



Параллель 55



Анастасия Ивершинь

Привет! Мы команда **Параллель 55**

Готовы научиться писать самые популярные комментарии? 🔥

15 мар в 21:26 Ответить Поделиться

♡ 46



Адель Чернятов

Поехали! 🚀

15 мар в 21:26 Ответить Поделиться

♡ 48

Модель повысит интерес пользователей к комментариям

Формирование мета-признаков

Длина текста
Настроение
Смайлы
Тэги
Связь с постом и другими
комментариями

Предобработка текста

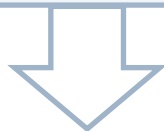
Очистка от символов
Обработка ссылок, смайликов
Разделение слов
Обработка сокращений

Финальная модель

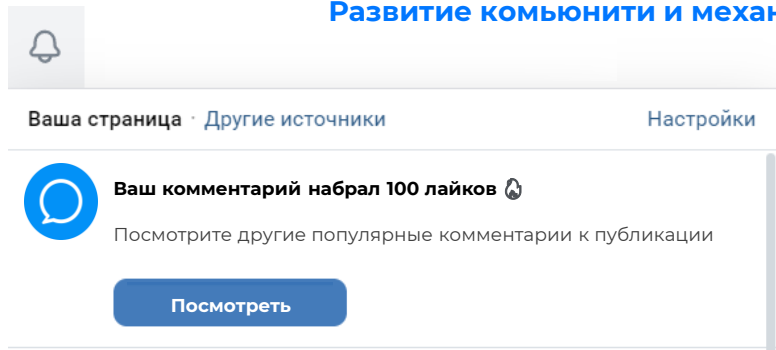
BERT + CatBoostRanker


0.89

NDCC

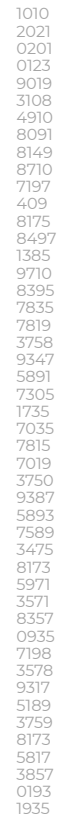


Развитие комьюнити и механизмов взаимодействия с комментаторами



 Сначала интересные ☒



[illegible]

Feature generation

Длина комментария

Число по отношению к длине комментария

• тэгов

#Сириус

• знаков ? и !

ЧТО?!

• смайлов



=D



:('



^_^ и др.

Флаг наличия в комментарии

Слов, написанных
ЗАГЛАВНЫМИ буквами

Цифр

[0-9]

Similarity

**Поста и
комментария**

Сходство Жаккара

Косинусное сходство



**Комментария с
другими
комментариями**

Косинусное сходство
(среднее, минимальное и
максимальное)



Также тестируется гипотеза о том, что менее популярными могут быть как комментарии, повторяющие смысл другого, так и комментарии «не по теме»

Sentiment

**Скор настроения
комментария**



**Флаг совпадения
настроения
поста и комментария**

Подход к моделированию

Выбор алгоритма

Критерии

- ✓ Масштабируемое решение
- ✓ Минимум экспертных корректировок результатов модели



Регрессия

LinearRegression
CatBoostRegression
XGBoostRegression
LSTM

Ранжирование

CatBoostRanker
XGBoostRanker
LambdaRankNN

Тюнинг бустингов проводился по max_depth

Метрики

NDCG

NDCG@3

Precision@3

Ключевая метрика

Дополнительные метрики

Пользователи обычно читают первые 2-3 комментария по порядку, поэтому важно правильно ранжировать именно их

Лучшая модель

BERT + CatBoostRanker

Метрики на кросс-валидации

0.89

NDCG

0.78

NDCG@3

0.33

Precision@3

Длина комментария

Число тэгов

Наличие цифр

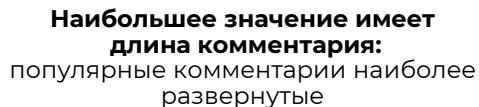
Jaccard сходство

Положительное настроение

Число ? и !

Слова КАПСОМ

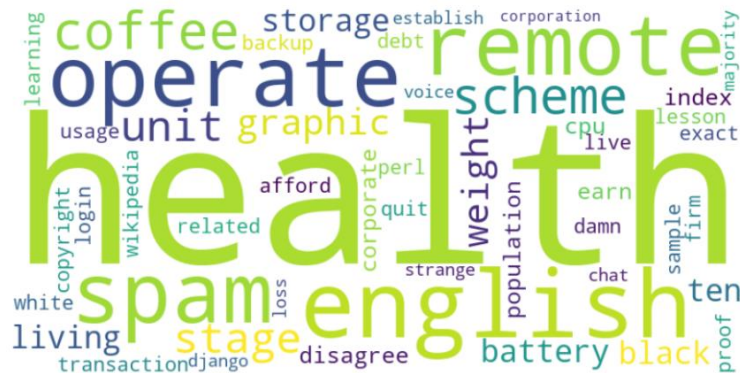
Косинусное сходство комментариев



О чем пишут в популярных комментариях?



А комментарии с таким содержанием не пользуются популярностью...



Взаимодействие с комментаторами – ключ к развитию комьюнити

Советы в процессе написания комментария

Опиши подробнее!

Мне очень понравился этот фильм



Поделись интересным фактом

Мне очень понравился этот фильм, смотрела не отрываясь из-за спецэффектов и классных актеров



Добавь #тэги

Мне очень понравился этот фильм, смотрела не отрываясь из-за спецэффектов и классных актеров! Кстати, фильм собрал уже больше 1 млн \$



Анастасия Ивершинь 🙌

Мне очень понравился этот фильм, смотрела не отрываясь из-за спецэффектов и классных актеров! Кстати, фильм собрал уже больше 1 млн \$ #Аватар #чтопосмотреть только что

Ответить Поделиться



Чтобы больше узнать о том, как писать популярные комментарии, воспользуйся нашим чек-листом!

Реакции на комментарии



Уведомления о лайках по контрольным точкам

Поддержка для топовых комментов

подсветка/анимация около сердечка, чтобы стимулировать кликать

Возможность создания голосования в комментариях

Соревновательный аспект

Если пользователей набрал 1000 лайков суммарно за период, то его последующие сообщения будут обрамлены рамочкой и выведены в топ

Инструмент продвижения личного бренда

Топ-комментаторы в различных тематиках получают продвижение

Параллель 55



**Ивершинь
Анастасия**

Team Lead

НГУ. Экономический факультет

ivershin_anastasia@mail.ru

+7 983 123 7442



**Бредихин
Арсентий**

Data Scientist

ЮГУ. Институт цифровой
экономики

bredihin.igorr@yandex.ru

+7 951 974 1803



**Сибгатуллина
Инна**

DL Engineer

Росбиотех. Инженерия и
технология пищевых производств

ophelials12@gmail.com

+7 968 719 4080



**Адель
Чернятов**

Data Scientist

КФУ. Прикладная математика

adelworkspace@mail.ru

+7 937 488 7147



**Полковникова
Елизавета**

Data Analyst

Финансовый Университет при
Правительстве РФ. Маркетинг
GeekBrains. Data Science курс

elizavetapolk@gmail.com

+7 960 141 7722



TO DO

Исправление орфографических ошибок, дублирующихся букв

Учесть неформальные сокращения, сленг

Провести более детальный анализ выбросов

Эмбединги Similarity для предобученной модели

Построить модель на эмбедингах (без PCA)

Оптимизация гиперпараметров

...