ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Департамент программной инженерии

СОГЛАСОВАНО
Доцент факультета компьютерных наук базовой кафедры «Системное программирование» НИУ ВШЭ, канд. физ.-мат. наук
_____ Д.Ю. Турдаков
«__» _____ 2017 г.

УТВЕРЖДАЮ
Академический руководитель образовательной программы «Программная инженерия» профессор департамента программной инженерии, канд. техн. наук
_____В.В. Шилов
«___»_____2017 г.

Отчет
по курсовой работе

Построение иерархии аспектов по пользовательским отзывам об электронных устройствах

по направлению подготовки бакалавров 09.03.04 «Программная инженерия»

Выполнила:
студентка группы БПИ143 образовательной программы
09.03.04 «Программная инженерия»
Репина А.А.

_____
Подпись, Дата

Москва, 2017

Реферат

Отчет 20 с., 5 рис., 1 табл., 14 источн., 4 прил.
Ключевые слова: аспекты; пользовательские отзывы; иерархия; электронные устройства; семантическое расстояние; характеристики.

В отчете представлены результаты курсовой работы на тему “Построение иерархии аспектов по пользовательским отзывам об электронных устройствах”, выполненной на основе приказа Национального исследовательского университета "Высшая школа экономики" № 2.3-02/0812-01 от 08.12.2016.

Объект исследования – аспектная иерархия пользовательских отзывов.
Предмет исследования – технология построения иерархии аспектов по пользовательским отзывам об электронных устройствах.
Цель исследования. Данная работа представляет собой подход организации различных аспектов продукта, относящегося к категории электронные устройства, в иерархию на основе знаний о потребительских отзывах. Основываясь на произвольной иерархии (построенной вручную), создается иерархическая организация опросов потребителей по различным аспектам продукта и совокупным мнениям потребителей по этим аспектам. При такой организации пользователь может получить обзор потребительских мнений в максимально короткий срок.
Научная новизна работы. Достоверность научных результатов подтверждена результатами экспериментальных исследований с прототипом программы построения иерархии аспектов.
Методы проведения исследования. В качестве основного логико-теоретического метода проведения исследования использовалось моделирование: разрабатывалась программа, наделенная исключительно той функциональностью, которая необходима для решения задач настоящей НИР.
Практическая значимость. С точки зрения практического применения, разработанный инструмент может дополняться функциональностью и использоваться для решения более широких задач, выходящих за рамки данной курсовой работы.
Результаты работы. По итогам выполнения исследования все поставленные цели были достигнуты.

Содержание

Определения

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями:

Парсер - скрипт или программа, которые используются для сбора информации с сайтов
PMI (Pointwise Mutual Information) - точечная взаимная информация являются мерой ассоциации, используемой в теории информации и статистике. В отличие от MI, которая основывается на PMI, PMI относится к отдельным событиям, тогда как MI относится к среднему значению всех возможных событий.
KL-divergence - это неотрицательнозначный функционал, являющийся несимметричной мерой удаленности друг от друга двух вероятностных распределений.
Context характеристика - KL-divergence между языковыми моделями.
Lexical характеристика - разница в длине слов между двумя аспектами.
Syntactic характеристика - средняя длина кратчайшего синтаксического пути между парами аспектов в дереве.
Smoothing – сглаживание данных.
N-gram - последовательность из n элементов, где элементы могут быть звуками, слогами, словами или буквами.
JSON - текстовый формат обмена данными, основанный на JavaScript.
Семантическое расстояние - насколько два аспекта близки друг к другу, это возможно определить с помощью набора характеристик.
Дерево иерархии - расположение элементов системы в порядке подчиненности (от высшего к низшему).
Аспект - слово или набор слов, главным словом в которых является существительное.
Идеальный аспект - слово-характеристика, полученная из технических описаний товаров на сайте ulmart.ru

Введение

С быстро растущей электронной торговлей большинство розничных веб-сайтов побуждают потребителей писать обзоры, чтобы выразить свои взгляды на различные аспекты продуктов. Огромная коллекция отзывов потребителей теперь доступна в Интернете. Эти обзоры стали важным ресурсом для потребителей и бизнеса. Потребители обычно ищут информацию о качестве в онлайн-опросах потребителей перед покупкой продукта, в то время как многие компании используют онлайн опросы в качестве важного ресурса в их разработке продукта, маркетинге и управлении взаимоотношениями с клиентами. Однако обзоры дезорганизованы, что приводит к трудностям в навигации информации и приобретении знаний. Пользователю нецелесообразно изучать обзоры потребительских мнений по различным аспектам продукта из тысяч источников. Среди аспектов продукта также неэффективно для пользователя просматривать данные и мнения потребителей по определенному аспекту. Таким образом, существует острая необходимость в организации опросов потребителей, чтобы превратить обзоры в полезную структуру знаний. Поскольку иерархия может улучшить представление и доступность информации, то видится разумным организовать аспекты продукта в иерархии и, соответственно, создать иерархическую организацию опросов клиентов.

