Обнаружение вопросов-дубликатов на Quora

Скорняков Кирилл, Козлов Илья

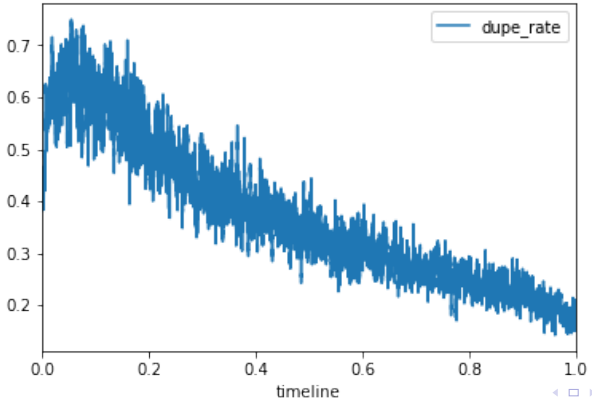
Москва, 2017

Постановка задачи

Даны пары вопросов с Quora, требуется определить являются ли вопросы в каждой паре дубликатами.
Метрика: logloss.
Трейн: 404 тыс. пар, 537 тыс. уникальных вопросов.
Тест: 2.3 млн. пар, 4.3 млн. уникальных вопросов.
Особенности: в тест были добавлены автоматически сгенерированные пары, достаточно шумная разметка, частота появления дубликатов зависела от времени.
Место на лидерборде: 11 из 3307.

Временная зависимость

Id вопроса сильно коррелирует с временной шкалой (можно поискать вопросы со спортивными событиями, праздниками).
На графике - зависимость частоты дубликатов от максиамльного id вопроса в паре.



Похожие соревнования

- **Avito Duplicate Ads Detection**
- Home Depot Product Search Relevance
- Crowdflower Search Results Relevance

Виды признаков

1. Различные функции похожести между двумя вопросами (косинус)
2. Текстовые характеристики каждого из вопросов (длина)
3. Bag-of-words/bag-of-ngrams представление текстов
4. Графовые характеристики каждого из вопросов (степень)
5. Графовые функции похожести (косинус)

Текстовые расстояния

1. Применяем к исходным текстам различные преобразования для получения набора трансформированных текстов
2. Считаем попарные расстояния между этими наборами
3. Эти расстояния - признаки для алгоритмов

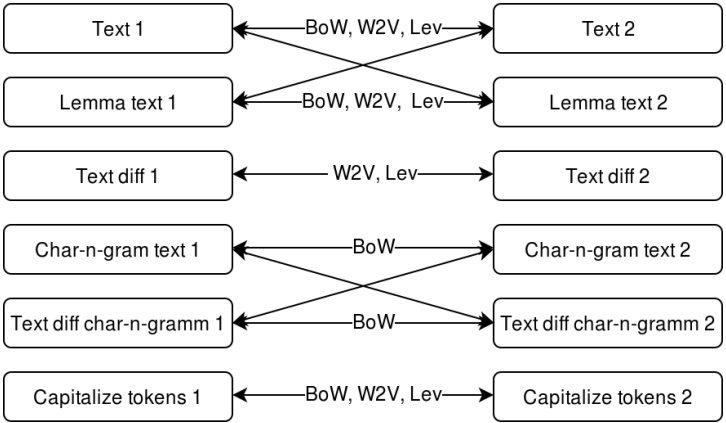
Виды преобразований текста для подсчета расстояний

1. Токенизация
2. Токенизация и лемматизация
3. Токенизация с обрезанием токенов до n букв
4. Токенизация с оставлением определенных частей речи
5. Токены разности текстов(левой, правой, симметричной)
6. Разбиение на буквенные n-граммы
7. Токены, начинающиеся с большой буквы
8. Токены/(буквенные n-граммы), с min df = 5, max df = 1500000

Текстовые расстояния

- BoW-based: косинус, жаккард, дайс, ассиметричный жаккард между bag-of-words, bag-of-ngrams(все тоже + svd)
- Расстояние Левенштейна по словам, символам(fuzzy-wuzzy)
- Normalized compression distance
- W2V-based(glove): косинус между взвешенным средним векторов вопросов(tf-idf, bm25), wmd, etc.

