Факультет
компьютерных наук

Образовательная
программа
“Прикладная математика и
информатика”

Москва
2023

«Разработка метода оценки сходства датасетов»
«Dataset Similarity Method Development»

Индивидуальный исследовательский проект

Автор: Копылов Олег Иванович, БПМИ-205
Руководитель: Малых Валентин Андреевич, кандидат технических наук,
инженер ключевых проектов ООО "Техкомпания Хуавэй"

Основные определения

Качество – в рамках этой работы под качеством подразумевается F1 score.

Сходство датасетов – свойство двух датасетов A и B, при котором модель, эффективно работающая на датасете A, также эффективно работает и на датасете B. В рамках данной работы сходство датасетов понимается двумя способами: сходство датасетов в первом смысле и во втором смысле.

Сходство датасетов в первом смысле – свойство двух датасетов A и B, при котором модель, обученная на датасете A и показывающая хорошие результаты при тестировании на A, также показывает хорошие результаты при тестировании и на B. В этом случае модель обучается только на датасете A, тогда как на датасете B происходит лишь тестирование.

Сходство датасетов во втором смысле – свойство двух датасетов A и B, которое состоит в следующем: **если** модель M после обучения на датасете A показывает хорошие результаты при тестировании на A, **то** эта же самая модель M, но уже обученная на датасете B, покажет хорошее качество при тестировании на датасете B.

Актуальность работы

Рассмотрим тривиальную, на первый взгляд, задачу: дан текстовый датасет, нужно подобрать наиболее подходящую для него модель из N возможных языковых моделей. Очевидное решение – сначала обучить на данном датасете все N моделей, а затем выбрать ту, которая показывает наилучшее качество при тестировании. Но такое решение, к сожалению, применимо *не всегда*.

Во-первых, данный датасет может быть слишком маленьким. Из-за этого обучаться на нём не целесообразно: ввиду слишком маленькой обучающей выборки модель не запомнит закономерность. Во-вторых, датасет может не обладать всей информацией, требуемой для тестирования или обучения на нём.

Если нельзя обучить модель непосредственно на данном датасете, значит необходимо обучить её на каком-то другом датасете. Его выбор имеет смысл производить не случайно, а на основании некоторого метода оценки сходства датасетов. При наличии сходства между двумя датасетами модель, хорошо работающая на одном из двух датасетов, также эффективна и на оставшемся.

Затем обучить N языковых моделей на новом подобранном датасете и после этого выбрать модель с наилучшим качеством, таким образом решив поставленную задачу.

Однако совершенно не понятно, по каким признакам можно определять сходство датасетов. Подобные исследования не проводились ранее.

Если коротко, возникают ситуации, когда для датасета нужно подобрать подходящую модель, но по каким-то причинам обучаться или тестироваться на этом датасете нельзя или нецелесообразно. Тогда модель надо обучить на каком-то другом датасете. Как его находить – абсолютно не понятно. Предлагается это делать с помощью метода оценки сходства датасетов, разработке которого и посвящена моя работа.

Сущность метода оценки сходства датасетов

Метод оценки сходства датасетов будет функцией, которая принимает два датасета и возвращает численное значение. Оно тем больше, чем лучше происходит переход с первого датасета на второй.

Возвращаемые методом численные значения должны коррелировать с изменением качества при переходе с одного датасета на другой.

Цель и задачи

Цель исследования – разработать метод оценки сходства текстовых датасетов.

Для достижения вышеуказанной цели поставлены следующие **задачи,** основные из которых выделены жирным шрифтом:

1) изучение научных работ о языковых моделях,
2) **поиск моделей и датасетов для проведения экспериментов,**
3) определение методики проведения экспериментов,
4) **проведение экспериментов,**
5) **визуализация результатов экспериментов,**
6) формулирование гипотез о том, на что должны опираться методы оценки сходства датасетов исходя из результатов пунктов 4-5,
7) **проверка сформулированных гипотез с помощью разработки соответствующих методов оценки сходства датасетов,**
8) **тестирование эффективности разработанных методов,**
9) **визуализация и оценка полученных результатов,**
10) анализ проделанной работы и формулирование итогового вывода,
11) формулирование направлений дальнейших исследований.