Чтобы автоматически получить иерархию аспектов из обзоров, можно было бы обратиться к традиционным методам генерации иерархии, которые сначала идентифицируют понятия из текста, а затем определяют отношения между родителем и ребенком. Тем не менее, основанные на шаблонах методы обычно страдают от несогласованности отношений «родитель-потомок» между понятиями, в то время как методы на основе кластеризации часто приводят к низкой точности. Таким образом, путем непосредственного использования этих методов для создания иерархии аспектов из обзоров потребителей полученная иерархия обычно неточна, что приводит к неудовлетворительной организации обзора. Данная работа предназначена для построения иерархии аспектов по пользовательским отзывам об электронных устройствах на основе произвольной иерархии, построенной вручную, что пользоволяет максимально быстро предоставить пользователю структурированные данные потребительских мнений. В широком смысле, с точки зрения актуальности и практического применения разработанный алгоритм может пополняться дополнительной функциональностью и использоваться для дальнейших, более глобальных исследований в данной области.

Цель исследования: Данная работа представляет собой подход организации различных аспектов продукта, относящегося к категории электронные устройства, в иерархию на основе знаний о потребительских отзывах. Основываясь на произвольной иерархии (построенной вручную), создается иерархическая организация опросов потребителей по различным аспектам продукта и совокупным мнениям потребителей по этим аспектам. При такой организации пользователь может получить обзор потребительских мнений в максимально короткий срок.

Задачи исследования:

- изучение наиболее эффективных и применимых на практике методов построения иерархии аспектов по пользовательским отзывам об электронных устройствах;
- создание метода построения иерархии аспектов по пользовательским отзывам об электронных устройствах с использованием анализа данных и методов машинного обучения;

- проведение эксперимента, позволяющего определить точность построения иерархии аспектов, с помощью предложенного метода.

Предметом исследования в данной работе является технология построения иерархии аспектов по пользовательским отзывам об электронных устройствах.

Методы исследования:

- изучение монографических публикаций и статей;
- сравнительный анализ;
- машинное обучение;
- анализ данных.

Новизна исследования и **достоверность исследования** определяются следующим:

Метод построения аспектной иерархии на основе пользовательских отзывов является довольно популярным для исследований на английском языке, однако русскоязычный вариант метода не нашел отражения в обнаруженных источниках, поэтому в силу малого количества исследований в данной области есть основания утверждать, что подобное исследование в русскоязычном формате проводится впервые. Во время проведения экспериментов погрешность может возникнуть на любом из этапов в связи с зашумленностью входных данных, которыми являются отзывы рядовых пользователей интернета. Под зашумленностью подразумевается наличие грамматических и пунктуационных ошибок, случайных повторов, логических несоответствий. Приняв во внимание данное обстоятельство, результаты исследования в целом считаются достоверными.

Теоретическая значимость работы обусловлена тем, что не существует русскоязычных аналогов методов, которые строят подобные иерархии на основе пользовательских отзыва, не смотря на всю актуальность. Таким образом, исследование в данной области позволит улучшить сложившуюся ситуацию.

Практическая ценность работы заключается в дальнейшем возможном использовании разработанного метода в области анализа текстов. На основе построенной иерархии можно проводить аспектный анализ эмоциональной окраски текста.

Основная часть

1. Обзор и анализ источников

Имеется несколько альтернативных исследований и программ, работающих с англоязычным сегментом интернета, например, HASM (Hierarchical Aspect-Sentiment Model), JST (Joint modeling of Sentiment and Topic), ASUM (Aspect Sentiment Unification Model), JST (Joint Sentiment-Topic). Однако похожего метода для обработки русскоязычных пользовательских отзывов и представления их в виде иерархии аспектов нет. Также в отличие от предыдущих работ, производится фокус на автоматическом создании иерархии аспектов для иерархической организации отзывов потребителей. Есть несколько взаимосвязанных работ по изучению онтологии, в которых сначала определяются понятия из текста, а затем определяются отношения между родителями и потомками между этими понятиями с использованием методов на основе шаблонов или кластеризации [8]. Шаблонные методы обычно определяли некоторые лексические синтаксические шаблоны для извлечения отношений, в то время как методы кластеризации в основном использовали иерархические методы кластеризации для построения иерархии [9]. В некоторых работах предлагается интегрировать методы на основе шаблонов и кластеров в общую модель, такую как вероятностная модель [10] и метрическая модель [11].

Данная работа была выполнена на основе статьи Дж. Ю, Ж. Джа, М. Венг, К. Венг, Т. Чуа, “Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews” [4], в которой описан подход к организации иерархий аспектов для пользовательских отзывов об электронных устройствах в англоязычном сегменте. Стоит отметить, что статья была написана неподробно, в связи с чем многие важные для понимания логические переходы упущены. Именно поэтому подход в статье был взят за основу в данной работе и доработан, и подробно описан в дальнейшем ниже с учетом всех недостатков. Для проведения исследования использовался логико-теоретический метод моделирование: была разработана программа на языке Python, в которой представлено алгоритмическое решение задачи.

