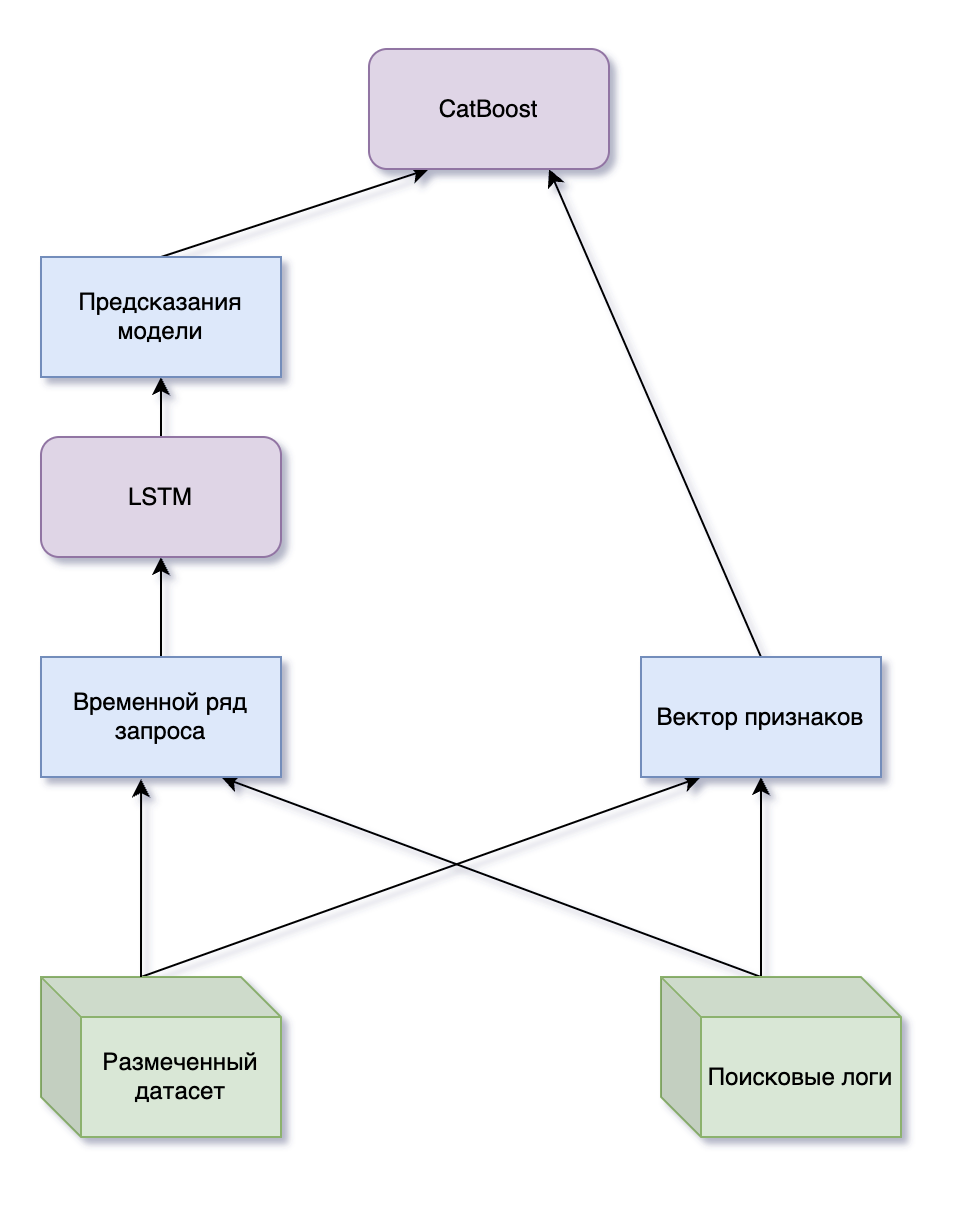


Рисунок 19 — Пайплайн классификации на ультрасвежесть

* 1. Обучение LSTM

По временному ряду необходимо предсказать вероятность отнесения данного запроса к ультрасвежим [5-11].

Пример временного ряда ультрасвежего запроса:

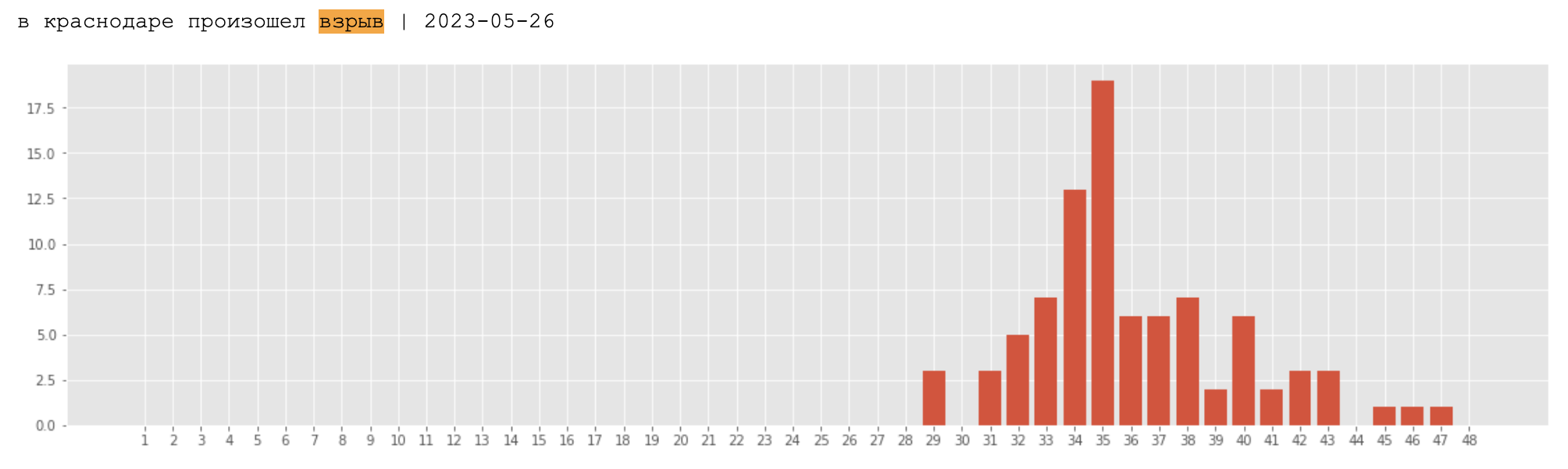


Рисунок 20 — временной ряд ультрасвежего запроса

Пример временного ряда новостного, но не ультрасвежего запроса про заранее спланированное спортивное событие (видим отсутствие всплеска):