Выбор моделей и обзор источников

Я изучил несколько статей и использовал описанные в этих статьях языковые модели при проведении своих экспериментов. **Обзор статей, посвященных этим моделям – на страницах 7-9 отчета, в разделе “Теоретическая часть 1”.**

Используются следующие модели:

1) **BERT**
2) **ROBERTA**
3) **ALBERT**
4) **ELECTRA**
5) **DistilBERT**
6) **MobileBERT**
7) **LaBSE**
8) **talkheads_ggelu_bert (BERT with Talking-Heads Attention and Gated GELU)**
9) **LAMBERT**
10) **tn_bert (A compressed BERT model using tensor networks)**

Я специально подобрал модели таким образом, чтобы они были похожи на BERT, принадлежали одному с ним семейству NLP моделей. Благодаря этому мы ожидаем, что модели будут вести себя схожим образом, то есть, например, изменение качества при переходе с одного датасета на другой будет приблизительно одинаковым для большинства моделей.

Серии экспериментов и используемые датасеты

Проведено две серии экспериментов.

Первая серия экспериментов посвящена бинарной классификации отзывов на позитивные и негативные. Рассматриваемых датасетов два: первый с отзывами об отелях, второй с отзывами о фильмах.

Вторая серия экспериментов посвящена бинарной классификации спам / не спам. В первом датасете содержатся SMS, а во втором – e-mail письма. Для всех них указано, является ли SMS или e-mail спамом.

Обучение и тестирование моделей, гитхаб-репозиторий

Обучение и тестирование моделей подробно описано в разделе “Практическая часть 3 – Обучение и тестирование моделей”, стр. 11.

Все файлы, относящиеся к работе, представлены в гитхаб-репозитории.

Каждой из 10 рассматриваемых моделей выделена своя папка, где представлены все связанные с обучением и тестированием ноутбуки.

В отдельные ноутбуки вынесены:

1) результаты экспериментов, их визуализация и выводы

   а) для первой серии экспериментов (определение позитивности отзыва),

   б) для второй серии экспериментов (определение спама),

   в) для датасетов, взятых из разных серий экспериментов.

2) гипотезы, соответствующие им методы оценки сходства датасетов и исследование эффективности этих методов.

Гипотезы

1) **сходство датасетов связано со схожестью их текстового содержания, которое отражается в схожести часто встречающихся в датасетах слов**;
2) используемые языковые модели берут от текстов эмбеддинги, поэтому **сходство датасетов может быть связано со сходством усредненного значения эмбеддингов** (для каждого датасета своё усредненное значение эмбеддинга).

Для проверки каждой из гипотез реализован свой метод, рассмотрим их.

Метод, разработанный в соответствии с первой гипотезой:

В каждом из датасетов найти топ из *num* наиболее часто встречающихся слов, исключив артикли, предлоги и т.п. Для каждого слова из первого топа найти наиболее схожее слово из второго топа; после чего запомнить число, соответствующее степени их схожести. Сложить все эти значения. Чем больше полученная сумма, тем более вероятно сходство датасетов.

Следует провести вычисления для различных значений *num* (число слов в топе), так как не очевидно, какое значение *num* даст наилучшие результаты.

Метод, разработанный в соответствии со второй гипотезой:

Сделать векторные представления от каждого текста из датасета (с помощью BERT) и усреднить. Для разных датасетов сравнить полученный эмбеддинг с помощью cosine_similarity. Чем больше полученное значение, тем больше сходство сравниваемых датасетов.

Методика исследования эффективности разработанного метода:

Подробно описана в разделе “Экспериментальная часть 4 – Методика исследования эффективности метода”, стр. 34.