2. Теоретическая часть

При разработке инструмента для проведения исследования одной из основных задач являлась создание эффективного по времени алгоритма построения иерархии аспектов. В результате был разработан алгоритм с вычислительной сложностью $O(n^2)$, где n – количество аспектов. Эффективность такого алгоритма обоснована тем, что величины-характеристики вычисляются для каждой пары аспектов (пара рассматривается только 1 раза, если пара «А Б» была, то пара «Б А» нас не интересует, ее вычислять не требуется), что необходимо для дальнейшего расчётов семантических дистанций и построения финальной иерархии. Описание алгоритма смотреть в пункте 3 настоящего отчета.

3. Описание эксперимента, анализ и оценка полученных результатов.

Программа предоставляет возможность построения иерархии аспектов по пользовательским отзывам об электронных устройствах. Для этого она содержит перечисленные ниже функции:

1) метод, принимающий в себя адрес сайта и возвращающий набор отзывов со всех страниц, находящихся по адресу;
2) метод, формирующий корпус аспектов, принимая в себя набор идеальных аспектов и отзывов;
3) метод, принимающий на входе корпус аспектов, корпус отзывов и корпус их предложений и возвращающий величины PMI для данных корпусов;
4) метод, принимающий на входе корпус аспектов, корпус отзывов и корпус их предложений и возвращающий величины Context для глобального и локального контекстов;
5) метод, принимающий на входе корпус аспектов и возвращающий величину Lexical для каждой пары аспектов;
6) метод, принимающий на входе корпус аспектов, корпус предложений и их синтаксических деревьев и возвращающий величину Syntactic для каждой пары аспектов;
7) метод, принимающий на вход корпус аспектов, корпус характеристик (вычисленных для каждой пары) и возвращающий семантическое расстояние для каждой пары аспектов;
8) метод, принимающий на входе корпус аспектов и корпус идеальных аспектов и возвращающий иерархию аспектов по пользовательским отзывам.

3.1. Описание парсинга входных данных

Для построения иерархии требуется получить пользовательские отзывы. Самым удобным способом их получения является парсинг вебстраницы и сохранение отзывов в базу данных.

Изначально в качестве источника был выбран yandex market, однако предоставляемое сервисом API не позволяло получить требуемое для работы количество информации за заданный промежуток времени. Далее был опробован способ парсинга 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 [12]. Всего было получено 24093 отзывов. Отзыв представляет собой объединение 3 частей: положительное резюме, отрицательное резюме и комментарий в свободном стиле, однако в работе данное разделение учтено не было, все остальные детали отзывов, а именно: вся информация о пользователе, рейтинг отзыва, количество согласных и несогласных людей, опыт использования пользователем продукта, были проигнорированы. Также была создана база данных с 415 идеальными аспектами – теми, что указываются, как описание характеристик того или иного товара. Все вышеуказанные данные и являются входными в созданной программе.

3.2. Описание использование ИСП РАН API

Для решения задач, связанных с анализом текстовых данных и их предобработкой, было принято решение об использовании API, предоставляемого ИСП РАН. Благодаря API решались следующие задачи: построение синтаксического дерева предложения, деление

текста на предложения и выявление частей речи. Для построения синтаксического дерева в метод, вызывающий API, передается предложение, а на выходе получается строка в формате JSON. Для деления текста отзыва на отдельные предложения вызывается соответствующий метод и в итоге получаются требуемые данные. Выявление частей речи в передаваемом тексте также осуществляется усилиями API. Подробная документация представлена на сайте: https://api.ispras.ru/texterra/v3.1/docs [2].

3.3. Описание выявления аспектов

Процессу выявления аспектов предшествовал пункт 3.1. В качестве входных данных имеется 415 идеальных аспектов и набор отзывов в количестве 24093. Так как отзывы, полученные с ulmart.ru, приходят зашумленными и содержат в себе большое количество случайных символов, то первым делом производится очистка входных данных. В процессе очистки из отзывов исчезают случайные пользовательские повторы, лишние знаки препинания, символы, распознавание которых программой невозможно. Все отзывы были приведены к единому стилю, под этим подразумевается нижний регистр и обязательное наличие в конце точки, что способствовала корректной работе API ИСП РАН [2] в дальнейших вычислениях. Далее производится выявление частей речи всех слов в каждом из отзывов. Данная обработка требуется для того, чтобы в дальнейшем осуществить поиск слова-существительного или набора слов, где главным словом также является существительное. Во втором случае также используется метод построения синтаксического дерева предложения, что позволяет найти нужные связи.

После создания набора всех аспектов производится выявление значимых аспектов в контексте данной работы. Для выполнения данной задачи были использованы возможности методов машинного обучения, а именно One-Class SVM. На вход модель получает набор из тренировочных данных, в результате полученный классификатор применяется к тестовым данным. Сначала все имеющиеся аспекты получают метку 1 или -1 в зависимости от того, присутствует ли данный аспект в списке идеальных аспектов. Далее данные делятся в соотношении 80 к 20 на тренировочные и тестовые. Набор из тренировочных данных очищается от тех аспектов, метка которых является -1. В итоге вся эта информация передается для обучения и предсказания классификатором меток для тестовых аспектов.

После данного шага было выявлено 421715 аспектов. Однако данное количество слишком велико для цели данной работы и содержит в себе большое количество мусора. 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, которыми являются самые частые слова и словосочетания. 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.