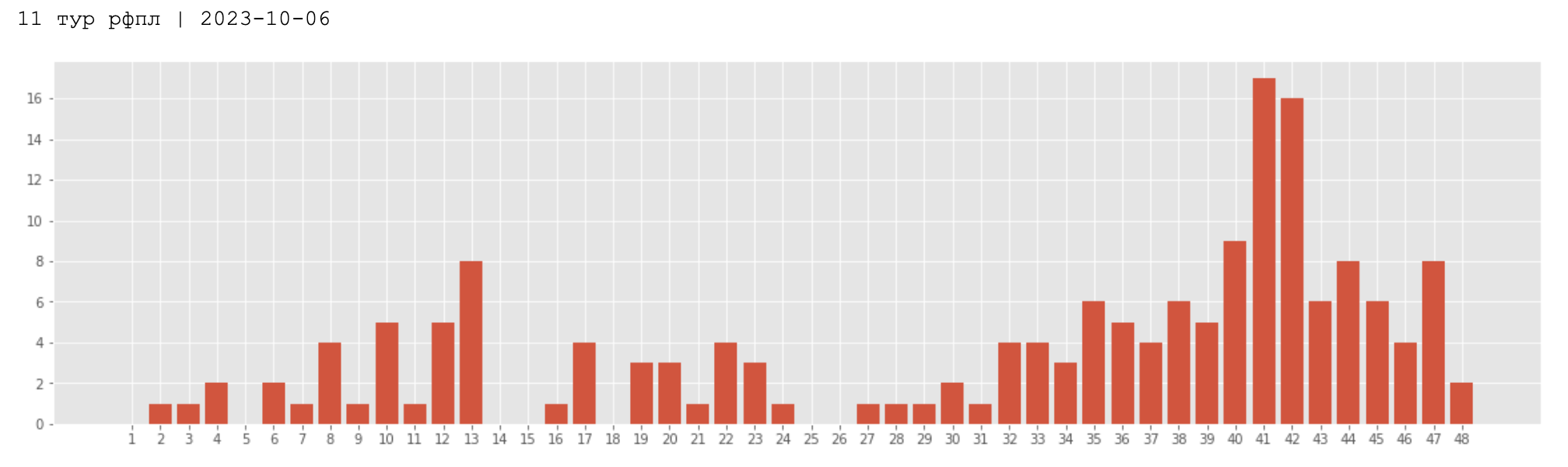


Рисунок 21 — временной ряд запроса, связанного с заранее спланированным событием

Такой запрос не будет являться ультрасвежем по нашему определению, так как о событии заранее известно.

Нормализовав такой временной ряд можно попробовать обучить первую модель нашего ансамбля для классификации запроса — LSTM. Ниже представлен листинг кода основного класса модели.

class LSTMClassifier(nn.Module):

"""Implementation of LSTM-based time-series classifier."""

def \_\_init\_\_(self, input\_dim, hidden\_dim, layer\_dim, output\_dim):

super().\_\_init\_\_()

self.hidden\_dim = hidden\_dim

self.layer\_dim = layer\_dim

self.rnn = nn.LSTM(input\_dim, hidden\_dim, layer\_dim, batch\_first=True)

self.fc = nn.Linear(hidden\_dim, output\_dim)

self.batch\_size = None

self.hidden = None

def forward(self, x):

h0, c0 = self.init\_hidden(x)

out, (hn, cn) = self.rnn(x, (h0, c0))

out = self.fc(out[:, -1, :])

return out

def init\_hidden(self, x):

h0 = torch.zeros(self.layer\_dim, x.size(0), self.hidden\_dim)

c0 = torch.zeros(self.layer\_dim, x.size(0), self.hidden\_dim)

return [t.cuda() for t in (h0, c0)]

В качестве оптимизитора градиентного спуска использовался Adam. Adam (Adaptive Moment Estimation) считается одним из лучших оптимизаторов в машинном обучении благодаря своей способности адаптивно изменять скорость обучения отдельных параметров модели. Эта адаптивность достигается путем комбинирования преимуществ двух других популярных методов — AdaGrad и RMSProp. Adam вычисляет адаптивные скорости обучения для каждого параметра, используя моменты первого и второго порядка (средние значения градиентов и их квадратов). Он корректирует шаги обучения, стабилизируя и ускоряя процесс оптимизации, что особенно полезно при работе с большими и шумными данными. В дополнение к этому, Adam требует минимальной настройки гиперпараметров и обеспечивает быструю сходимость, что делает его предпочтительным выбором для широкого спектра задач машинного обучения, включая глубокие нейронные сети.

Поверх выхода нейронной сети применялась функции активации сигмойды т.к. решаем задачу бинарной классификации.