Краткое описание:

1) Рассматривается множество пар датасетов, составляемое двумя способами (для каждого множества получается свое значение коэффициента корреляции из пункта 4),
2) Для каждой из рассматриваемых пар вычисляется среднее арифметическое (по всем N=10 рассматриваемым языковым моделям), а также медианное изменение качества при переходе с первого датасета на второй,
3) Для каждой из рассматриваемых пар вычисляется значение, возвращаемое методом оценки сходства датасетов,
4) Вычисляется значение коэффициента корреляция между значениями из двух предыдущих пунктов.

При оценке эффективности метода учитывается корреляция, получающаяся в каждом из двух способов составления множества из пункта 1. Рассмотрим, как это происходит.

Критерии эффективности разработанного метода

В таблице 1 по значению коэффициента корреляции Пирсона определяется степень корреляции. В таблице 2 определяется эффективность метода в зависимость от того, какая степень корреляции получается в каждом из двух способов составить множество рассматриваемых пар датасетов.

Таблица 1. Описание корреляции

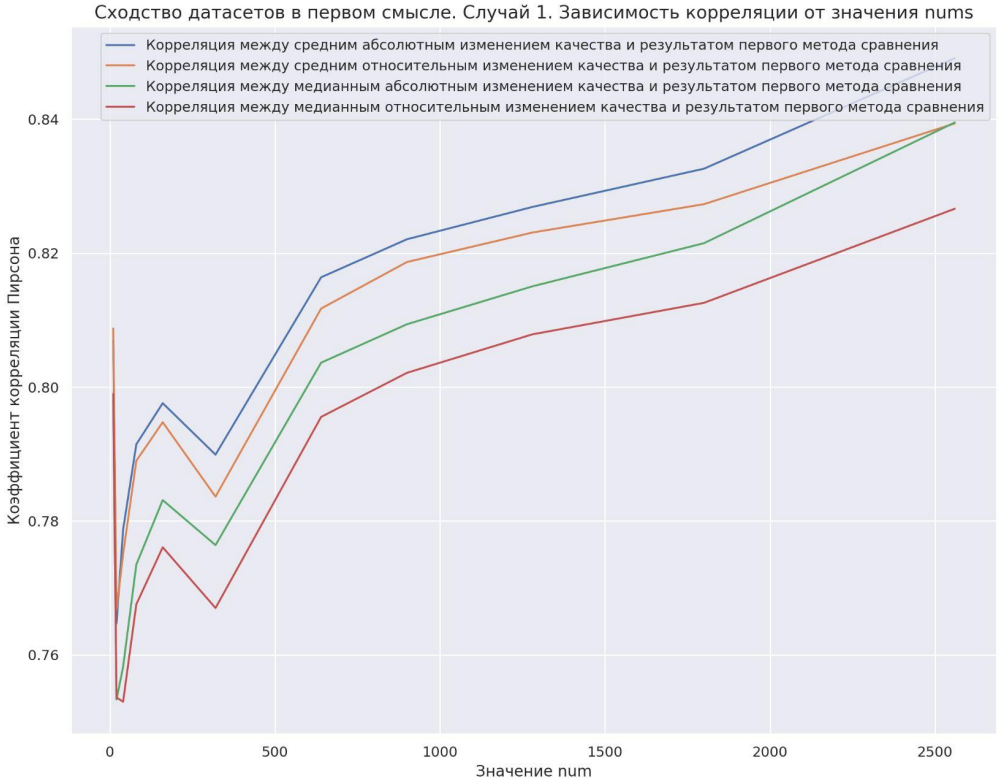
Значение коэффициента корреляции Пирсона	**Описание корреляции**
менее 0.05	отсутствие корреляции
от 0.05 до 0.3	очень слабая
от 0.3 до 0,5	слабая
от 0.5 до 0,7	средняя
от 0.7 до 0.9	высокая
более 0.9	очень высокая

