ФКН ВШЭ
Автоматическая Обработка Текста
Осенний семестр 2023

КТ 2

Выполнили: Успенский Д. А.
Беляев И. А.
Карбаев С. А.
Группа: 208

Содержание

ФКН ВШЭ
Автоматическая Обработка Текста
Осенний семестр 2023

КТ 2

Выполнили: Успенский Д. А.
Беляев И. А.
Карбаев С. А.
Группа: 208

1 Краткое описание контрольной точки 1

Давайте немного вспомним какая наша цель.

Мы участвуем в соревновании по оценке семантической текстовой связи между предложениями. В датасете есть два предложения и оценка их семантического сходства, полученная с помощью асессоров. Метрикой качества является коэффициент ранговой корреляции Спирмена.

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{(R(X))}\sigma_{(R(Y))}}$$

- ρ обозначает коэффициент корреляции Пирсона, но применяется к ранговым переменным
- $cov(R(X), R(Y))$ является ковариацией ранговых переменных
- $\sigma_{(R(X))}$ и $\sigma_{(R(Y))}$ являются стандартными отклонениями ранговых переменных

Коэффициент отражает, насколько хорошо предсказанные системой скоры согласуются с суждениями человека.

Для бейслайн решения мы использовали TF-IDF вместе с RandomForestClassifier. Получили корреляцию 0.435.
Авторы соревнования предложили использовать в качестве бейслайна долю общих слов в предложениях. Такой подход дал корреляцию 0.63.

Наша основная модель это Bert. Мы вычисляли для каждого предложения его Bert-эмбеддинги, затем оценивали схожесть между предложениями как косинусное сходство этих эмбеддингов. Его выбрали из-за простоты реализации, а также потому, что он улучшает анализ слов относительно базового решения организаторов.

2 Новая модель

Для улучшения метрики качества мы стали пробовать другие Bert-base модели. Рассмотрим подробнее каждую из них.

- **ALBERT**
  Модель ALBERT была предварительно обучена на BookCorpus, наборе данных, состоящем из 11 038 неопубликованных книг и English Wikipedia (исключая списки, таблицы и заголовки).
  Модель выдала корреляцию 0.08%.
- **RoBERTa**
  Модель RoBERTa была предварительно обучена на объединении пяти наборов данных:
  - BookCorpus — набор данных, состоящий из 11 038 неопубликованных книг;
  - English Wikipedia (за исключением списков, таблиц и заголовков);
  - CC-News - набор данных, содержащий 63 миллиона новостных статей на английском языке, просканированных в период с сентября 2016 года по февраль 2019 года;
  - OpenWebText, воссоздание набора данных WebText с открытыми источниками, используемый для обучения GPT-2;
  - Stories, содержащий подмножество данных CommonCrawl.

  Модель показала корреляцию 0.08%.
- DistilBERT
  Как и ALBERT, был предобучен на BookCorpus и English Wikipedia.
  Данная модель показала качество 0.797%.

По итогу мы использовали модель DistilBert из-за маленькой выборки входных данных. Она является более облегченной версией Bert, которая, практически, не теряет своей эффективности, но ощутимо увеличивает скорость.

Использование небольшой модели снизило переобучение, которое наблюдалось при использовании крупных Bert-base моделей.

3 Результаты на соревновании

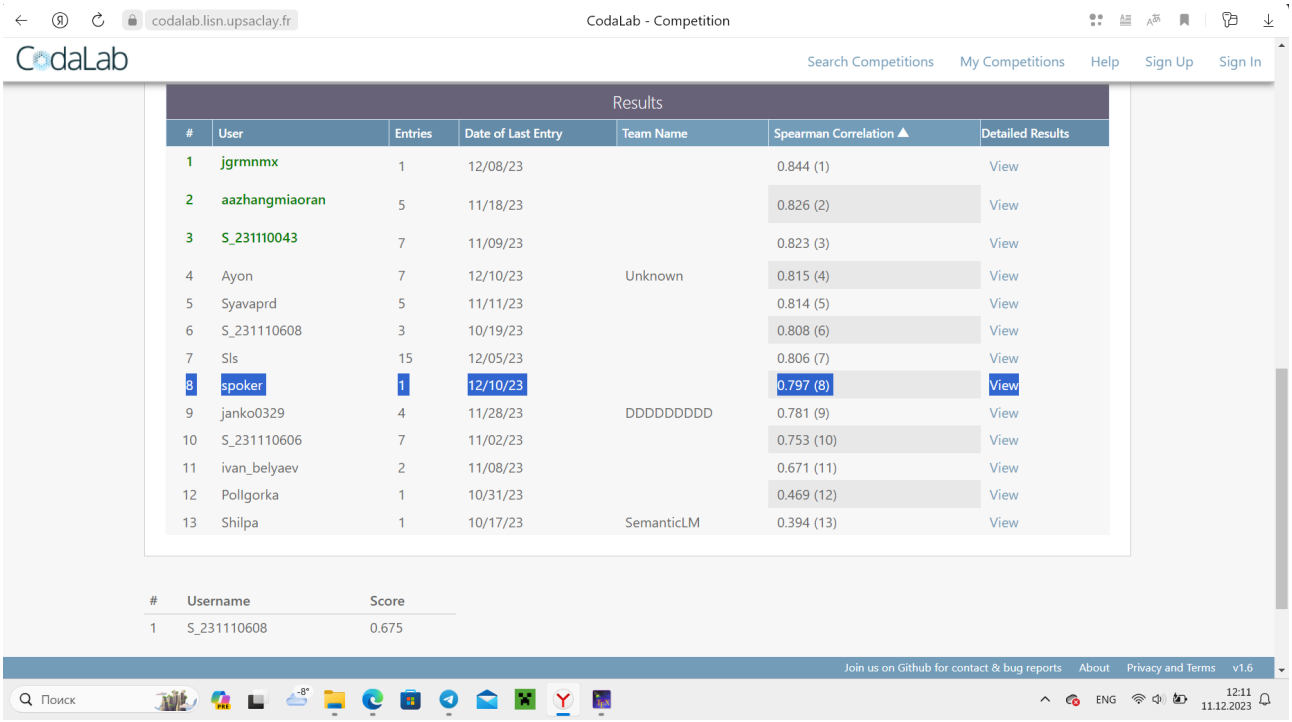

Рис. 1: Leaderboard

Выделение синим - наш текущий результат 0.797 (User spoker)

Наш прошлый результат - 0.671 (User ivan belyaev)

4 Итоги

В конечном итоге у нас получилось достичь около 80% точности ответа. Как ни странно, облегчение модели помогло нам улучшить результат.

Большинство участников, которые нас превзошли, получили результаты буквально на пару процентов выше. От лучшего результата мы отстаем не более чем на 5%