1. При такой организации пользователь может получить обзор потребительских мнений в максимально короткий срок.
2. Чтобы не было недопонимания, я вынесла несколько определений на слайд. Важно понять разницу между аспектом и идеальным аспектом, т.к. в дальнейшем я буду использовать данные термины.
3. Метод построения аспектной иерархии на основе пользовательских отзывов является популярным для исследований на английском языке. Имеется немало альтернативных исследований и программ, работающих с англоязычным сегментом интернета.

Однако русскоязычный вариант метода не нашел отражения в обнаруженных источниках, поэтому есть основания утверждать, что подобное исследование в русскоязычном формате проводится впервые.

Работа выполнена на основе статьи, в которой описан подход к организации иерархий аспектов для пользовательских отзывов об электронных устройствах в англоязычном сегменте. Статья написана неподробно. Подход в статье был взят за основу и доработан, подробно описан в дальнейшем с учетом всех недостатков.

1. Технологии
2. Для построения иерархии требуется получить пользовательские отзывы.

Самым удобным способом получения является парсинг вебстраницы и сохранение отзывов в базу данных.

Источник - yandex market. API не позволяло получить требуемое для работы количество информации за заданный промежуток времени. Парсинг данных, но yandex market постоянно блокировал запросы от программы, в связи с чем было принято решение о поиске альтернативы.

Для работы был выбран сайт ulmart.ru, а именно раздел электронных товаров.

Для получения отзывов со всех страниц товаров данного сайта был написан парсер, выявляющие отзывы и добавляющий их в базу данных. Парсер был выполнен с использованием библиотеки BeautifulSoup 4 версии. Получено 24093 отзывов.

1. Имеется большое количество категорий, все они отражены на экране. Отзывы пользователей о товарах именно из этих категорий были получены методом парсинга.
2. Отзыв - объединение 3 частей: положительное резюме, отрицательное резюме и комментарий в свободном стиле, однако данное разделение учтено не было

Все остальные детали отзывов были проигнорированы.

Создана БД с 415 идеальными аспектами – описание характеристик того или иного товара.

1. О получении идеальных аспектов уже было сказано, это верхняя ветвь схемы на слайде. Для них была построена иерархия вручную.

Теперь более подробно про нижнюю ветвь схемы. На ней представлены шаги, каждый из которых я более подробно объясню далее. Суть алгоритма - из полученных отзывов выявлено 45435 ключевых аспектов, для них рассчитаны семантические дистанции для каждой из пар и в итоге построена иерархия аспектов для отзывов пользователей об электронных устройствах на основе идеальной иерархии, построенной вручную.

1. Так как отзывы, полученные с ulmart.ru, приходят зашумленными и содержат в себе большое количество случайных символов, то первым делом производится очистка входных данных.

Удаление пользовательских повторов, знаков препинания -> нижний регистр и обязательное наличие в конце точки -> выявление частей речи всех слов в каждом из отзывов -> поиск существительных (метод построения синтаксического дерева предложения, что позволяет найти нужные связи).

Выявление значимых аспектов в контексте данной работы. Использованы возможности методов машинного обучения, а именно One-Class SVM. все имеющиеся аспекты получают метку 1 или -1 в зависимости от того, присутствует ли данный аспект в списке идеальных аспектов -> данные делятся в соотношении 80 к 20 на тренировочные и тестовые -> тренировочные очищаются от тех аспектов, метка которых является -1 -> обучение и предсказание классификатором меток для тестовых аспектов.

421715 аспектов. Данное количество слишком велико для цели данной работы. Поэтому следующим шагом была группировка, под этим подразумевается удаление всех дублей, что существенно снизило количество аспектов и их стало 45435.

В дальнейшем для ускорения разработки программы количество аспектов было еще более сужено до 1000 экземпляров (частые). Последующие результаты и числа будут актуальны в рамках данной 1000. Данный набор перекликается по содержанию с 415 идеальными аспектами, более 50% аспектов из них содержатся в суженом аспектном наборе.