Таблица 2. Критерии эффективности метода

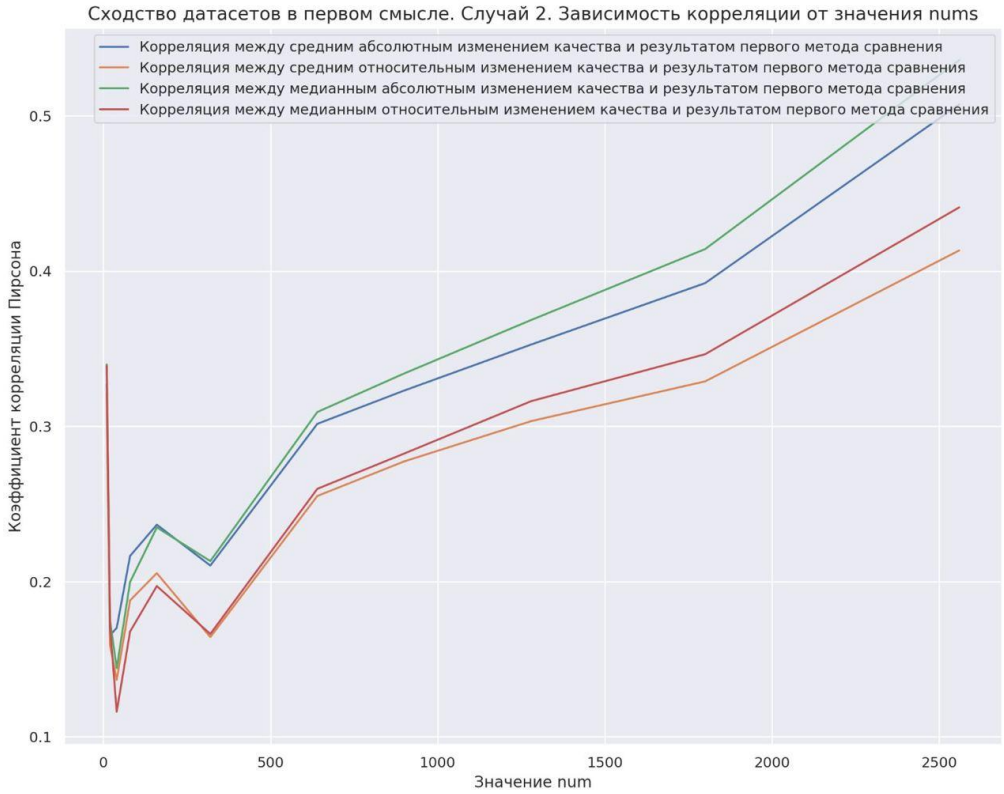
Эффективность метода	**Корреляции для рассматриваемых двух случаев**
абсолютно неэффективный	обе корреляции не выше слабой
слабо эффективный	одна корреляция очень слабая, другая средняя, высокая или очень высокая
условно эффективный	обе корреляции средние; одна корреляция слабая, другая корреляция средняя, высокая или очень высокая
эффективный	одна корреляция средняя, другая высокая или очень высокая; одна корреляция высокая, другая высокая или очень высокая
абсолютно эффективный	обе корреляции очень высокие

Результаты для первого метода и сходства датасетов в первом смысле

Случай 1: рассматриваются все возможные пары датасетов, в том числе пары с совпадающими датасетами.



