# Поиск семантического сходства текстов в задаче сопоставления видом деятельности компаний с нормативами обращения ТКО субъекта РФ

*Работа выполнена Ветровой Александрой*

## Введение

Все интеллектуальные системы обработки текста на современном этапе развития нашли своё применение в различных сферах и решают конкретные задачи. Так, например, задачу поиска информации помогают решать информационно-поисковые системы; переводить слова и тексты с одного языка на другой помогают системы машинного перевода; практически в каждом современном устройстве связи стоит система распознавания речи; задачу сокращения объема текста и краткого изложения содержания решают системы автоматического аннотирования и реферирования. Примеров много.

В настоящее время поиск сходства между текстами имеет большое практическое применение. Однако основных категорий обнаружения сходства можно назвать три:

- сравнение на основе слов

- линейный поиск на основе пунктов (простое сравнение очередного рассматриваемого значения на момент сходства)

- стилистический анализ

В данной работе используется метод сравнения на основе слов, включающий в себя элементы стилистического анализа - учитывание смысла, контекста и свойства слов при переводе в векторы.

## Постановка задачи

Решается задача семантического сходства словосочетаний из двух файлов с целью сведения файлов в один реестр для дальнейшего использования.

## Описание данных

В работе использовались 2 датасета: «Рубрики» (далее рубрики) – датасет, предоставленный 2ГИС (в файле представлена градация различных видов деятельности компаний. Каждую организацию в Российской Федерации следует отнести к одной или нескольким из указанных групп):

| Категория | Рубрика 2го уровня | Рубрика | Тематика рубрики | Код рубрики | Категория Пензенская облас |
|---|---|---|---|---|---|
| Автосервис / Автотовары | Автосервис | Автомойки | Автомойки | 405 | p_10 |
| Автосервис / Автотовары | Автосервис | Заправочные станции | Автозаправочные станции | 18547 | p_7 |
| Автосервис / Автотовары | Автотовары | Автоаксессуары | Промтоварный магазин | 425 | p_5 |

*Рис 1.1 Пример предоставления данных в файле «Рубрики» по разделу «Автосервис / Автотовары»*

| Категория | Рубрика 2го уровня | Рубрика | Тематика рубрики | Код рубрики | Категория Пензенская облас |
|---|---|---|---|---|---|
| Медицина / Здоровье / Красота | Медицинские услуги | Челюстно-лицевая хирургия | Поликлиники, медицинские центры | 110653 | p_31 |
| Медицина / Здоровье / Красота | Медицинские услуги | Школы для будущих мам | Медицинские центры | 10135 | p_31 |
| Медицина / Здоровье / Красота | Медицинские учрежде | Больницы | Больницы | 201 | p_31 |

*Рис 1.2 Пример предоставления данных в файле «Рубрики» по разделу «Медицина / Красота / Здоровье»*

Второй датасет – нормативы производимого организациями мусора по категориям: «Нормативы ТКО» (далее нормативы) в зависимости от субъекта РФ. Например, тренировочные данные для проекта – нормативы Пензенской области:

| Код | Наименование категории объектов |
|---|---|
| p_1 | Научно-исследовательские, проектные институты и конструкторские бюро |
| p_2 | Банки, финансовые учреждения |
| p_3 | Отделения связи |
| p_4 | Административные, офисные учреждения |
| p_5 | Предприятия торговли |

*Рис 1.3 Пример предоставления данных в файле «Нормативы ТКО»*

В файле нормативы были проставлены вручную коды нормативов. Далее коды так же механическим способом были соотнесены с рубриками (крайний справа столбец см. Рис.1.1 и 1.2).

Помимо нормативов Пензенской области представлены также нормативы Ленинградской области и республики Бурятия.

