Федеральное государственное бюджетное образовательное учреждение высшего образования "МИРЭА - Российский технологический университет"

Детский технопарк "Альтаир"

**Разработка программного обеспечения**

**для визуализации тезауруса**

Шевцов Владислав Сергеевич,

Хрулёв Анатолий Сергеевич

10 «А» класс ГБОУ г. Москвы №1542

Руководитель: Русаков Алексей Михайлович

Преподаватель детского технопарка «Альтаир»

**Москва, 2022**

# Оглавление

[Оглавление 2](#_Toc94643710)

[Введение 3](#_Toc94643711)

[Цель 3](#_Toc94643712)

[Задачи 3](#_Toc94643713)

[Методы и этапы 4](#_Toc94643714)

[Актуальность 4](#_Toc94643715)

[Использованные программного решения 4](#_Toc94643716)

[Обзор проектного решения 5](#_Toc94643717)

[Результаты тестирования и реальное применение 7](#_Toc94643718)

[Итоги и перспективы 8](#_Toc94643719)

[Список литературы](#_Toc94643720) 11

## Введение

В области искусственного интеллекта и компьютерной лингвистики существует огромное количество задач, где активно для их решения используются электронные справочники и тезаурусы. Эти задачи связаны с поиском и обработкой информации на естественном языке. Компьютеры прекрасно справляются с точными математическими вычислениями, а люди способны быстро выявлять закономерности. Объединить эти два важных аспекта поможет механизм визуализации, точно и аккуратно отображающий данные на экране компьютера в некотором визуальном представлении, которое позволит человеку быстро понять их суть, найти информацию, и решить поставленную задачу.

## Цель

Разработать программное средство для визуализации тезауруса.

## Задачи

- Обзор и анализ современных математических методов

обработки текстовой информации.

- Обзор и анализ имеющегося программного обеспечения

для обработки текстовой информации;

- Выбор и обоснование инструментальных средств для

решения поставленной задачи;

- Разработка и реализация алгоритма визуализации тезауруса;

- Разработки структуры программного обеспечения; подготовка документации.

## Методы и этапы

|  |  |
| --- | --- |
| Изучение материалов по данной теме | Определение характеристики предметной области;  Обзор и анализ программных решений, пригодных для анализа и визуализации текстовой информации;  Обзор и анализ современных моделей систем визуализации тезауруса и математических методов анализа текста;  Анализ моделей извлечения ключевых слов;  Анализ моделей визуализации тезауруса. |
| Программирование | Разработка структуры программного комплекса;  Обоснование выбора инструментальных средств для разработки ПО;  Обоснование выбора инструментальных средств для разработки ПО;  Описание основных алгоритмов работы ПО;  Разработка и отладка программного обеспечения;  Тестирование и отладка программного обеспечения. |
| Внедрение | Тестирование разработанной программы и попытка внедрения в маркетинговые сервисы. |

## Актуальность

Обработка естественного языка является бурно развивающейся областью исследований с широким спектром практических реализаций. В связи с тем, что в настоящее время наблюдается большой рост требований в естественно-языковых интерфейсах, способных на автоматическую обработку текстов и документов, а также содержательного и исчерпывающего поиска и их классификации, стоит большая необходимость в систематизированных умениях и знаниях в компьютерной лингвистике, задачей которой становится обслуживание потребностей пользователя в этой области.

## Использованные программные решения

Программное обеспечение (Среда разработки – Visual Code, Языки программирования – Python и JavaScript, JavaScript-библиотеки jQuery и Django), компьютер под управлением OS Windows. Для получения наилучшего результата было произведено ознакомление с технической частью проектов (к примеру, морфологический анализатор pymorphy).

## Обзор проектного решения

Программное средство, которое реализуется в этом проекте, будет решать проблему визуализации текстовой информации (тезауруса). Исходя из полученный данных, программа будет выводить информацию в виде графа, показывающие смысловые связи между словами.

Программные комплексы для визуализации информации позволяют представить данные в виде графиков, блок-схем, таблиц, графов. Такой подход позволяет взглянуть на общую картину полученных данных, структурировать их, в итоге сэкономив ресурсы и время на аналитику.

