Анализ веб-документов

Команда «Вялые Питоны»

- Резникова Александра
- Миклин Артём
- Вальков Павел
- Конов Михаил

Kaggle Leaderbord #6, score = 0.69590

История сабмитов. Начало

Линейная модель из дз

- Топ 15 пересечений по заголовкам

  score 0.59428

- Топ 15 пересечений по заголовкам + удаление стоп слов + лемматизация(nltk)

  Cross-validation score 0.65

  Score 0.66386

История сабмитов. Random forest #1

- Расстояние Jackard по нормализованным заголовкам и текстам

  CountVectorizer по словесным 2-grams и символьным (3-4)-grams, по ним взяты min, mean, median.

  Другие ngrams был отсеяны по feature importance.

  Score 0.68951

История сабмитов. Random forest #2

- К предыдущей модели добавлены фичи по сырым текстам

  CountVectorizer по словесным 2-grams и символьным (3-4)-grams, по ним взяты min, mean, median.

  Score 0.69186

История сабмитов. Random forest #3

- К предыдущей модели добавлены фича tf-idf.

  Для каждой группы были найдены топ 15 слов, характеризующие их тематику.

  Для всех документов было посчитано количество совпадений с тематическими словами.

  Score 0.69415

История сабмитов. Xgboost

- Те же фичи, что и для random forest

Score 0.69590

История сабмитов. SVD

- Топ 15 пересечений по заголовкам + удаление стоп слов + лемматизация(nltk) + tf-idf по словам + двухмерная декомпозиция SVD

Score 0.68543