Основные параметры модели:

* Размер скрытого полносвязного слоя идущего сразу после LSTM — 256 (hidden\_dim в коде).
* Количество эпох — 100
* Learning rate — 0.0005

При обучении использовался шедулер, засчет которого можно было циклически изменять learning\_rate, чтобы не застревать в точках локального минимума и с наилучшем приближением придти к глобальному экстремуму. Ниже представлен график коэффицента, на который с течением времени будет убножаться learning rate:

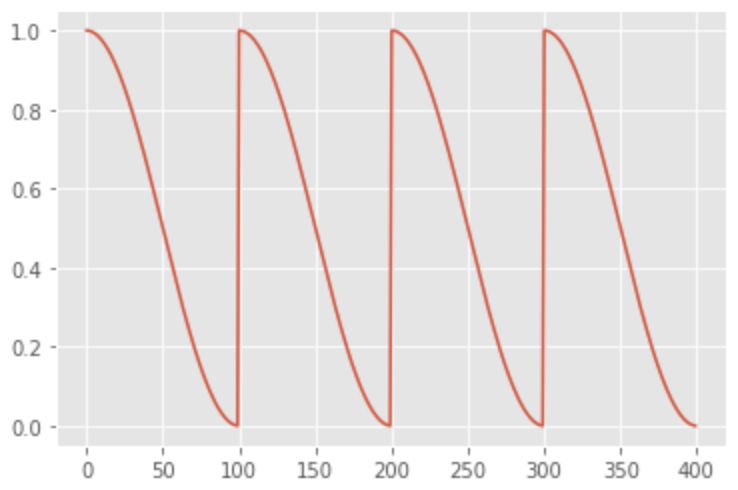


Рисунок 22 — коэффициент умножения скорости обучения

Итоговые метрики при пороге 0.403 при бейзлайне (предсказание ультрасвежести на всех запросах) precision = 0.05, recall = 1:

|  |  |  |
| --- | --- | --- |
| Метрика | Precision | Recall |
| Значение | 0.14 | 0.96 |

Стоит отметить, что бизнес требованием к данной задаче являлась высокая полнота классификации запросов [4]. Идея в том, чтобы как можно более честно брать ультрасвежие запросы из потока, без смещений на модели. В случае низкого recall часть запросов бы терялось, потому что их не различила бы модель, а возможно именно на этих запросах у поисковой системы низкое качество.

Учитывая, что раньше лишь 0.05 запросов по разметке было ультрасвежими делаем вывод, что засчет данной модели необходимо в 3 раза меньше асессорской разметки, чтобы набрать такое же количество ультрасвежих запросов за определенный день. Поэтому обучение модели можно считать успешным.

* 1. Обучение градиентного бустинга

В качестве фреймворка градиентного бустинга был выбран CatBoost, на данный момент это лучший технологический фреймворк, который позволяет гибко обучить модель и настроить визуализацию обучения, необходимую тебе.

Датасет на котором обучалась модель содержал разные нейросетевые признаки запроса такие как: вероятность отнесения запроса, к новостным или спортивным, является ли запрос медицинским и т.д. Всего более 300 колонок с разными текстовыми признаками. Временной ряд каждого запроса был обработан LSTM, ее выход был добавлен в качестве дополнительного признака.

За отложенное нам время нам удалось разметить почти 20 тысяч разных запросов. В датасете была всего 1 тысяча ультрасвежих запросов, поэтому необходима была стратификация по предсказываемому параметру перед делением данных на тренировочную и тестовую части. Несмотря на стратификацию обучение катбуста и его финальные метрики очень сильно зависили от рандома разбиения датасета на тестовую и тренировочную части. Становилось ясно, что необходимо кросс-валидация — метод оценки производительности модели, состоящий в том, чтобы разбить обучающую выборку на несколько фолдов (частей), проверить качество модели на каждом фолде, используя остальные фолды для обучения, и усреднить результаты. Результаты оценок на всех фолдах затем усредняются, чтобы получить общую оценку производительности модели. Этот метод позволяет максимально использовать имеющиеся данные и обеспечивает более надёжную и стабильную оценку качества модели.

Т.к. выборка несбалансирована, метрика точности классификации не подходит т.к. ее значение будет не информативным, ведь точность бейзлайна (предсказание константы 0) достигало бы уже 95% т.к. в исходном датасете всего 5% ультрасвежих запросов.

Продуктово нам было необходимо, чтобы модель как можно меньше пропускала ультрасвежих запросов для более точного и объективного подсчета метрики, поэтому было принято решение максимизировать precision при recall >= 0.95.

Ниже представлен график обучение градиентного бустинга на каждом из фолдов кросс-валидации (всего их было 50). Несмотря на то, что на графике отображено большое количество линий, засчет усреднения каждой из значений кривой при определенных количествах обученных деревьях мы можем получить обобщенное значение метрики при указанном количестве обученных деревьях.

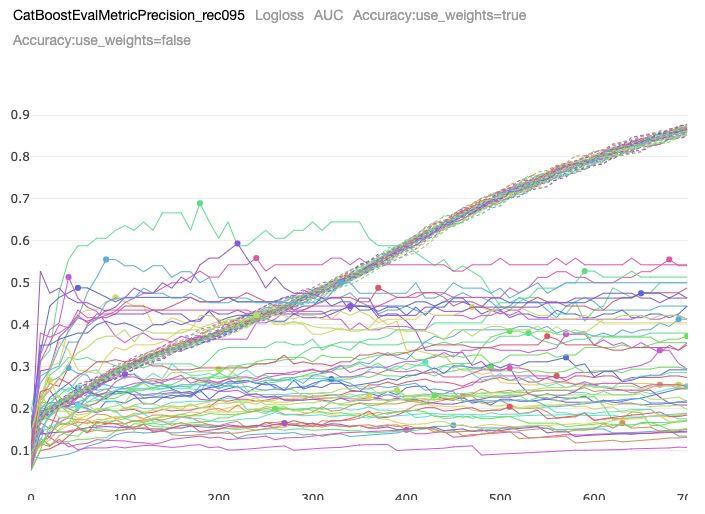


Рисунок 23 — график precision модели при recall >= 0.95 на каждом фолде кросс-валидации

Такие графики генерирует фреймворк CatBoost при указании необходмой конфигурации обучения. По оси Ox показывается количество обученных деревьев по оси Oy значение выделенной жирным метрики (на Рисунке 20 выбрана самописная метрика — значение precision при кecoll не меньше 0.95). Сплошной линией показывается значение метрик на тестовом датасете, пунктирной — на тренировочном. Точкой отмечено количество деревьев в градиентном бустинге при котором значение метрик максимальной на данном фолде. Как только значение метрик на тестовом множестве перестает падать градиентный бустинг переобучается и обучение следует остановить, но на каждом фолде необходимо обучить разное количество деревьев, прежде чем модель начнет переобучаться, поэтому было зафиксировано, что на каждом фолде мы последовательно обучаем 700 деревьев.

