**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**Факультет прикладной математики – процессов управления**

**отчет**

**по лабораторной работе**

**по дисциплине «Функциональное программирование»**

**на тему «Параллельная обработка»**

|  |  |  |
| --- | --- | --- |
| Студент гр. 22.Б15 |  | Агишев А.Б. |
| Преподаватель |  | Киямов Ж.У. |

**Санкт-Петербург**

**2023 г.**

**Содержание**

[1. Цель работы 2](#_Toc150296671)

[2. Задача 2](#_Toc150296672)

[3. Теоретическая часть 2](#_Toc150296673)

[4. Алгоритм метода 3](#_Toc150296674)

[5. Описание программы 4](#_Toc150296675)

[6. Рекомендации пользователю 6](#_Toc150296676)

[7. Рекомендации программисту 6](#_Toc150296677)

[8. Контрольный пример 6](#_Toc150296678)

[9. Заключение 7](#_Toc150296679)

# **Цель работы**

Разработать программу, которая сначала соберет текстовые данные из нескольких источников социальных сетей, затем проведет их предварительную обработку и, наконец, выполнит анализ, чтобы выявить популярные темы и тренды.

# **Задача**

1. Приобрести и закрепить знания в области сбора данных из социальных сетей, анализа текста и параллельных вычислений.
2. Собрать базу Telegram каналов и VK групп.
3. Разработать программное обеспечение для анализа данных.
4. Оптимизировать вычисления и алгоритмы для ускорения обработки данных.
5. Создать интерфейс для программы.

# **Теоретическая часть**

*Параллельные вычисления*:

Параллельные вычисления — это метод организации вычислительного процесса, при котором задачи выполняются одновременно с использованием множества вычислительных ресурсов. Это позволяет ускорить обработку данных и повысить эффективность вычислительных задач.

*Парсинг данных*:

Парсинг данных — это процесс извлечения и структурирования информации из исходного набора данных, чаще всего в текстовом формате. Цель парсинга заключается в обработке и анализе данных для последующего использования в приложениях, базах данных или других системах.

*Анализ слов в виде облака слов*:

Анализ слов в виде облака слов представляет собой визуализацию частоты использования слов в тексте. Ключевые слова извлекаются и представляются графически, где размер и расположение слова зависят от его частоты в тексте. Этот метод обеспечивает быстрый обзор ключевых тем и концепций в текстовых данных.

*Выделение топиков через gensim*:

Выделение топиков с использованием библиотеки gensim включает в себя создание корпуса текстов и применение модели LDA (Latent Dirichlet Allocation) для выявления тем в текстовых данных. Gensim предоставляет инструменты для построения корпусов, обучения моделей и извлечения семантических тематик из текстов.

# **Алгоритм метода**

1. Запуск интерфейса в отдельном потоке.
2. Параллельный парсинг текстовой информации из различных Telegram каналов, а также VK групп.
3. Параллельная обработка текста от лишней информации из разных источников.
4. Анализ текстовой информации с целью выявления трендов и топиков.
5. Анализ текстовой информации для построения облака слов.

# **Описание программы**

В программе используется 2 класса: 1 связанный с интерфейсом программы, 1 связанный с обработкой текста. В таблице 5.1 представлено описание классов.

*Таблица 5.1. Описание классов*

|  |  |  |
| --- | --- | --- |
| Имя класса | Наследование | Описание |
| DataAnalysis | — | Парсинг данных из различных источников, обработка текста, анализ текста |
| Interface | — | Создание интерфейса, распараллеливание обработки текстовой информации |

Описание переменных класса *«DataAnalysis»* представлено в таблице 5.2.

*Таблица 5.2. Описание переменных класса «DataAnalysis»*

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| List[] | vk\_data | Текстовые данные, полученные из VK |
| List[] | telegram\_data | Текстовые данные, полученные из Telegram |
| List[] | processed\_texts | Список обработанных строк |
| Mystem | mystem | Объект класса Mystem, используемый для лемматизации текстов |
| set | russian\_stopwords | Множество стоп-слов русского языка |

Описание функций класса *«DataAnalysis»* представлено в таблице 5.3.

*Таблица 5.3. Описание функций класса «DataAnalysis»*

|  |  |  |
| --- | --- | --- |
| Входные данные | Имя | Описание |
| — | \_\_init\_\_ | Инициализация экземпляра класса анализа данных |
| — | get\_telegram\_data | Извлечение данных из Telegram каналов |
| — | get\_vk\_data | Извлечение данных из VK групп |
| Название социальной сети | processing | Выполнение обработки текста на основе социальной сети |
| — | parallel\_parsing | Распараллеливание парсинга данных из различных источников |
| — | parallel\_processing | Распараллеливание обработки данных из различных источников |
| — | analyze\_topics | Выполнение тематического анализа с помощью LDA |
| — | plot\_wordcloud | Генерация и вывод облака слов |

Описание функций класса *«Interface»* представлено в таблице 5.4.

*Таблица 5.4. Описание функций класса «Interface»*

|  |  |  |
| --- | --- | --- |
| Входные данные | Имя | Описание |
| Объект интерфейса | \_\_init\_\_ | Инициализация переменных, настройка параметров интерфейса, запуск корректировки интерфейса |
| — | create\_interface | Корректировка интерфейса |
| — | start\_threading | Запуск интерфейса в отдельном потоке |
| — | start\_processing | Запуск обработки текстовых данных |

# **Рекомендации пользователю**

Кнопка *«start»* отвечает за запуск обработки текстовых данных. По окончании работы программы будет выведено изображение с облаком слов, а также файл *«topics.txt»*, описывающий наиболее популярные топики. Для выхода из программы нажмите крестик в правом верхнем углу программы.

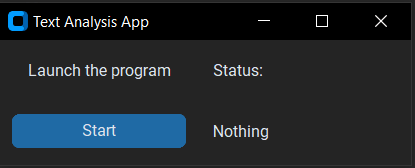
# **Рекомендации программисту**

Для запуска программы необходим Python, а также 64-битная операционная система Windows, или Linux, или macOS. Необходимые библиотеки: asyncio версии 3.4.3, matplotlib версии 3.7.1, customtkinter версии 5.2.0, requests версии 2.28.1, Pyrogram версии 2.0.106, wordCloud версии 1.8.1, pymystem3 версии 0.2.0, nltk версии 3.2.2, genism версии 1.12.0. Для работы с кодом необходим PyCharm версии не ниже 2022.2.1.

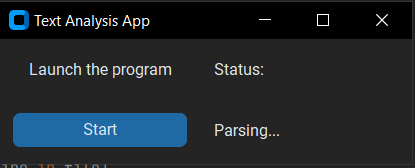
Минимальное необходимое место на диске: 0.5 МБ. Минимальное необходимое количество оперативной памяти: 100 МБ.

# **Контрольный пример**

В данном разделе представлен контрольные пример, демонстрирующий способность выполнять комплексную параллельную обработку текстовых данных.



*Рисунок 8.1. Интерфейс программы*



*Рисунок 8.2. Интерфейс программы во время обработки*



*Рисунок 8.3. Вывод статистики топиков*

**

*Рисунок 8.4. Облако слов*

# **Заключение**

В рамках представленной работы был разработан и успешно реализован алгоритм для комплексной параллельной обработки, анализа и интерпретации текстовых данных. Этот алгоритм направлен на изучение ключевых метрик, связанных с распознаванием трендов и тем в социальных сетях.