Автоматическая суммаризация отзывов

Софья Генералова, Анна Запорощенко

Мотивация (2-3 предложения)

Совершая ту или иную покупку, потребитель зачастую стоит перед сложным выбором. Описания от производителей редко вызывают доверие, и покупатель вынужден просматривать десятки отзывов, чтобы получить объективное представление о товаре. Решить эту проблему может инструмент, который соберет важные характеристики товара в единый текст.

Команда и роли

Анна Запорощенко - разработчик (бейзлайн), аналитик (анализ и предобработка первичных данных, результаты бейзлайна, результаты) и менеджер

Софья Генералова - разработчик (основной алгоритм), аналитик (результаты), менеджер поменбше

Данные (откуда, сколько)

https://www.kaggle.com/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews_Jun19.csv

- 10000 отзывов об 1400 отелях;
- в заголовок часто вынесены характеристики, сами тексты ими насыщены;
- узкая тематика: Accommodation & Food Services;
- есть вынесенное название отеля;
- есть оценка пользователя.

https://www.kaggle.com/snap/amazon-fine-food-reviews

- 500000 отзывов обо всякой еде;
- хорошие тексты, есть краткое саммари от пользователя;
- есть оценка пользователя.

Запасной датасет на случай, если после аналитики окажется, что первого не хватает.

Бейзлайн

1) sentiment classification с помощью CoreNLP, выделение отрицательных и положительных предложений
2) Rule-Based Filtering: убираем предложения неподходящей длины или с личными местоимениями (*he*, *she*, etc. but not *them*, *it*)
3) подсчет метрики близости (косинусная мера) на TF-IDF Weighted Embeddings
4) применение алгоритма K-means classification score.

Метрики

- ручная оценка текстов;
- + для реальной модели: доля выделенных ключевых слов в сгенерированном тексте

План

1) анализ данных: достаточное количество отзывов на один продукт, подходящая длина отзывов, эмоциональная окрашенность
2) реализация бейзлайна
3) экстракция положительных/отрицательных словосочетаний с помощью Stanford CoreNLP
4) кластеризация, выбор наиболее разнообразных “ключевых слов” для каждого продукта
5) * NER: выделение наименования продуктов для включения в список ключевых слов (в случае второго датасета)
6) дообучение GPT на отзывах
7) применение conditional GPT с ключевыми словами
8) оценка полученных текстов

Наша доска: https://trello.com/b/JWGKtWqz/reviews-summarization

Литература, что оттуда взяли

Stanford CoreNLP (sentiment analysis + NER):
https://stanfordnlp.github.io/CoreNLP/annotators.html

Идея бейзлайна: http://www.kiraradinsky.com/files/www19_slava.pdf

Про conditional text generation:
https://towardsdatascience.com/conditional-text-generation-by-fine-tuning-gpt-2-11c1a9fc639d

https://github.com/minimaxir/gpt-2-keyword-generation