Ниже представлен график метрики LogLoss, где каждый цвет также определяет определенный фолд кросс-валидации, обозначения, описанные выше для предыдущего графика также верны. Данный график также позволяет строить фреймворк CatBoost, необходимо лишь задать нужные параметры для его построения. График наглядно показывает, как сильно итоговое значение метрики качества зависит от разбиения, это лишний раз подтверждает необходимость кросс валидации.

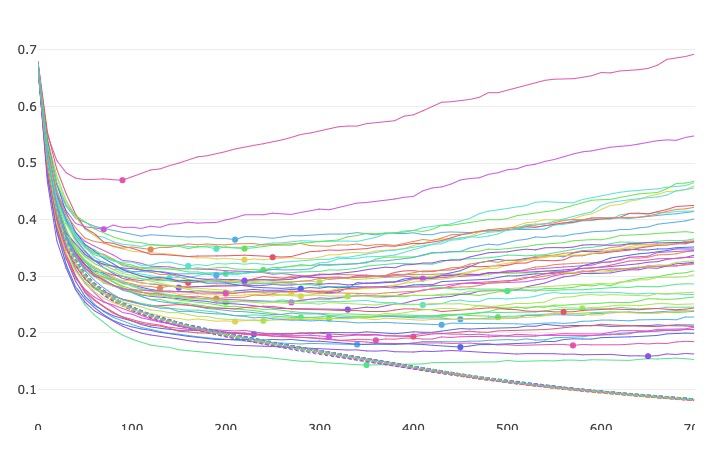


Рисунок 25 — значение LogLoss на каждом фолде кросс-валидации

Чтобы выбрать итоговое количество деревьев необходимо выбрать количество итераций обучения деревьев, при котором среднее значение целевой метрики precision (Рисунок 20) больше усредненного значения метрики на любом другом количестве итераций обучения деревьев.

Итоговые метрики градиентного бустинга:

|  |  |  |
| --- | --- | --- |
| Метрика | Precision | Recall |
| Значение | 0.3 | 0.95 |

Взятое количество деревьев — 410, глубина каждого дерева — 5. Порог при котором посчитаны метрики — 0.29 (градиентный бустинг предсказывает вероятность класса ультрасвежести для запроса).

Сравнивая разные подходы предсказания были получены следующие метрики (recall в каждом случае был не меньше 0.95):

1. LSTM, обученная на временном ряде запроса показывала precision = 0.15.
2. CatBoost, обученный на нейросетевых признаках, используемых в продакшене Яндекс Поиска показал precision = 0.25.
3. Если добавить предсказания LSTM, как дополнительный признак и обучить CatBoost, то precision = 0.3, что является наилучшим показателем.
   1. Поиск начала события

После того, как мы научились определять запросы, на которых хотим замерять качество необходимо реализовать итоговый процесс подсчета метрики. Для каждого ультрасвежего запроса необходимо определить момент начала этого события, точнее, момент, когда по данному событию был создан первый релевантный документ. Начиная с этого момента, мы будем считать, что миру стало известно о данном событии, а значит, если это событие громкое, то в скором времени ожидается пик запросов по данному событию, на таких запросах мы хотим померить качество.

Чтобы определить, когда был создан первый релевантный документ к этому запросу необходимо взять из базы индексированных страниц поиска страницы, созданные в день запроса, а далее в автоматическом режиме определить, какие из этих страниц релевантны запросу. Эту задачу мы реализовали с помощью архитектуры BERT нескольких типов: запросный BERT, и документный BERT. На выходе каждой модели мы получали вектор признаков, описывающий запрос и страницу соответственно. Значение скалярного произведения будет являться мерой схожести запроса и документа (Рисунок 22). Подобрав порог скалярного произведения, можно автоматически отбирать страницы, релевантные запросу.



Рисунок 26 — архитектура процесса, определяющего схожесть между запросом и страницей.

Страницы, релевантные запросу отправляются на запросно-документную асессорскую разметку (Рисунок 22), где асессоры отвечают на вопрос «Содержит ли данная страница актуальную новость по указанному запросу?».