3.4. Описание используемых для вычисления семантического расстояния характеристик

3.4.1. PMI

PMI включает в себя 2 характеристики: для отзывов и для предложений.

Для вычисления обоих производится итерация по всем парам аспектов, далее производится поиск количества отзывов/предложений, где есть оба аспекта из пары. Значение PMI считается как логарифм от количества отзывов/предложений, где есть оба аспекта из пары, деленный на произведение их предвычисленного количества по-отдельности.

$$PMI(a_x, a_y) = \log\left(\frac{Count(a_x, a_y)}{Count(a_x)Count(a_y)}\right)$$

Разница в характеристиках заключается в корпусе отзывов/предложений, который передается для дальнейшей обработки в метод расчёта PMI.

Пример:

Пусть имеется пара аспектов: «компьютер» и «экран». А также 3 отдельных отзыва, полученных с сайта ulmart.ru:

1) «Обычный офисный компьютер, начального уровня. Очень шумный. Постоянно присутствует гул процессорного вентилятора, скорость которого система не желает регулировать.»

2) «Хорошая производительность, чисто рабочая машина. Куплено уже 4 компьютера 3,3 ггц с 8 гб памятью. 1С, офис, тимвьюер и другое рабочее ПО одновременно держит очень неплохо.»

3) «В компьютерах разбираюсь плохо, но нужен был компьютер для работы. Купила данный компьютер, и нисколько не жалею, компьютер мощный, быстрый. В контакт центре очень вежливый молодой человек помог все подключить, экран отличный, все работает) спасибо»

Результат PMI review будет равен log(1/3*1) = -1.0986122886681098.

Результат PMI sentence будет равен log(1/4*1) = -1.3862943611198906.

3.4.2. Contextual

Данная характеристика состоит из двух значений: локального и глобального.

Для получения локального контекста для каждой пары аспектов были предварительно вычислены значения 4 окружающих каждый из аспектов слов в отзывах: двух справа и двух слева. Из данных предварительных значений была сформирована база данных, которая позволила повысить скорость дальнейших вычислений.

Для вычисления глобального контекста для каждого из аспектов в паре был сформирован корпус отзывов, где он встречался. Далее, как для глобального, так и для локального контекста были загружены предвычисленные данные из баз данных в память, для получения результатов, поддающихся анализу, был произведен smoothing данных методом Add-1 Smoothing [1].

После было осуществлено вычисления значения, именуемого kl-divergence для локального и глобального контекстов, которое означает, насколько контексты аспектов в паре удалены друг от друга.

Пример:

Пусть имеется пара аспектов: «компьютер» и «экран». А также 3 отдельных отзыва, полученных с сайта ulmart.ru:

1) «Обычный офисный компьютер, начального уровня. Очень шумный. Постоянно присутствует гул процессорного вентилятора, скорость которого система не желает регулировать.»

2) «Хорошая производительность, чисто рабочая машина. Куплено уже 4 компьютера 3,3 ггц с 8 гб памятью. 1С, офис, тимвьюер и другое рабочее ПО одновременно держит очень неплохо.»

3) «В компьютерах разбираюсь плохо, но нужен был компьютер для работы. Купила данный компьютер, и нисколько не жалею, компьютер мощный, быстрый. В контакт центре очень вежливый молодой человек помог все подключить, экран отличный, все работает) спасибо»

Локальным контекстом для аспекта «компьютер» будет набор: Обычный офисный начального уровня уже 4 3,3 ггц _BEGIN_SENTENCE_ В разбираюсь плохо нужен был для работы Купила данный и нисколько не жалею мощный, быстрый.

Локальным контекстом для аспекта «экран» будет набор: все подключить отличный все.

Глобальным контекстом для аспекта «компьютер» будет набор отзывов 1-3.

Глобальным контекстом для аспекта «экран» будет отзыв 2.

Далее для каждого из контекстов, преобразованных в вектора с помощью *vectorizer.fit_transform*, был произведен smoothing данных. Результатом работы метода smoothing'а данных является n-gram, каждое значение в котором, если он не пуст, делится на количество слов в контекстах аспектов. Далее для каждой пары аспектов производится вычисления значения kl-divergence между n-gram'ами с помощью *stats.entropy*. Данное значение и является результатом характеристики context. Разница лишь в том, где ищутся контексты для аспектов.

3.4.3. Syntactic

Для вычисления данной величины были предварительно сохранены в базу данных синтаксические деревья, полученные с помощью API ИСП РАН [2], для всех предложений из отзывов. Далее для каждой из пар аспектов были найдены предложения, где встречаются они оба. Во всех таких предложениях с помощью вышеуказанной базы данных был вычислен кратчайший путь от одного аспекта к другому. Далее для каждой из пар был вычислен средний такой путь через деление суммы путей на их количество.

Пример:

Пусть имеется пара аспектов: «компьютер» и «экран». А также отзыв, полученный с сайта ulmart.ru: «В компьютерах разбираюсь плохо, но нужен был компьютер для работы. Купила данный компьютер, и нисколько не жалею, компьютер мощный, быстрый. В контакт центре очень вежливый молодой человек помог все подключить, экран отличный, все работает) спасибо». С помощью ИСП РАН API вычисляются синтаксические деревья. Для первого предложения из данного примера дерево находится в Приложении В данного отчета, его читаемый вариант можно увидеть на Рисунке 1.

