|  |  |
| --- | --- |
| Подпись и дата |  |
| Инв. № дубл. |  |
| Взам. инв. № |  |
| Подпись и дата |  |
| Инв. № подл. | RU. 17701729. 503200-0134 1-1ЛУ |

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

|  |  |  |
| --- | --- | --- |
| 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» |   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.В.Шилов  «\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2017 г. |

**Построение иерархии аспектов по пользовательским отзывам об электронных устройствах**

**Руководство оператора**

**ЛИСТ УТВЕРЖДЕНИЯ**

RU. 17701729. 503200-0134 1-1ЛУ

Исполнитель: студентка группы БПИ143

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/А.А.Репина/

«\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2017 г.

2017

|  |  |
| --- | --- |
| Подпись и дата |  |
| Инв. № дубл. |  |
| Взам. инв. № |  |
| Подпись и дата |  |
| Инв. № подл. | RU. 17701729. 503200-01 34 |

**УТВЕРЖДЕНО**

**RU. 17701729. 503200-01 34 1-1-ЛУ**

|  |  |
| --- | --- |
|  |  |

**Построение иерархии аспектов по пользовательским отзывам об электронных устройствах**

**Руководство оператора**

RU. 17701729. 503200-01 34 1-1

Листов 10

2017

СОДЕРЖАНИЕ

[1. НАЗНАЧЕНИЕ ПРОГРАММЫ 4](#_Toc477632507)

[1.1 Назначение программы 4](#_Toc477632508)

[1.2 Информация о функциях и принципе эксплуатации программы 4](#_Toc477632509)

[2. УСЛОВИЯ ВЫПОЛНЕНИЯ ПРОГРАММЫ 5](#_Toc477632510)

[2.1 Требования к составу и параметрам технических средств 5](#_Toc477632511)

[2.2. Условия необходимые для выполнения программы 5](#_Toc477632512)

[3. ОБРАЩЕНИЕ К ПРОГРАММЕ 6](#_Toc477632513)

[3.1. Характеристики для вычисления семантического расстояния 6](#_Toc477632514)

[3.1.1. PMI 6](#_Toc477632515)

[3.1.2. Context 6](#_Toc477632516)

[3.1.3. Lexical 6](#_Toc477632517)

[3.1.4. Syntactic 6](#_Toc477632518)

[3.2. Вычисление семантического расстояния 6](#_Toc477632519)

[3.3. Построение иерархии 6](#_Toc477632520)

[3.4. Парсинг сайта 7](#_Toc477632521)

[4. ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ 8](#_Toc477632522)

[ТЕРМИНОЛОГИЯ 9](#_Toc477632524)

[ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ 11](#_Toc477632525)

# НАЗНАЧЕНИЕ ПРОГРАММЫ

1.1 Назначение программы

Данная работа представляет собой подход организации различных аспектов продукта, относящегося к категории электронные устройства, в иерархию на основе знаний о потребительских отзывах. Основываясь на производной 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.

1.2 Информация о функциях и принципе эксплуатации программы

Программа должна предоставлять возможность построения иерархии аспектов по пользовательским отзывам об электронных устройствах. Для этого она должна предоставлять возможность выполнения перечисленных ниже функций:

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.

# УСЛОВИЯ ВЫПОЛНЕНИЯ ПРОГРАММЫ

2.1 Требования к составу и параметрам технических средств

Для надежной и бесперебойной работы методов рекомендуется следующий состав аппаратурных средств:

1. персональный компьютер, оснащенный процессором с тактовой частотой 400 MГц и выше или аналогичный;
2. 128 ГБ оперативной памяти или больше (в связи с большими объемами анализируемых данных);

2.2. Условия необходимые для выполнения программы

Для работы программы необходим следующий состав программных средств:

1. Установленный Python 2.7 или Python 3.2+.