Стоит заметить, что данные могут быть абсолютно разными, иметь разный формат. Тут встает вопрос о методах визуализации, а также о корректном представлении данных, которые нужно визуализировать. Подготовка и представление данных, готовых для качественной визуализации остается открытым вопросом. Требуются программные системы, которые могут, как и анализировать данные, выделяя ключевые слова и формируя их в словарь (тезаурус), так и графической визуализации данных.

**Библиотека rutermextract**

Эта библиотека была вдохновлена topia.termextract, которая работала только с англоязычными текстами. Эта же библиотека используется для извлечения ключевых слов из текстов, которые написаны на русском языке. Для анализа морфологии используется pymorphy2. Реализована данная библиотека для языка Python.

Библиотека мажет быть использована для автоматического присвоения тегов, а также для вычисления похожести текстов на основе извлеченных слов. Ключевые слова, которые были извлечены, приводятся в нормальную форму и упорядочиваются от более важных в контексте анализируемого текста, к менее значащим.

Библиотека извлекает ключевые слова на основе заранее заданных правил. К сожалению, на данный момент это единственный возможный вариант, поскольку для русского языка не существует открытого синтаксического корпуса, который можно использовать для обучения синтаксических моделей.

**Библиотека spaCy**

spaCy - это библиотека для продвинутой обработки естественного языка в Python и Cython. Он основан на самых последних исследованиях и был разработан с самого первого дня для использования в реальных продуктах. spaCy поставляется с предварительно обученными статистическими моделями и векторами слов, и в настоящее время поддерживает токенизацию для 50+ языков. В нем реализована современная скорость, сверточные модели нейронных сетей для тегирования, анализа и распознавания именованных объектов, а также простая интеграция с глубоким обучением. Это коммерческое программное обеспечение с открытым исходным кодом, выпущенное под лицензией MIT.

**Библиотека FlashText**

Библиотека реализована на чистом Python. FlashText принимает список ключевых слов, который использует для создания словаря на нагруженном дереве, также известного как префиксное дерево. От программиста требуется передать интересующую строку и уточнить, нужно выполнить замену или поиск. Если выбрана замена, библиотека создаст новую строку, в которой ключевые слова будут заменены. В случае поиска FlashText возвратит список ключевых слов, найденных в строке. Все это произойдет всего за один проход по входной строке.

**Коллекция библиотек Igraph**

Представляет из себя пакет для анализа графов общего назначения. Реализована на языке программирования C, так же существует для Python и R, реализованная в виде пакетов. Данный продукт широко используется в научных исследованиях и смежных областях. Igraph был разработан Габором Чарди и Тамасом Непушем. Имеет одну из самых эффектных графовых визуализаций, которая была сделана одним из разработчиков данной коллекции библиотек. Существенным недостатком является очень плохо структурированная документация к Python API.

**Библиотека Matplotlib**

Представляет из себя библиотеку для визуализации и построения графиков, как и в 2D, так и в 3D. Реализована на Python. Matplotlib может использоваться в скриптах и оболочках Python, IPython, в серверах веб-приложений, а так же в виде набора инструментов графического интерфейса пользователя.

Для простой прорисовки pyplot модуль предоставляет интерфейс, похожий на MATLAB, особенно в сочетании с IPython. Библиотека позволяет иметь полный контроль над стилями линий, свойствами шрифта, свойствами осей и прочих. Управление производится через объектно-ориентированный интерфейс или через набор функций, знакомых пользователям MATLAB.

По заверению разработчиков, Matplotlib старается сделать легкие вещи легкими, а сложные - возможными. Есть возможность создавать графики, гистограммы, спектры мощности, гистограммы, диаграммы ошибок, диаграммы рассеяния. С помощью всего лишь нескольких строк кода.

Лицензия Matplotlib основана на лицензии Python Software Foundation (PSF). Есть открытый репозиторий на GitHub, где так же ведется трекинг проблем и pull requests.