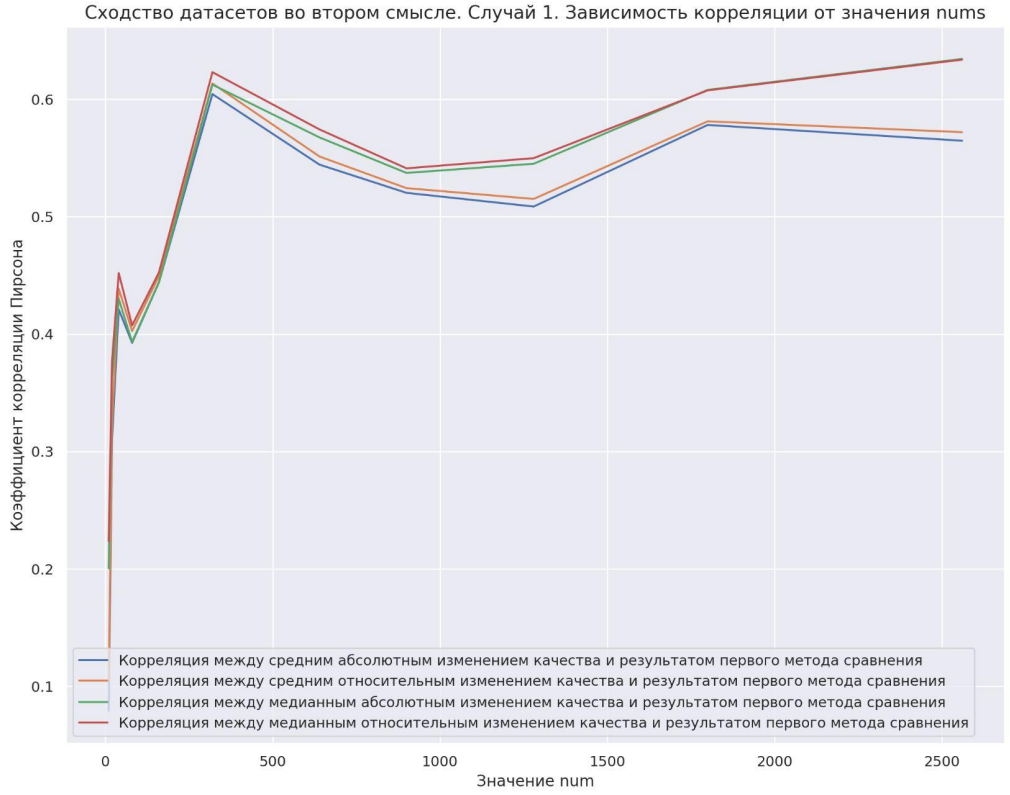
Случай 2: рассматриваются все возможные пары **различных** датасетов.



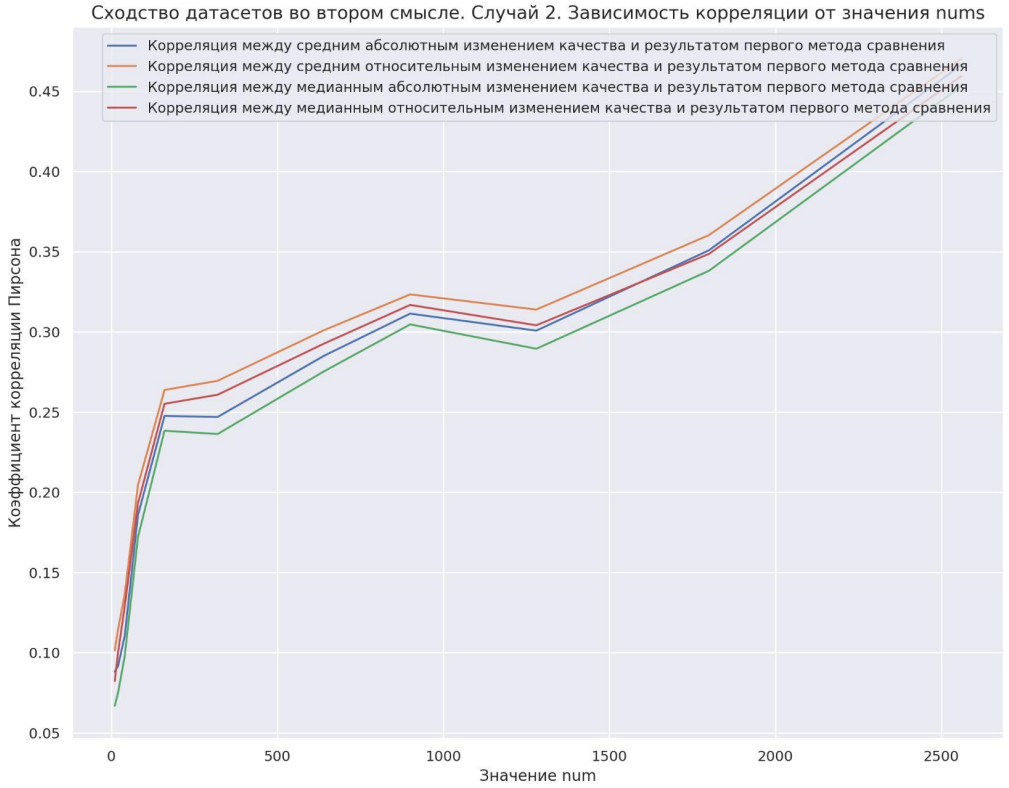
Подводя итог, при оптимальном значении num получается высокая корреляция для случая 1 и средняя корреляция для случая 2. **Первый метод получился эффективным для оценки сходства датасетов в первом смысле.**