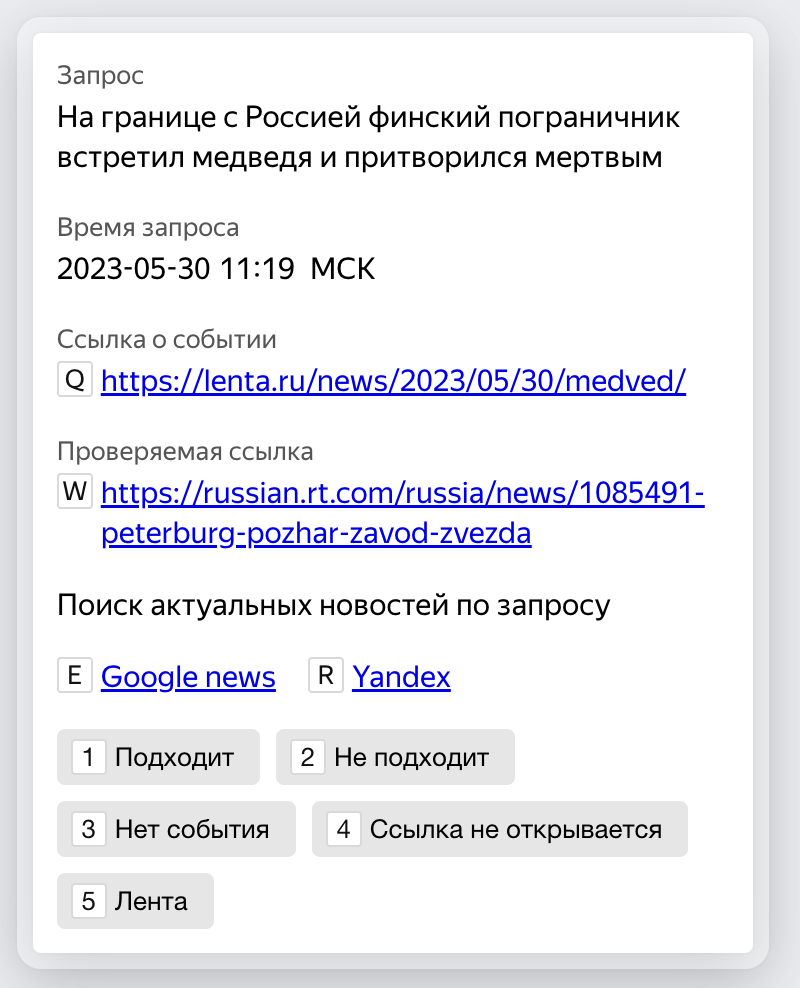


Рисунок 27 — асессорская разметка на соответствие страницы ультрасвежему запросу

Таким образом, разметив асессорами первые 100 созданных релевантных по BERT-у страниц мы сможем найти страницу, которая первая описала произошедшее событие. Время создания такой страницы и будет временем начала события.

* 1. Итоговый подсчет метрики

Итак, у нас есть ультрасвежий запрос, с ним связанно какое-то внезапное громкое новостное событие, мы знаем, в какой момент времени появилась первая новость, описывающая это событие. После момента начала данного события в первые два часа обычно происходит пик запросов пользователей, интересующихся данной новостью. Нам необходимо каким-то образом померить качество выдачи этих запросов в первые два часа после происшествия события.

В Яндексе есть множество внутренних замеров и метрик, измеряющих релевантность выдачи. Мы знаем, что показатель релевантности выдачи Яндекс Поиска очень высокий и на запрос «милые котики» мы не будем показывать собачек или не милых котиков. Поэтому можно априори считать, что первые топ-5 страниц по запросу имеют высокую релевантность, если запрос не являлся бредом или случайным набором букв. Поэтому, зная время начала события по запросу и зная время создания показанных страниц, можно оценить их актуальность: страница в топ-5 выдачи по запросу будет актуальна, если она будет создана после момента начала события. Например: по запросу «землетрясение в Турции», заданному 6 февраля 2023 году можно показать страницу из википедии с землетрясением 2014 года, эта страница будет релевантной, но не актуальной. Пользователь хочет увидеть информацию про актуальное ЧП, ведь запрос был задан именно в день трагедии, а значит актуальная страница должна быть создана после события.

Итак, зная время происшествия события по запросу, время создания страниц на выдачи, мы можем определить актуальность новостной выдачи, как долю страниц в топ-5, созданных после происшествия события (Рисунок 24).



Рисунок 28 — подсчет метрики на одном запросе

Итоговые процесс вычислений после получения множества ультрасвежих запросов можно визуализировать следующим образом:



Рисунок 29 — процесс вычисления метрики после получения множества ультрасвежих запросов

Таким, образом, если доля актуальной выдачи в Яндекс Поиске в первые два часа после огласке в СМИ события — низкая, значит Яндекс Поиск не удовлетворяет потребности пользователя и обладает недостаточной скоростью реакции на новостные внезапные события. Если на большинстве запросов Яндекс Поиск показал актуальные страниц, созданные после момента происшествия событий — значит, метрика скорости реакции будет показывать высокие значения.

Итоговый процесс был реализован в сервисе Nirvana — внутренний сервис Яндекса для реализации различных алгоритм обработки данных и обучения моделей.

Последовательность действий в итоговом процессе обработки данных для построения метрики скорости реакции:

