Ранжирование комментариев поста

Участники:

- **Калинин Дмитрий Витальевич**
- **Горелов Василий Сергеевич**
- **Фофанов Максим Александрович**
- **Уваров Дмитрий Александрович**
- **Киселев Михаил Николаевич**

https://github.com/dmvmd/CupIT2023-Data-science

Проблема

Ранжирование комментариев на форумах происходит по количеству оценок. Но так пользователь не увидит недавний комментарий без оценок, который может быть наиболее полезный.

Задача

Предложить модель, которая основываясь на содержании топика и комментариев, выстроит наиболее подходящее ранжирование.

Существующие решения:

Ranking Comments on the Social Web
Chiao-Fang Hsu, Elham Khabiri, and James Caverlee

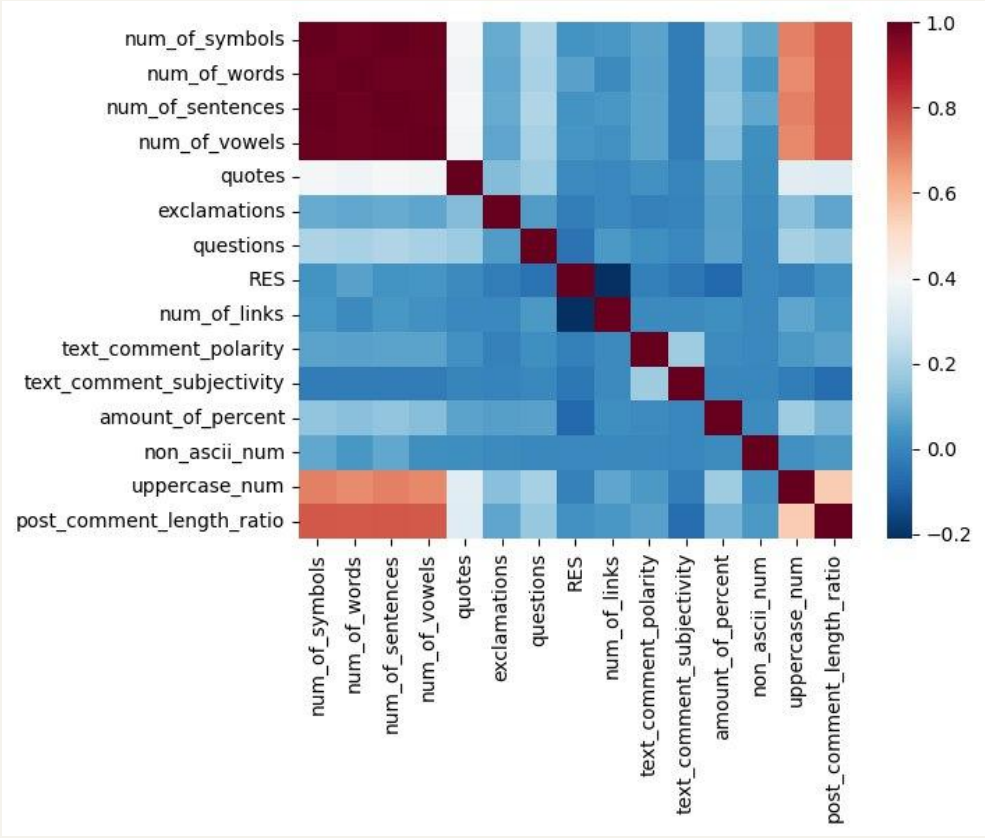
Подход авторов существенно опирается на метрики пользовательского взаимодействия.
Данный подход для нас неприменим в силу отсутствия информации о пользователях.

Relevance-Based Ranking of Video Comments on YouTube
Andrei Serbanoiu, Traian Rebedea

В статье используется подход на основе нейросетей, что требует большого количества исходных данных для обучения. В рамках нашей задачи такой подход не приведет к хоть сколько-нибудь положительным результатам из-за небольшого объема датасета.

Извлечение признаков из текста

- Количество предложений
- Количество символов
- Количество слов
- Количество гласных
- Количество цитат
- Количество восклицательных знаков
- Количество вопросительных знаков
- Метрика сложности текста - RES
- Количество ссылок
- Отношение полярности поста и комментария
- Отношение субъективности поста и комментария
- Количество чисел записанных как процент
- Количества не ASCII символов
- Количество заглавных букв
- Отношение длины поста к длине комментария



Модель для ранжирования catBoost Ranking

CatBoostRanker – это модель ранжирования, которая была разработана компанией Yandex и использует алгоритм градиентного бустинга на основе деревьев решений (Gradient Boosting Decision Trees, GBDT).
В качестве метрики модели используется NDCG.

Модель для классификации binary classification

Модель бинарной классификации мы использовали в следующем смысле: давали на вход соединенные признаки двух комментариев и получали на выходе 1, если первый комментарий стоит выше по рейтингу, 0 если ниже.

Подходы

Обучаем модель CatBoostRanker. Далее передаем в нее признаки комментария и получаем ранг комментария. После этого мы сортируем комментарии в каждом посте по его рангу. Для ранжирования комментариев с одинаковым рангом мы случайно прибавляли или вычитали маленькое число у одного из рангов этих двух комментариев.

NDCG score на валидационных данных: 0.88

Обучаем модель CatBoostClassifier. Потом, основываясь на их относительное положение, строим матрицу смежности и проверяем есть ли в них цикл. Если цикла нет, то с помощью топологической сортировки получаем ранги. Если цикл есть, то строим минимальное остовное дерево и рандомно сортируем комменты стоящие на одном уровне дерева.

NDCG score на валидационных данных: 0.5

Обучаем модель CatBoostRanker. После обучения первой модели CatBoostRanker мы заметили проблему, что он часто выдает различным комментариям одинаковые ранги, поэтому мы решили использовать, ранее обученную модель CatBoostClassifier для разделения комментариев с одинаковым рангом.

NDCG score на валидационных данных: 0.91

Результаты

После реализации трех подходов, мы получили, что модель для ранжирования от CatBoost, объединенная с моделью бинарной классификации дают наилучшие результаты.
После рассмотрения влияния на решение моделей мы пришли к тем выводам, что чтобы попасть в топ рейтинга комментариев Вам стоит:

1. Писать более разборчивые и длинные предложения, которые выражают вашу мысль
2. Используйте ссылки на сторонние сайты для ответы на вопросы
3. Приводите статистику, чтобы показать объективность ваших рассуждений
4. Не бойтесь задавать вопросы
5. Не стыдитесь использовать цитаты из других источников, не забывая оформлять цитату
6. Не стесняйтесь выражать эмоции, строить восклицательные предложения и использовать emoji