1. Для семантического расстояния 6 характеристик, о каждой из них я расскажу подробнее.
2. PMI характеристика содержит в себе 2 характеристики. Разница в характеристиках заключается в корпусе отзывов/предложений, который передается для дальнейшей обработки в метод расчёта PMI. Формула для расчета на экране, ах и ау – это аспекты из пары.
3. 2 аспекта в паре компьютер и экран и 3 изначальных отзыва пользователей до какой-либо обработки. Количество отзывов, где присутствуют оба аспектов всего 1. Количество отзывов, где есть аспект компьютер = 3, а экран = 1. Эти данные позволяют нам произвести вычисление характеристики pmi review. В случае с pmi sentence расчеты производятся аналогичным образом, только все количества мы берем относительно предложений, а не отзывов.
4. Contextual характеристика также содержит в себе 2 характеристики. Для локальной характеристики контекст - набор 2 левых и 2 правых слов от аспекта во всех отзывах, где он встречается. Для глобальной характеристики контекст - набор отзывов, где встречается конкретный аспект.
5. Количество слов в локальном контексте для аспекта всегда кратно 4, если слово стоит на границе, не имеет 2 слов слева или справа, то тогда подставляются слова \_BEGIN\_SENTENCE\_ и \_END\_SENTENCE\_. Для глобального контекста просто ищем отзывы, где встречается наше аспект.

Каждый контекст -> вектор с помощью vectorizer.fit\_transform -> smoothing -> n-gram, каждое значение в котором, если он не пуст, делится на количество слов в контекстах аспектов -> для каждой пары аспектов kl-divergence между n-gram’ами. Данное значение и является результатом характеристики context. Разница лишь в том, где ищутся контексты для аспектов.

1. Далее syntactic feature. Для каждого из отзывов для ускорения работы программы с помощью API ИСП РАН для отзывов строятся синтаксические деревья. Для каждой пары аспектов выполняется поиск кратчайшего расстояния между ними в дереве.
2. Возьмем отзыв, который на слайде. Для примера приведен графический вариант синтаксического дерева для 1 предложения данного отзыва. Для общего дерева отзыва выполняется поиск кратчайшего пути между указанными аспектами. Если количество отзывов, где присутствует пара аспектов больше одного, то берется среднее арифметическое длин кратчайших путей, в противном случае просто берем данный путь. Это и будет значением величины syntactic.
3. Lexical характеристика самая простая в вычислении. Формула на слайде.
4. Пример на слайде
5. Все характеристики нужны для вычисления семантического расстояния между аспектами.
6. Для вычисления семантической дистанции было выбрано два способа расчётов, один приведен в статье, которая легла в основу данной работы, а второй является примером использования методов машинного обучения на практике.

В первом способе производится вычисление вектора w, представляющего собой набор характеристик (PMI, Lexical, Syntactic, Contextual) идеальных аспектов. Для каждой пары аспектов семантическое расстояние равно формуле со слайда.

1. Для 1 подхода было решено также изучить влияние каждой из характеристик на итоговый результат для семантического расстояния. В итоге была получена таблица, которую вы видите на слайде. По очереди применялся алгоритм вычисления семантической дистанции между аспектами в парах для разных наборов характеристик. Также было попробовано добавление еще одной 7 характеристики, чье значение константно равнялось 1.
2. Второй способ представляет собой пример использования *RandomForestRegressor*, который являет частью пакета библиотеки *sklearn.ensemble*. В качестве входных данных для обучения модели поступают результаты идеальных аспектов, а именно их предвычисленные реальные длины путей для каждой пары аспектов, а также набор из вышеупомянутых 6 характеристик для каждой пары аспектов.

Обучившись модель получает задание предсказать результаты уже для реальных данных.

1. Проведенное изучение двух подходов к вычислению семантического расстояния показало наличие, как преимуществ, так и недостатков у обоих. Дальше вся инфа на слайде.
2. В итоге мы получаем иерархию аспектов. На слайде приведена часть иерархии, т.к. дерево велико (много аспектов).