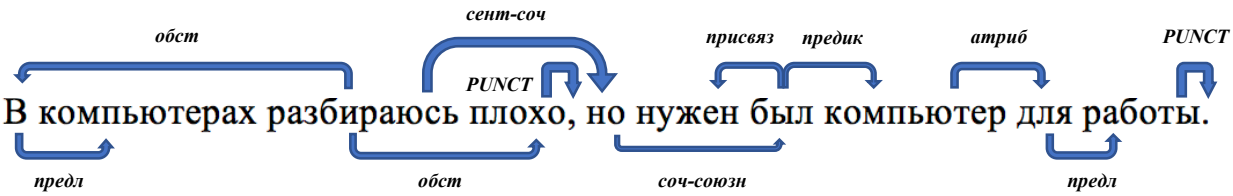

Рисунок 1. Наглядная демонстрация синтаксического разбора предложения (Стрелками представлены отношения родитель-ребенок)

Для данного дерева 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.

3.4.4. Lexical

Для вычисления данной величины были проитерированы все пары аспектов и вычислена разница между их длинами.

Пример:

Пусть имеется пара аспектов: «компьютер» и «экран».

Результат Lexical будет равен abs(len(компьютер)-len(экран)) = 4.

3.5. Описание вычисления семантического расстояния

Для вычисления семантической дистанции было выбрано два способа расчётов, один из которых был 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 $w \in \mathbb{R}^{1X6}$, представляющего собой набор характеристик (PMI, Lexical, Syntactic, Contextual) идеальных аспектов. Данный вектор вычисляется по формуле:

$$w* = (f^T f + \eta * \mathrm{I})^{-1}(f^T \mathrm{d}),$$

где f – вектор характеристик, η – константное значение, равное 0.4, I – единичная матрица, d – вектор, содержащий в себе все кратчайшие пути между идеальными аспектами в построенном вручную дереве. Далее для каждой пары аспектов семантическое расстояние равно:

$$distance = w[0] * pmi_{review} + w[1] * pmi_{sentence} + w[2] * lexical + w[3] * syntactic + w[4] * context_{local} + w[5] * context_{global},$$

где w - вектор, вычисленный на шаг раньше, а другие множители являются значениями характеристик для конкретной пары аспектов, которые получаются путем извлечения из баз данных.

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.

Проведенное изучение двух подходов к вычислению семантического расстояния показало наличие, как преимуществ, так и недостатков у обоих. Для способа 1 данные расстояний угадываются верно существенно реже, нежели чем в способе 2, однако способ 2 сильно зависит от тренировочной выборки и поэтому не имеет возможности выдавать результаты, выходящие за пределы предложенных ему для обучения, что является существенным минусом, коим не страдает 1 способ. Также за счет большего диапазона значений 1 способ является более удобным в плане построения дальнейшей иерархии аспектов, однако для полноты картины в пункте 3.6. были использованы и результаты 2 подхода для сравнения. В качестве преимущества второго способа можно указать скорость его работы, которая в разы выше нежели чем у 1 подхода.

Для 1 подхода было решено также изучить влияние каждой из характеристик на итоговый результат для семантического расстояния.

Характеристики	Минимальное семантическое расстояние	Максимальное семантическое расстояние
Все - PMI	3.22577963930419	174.024996616039
Все - Context	1.122215	206.717428100923
Все - Lexical	1.27966383810979	269.474891065696
Все - Syntactic	2.10356463930419	276.680485058915
Все + Bayes (Дополнительная характеристика)	10.2304768970291	436.156013125701
Все	3.22577963930419	277.802700058915

Как видно из приведенной таблицы, каждая из характеристик влияет на итоговый результат, что говорит о необходимости использования каждой из них.

3.6. Описание построения иерархии

Для построения иерархии аспектов было вручную построено дерево сначала для идеальных аспектов. Далее алгоритм для каждого из свободных листов дерева (родителей) выполняет поиск подходящих продолжений (детей) на основе того, насколько семантическая дистанция отличается от среднего значения дистанции для изначального дерева идеальных аспектов. Каждый новый ребенок запоминается в базу данных вместе со своим родителем и добавляется в список потенциально возможных родителей для новых аспектов.

Пример иерархии находится в Приложении Б данного отчета.

Заключение

Был изучен метод построения иерархии аспектов по пользовательским отзывам об электронных устройствах в англоязычном сегменте. Во время проведения исследования метод построения такой иерархии аспектов был модифицирован для русского языка. На основе произвольной иерархии можно создать иерархическую организацию обзоров потребителей, а также мнения потребителей по аспектам. С такой организацией пользователь может легко проанализировать набор потребительских мнений, а также искать отзывы потребителей по любому конкретному аспекту путем навигации по иерархии. Была создана программа, не имеющая аналогов в свободном доступе, являющаяся незаменимым инструментом для проведения экспериментов настоящей НИР. С помощью нее были достигнуты все поставленные цели исследования, а также повысился навык программирования у исполнителя.

Был проведен сравнительный анализ результатов эксперимента по вычислению семантического расстояния между аспектами в парах. Проведенное изучение двух подходов к вычислению семантического расстояния показало наличие, как преимуществ, так и недостатков у обоих. Для одного из подходов было изучено влияние каждой из характеристик на итоговый результат для семантического расстояния. Выявлено, что каждая из характеристик влияет на итоговый результат, что говорит о необходимости использования каждой из них.

При дальнейшем исследовании в данной области необходимо добавить большое число характеристик, влияющих на расчет семантического расстояния, расширить корпус входных данных, возможно используя отзывы пользователей об электронных устройствах с других ресурсов. Также можно использовать и другие части отзывов пользователей об электронных устройствах, которые не были использованы в рамках данной работы и о которых было сказано в пункте 3.1.

Список использованных источников

1. Сайт материалов университета Иллинойс [Электронный ресурс]. URL: https://courses.engr.illinois.edu/cs498jh/Slides/Lecture03.pdf (дата обращения: 20.12.2016);
2. Сайт API Института системного программирования Российской академии наук [Электронный ресурс]. URL: https://api.ispras.ru/ (дата обращения: 22.11.2016);
3. С. Ким, Д. Цанг, Ж. Чен, Э. Оу, Ш. Лиу, “A Hierarchical Aspect-Sentiment Model for Online Reviews”, Департамент компьютерных наук, Корея, 2010;
4. Jianxing Yu, Zheng-Jun Zha, Meng Wang, Kai Wang, Tat-Seng Chua, “Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews”, School of Computing, National University of Singapore, Institute for Infocomm Research, Singapore;
5. Jianxing Yu, Zheng-Jun Zha, Meng Wang, Tat-Seng Chua, “Hierarchical Organization of Unstructured Consumer Reviews”, School of Computing National University of Singapore;
6. В. Проноза, Е.В. Ягунова, “Аспектный анализ отзывов о ресторанах для рекомендательных систем е-туризма”, Санкт-Петербургский государственный университет, Санкт-Петербург, 2010;
7. А.А. Бреслав, А.П. Лукьянова, М.А. Коротков, “Построение иерархии классов по текстовым описаниям”, Санкт-Петербургский государственный политехнический университет, Санкт-Петербург, 2011;
8. K. Murthy, T.A. Faruquie, L.V. Subramaniam, K.H. Prasad, M. Mohania, “Automatically Generating Term-frequency-induced Taxonomies”, ACL, 2010.
9. S. Roy and L.V. Subramaniam, “Automatic Generation of Domain Models for Call Centers from Noisy Transcriptions”, ACL, 2009;
10. R. Snow and D. Jurafsky, “Semantic Taxonomy Induction from Heterogenous Evidence”, ACL, 2006;
11. H. Yang and J. Callan, “A Metric-based Framework for Automatic Taxonomy Induction”, ACL, 2009;
12. Сайт документации библиотеки BeautifulSoup [Электронный ресурс]. URL: https://www.crummy.com/software/BeautifulSoup/ (дата обращения: 05.03.2017);
13. Сайт документации языка Python [Электронный ресурс]. URL: https://docs.python.org/3/ (дата обращения: 20.01.2017);
14. D. Turdakov, N. Astrakhantsev, Y. Nedumov, A. Sysoev, I. Andrianov, V. Mayorov, D. Fedorenko, A. Korshunov, S. Kuznetsov, “Texterra: A Framework for Text Analysis”, Proceedings of the Institute for System Programming, vol. 26, issue 1, 2014, pp. 421-438.

ПРИЛОЖЕНИЕ А. Пример входных данных программы.

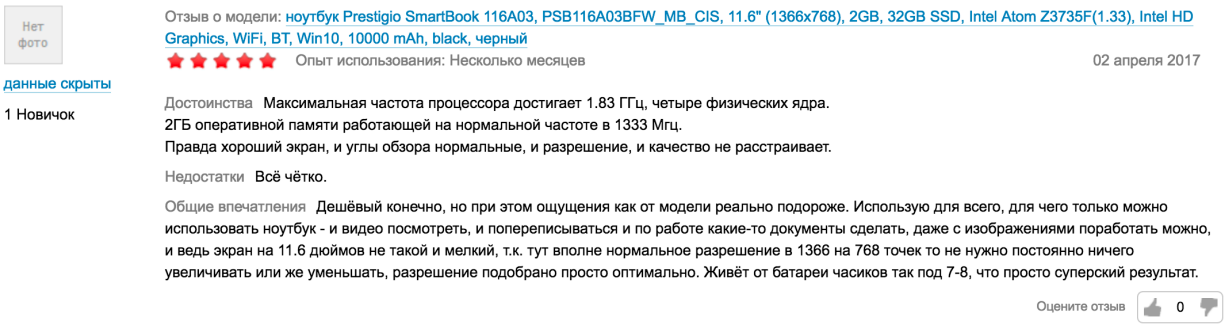

Рисунок 2. Пример отзыва пользователя из категории «Ноутбуки»