1. **Агрегация поисковых логов за предыдущий день.** В поисковых логах обычно хранится широкий спектр данных, связанных с пользователями и их поисковыми запросами. Эти данные используются для различных целей, включая улучшение алгоритмов поиска, персонализацию пользовательского опыта, таргетированную рекламу, анализ пользовательского поведения и трендов, а также для обеспечения безопасности и предотвращения мошеннических действий
2. **Извлечение необходимых нейросетевых признаков для запроса.** Для обработки поисковых запросов на практике используются различные нейросетевые признаки, которые помогают моделям эффективно понимать контекст и намерения пользователя. Это включает признаки самих запросов (например, токены, n-граммы, длина запроса), пользовательские данные (история запросов, демографические данные, устройство), контекст (временные метки, местоположение), взаимодействия (клики п
3. **Обработка данных для построения временного ряда задавания запроса в конкретный день**. Временной ряд задавания поискового запроса представляет собой последовательность поисковых запросов, сделанных пользователями в хронологическом порядке. Этот ряд фиксирует, когда именно были выполнены поисковые запросы, что позволяет анализировать изменения и паттерны во временных данных. Например, если наблюдать за временным рядом поисковых запросов, связанных с праздниками, можно заметить определенный всплеск активности перед праздничными датами.
4. **Применение LSTM для классификации временного ряда запроса**. Для классификации временных рядов данные разбиваются на окна фиксированной длины и подаются на вход модели, где последовательность временных шагов обрабатываются через один или несколько слоев LSTM. Затем полученные представления передаются полносвязным слоям, которые осуществляют финальную классификацию. В результате модель способна предсказывать метки или классы новых временных рядов, эффективно учитывая как краткосрочные, так и долгосрочные паттерны в данных.
5. **Применение градиентного бустинга над нейросетевыми признаками поискового запроса и выходом LSTM**. Устройство градиентного бустинга рассказано в исследовательской части.
6. **Отправка запросов на разметку асессорам**. Полученные запросы с определенной точностью (precision) являются ультрасвежими. Для того чтобы увеличить точность необходимо применить асессорскую разметку чье качество постоянно отслеживается на дашбордах измерения качества поисковых разметок.
7. **Агрегация результатов асессоров**. На данном шаге необходимо из 3 оценок асессоров (каждый запрос размечается в пересечении 3) выбрать одну, наиболее верную. Для этого используется метод majority vote, что переводится как решение большинством голосов. После получения оценки для кжадого запроса необходимо отобрать те запросы, которые являются ультрасвежими и в дальнейшем обрабатывать только их.
8. **Получение выборки релевантных страниц к запросу, созданных в интернете в этот день**. Для этого мы считаем документный BERT эмбеддинг и запросный эмбеддинг модели BERT. Эмбеддинг - это способ преобразования чего-то абстрактного, например слов или изображений в набор чисел и векторов . Эти числа не случайны; они стараются отражают суть или семантику нашего исходного объекта. Скалярное произведение таких эмбеддингов даст нам оценку релевантности страницы и запроса. Необходимо отобрать топ 100 страниц по подобранному порогу релевантности.
9. **Предсказания релевантности страниц запросу**. Т.к. скалярно произведение эмбеддингов обеспечивают не 100% точность. Каждая релевантная запросу страница размечается асессорами для увелечения итогового precision. Стоит отметить, что такое комбинированное использование моделей машинного обучения и ручного труда довольно эффективно. Ведь можно обучить модель с недостаточным precision (при этом значимо выше, чем у бейзлайна), но высоким recall, а засчет асессорской разметки у результирующего множества повысить precision и отобрать только верно отнесенные объекты к необходимому классу.
10. **Разметка на релевантность асессорами**. Асессоры значительно повышают precision предсказаний, засчет чего удается выявить релевантные запросу страницы.
11. **Определение момента начала события**. Имея множество релевантных страниц запросу необходимо выбрать самую первую т.е. ту, которую поисковые роботы яндекса обошли раньше всех. По нашим метрикам поисковые роботы очень быстро обходят недавно созданные новостные страницы, что позволяет с точностью до минут определить точное время создания страниц. Поисковый робот, также известный как веб-краулер или бот, — это автоматизированная программа, разработанная для систематического сканирования и индексирования веб-страниц в интернете. Он работает по алгоритму, который позволяет ему перемещаться от одной страницы к другой через гиперссылки, собирая информацию и метаданные о содержимом веб-страниц. Эти данные затем используются поисковыми системами, такими как Яндекс, для создания индексируемой базы данных, которая позволяет быстро и эффективно находить релевантные результаты по запросам пользователей. Основная цель поискового робота — обеспечить актуальность и полноту индекса поисковой системы, способствуя улучшенной навигации и доступу к информации для конечных пользователей.
12. **Извлечение выдачи из поисковых логов на ультрасвежих запросах**. По логам Поиска необходимо понять, какие именно страницы показывались пользователю по запросу, который требует актуальных вновь созданных новостных страниц.
13. **Извлечение времени обхода показанных страниц.** Как было описано ранее время обхода вновь созданных страниц позволяет с точностью до нескольких минут определить время создания страницы
14. **Определить актуальность пользовательской выдачи.** Как и было сказано ранее Яндекс Поиск показывает очень высокие метрики релевантности выдачи. Поэтому, при условии релевантности выдачи, которое всегда верно для абсолютного большинства поисковых запросов, зная точное время происшествия события по запросу и дату создания показанного в топ-5 документа можно определить, является ли он актуальным в указанное время или нет. Документ будет актуальным, если он релевантен запросу (достигается сложными поисковыми алгоритмами ранжирования) и отражает последние актуальные события по запросу (считаем, что отражает ведь страница про искомое событие была опубликована после момента его происшествия). Таким образом мы можем в автоматическом режиме получить долю актуальных страниц на выдаче по текущему запросу.
15. **Рассчет итогового значение метрики.** Общая доля актуальной выдачи усредняется по всем запросам и получается одно число — значени метрики скорости реакции.

Последовательность действий, описанная выше реализована мной в коде и внедрена в Яндекс, как регулярный процесс для подсчета качества поисковых алгоритмов ранжирования на запросах, связанных с различными новостями. Процесс рассчета запускается ежедневно в определенное время, запуск асессорских разметок и применение моделей также автоматизировано и отлажено, в результате чего имеем регулярный процесс, который строит точку на общем графике качества поисковой выдачи. Для более широкой информативности построен дашборд, где указаны все ультрасвежие запросы в определенный день, которые участвовали в расчете метрики скорости реакции. Для кажого такого запроса метрика сокрости реакции посчитана отдельно и покрашена определенным цветом для наглядного выделения запросов, для которых показатели метрики низкие или наоборот — высокие. Данный процесс играет важную роль в измерении качества Поисковых алгоритмов ранжирования и позволяет детально увидеть запросы, на которых Поиск срабатывает недостаточно хорошо, что является отправной точкой для дальнейшей работы по улучшению качества ранжирования.

ЗАКЛЮЧЕНИЕ

В результате выполнения Выпускной квалификационной работы бакалавриата был разработан процесс, вычисляющий метрику скорости реакции поисковой системы Яндекс Поиск. Метрика отражает качество выдачи на запросах, заданных в первое время после момента происшествия новостного события. В ходе работы была также исследована динамика таких запросов, обучены модели машинного и глубокого обучения для того чтобы классифицировать такие запросы из общего потока запросов. Создано несколько проектов по разметке данных, разработан пайплайн предсказания таких запросов с помощью рекуррентной нейронной сети LSTM, обученной по временному ряду задавания запроса и модели машинного обучения — градиентного бустинга, обученного на выходе LSTM и нейросетевых поисковых признаках.

Благодаря разработанному мной модулю метрики скорости реакции поисковой системы Яндекс Поиск разработчики понимают текущее качество выдачи на таких запросах, а также появляется возможность более детально анализировать причины низких значений в отдельные дни на некоторых поисковых запросах. По итогу разработки данной метрики появилась возможность проводить AB эксперименты, которые способны увеличить метрику скорости реакции Яндекс Поиска на новостные события. В результате чего создается огромный потенциал к улучшению алгоритмов ранжирования Яндекс Поиска, а также повышению качества пользовательского опыта использования поисковой системы.

