МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК

КАФЕДРА ИНФОРМАЦИОННЫХ СИСТЕМ

ОТЧЁТ

**по учебной практике**

Место прохождения практики

Кафедра Информационных Систем ТюмГУ

|  |
| --- |
| **Выполнил:**  Студент группы 25ИСИТ169  Зухритдинов А.Н.  (Подпись)  **Научный руководитель:**  к.ф.н., доцент кафедры ИС  Бидуля Ю.В.  (Подпись) |

Тюмень – 2018

Оглавление

[Введение 3](#_Toc2197795)

Введение

Анализ данных – область математики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Информационная система (ИС) – это система, включающая совокупность компонентов, отвечающих за выполнение ряда функций: извлечение, обработка, хранение, а также организационных ресурсов: человеческие, финансовые, технические, программные и т.д.

Курсовая работа включает в себя разработку приложения для обработки отзывов покупателей интернет-магазинов, пользователей интернет ресурсов, что предоставляют услуги и сервисы, переработки текста, и ведение анализа данных на имеющихся ресурсах с помощью применения готовых модулей для парсинга, сбора данных из ресурсов на языке Python и использование алгоритмов обработки текста и принятия решений, такие как «токенизация» и «стемминг». Данная задача считается релевантной из задач обработки данных и принятия решений на основании данных. Стемминг – это процесс нахождения основы слова. Это, в свою очередь, помогает избежать принятие неадекватных, нерелевантных решений. Токенизация – выделение слов из приложений, это считается одним из фундаментальных апсектов при обработке текста. выделения сущностей тут не обойтись, т.е. здесь имеет смысл обозначения слов по их категориям, частям речи, чтобы легче было структурировать данные. Большая часть интернета плохо структурирована, так же, как сам естественный язык пользователей. Разбором алгоритмов и методов обработки естественного языка занимается особое направление искусственного интеллекта: NLP (Natural Language Processing). Пайплайн – это разработка программных продуктов; в программировании – разбор алгоритма по этапам. Пайплайн «NLP»:

* Разбор предложений по специальным знакам. В основном, это точки, восклицательные знаки и т.п. Это помогает эффективнее разбирать слова по частям речи и составлять дерево сущностей;
* Токенизация включает в себя выделение слов по атрибутам, т.е. это своего выделение слов и сохранение их в словари;
* Определение частей речи. После данного этапа у каждого токена(слова) появляется значение, наименование его части речи;
* Лемматизация – приведение к начальной форме слов: у глаголов к инфинитиву, у существительных к именительному падежу ед. числа и т.п;
* Выделение и удаление «стоп-слов». Данный шаг включает в себя удаление разного рода частей речи (союзы, предлоги и т.п.), чтобы и смысл не менялся, и для увеличения эффектности алгоритма, т.е. для отсечения «воды»;
* Парсинг зависимостей. Создание дерева зависимостей между токенами, с обозначением их связей. Целью алгоритма является дерево токенов, у которых есть единственный родитель и связи между ними;
* Поиск групп существительных. Каждое сочетание токенов принимается за существительное и группируются, если это возможно.
* Распознавание именованных сущностей. Каждый токен, в соответствии с NER-моделями (Named Entity Relations), ставится в соответствие по его признакам. Данный этап фильтрует слова, дав возможность отличить

В первую очередь стоило бы отметить то, что для данной работы должна быть представлена выборка из немалого количества отзывов, товаров и услуг.)