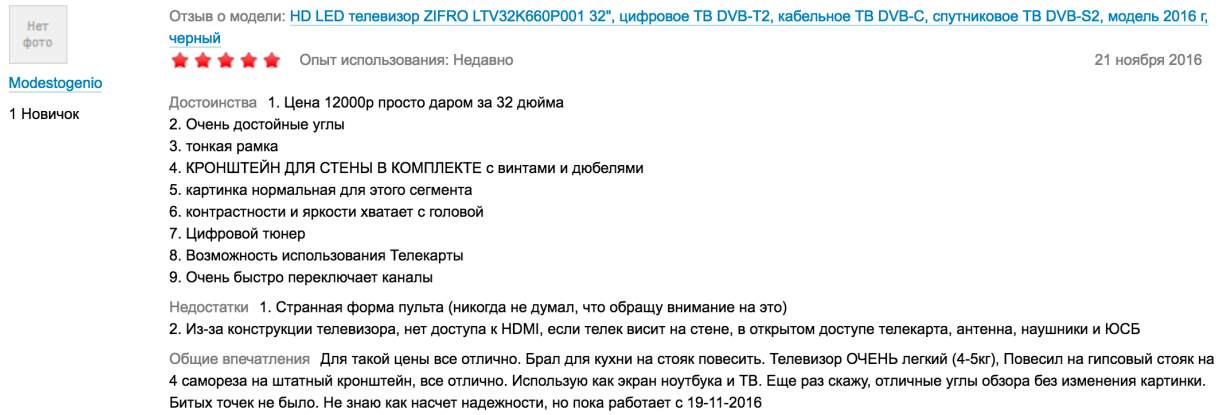

Рисунок 3. Пример отзыва пользователя из категории «Телевизоры»

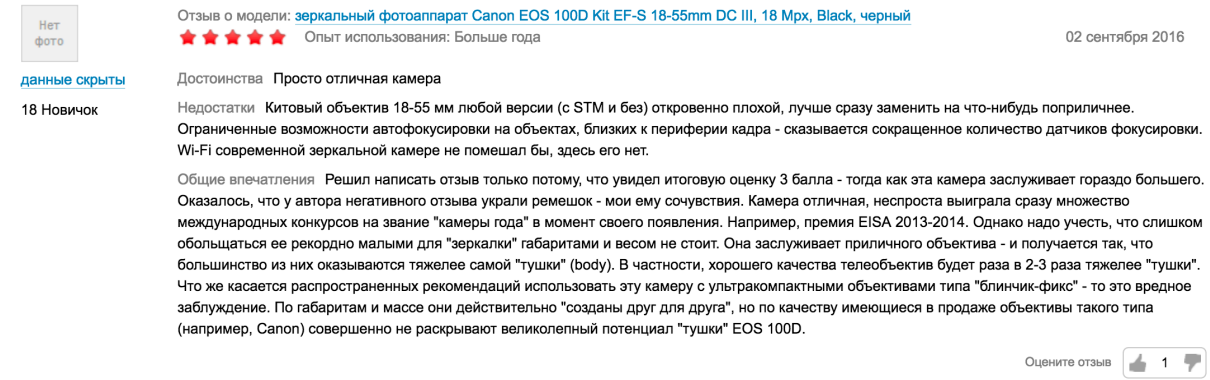

Рисунок 4. Пример отзыва пользователя из категории «Фотоаппараты»

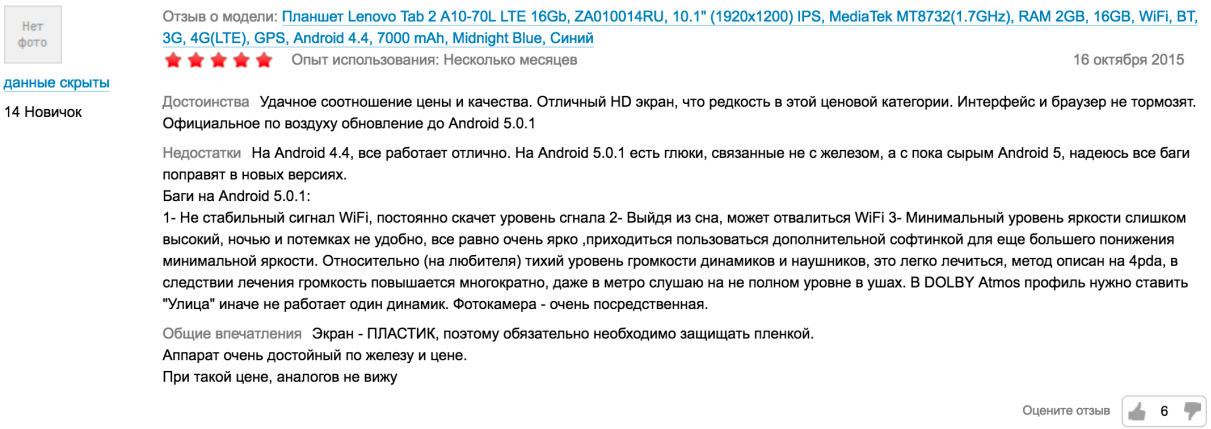

Рисунок 5. Пример отзыва пользователя из категории «Планшеты»

ПРИЛОЖЕНИЕ Б. Пример выходных данных программы.

