Автоматизированный анализ мнений о товарах

Сергей Ермаков

Проблема

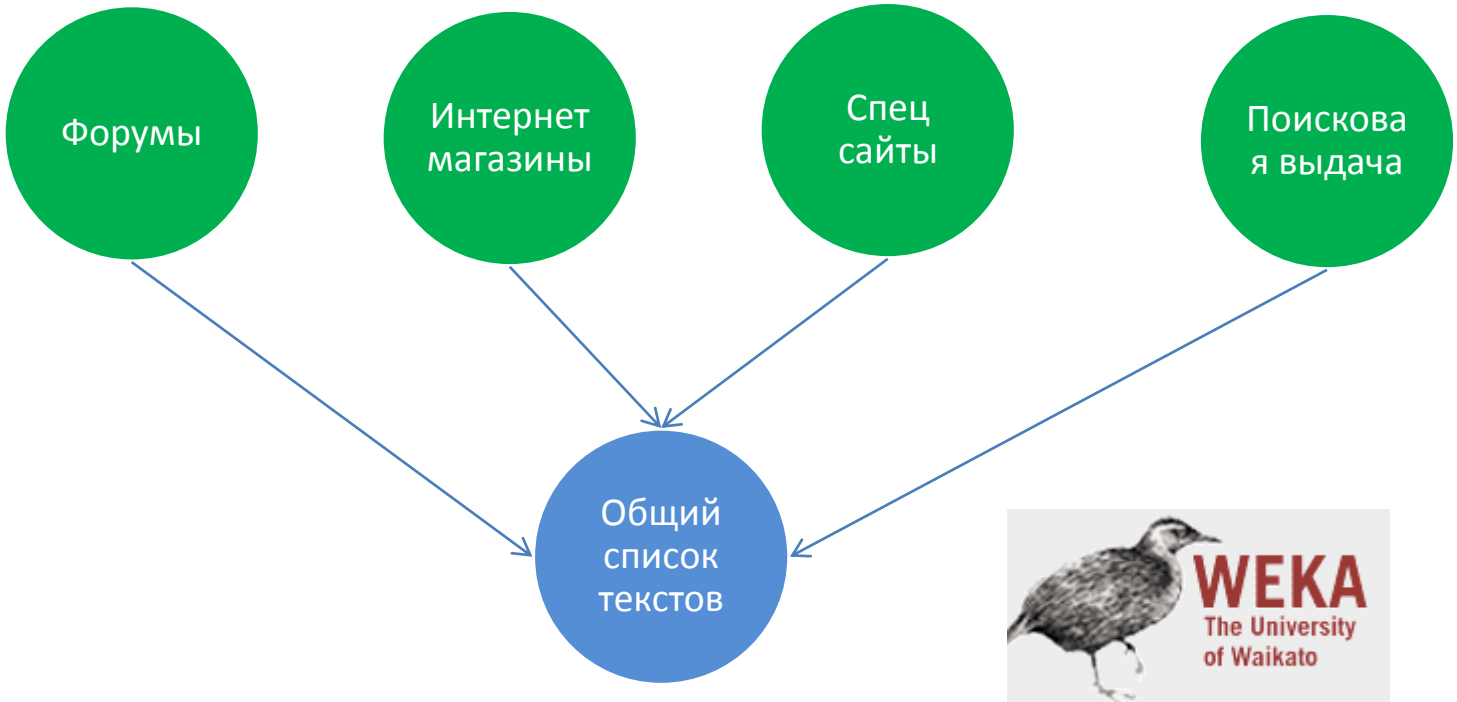
- При выборе покупки в данный момент мы ищем отзывы и комментарии о данном продукте
- Существуют сервисы, которые помогают нам в этом
- Для получения полной картины необходимо обработать большое количество информации


Or

?

Яндекс
маркет

Сбор данных



- Анализ HTML
- Выделение комментариев среди частей страницы

Feature Extraction

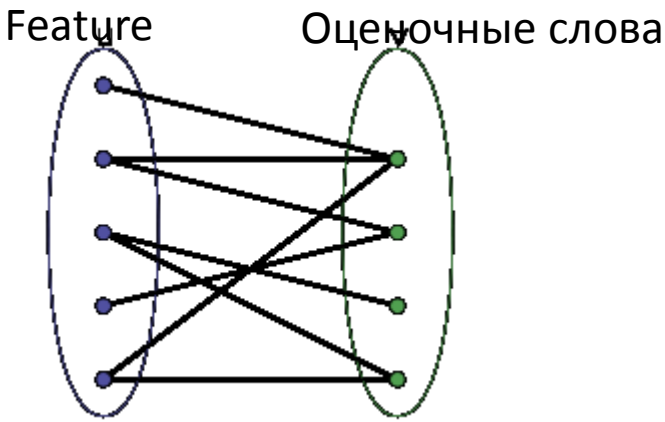
- Перед выделением фич, нужно провести нормализацию орфографии
- Исходные отзывы разбиваются на фрагменты соответствующие одной feature
- Для feature extraction необходим синтаксический парсер, но простые отношения можно определять по словарю, остальные по Wikipedia

Ранжирование фрагментов

- Первоначально отзывы оцениваются по тому ресурсу, откуда они были получены (рейтинг сайта)
- Фрагменты оцениваются на основе
  - Длины предложений
  - Положения
  - Количества знаков восклицания
  - наличие слов с высоким индексом TF-IDF

Sentiment Analysis

- На основе двудольного графа



- Планируется большую часть словаря извлекать бутстрепом с помощью шаблонов и структуры ресурса

Результаты

- В результате проведения анализа тональности каждый фрагмент оценен
- Оценки по каждой характеристике усредняются с учетом ранга фрагмента
- В итоге мы получаем категоризированную оценку товара

Текущее состояние

- Основной акцент на feature extraction и sentiment analysis.
- Пока на базе существующих размеченных корпусов (дорожки РОМИП)
- Реализованы вспомогательные части: нормализатор орфографии, подключен словарь Зализняка, построен индекс, ведется работа над построением тонального словаря.

Спасибо за внимание