Результаты для первого метода и сходства датасетов во втором смысле

Случай 1: рассматриваются не все возможные пары датасетов, а половина из них.



Случай 2: рассматриваются все возможные пары различных датасетов. Иными словами, добавлен переход в обратную сторону.



Подводя итог, при оптимальном значении num получается стойкая средняя корреляция для случая 1, но для случая 2 корреляция получается ниже среднего. **Первый метод получился условно эффективным для оценки сходства датасетов во втором смысле.**

Результаты для второго метода и сходства датасетов в первом смысле

Случай 1: рассматриваются все возможные пары датасетов, в том числе пары с совпадающими датасетами.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.719517
correlation_for_median_absolute_quality_difference_and_second_method_result	0.700215
correlation_for_mean_relative_quality_difference_and_second_method_result	0.702173
correlation_for_median_relative_quality_difference_and_second_method_result	0.689387

Случай 2: рассматриваются все возможные пары *различных* датасетов.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.137801
correlation_for_median_absolute_quality_difference_and_second_method_result	0.118253
correlation_for_mean_relative_quality_difference_and_second_method_result	0.072547
correlation_for_median_relative_quality_difference_and_second_method_result	0.082060

Подводя итог, получается стойкая средняя корреляция для случая 1, но для случая 2 корреляция получается очень низкая.

Второй метод получился слабо эффективным для оценки сходства датасетов в первом смысле.

Результаты для второго метода и сходства датасетов во втором смысле

Случай 1: рассматриваются не все возможные пары датасетов, а половина из них: есть пара, соответствующая первой серии экспериментов, есть пара, соответствующая второй серии экспериментов, также есть пары, соответствующие переходам с датасета первой серии экспериментов на датасет второй серии экспериментов. Среди рассматриваемых пар выполнено, что если есть пара (A, B), то пары (B, A) нет среди рассматриваемых.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.039861
correlation_for_median_absolute_quality_difference_and_second_method_result	0.121839
correlation_for_mean_relative_quality_difference_and_second_method_result	0.070966
correlation_for_median_relative_quality_difference_and_second_method_result	0.149076

Случай 2: рассматриваются все возможные пары различных датасетов. Иными словами, к случаю 1 добавлен переход в обратную сторону.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	-1.519744e-17
correlation_for_median_absolute_quality_difference_and_second_method_result	2.017388e-17
correlation_for_mean_relative_quality_difference_and_second_method_result	2.364859e-03
correlation_for_median_relative_quality_difference_and_second_method_result	1.037911e-02

Подводя итог, получается очень слабая корреляция для случая 1, для случая 2 корреляция получается чрезвычайно близкой к нулю.

Второй метод является абсолютно неэффективным для оценки сходства датасетов во втором смысле.

Результаты и выводы:

Для каждой из N = 10 рассматриваемых языковых моделей проведено две серии экспериментов (в серии экспериментов два датасета, которые решают близкие задачи бинарной классификации). Результаты экспериментов визуализированы и проанализированы. Сформулировано две гипотезы относительно того, какие признаки могут быть связаны со сходством датасетов.