**Библиотека NetworkX**

Является пакетом Python для создания, управления и изучения структуры, динамики и функций сложных сетей. Имеет классы для графов и орграфов, может преобразовывать графики из одного формата в другой, возможность строить случайные графы или строить их постепенно, представлять визуализации так и в 2D, так и в 3D, и другие функции.

NetworkX подходит для визуализации больших графиков, где может находиться до 10 миллионов узлов и 100 миллионов ребер. Библиотека представляет собой достаточно эффективную, масштабируемую, легко переносимую среду для анализа данных. Распространяется по лицензии BSD. Имеет постоянно обновляемый репозиторий на GitHub.

**Модуль Graph-tool**

Представляет из себя модуль для Python для статического анализа графов и орграфов. Основные структуры данных и алгоритмы реализованы на C++, что может сопоставимый по скорости и потреблению памяти с библиотекой на чистом C/C++.

Имеется расширенный набор функций, таких как настройка произвольных вершин и ребер и их свойств, фильтрация данных «на лету», введение статистики. Инструмент имеет собственные алгоритмы компоновки и интерактивные процедуры рисования широкого применения. Ко всему этому имеется подробная документация, но только на английском языке.

**Библиотека D3.js**

Библиотека на Java Script для динамической интерактивной визуализации данных в браузерах. Данную библиотеку применяют для создания интерактивной графики для интернет-сайтов, информационных панелей и составления карт, а также для использования графики в печатных публикациях. Визуализацию легко настроить и использовать.

Распространяется данная библиотека по лицензии BSD. Присутствует репозиторий на GitHub. Есть также русскоязычная документация, переводимая сообществом.

Рассмотренные программные решения показали, что существует возможность не только определения тематики текста и извлечения ключевых слов, но и визуализировать данные в удобном для нас виде.

## Результаты тестирования и реальное применение

При открытии приложения, которое представляет из себя веб-сайт, пользователя будет встречать поле для ввода. В него следует вставить исследуемый текст. После вставки и нажатия кнопки анализировать, исследуемый текст отправляется на сервер, где выполняется его анализ. Затем возвращаются готовые данные, которые визуализируются на клиенте в браузере в виде графа

При вставке текста в текстовое и нажатии кнопки «Анализировать», текст отправляется к анализатору, производя анализ и извлечение ключевых слов в массив, с добавлением параметров из словаря и данных, полученных в результате анализа. После этого всё это формируется в файл формата csv и отправляется обратно к пользователю. В конечном итоге производится построение результатов по полученным данным и заданным параметрам.

После проведения анализа, появляется кнопка «Перестроить», которая позволяет перестроить полученный граф (или графы) в новый вид, не отправляя текст на новый анализ и не ожидая результатов, что экономит время пользователя и производит меньшую нагрузку на ЭВМ.

**Выводы**

Были рассмотрены структуры, интерфейс и основные алгоритмы ПО. Также было проведено тестирование разработанной программы, которое показало работоспособность данного программного обеспечения и соответствие поставленным задачам в рамках индивидуального проекта.

## Итоги и перспективы

Было разработано программное обеспечение для визуализации тезауруса. Также было проведено тестирование разработанной программы, которое показало работоспособность данного программного обеспечения и соответствие поставленным задачам в рамках индивидуального проекта продвинутого уровня.

Данный проект, в силу расширяемости своего функционала, пригодится в любом маркетинговом сервисе. С помощью этого приложения пользователь может эффективно организовывать поиск и извлечение ключевых слов из текста, формировать из них тезаурус и выполнять виртуализацию в виде графа.

## Список литературы

[Учебник] Шихи Д. Серия: Структуры данных в Python. Начальный курс. Изд-во ДМК-Пресс. 2021.

[Электронный ресурс] Язык программирования Python 3 для начинающих и чайников] URL: https://pythonworld.ru/

01.02.2022. URL: <https://quasar.dev/>

[Учебник] Заяц А. М., Васильев Н. П. Проектирование и разработка WEB-приложений. Введение в frontend и backend разработку на JavaScript и node. js. 2019.

[Учебник] Браун И. Веб-разработка с применением Node и Express. Полноценное использование стека JavaScript. 2-е издание. – Издательский дом Питер, 2021.