## Обзор литературы

Был проведен анализ литературы и выявлены основные методы поиска сходства текстов. Наиболее часто в подобных задачах использовались:

- Сходство Жаккара
- Расстояние Левенштейна
- Косинусное сходство

На рисунке 2 показаны результаты работы моделей, оцененные метрикой Accuracy, при использовании различных способов определения сходства векторов. Как стоит заметить, косинусное расстояние является одной из наиболее успешных метрик для вычисления расстояния.

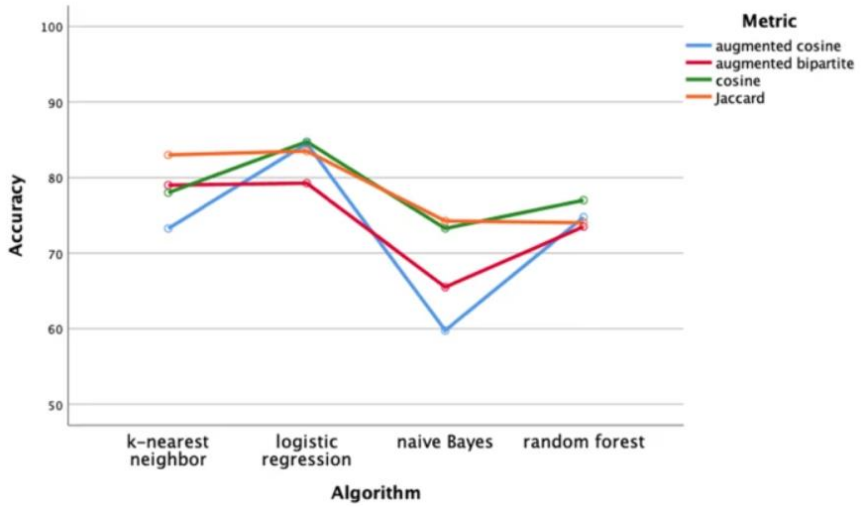



*Рис 2 Графики качества моделей при использовании различных метрик на основе сходства*

Для поиска на основе вектора также можно использовать один из нескольких методов построения векторов. В данном случае рассматривались и использовались:

1. TF-IDF
2. Doc2Vec
3. BERT

## Выбор метода

В ходе дальнейшего анализа литературы было выяснено, что чаще всего в задачах поиска схожих текстов используется именно косинусное расстояние, т.к. оно помогает вычислить сходство документов вне зависимости от их размера. При этом, если вектора правильно закодировать, косинусное расстояние может быть использовано как мера поиска семантического сходства, что и является целью работы.

Косинусное сходство вычисляет косинус угла между двумя векторами:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}};$$

*Рис 3. Формула расчёта косинусного расстояния*

Меньший угол между векторами (рассчитанный с помощью косинусного сходства) означает, что они более выровнены. Для плотных векторов это коррелирует с большим семантическим сходством

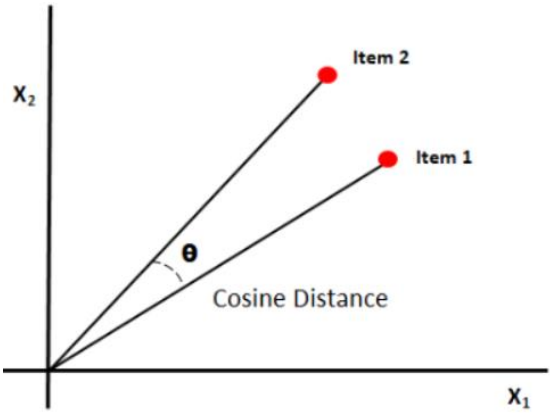



*Рис 4. Совпадающие по направлению вектора*

## Выбор метрики

Классически в задачах Texts similarity используют гармоническое среднее точности и полноты – меру F1.