Благодаря разработанному мной модулю метрики скорости реакции поисковой системы Яндекс Поиск разработчики понимают текущее качество выдачи на таких запросах, а также появляется возможность более детально анализировать причины низких значений в отдельные дни на некоторых поисковых запросах. По итогу разработки данной метрики появилась возможность проводить AB эксперименты, которые способны увеличить метрику скорости реакции Яндекс Поиска на новостные события. В результате чего создается огромный потенциал к улучшению алгоритмов ранжирования Яндекс Поиска, а также повышению качества пользовательского опыта использования поисковой системы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Sklearn documentation [Электронный ресурс] // https://sklearn-crfsuite.readthedocs.io/en/latest/ (Дата обращения: 20.05.2024).
2. Рекуррентная нейронная сеть [Электронный ресурс] // ru.wikipedia.org URL: https://ru.wikipedia.org/wiki/Рекуррентная\_нейронная\_сеть / (дата обращения: 29.05.2024)
3. Долгая краткосрочная память [Электронный ресурс] // ru.wikipedia.org URL: <https://ru.wikipedia.org/wiki/Долгая_краткосрочная_> / (дата обращения: 29.05.2024)
4. Основные метрики задач классификации в машинном обучении [Электронный ресурс] // https://webiomed.ai/blog/osnovnye-metriki-zadach-klassifikatsii-v-mashinnom-obuchenii/ (дата обращения: 04.04.2020).
5. Streamlit 0.82.0 documentation [Электронный ресурс] // https://docs.streamlit.io/en/stable/ (Дата обращения: 02.05.2024).
6. Зализняк А.A. Грамматический словарь русского языка. — М., Русский язык, 1980. – 880 с
7. Automatic morphological analysis for Russian: a comparative study [Электронный ресурс] // http://www.dialog-21.ru/media/3473/dereza.pdf (Дата обращения: 20.05.2024).
8. Официальный сайт AOT [Электронный ресурс] // http://aot.ru/ (Дата обращения: 10.05.2020).
9. TreeTagger [Электронный ресурс] // https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Дата обращения: 10.05.2024).
10. Открытый корпус [Электронный ресурс] // http://opencorpora.org/ (Дата обращения: 10.05.2024).
11. MyStem [Электронный ресурс] // https://yandex.ru/dev/mystem/ (Дата обращения: 10.05.2024).
12. Streamlit 0.82.0 documentation [Электронный ресурс] // docs.streamlit.io URL: https://docs.streamlit.io/en/stable / (дата обращения: 29.05.2024)

ПРИЛОЖЕНИЕ A  
ГРАФИЧЕСКИЕ МАТЕРИАЛЫ

ПРИЛОЖЕНИЕ Б  
ТЕХНИЧЕСКОЕ ЗАДАНИЕ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

|  |  |  |  |
| --- | --- | --- | --- |
|  | Утверждаю  Заведующий кафедрой ИУ-5 |  | Согласовано  научный руководитель |
|  | \_\_\_\_\_\_\_\_\_\_\_В.И.Терехов  "\_\_"\_\_\_\_\_\_\_\_\_\_2024 г. |  | \_\_\_\_\_\_\_\_\_\_Ю.Е. Гапанюк  "\_\_"\_\_\_\_\_\_\_\_\_\_2024 г. |

**Модуль вычисления скорости реакции поисковых систем**

Техническое задание

(вид документа)

писчая бумага

(вид носителя)

5

(количество листов)

ИСПОЛНИТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Сафонов Федор Алексеевич

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2024 г.

Москва - 2024

1. **Введение**

Поисковая система  — это специальная программа, которая в ответ на запрос пользователя через веб-интерфейс (сайт) выдает список ресурсов, отсортированных по релевантности этому запросу.

Чтобы отслеживать качество работы различных аспектов поисковой системы необходимы соответствующие метрики. С помощью метрик разработчики понимают, насколько хорошо работают их алгоритмы и модели машинного обучения. Метрики являются основным критерием качества уже выполненной работы. Благодаря метрикам разработчики могут быстро заметить какую-то неисправность в алгоритмах ранжирования и немедленно исправить ее.

Большая доля запросов в Яндекс Поиске касаются новостей, происшествий и других внезапных событий, произошедших в мире. Такие запросы обладают особой спецификой, т.к. на них крайне важно показать недавно созданные страницы в интернете, ведь именно на них отражены последние актуальные новости, которые нужны пользователю.

Метрика скорости реакции показывает, насколько хорошо Яндекс Поиск сработал на запросах про внезапные новостные события. Имея такую метрику, можно сделать вывод о текущем качестве ранжирования и спланировать дальнейшую работу по улучшению данной метрики.

1. **Основания для разработки**

Основанием для разработки является рабочая задача в компании Яндекс, отделе Поисковых систем. Задание было поставлено руководителем группы отдела ранжирования.

**3. Назначение разработки**

Разрабатываемая метрика предназначена для внутренних разработчиков Яндекс Поиска, для отслеживания и улучшения качества работы алгоритмов ранжирования на пользовательских новостных запросах.

**4. Требования к программе или программному изделию**

4.1 Требования к функциональным характеристикам

Программа должна выполнять следующие функции:

4.1.1 Загрузка запросов и логов за соответствующий день

4.1.2 Выделение набора признаков для классификации

4.1.3 Классификация запроса на ультрасвежесть

4.1.4 Определение времени происшествия новостного события по данному запросу

4.1.5 Оценка качества ранжирования выдачи по запросам, заданным после момента происшествия новостного события

4.1.6 Итоговое вычисление метрики, определяющее качество ранжирования за текущий день.