Пример иерархии:

Зеркальные фотоаппараты

- Общая информация
    - Категория
    - Тип
    - Цвет
    - Производитель
    - Цена
- Характеристики
    - Тип матрицы
    - Кроп-фактор
    - Физический размер матрицы
    - Разрешение матрицы
    - Формат файлов
    - Разрешение
    - Съемка видео
    - Скоростная съемка
    - Таймер
    - Разъемы
    - Карты памяти
    - Элемент питания
    - Дисплей
        - Тип дисплея
        - Размер дисплея
    - Вспышка
        - Подключение внешней вспышки
    - Объектив
        - Объектив в комплекте
        - Оптический зум
        - Тип фокусировки
        - Расстояние фокусировки
    - Дополнительно
        - Гнездо для крепления штатива
        - Особенности
    - Габариты устройства
        - Размеры
        - Вес
    - Упаковка
        - Размер упаковки

ПРИЛОЖЕНИЕ В. Синтаксическое дерево.


```
{
  "text": "В компьютерах разбираюсь плохо, но нужен был компьютер для работы.",
  "annotations": {
    "syntax-relation": [
      {
        "start": 0,
        "end": 1,
        "value": {
          "parent": {
            "start": 14,
            "end": 24
          },
          "type": "обст"
        }
      },
      {
        "start": 2,
        "end": 13,
        "value": {
          "parent": {
            "start": 0,
            "end": 1
          },
          "type": "предл"
        }
      },
      {
        "start": 14,
        "end": 24,
        "value": {}
      },
      {
        "start": 25,
        "end": 30,
        "value": {
          "parent": {
            "start": 14,
            "end": 24
          },
          "type": "обст"
        }
      },
      {
        "start": 30,
        "end": 31,
        "value": {
          "parent": {
            "start": 25,
            "end": 30
          },
          "type": "PUNCT"
        }
      },
      {
        "start": 32,
        "end": 34,
        "value": {
          "parent": {
            "start": 14,
            "end": 24
          },
          "type": "сент-соч"
        }
      },
      {
        "start": 35,
        "end": 40,
        "value": {
          "parent": {
            "start": 41,
            "end": 44
          },
          "type": "присвяз"
        }
      },
      {
        "start": 41,
        "end": 44,
        "value": {
          "parent": {
            "start": 32,
            "end": 34
          },
          "type": "соч-союзн"
        }
      },
```

```
{
 "start": 45,
 "end": 54,
 "value": {
  "parent": {
   "start": 41,
   "end": 44
  },
  "type": "предик"
 }
},
{
 "start": 55,
 "end": 58,
 "value": {
  "parent": {
   "start": 45,
   "end": 54
  },
  "type": "атриб"
 }
},
{
 "start": 59,
 "end": 65,
 "value": {
  "parent": {
   "start": 55,
   "end": 58
  },
  "type": "предл"
 }
},
{
 "start": 65,
 "end": 66,
 "value": {
  "parent": {
   "start": 59,
   "end": 65
  },
  "type": "PUNCT"
 }
}
]
}
}
```

ПРИЛОЖЕНИЕ Г. Код программы.

Программа состоит из 19 основных классов и 35 баз данных на языке Python. Текст программы на исходном языке находится в директории документация на носителе информации типа компакт-диск в связи с большим количеством строк кода. Также с ним можно ознакомиться в репозитории проекта по адресу: https://github.com/arepina/courseWork2016.

Описание и функциональное назначение классов и интерфейсов

Класс или интерфейс	Назначение
Aspects	Отвечает за поиск аспектов в отзывах пользователей и морфологический разбор предложений
Context	Отвечает за расчет характеристик Context Global и Context Local для пар аспектов
DB	Отвечает за создание и взаимодействие со всеми базами данных
FrequentAspects	Отвечает за поиск 1000 наиболее популярных аспектов
HierarchyBuilder	Отвечает за построение иерархии аспектов по пользовательским отзывам о электронных устройствах
IdealAspectsDB	Отвечает за заполнение базы данных с идеальными аспектами
Lexical	Отвечает за расчет характеристики Lexical для пар аспектов
Main	Является точкой управления программой. Отвечает за запуск работы всех классов
OneClassSVM	Отвечает за One Class SVM тренировку и обучение
PMI	Отвечает за расчет характеристик PMI Review и PMI Sentence для пар аспектов
SemanticDistanceLearning	Отвечает за вычисление семантического расстояния
RandomForest	Отвечает за вычисление семантического расстояния методами машинного обучения
Sentence	Отвечает за разделение пользовательских отзывов на предложения
Splitter	Отвечает за очистку отзывов и предложений от мусора
Syntactic	Отвечает за расчет характеристики Syntactic для пар аспектов
Unnecessary	Отвечает за удаление ненужных частей предложений
CategoryNames	Отвечает за создание списка категорий сайта ulmart.ru
DataBase_Ulmart	Отвечает за формирование базы отзывов
Parser	Класс-парсер сайта ulmart.ru