$$F = (\beta^2 + 1)\frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

*Рис 5. Формула расчёта F1*

## EDA

В процессе исследовательского анализа данных выяснилось, что рубрики распределены неравномерно. Большая часть строк отнесена к рубрике «Медицина / Здоровье / Красота», второе место занимает «Оборудование / Инструмент». Меньше всего представлены среди рубрик «Аварийные / справочные / экстренные службы». На рисунке 5 представлена гистограмма распределения рубрик:

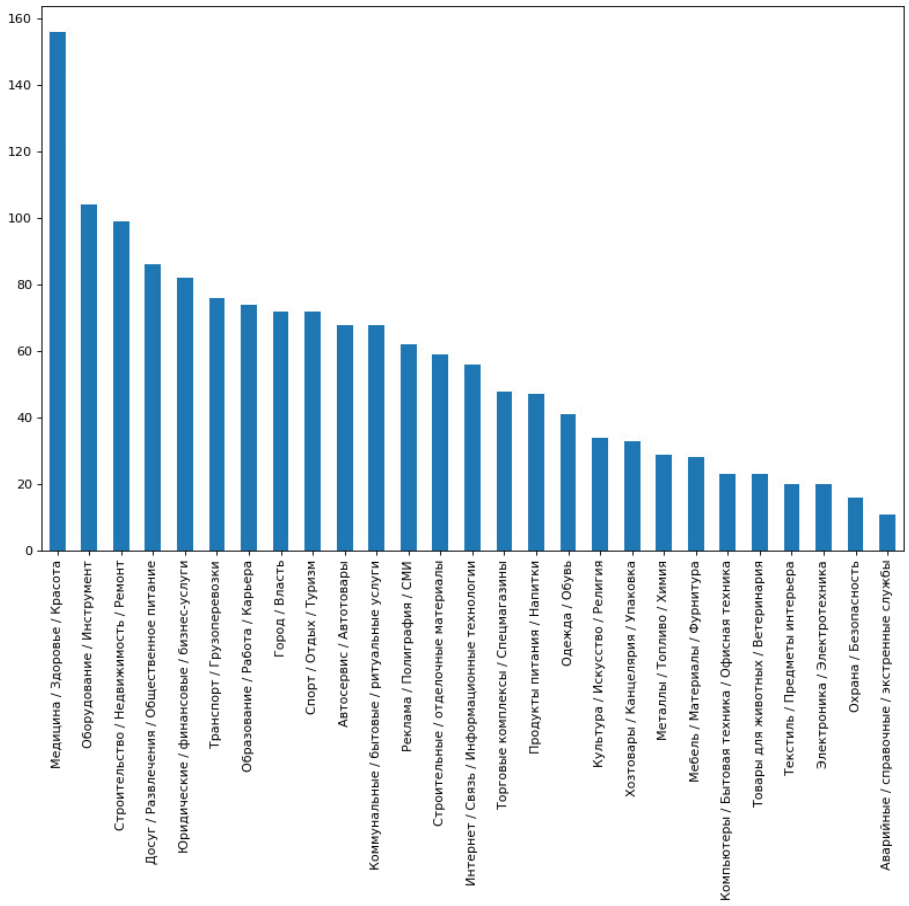



*Рис 6. Гистограмма распределения рубрик*

Также стоит отметить, что для каждого из трех вариантов нормативов, что представлены для данного проекта, сами наименования нормативов незначительно, но отличны друг от друга. Так, например, выведя 2 наиболее часто используемых норматива и самый редко встречающийся в файле рубрик, можно заметить, что смысл нормативов схож, однако написание их различно:

| | Код | Наименование категории объектов | Количество строк категории в файле 2ГИС |
|---|---|---|---|
| **3** | p_4 | Административные, офисные учреждения | 520 |
| **4** | p_5 | Предприятия торговли | 515 |
| **31** | p_32 | Бани, сауны | 1 |

*Рис 7.1 Таблица наиболее часто и редко встречающихся нормативов Пензенской области в файле рубрик*

| | Код | Наименование категории объектов | Количество строк категории в файле 2ГИС |
|---|---|---|---|
| **1** | L_2 | Офисные учреждения, служебные помещения, банки... | 543 |
| **3** | L_4 | Промтоварные магазины, аптеки | 477 |
| **18** | L_19 | Бани, сауны | 1 |

*Рис 7.2 Таблица наиболее часто и редко встречающихся нормативов Ленинградской области в файле рубрик*

| | Код | Наименование категории объектов | Количество строк категории в файле 2ГИС |
|---|---|---|---|
| **0** | b_1 | научно-исследовательские, проектные институты ... | 501 |
| **1** | b_2 | продовольственные магазины, промтоварные магаз... | 492 |
| **16** | b_17 | бани, сауны | 1 |

*Рис 7.3 Таблица наиболее часто и редко встречающихся нормативов республики Бурятия в файле рубрик*

Как можно заметить, единственное совпадение из трех примеров – норматив «Бани, сауны».

Также интересно отметить, что в файлах нормативов (см. в папке) разни́тся деление предприятий торговли. Например, в Пензенской области такие предприятия имеют только один норматив – «Предприятия торговли», в файле республики Бурятия – 2 разных норматива: супермаркеты и рынки имеют свою категорию. Что же касается Ленинградской области, 3 категории отданы под торговлю: помимо супермаркетов также разделены продовольственные и промтоварные магазины.

Соответственно, по этой причине нельзя сказать, что данная задача – это классификация, т.к. при постоянных признаках в данном случае меняется таргет.

## Предобработка данных

В процессе предобработки данных были заполнены пропущенные значения (ввиду того, что строки удалять никак нельзя). Помимо стандартного очищения данных и заполнения пропусков было использовано несколько методов векторизации текстов:

1. TF-IDF (с предшествующими ими очищением текста от знаков, токенизацию и лемматизацию)
2. Doc2Vec
3. Предобработанная модель BERT как наиболее популярная на сегодняшний день модель трансформатора.

Также в ходе анализа подтвердилось, что предобработанные модели работают эффективнее, чем непредобработанные.

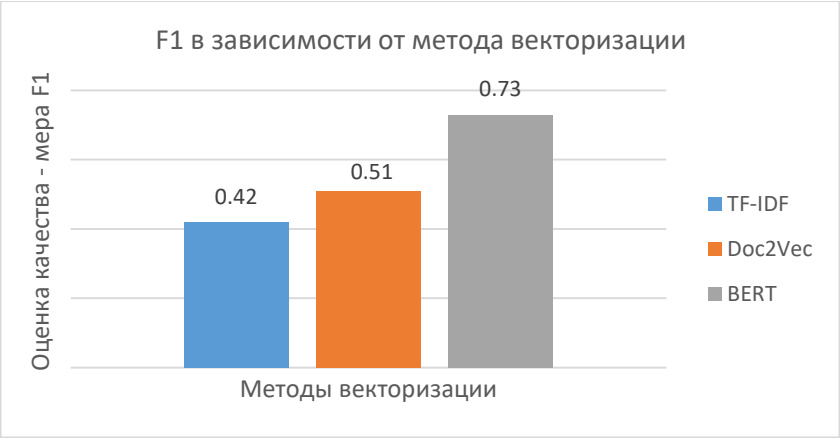
В процессе предобработки данных выяснилось, что наиболее затратная по времени – предобработка данных при помощи первого метода: TF-IDF с сопутствующей лемматизацией. В целом же распределение по времени выглядит так:


*Гистограмма 1. Время, потраченное на предобработку данных в зависимости от метода векторизации*

Doc2Vec в условиях ограниченного времени может стать оптимальным вариантом, если F1-мера достаточно высока для задачи.