**5. Требования к программной документации**

Для представления заказчику разрабатываются следующие документы:

1. Техническое задание

2. Программа и методика испытаний

**6. Стадии и этапы разработки**

График выполнения отдельных этапов работ приведен в соответствии с приказом об организации учебного процесса в 2023/2024 учебном году.

Таблица 1: Этапы разработки

| № п/п | Наименование этапа и содержание работ | Сроки исполнения |
| --- | --- | --- |
| 1 | Разработка и утверждение ТЗ | Март 2024г. |
| 2 | Исследование предметной области | Март-Апрель 2024г. |
| 3 | Разработка архитектуры программного обеспечения | Апрель 2024 г. |
| 4 | Создание программ | Апрель — Май 2024 г. |
| 5 | Тестирование и отладка | Май 2024 г. |
| 6 | Оформление документации | Май — Июнь 2024 г. |
| 7 | Защита работы | Июнь 2024 г. |

**8. Порядок контроля и приёмки**

Приём и контроль программного изделия осуществляется в соответствие с методикой испытаний (см. документ «Программа и методика испытаний»).

ПРИЛОЖЕНИЕ В  
ПРОГРАММА И МЕТОДИКА ИСПЫТАНИЙ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

|  |  |  |  |
| --- | --- | --- | --- |
|  | Утверждаю  Заведующий кафедрой ИУ-5 |  | Согласовано  научный руководитель |
|  | \_\_\_\_\_\_\_\_\_\_\_В.И.Терехов  "\_\_"\_\_\_\_\_\_\_\_\_\_2024 г. |  | \_\_\_\_\_\_\_\_\_\_Ю.Е. Гапанюк  "\_\_"\_\_\_\_\_\_\_\_\_\_2024 г. |

**Модуль вычисления скорости реакции поисковой выдачи**

Программа и методика испытаний

(вид документа)

писчая бумага

(вид носителя)

5

(количество листов)

ИСПОЛНИТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Сафонов Федор Алексеевич

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2024 г.

Москва - 2024

Аннотация

В данном документе описываются последовательность и методы проведения испытаний при тестировании программного изделия, состав и структура технических и программных средств, необходимых для проведения испытаний, а также приводятся требования к предъявляемой документации, характеристикам программы применительно к условиям эксплуатации и требования к информационной и программной совместимости. Описывается тестовый пример и реакция системы на него.

Объект испытаний

Веб-приложение приложение, в котором приведена методика расчета метрики скорости реакции поисковых систем.

Сокращенное наименование: программа методики расчета метрики скорости реакции.

Цель испытаний

Цель испытания – проверка функционирования всех указанных в техническом задании функций программы.

Состав предъявляемой документации

На испытания программного продукта предъявляются следующие документы:

1) Техническое задание.

2) Программа и методика испытаний.

Технические требования

* 1. **Требования к программной документации**

Комплектность программной документации должна удовлетворять разделу данного документа "Состав предъявляемой документации".

* 1. **Требования к техническим характеристикам**
     1. **Требования к составу аппаратного обеспечения**

Для работы программы методики расчета метрики скорости реакции требований к составу аппаратного обеспечения не предъявляется.

* + 1. **Требование к составу программного обеспечения**

Для работы программы методики расчета метрики скорости реакции требований к составу программного обеспечения не предъявляется.

Порядок проведения испытаний

Испытания данного программного продукта будут проводиться в следующем порядке:

1) Запуск программы методики расчета метрики скорости реакции

2) Тестирование функционала программы методики расчета метрики скорости реакции

* 1. **Требования к составу аппаратного обеспечения**

Требования к составу аппаратного обеспечения учитываются согласно пункту 4.2.1.

* 1. **Требования к составу программного обеспечения**

Требования к составу программного обеспечения учитываются согласно пункту 4.2.2.

1. Методы испытаний

|  |  |  |  |
| --- | --- | --- | --- |
| N | N пункта ТЗ | Выполняемые действия | Результат |
| 1 | 4.1.1.  Загрузка запросов и логов за соответствующий день. | Пользователь:  Открывает приложение, заходит на вкладку «Понятие ультрасвежести». | Виден пример запроса, история задавания которого выгружена из логов Яндекс Поиска. |
| 2 | 4.1.2. Выделение набора признаков для классификации. | Пользователь:  Заходит на вкладку «Архитектура пайплайна». | Открывается окно, показывающее схематично иллюстрирующее процесс извлечения признаков. |
| 3 | 4.1.3. Классификация запроса на ультрасвежесть. | Пользователь:  Заходит на вкладку «Обучение LSTM» и вкладку «Обучение CatBoost». | Открывается окно, пользователь видит процесс обучения моделей, благодаря котором запросы классифицируются на ультрасвежие и нет. |
| 4 | 4.1.4. Определение времени происшествия новостного события по данному запросу. | Пользователь:  Заходит на вкладку «Поиск начала события». | Открывается окно, пользователь видит интерфейс разметки данных, а также архитектуру процесса, который предварительно выявляет подходящие запросу страницы. |
| 5 | 4.1.4.  Оценка качества ранжирования выдачи по запросам, заданным после момента происшествия новостного события. | Пользователь:  Заходит на вкладку «Финальный подсчет метрик». | Открывается окно, пользователь видит процесс рассчета метрики на примере одного ультрасвежего запроса. |
| 6 | 4.1.5. Итоговое вычисление метрики, определяющее качество ранжирования за текущий день. | Пользователь:  Заходит на вкладку «Финальный подсчет метрик». | Открывается окно, пользователь видит итоговую схему рассчета метрики за определенный день на выявленных ультрасвежих запросах. |

1. Результаты испытаний

Основой испытаний является демонстрация работы основных функций

приложения: визуализации динамики ультрасвежего запроса, запуск разметок для ее классификации, обучение моделей LSTM и CatBoost, итоговый расчет метрики скорости реакции для поисковой системы.

прошедших турниров, ознакомление с данными предстоящих турниров.

Испытание считается пройденным успешно, если в процессе

демонстрации пользователь смог посмотреть визуализации всех процессов, описываемых выше.