Команда
IMPORT THIS

2 место на паблик лидерборде
f1score = 0.754

Состав команды:

Максим Кулис

Дмитрий Ломоносов

Дмитрий Андреев

Идея 1: мера Жаккара на заголовках

Используется модель линейной/логистической регрессии

- f1score на валидации ДО: **0.598**
- f1score на валидации ПОСЛЕ: **0.631**
- f1score на лидерборде: **0.569**

Идея 2: предобработка заголовка

Основную долю заголовков составляют цифры и символы русского и английского алфавитов. Остальные символы были отброшены

Слова, составляющие заголовки, обрабатывались стеммером **SnowballStemmer** из модуля **nltk**

Были отброшены слова из списка **stopwords**, взятого так же из модуля **nltk**

- f1score на валидации: **0.696**
- f1score на лидерборде: **0.647**

Идея 3: tf-idf и косинусное расстояние

Заголовки представляются в векторном виде с применением tf-idf преобразования

В качестве признаков берутся N минимальных косинусных расстояний между заголовками документа и других документов из его группы

- f1score на валидации: **0.727**
- f1score на лидерборде: **0.699**

Идея 4: другие модели

Были обучены модели

- Linear Regression
- Logistic Regression
- Support Vector Classification
- Random Forest Classifier

По результатом лучших моделей обучена новая, выбранная финальной

- f1score на валидации: **0.724**
- f1score на лидерборде: **0.725**

Идея 5: признаки, характеризующие группу

К уже имеющимся фичам были добавлены их средние в пределах группы

- f1score на валидации: **0.735**
- f1score на лидерборде: **0.754**

Спасибо за внимание!