## Выбор расчёта сходства

Как было отмечено выше, выбор расчёта меры сходства между текстовыми данными сделан в пользу косинусного расстояния. Для этого была написана функция, сравнивающая попеременно каждый из векторов двух списков при помощи косинусного сходства, далее выбиралось наибольшее значение и по индексу подтягивались текстовые значения из двух таблиц. На основе полученных результатов рассчитана метрика F1 для каждого из типов предобработки:




*Гистограмма 2. Среднее точности и полноты для разных вариантов предобработки данных*

## Результаты вычислений

На основании полученных результатов было принято решение о дополнительной мере расчёта близости текстов: написана нейросетевая модель со слоями долгой краткосрочной памяти LSTM.

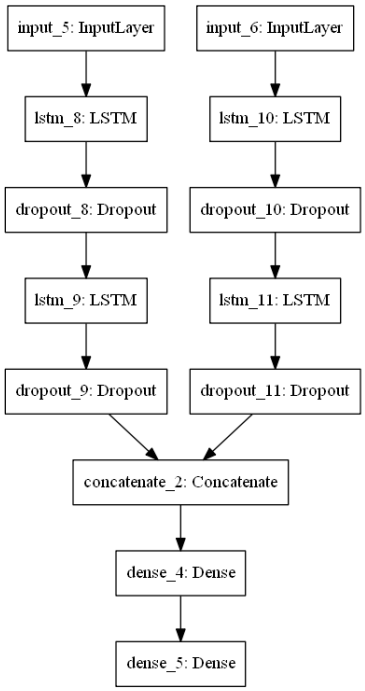



*Рис 8. Структура модели нейронной сети*

После обучения модели на 100 эпохах F1 на тестовой части данных используемого датасета Пензенской области составило более 0,90. На новых данных – на нормативах Ленинградской области и республики Бурятия 0,85 и 0,83 соответственно. Результат обучения виден на рисунке 7:

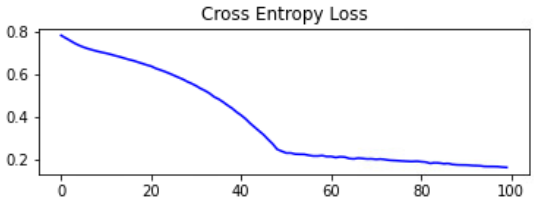



*Рис 9. Процесс обучения модели*

## Выводы

В данном проекте главной задачей был поиск семантического сходства словосочетаний для объединения данных двух реестров с дальнейшим использованием общего реестра.

Цель написания проекта – оптимизировать временные затраты аналитического отдела, значительно сократив время объединения рубрик и новых нормативов различных субъектов РФ.

По результатам работы можно сказать, что качество предсказания на новых нормативах составляет более 80%. При этом время выполнения задачи сокращается в 32 раза.

В качестве продолжения работы можно углубить нейронную сеть, задействовать другие текстовые столбцы из файла рубрик.

Таким образом, можно сказать, что данный проект рекомендуется использовать в аналогичных задачах в перспективе. Проект планируется реализовать в аналитическом отделе в течение ближайших 3 месяцев.

## Список используемой литературы


1. A Glance at Text Similarity. How To Compute the Similarity between Documents// (https://towardsdatascience.com/a-glance-at-text-similarity-52e607bfe920)
2. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ – Петербург, 2007. – С. 194 – 204.
3. О методе определения текстовой близости основанном на семантических классах / С. Х. Г. Бермудес, С. У. Керимова / Инженерный вестник Дюна, №4 / Ростов-на-Дону, 2016.
4. Semantic Similarity Using Transformers. Compute Semantic Textual Similarity between two texts using Pytorch and SentenceTransformers /Raymond Cheng/ 2021. // https://towardsdatascience.com/semantic-similarity-using-transformers-8f3cb5bf66d6
5. Semantic Textual Similarity Methods, Tools, and Applications: A Survey / Goutam Majumder, Partha Pakray, Alexander Gelbukh, David Pinto / Comp. y Sist. vol.20 no.4 Ciudad de México oct./dic. 2016