В соответствии с выдвинутыми гипотезами реализованы методы оценки сходства датасетов. По заранее сформулированным критериям проверена эффективность этих методов. По результатам проверки подтверждена первая гипотеза и опровергнута вторая.

Лучший из разработанных методов оценки сходства датасетов получился *эффективным* для оценки сходства датасетов в первом смысле и *условно эффективным* для оценки сходства во втором смысле. Данный метод опирается на гипотезу о схожести наиболее часто встречающихся слов.

Таким образом, поставленные задачи выполнены в полном объеме, цель работы достигнута.

Направления дальнейшего исследования:

1) **исследовать эффективность хорошо зарекомендовавшего себя метода *на других задачах*, отличных от бинарной классификации**: например, на задаче многоклассовой классификации.

2) **провести дополнительное исследование для неподтвердившейся гипотезы номер 2**.

   Как я объяснил в “Экспериментальной части”, разработанный в соответствии с гипотезой номер 2 метод является симметричным: если методу в качестве аргументов передать на вход датасеты в другом порядке, то возвращаемое методом численное значение не изменится. С другой стороны, сходство датасетов в первом смысле не обязательно является симметричным, а сходство датасетов во втором смысле – точно не является симметричным в общем случае.

   Разработанный метод является симметричным потому, что в нём усредненные векторы сравниваются с помощью cosine similarity. Если же сравнивать методы неким *несимметричным* образом, то результаты могут стать совершенно другими. Это, в свою очередь, может повысить эффективность метода, построенного в соответствии со второй гипотезой.

Список источников

[1]. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // arxiv.org. 2018. URL: https://arxiv.org/abs/1810.04805. DOI: 10.48550/ARXIV.1810.04805.

[2]. Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. // arxiv.org. 2018. URL: https://arxiv.org/abs/1804.07461. DOI: 10.48550/ARXIV.1804.07461.

[3]. Rajpurkar, Pranav and Zhang, Jian and Lopyrev, Konstantin and Liang, Percy. SQuAD: 100,000+ Questions for Machine Comprehension of Text. // arxiv.org. 2016. URL: https://arxiv.org/abs/1606.05250. DOI: 10.48550/ARXIV.1606.05250.

[4]. Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. RoBERTa: A Robustly Optimized BERT Pretraining Approach. // arxiv.org. 2019. URL: https://arxiv.org/abs/1907.11692. DOI: 10.48550/ARXIV.1907.11692.

[5]. Lai, Guokun and Xie, Qizhe and Liu, Hanxiao and Yang, Yiming and Hovy, Eduard. RACE: Large-scale ReAding Comprehension Dataset From Examinations. // arxiv.org. 2017. URL: https://arxiv.org/abs/1704.04683. DOI: 10.48550/ARXIV.1704.04683.

[6]. Clark, Kevin and Luong, Minh-Thang and Le, Quoc V. and Manning, Christopher D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. // arxiv.org. 2020. URL: https://arxiv.org/abs/2003.10555. DOI: 10.48550/ARXIV.2003.10555.

[7]. Feng, Fangxiaoyu and Yang, Yinfei and Cer, Daniel and Arivazhagan, Naveen and Wang, Wei. Language-agnostic BERT Sentence Embedding. // arxiv.org. 2020. URL: https://arxiv.org/abs/2007.01852. DOI: 10.48550/ARXIV.2007.01852.

[8]. Ссылка на TensorFlow Hub: https://tfhub.dev/.

[9]. Ссылка на посвященный курсовой работе гитхаб репозиторий: https://github.com/Oleg13Kopylov/comparing_datasets.

Факультет
компьютерных наук

Образовательная
программа
“Прикладная математика и
информатика”

Москва
2023

«Разработка метода оценки сходства датасетов»
«Dataset Similarity Method Development»

Индивидуальный исследовательский проект

Автор: Копылов Олег Иванович, БПМИ-205
Руководитель: Малых Валентин Андреевич, кандидат технических наук,
инженер ключевых проектов ООО "Техкомпания Хуавэй"