Talent Case Contest 2023

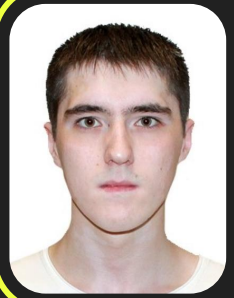
Решение команды “Кринженеры”

Состав команды

Гусев Иван

Java/Kotlin-backend developer

C++ enjoyer



Галимуллин Данис

C++/Kotlin/Python Android Developer Data Scientist

Симоненко Никита

C/C++ system programmer.

Rust enjoyer.

Руденко Юрий

JS/TS

React/Next

Front-end Developer

Новиков Михаил

C/C++, Python developer.

All enjoyer.

Постановка задачи

Дан датасет, содержащий строки на русском языке. Необходимо реализовать не менее 2-х алгоритмов выявления рерайта, учитывая, что запрещено использовать нейронные сети и готовые модели машинного обучения.

Этапы решения задачи

1) Произвести анализ датасета
2) Привести данные к практичному представлению (форматировать и структурировать)
3) Разработать и применить алгоритмы группировки строк
4) Провести оценку результатов работы разработанных решений

Анализ датасета

В результате анализа датасета были выявлены следующие способы рерайтинга:

1) Перестановка слов
2) Добавление(удаление) слов
3) Орфографические ошибки
4) Синонимичные конструкции

Дополнительные сложности (похожие строки, отличающиеся по смыслу):

1) Строки, отличающиеся только местоимениями
2) Строки, которые отличаются на 1 слово

Прочее:

1) Присутствуют строки, не имеющие рерайта
2) Присутствуют строки, имеющие более 1 рерайта

Форматирование данных

Предобработка датасета:

1) Приведение строк к нижнему регистру
2) Удаление знаков препинания
3) Замена “ё” на “е”
4) Удаление лишних пробелов

Входные и выходные данные представлены в формате JSON.

Решение №1 - Фильтрация

Идея решения заключается в поэтапном отборе групп(фильтрации датасета) с помощью перечисленных ниже алгоритмов:

1) Расстояние Карловского (TS = 0.89, CP = True)
2) Сравнение множеств слов (TS = 0.99, CP = False)
3) Расстояние Хэмминга (TS = 0.9, CP = False)
4) Косинусное сходство с векторизацией TF-IDF (TS = 0.7, CP = True)
5) Сходство Джаро-Винклера (TS = 0.9, CP = True)
6) Сравнение наборов местоимений

где TS - пороговое значение метрики, CP - необходимость сравнения наборов местоимений.

*Значения параметров алгоритмов подобраны экспериментально

Решение №2 - Равное голосование

Решение основано на пошаговой итерации по датасету с попарным сравнением строк на факт рерайта. В ходе сравнения применяются следующие алгоритмы и метрики:

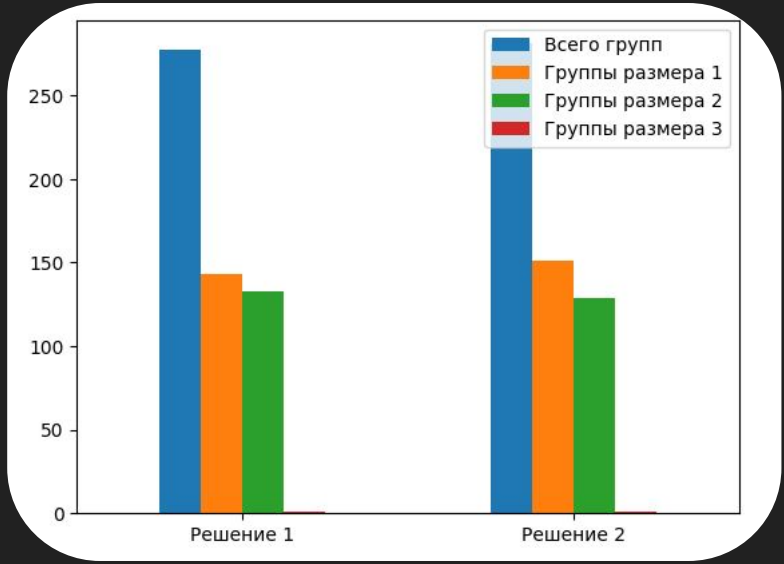
1) Косинусное расстояние
2) Расстояние Дамерау-Левенштейна
3) L2-расстояние
4) N-граммы в сочетании с индексом Тверского
5) Расстояние Хэмминга
6) Проверка наборов местоимений

В случае, если хотя бы одна метрика признала факт рерайта при совпадении по наборам местоимений, то из данных строк формируется группа. Для метрик было экспериментальным образом подобрано пороговое значение 0,9.

Сравнение решений

Различия между первым и вторым решениями по количеству найденных групп и размерам групп сравнительно небольшие. Данное различие обусловлено следующими особенностями алгоритмов:

- Решение №1: Гибкое решение, которое группирует строки, которые очень похожи друг на друга по смыслу и структуре.
- Решение №2: Более строгое решение, рассчитанное на поиск строк с минимальными отличиями по смыслу и структуре. Проигрывает по скорости решению №1.



Выводы

В ходе выполнения проекта было реализовано 2 решения, которые хорошо справляются с группировкой строк на исходном датасете. Данные алгоритмы имеют собственную специализацию, за счет чего есть некоторая гибкость в потенциальном практическом применении. Решения основаны на применении различных методов оценки схожести строк, что можно было бы потенциально выполнять параллельно, тем самым повысив производительность.