Текстовые расстояния



Признаки для текста

- Длина в словах/символах
- Число слов различных частей речи
- Сумма tf-idf(bag-of-words, bag-of-ngram etc.) для текста

Строим граф, где вершины это вопросы, а ребра - есть ли такая пара вопросов в датасете.

$$A = \begin{cases} 1, & \text{if (i, j) in sample;} \\ 0, & \text{otherwise} \end{cases}$$

$$A = \begin{array}{c} \\ i \end{array} \begin{array}{c} \begin{array}{cccc} & & j & \end{array} \\ \begin{pmatrix} 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 \end{pmatrix} \end{array}$$

Графовые признаки

Считаем обычные графовые similarity:

- Косинус
- Максимальная, минимальная степень вершин
- min, max, PageRank

Графовые признаки

Графовые признаки работают очень хорошо, попробуем добавить веса вершинам:

- Средний жаккард/левенштейн между вершиной и всеми ее соседями
- Среднее oof предсказаний хорошей одиночной модели для всех пар с данной вершиной

Дальше считаем min, max, cosine.

Автоматически сгенерированные вопросы в тесте

Такие пары не учитывались при подсчете лидерборда, но они входят в пары с настоящими вопросами. Удаление вопросов из одного слова при подсчете графовых фичей заметно ухудшало качество - это leak.

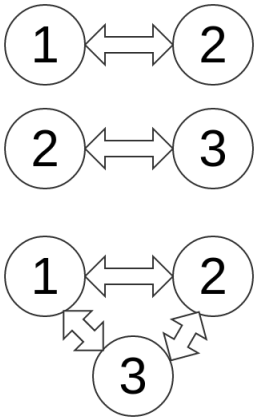
	question1	question2
520	Which is to quit smoking?	How
975	Is USMLE the life expectancy of a person with Hashimoto's thyroiditis disease?	What is
1138	What	How could best control my emotions and my negative thoughts?
1442	What	What is vs macro in programming?
1454	What	Why is s Raghuram Rajan resigning?
1986	What is	Why learn Babur not considered as an invader of India?
2558	What is	What is your required to accept personal checks as payment?
2934	What	Did Hindus and Muslims live peacefully before the British?
3182	How	What kinds of flowers do girls four besides roses?
3612	What	Is Arnab Goswami a title genius?

Текстовые sparse признаки

Также хорошо работало добавление симметричной разности и пересечения вопросов(буквенные и словесные n-граммы).

Кросс-валидация

Фолды строились так, чтобы между ними не было общих вопросов.



Различия трейна и теста

Среднее значение целевой переменной:

- Трейн: 0.37
- Тест: 0.17

Сглаживание различий в среднем значений целевой переменной

- **Scale_pos_weight в xgb**
- Oversampling
- Трансформация ответов

Трансформация ответов

Предполагаем, что распределение меток на тесте $P(y_{test} = i) = \gamma_i P(y_{train} = i)$, тогда:

$P(y_{test} = 1|x) = \frac{\gamma_1 p}{\gamma_1 p + \gamma_0 (1-p)}$, $p = P(y_{train} = 1|x)$

В итоге можно трансформировать ответы следующей функцией:

$f(x) = \frac{\gamma_1 x}{\gamma_1 x + \gamma_0 (1-x)}$

Трансформация ответов

На самом деле исходное предположение не совсем верно, лучше разбивать на 3 распределения(по словам победитлей):

- Оба вопроса встретились 1 раз
- Один из вопросов встретился 1 раз, 2ой больше
- Оба вопроса встретились больше одного раза

Стекинг

Разбиение пространства признаков

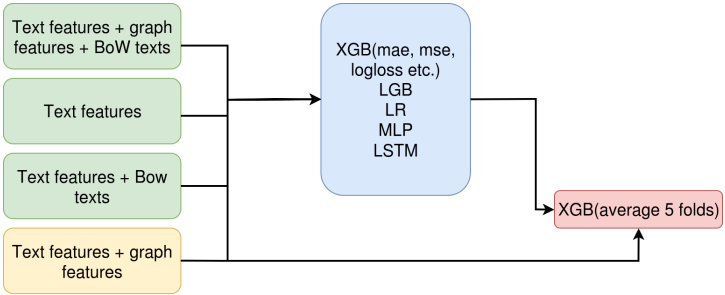
- Обычные фичи
- Обычные фичи + текстовые фичи
- Фичи без графовых признаков
- Только текст(для LSTM)

Алгоритмы

- Xgb(разные функции потерь, logloss, huber, mae, mse etc.)
- Lgb
- LogisticRegression
- LSTM
- mlp

2ой уровень

Средний xgb по 5 фолдам на исходных признаках (все кроме sparse текстов) + oof моделей первого уровня.



Что не сработало

- Embeding графа
- Делать свои пары (шинглы)

Вопросы

Вопросы?