# ОБРАЩЕНИЕ К ПРОГРАММЕ

3.1. Характеристики для вычисления семантического расстояния

3.1.1. PMI

Характеристика PMI включает в себя PMI review и PMI sentence. В метод расчета PMI review calculate\_pmi(corpus\_review, corspus\_aspects, db) передается корпус отзывов, корпус аспектов и объект класса DB для возможности сохранения вычисленных результатов в базу данных. В метод расчета PMI sentence calculate\_pmi(corpus\_ sentence, corspus\_aspects, db) передается корпус предложений из отзывов, корпус аспектов и объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.1.2. Context

Характеристика Context включает в себя Context local и Context global. В метод расчета Context local local\_context(corpus\_context\_local, corspus\_aspects, db) передается корпус наборов 4-словных контекстов, корпус аспектов и объект класса DB для возможности сохранения вычисленных результатов в базу данных. В метод расчета Context global global\_context(corpus\_context\_global, corspus\_aspects, db) передается передается корпус отзывов, корпус аспектов и объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.1.3. Lexical

В метод расчета Lexical process(corspus\_aspects, db) передается корпус аспектов и объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.1.4. Syntactic

В метод расчета Syntactic process(corpus\_aspects, corpus\_syntactic, corpus\_sentence, db) передается корпус аспектов, набор предвычисленных синтаксических деревьев предложений, корпус предложений, а также объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.2. Вычисление семантического расстояния

В метод расчета сематического расстояния process\_semantic\_distance\_learning(corpus\_aspects, corpus\_features, db) передается корпус характеристик для каждой пары аспектов и корпус аспектов, а также объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.3. Построение иерархии

Для построения иерархии аспектов по пользовательским отзывам об электронным устройствам в метод расчета build\_hierarchy(corspus\_ideal\_aspects, avg\_distance, corpus\_aspects, corspus\_semantic\_distance, db) передается корпус идеальных аспектов, среднее значение семантичсекой дистанции для них, корпус аспектов и корпус семантичсеких расстояний для них, а также объект класса DB для возможности сохранения вычисленных результатов в базу данных.

3.4. Парсинг сайта

В метод получения отзывов об электронных устройствах process(url, db) передается адрес веб страницы, с которой будет начат парсинг, а также объект класса DB для возможности сохранения вычисленных результатов в базу данных.

# ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ

Тип входных данных в методе calculate\_pmi – 2 объекта list и объект класса DB как для PMI review, так и для PMI sentence

Тип входных данных в методах local\_context и global\_context - 2 объекта list и объект класса DB

Тип входных данных в методе process класса Lexical- объект list и объект класса DB

Тип входных данных в методе process класса Syntactic - 3 объекта list и объект класса DB

Тип входных данных в методе process\_semantic\_distance\_learning - 2 объекта list и объект класса DB

Тип входных данных в методе build\_hierarchy - 3 объекта list, float и объект класса DB

Тип входных данных в методе process класса Parser - строка и объект класса DB

Все вышеуказанные методы записывают результаты своей работы с соответствующие базы данных. Если тип входных данных не соответствует с указанными в этом пункте, то при вызове соответствующих методов возникнет исключительная ситуация.

# ПРИЛОЖЕНИЕ 1

# ТЕРМИНОЛОГИЯ

Ниже приведен список необходимых терминов для ознакомления.

**Парсер -** скрипт или программа, которые используются для сбора информации с сайтов

**PMI(pointwise mutual information )** - точечные взаимные сведения или 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.

**Context** - KL-расхождение между языковыми моделями.

**Lexical** - разница в длине слов между двумя аспектами.

**Syntactic** - средняя длина кратчайшего синтаксического пути между парами аспектов в дереве.

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.

**KL-divergence** -  это неотрицательнозначный функционал, являющийся несимметричной мерой удаленности друг от друга двух вероятностных распределений.

**ПРИЛОЖЕНИЕ 2**

**ИСТОЧНИКИ, ИСПОЛЬЗОВАННЫЕ ПРИ РАЗРАБОТКЕ**

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. Дж. Ю, Ж. Джа, М. Венг, К. Венг, Т. Чуа, “Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews”, 2011;
5. В. Проноза, Е.В. Ягунова, “Аспектный анализ отзывов о ресторанах для рекомендательных систем е-туризма”, Санкт-Петербургский государственный университет, Санкт-Петербург, 2010;
6. А.А. Бреслав, А.П. Лукьянова, М.А. Коротков, “Построение иерархии классов по текстовым описаниям“, Санкт-Петербургский государственный политехнический университет, Санкт-Петербург, 2011;
7. Сайт документации языка Python [Электронный ресурс]. URL: https://docs.python.org/3/ (дата обращения: 20.01.2017).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ | | | | | | | | | | |
| Изм. |  | Номера листов (страниц) | | | | Всего  листов  (страниц) в  документе | №  документа | Входящий №  сопроводительн  ого документа и  дата | Подпись | Дата |
| измененн  ых | | замененны  х | новых | аннулиро  ванных |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |