Частотный анализ русского текста и облако слов

на языке программирования Python

Штурмина Инна Сергеевна
9 класс, МАОУ «Лицей №38»
Научный руководитель: Попова Наталья Львовна
Учитель информатики высшей квалификационной категории;
Штурмин Сергей Александрович
Заместитель генерального директора АО "ГНИВЦ"

В работе изучен такой метод анализа текста как «частотный анализ», была написана программа на языке программирования Python, позволяющая анализировать текст разной направленности и разного объема методом частотного анализа за несколько минут и визуализировать результат работы в «облако слов».

В век технологий человек стремится максимально освободить свое время, делегировав все задачи, которые могут сделать за него машины и механизмы. В последнее время стали набирать популярность сервисы, позволяющие прослушать нужный текст, будь то книга, учебник или инструкция, дабы не тратить на это драгоценные минуты. Таким образом появилась идея создать программу на языке программирования Python, которая может проанализировать текст, выделить слова его характеризующие и визуализировать результат в форме «облака слов».

Частотный анализ, идея которого используется для создания алгоритма – это метод криптоанализа, основывающийся на теории о существовании определенного числа встречаемости символа и последовательности символов в тексте. Частотный анализ предполагает, что частота появления заданной буквы алфавита в достаточно длинных текстах одна и та же для разных текстов одного языка. Например, пара стоящих рядом букв «ся» в русском языке более вероятна, чем «цы», а «оь» в русском языке не встречается никогда (зато часто встречается, например, в чеченском). Анализируя достаточно длинный текст, зашифрованный методом замены, можно по частотам появления символов произвести обратную замену и восстановить исходный текст.

Для разработки приложения был использован принцип работы метода частотного анализа, с помощью которого программа определяет количество повторяющихся слов, а не букв. В некоторых источниках данный метод анализа текста называется статистическим.

Алгоритм был реализован на языке программирования Python. Это обусловлено возможностью использования существующих библиотек для решения задач: библиотеки NLP (natural language processing), библиотеки NLTK для анализа текста, библиотеки wordcloud для построения облака слов. Используется интерактивная веб-платформа для разработки python-скриптов Jupyter Notebook.

Алгоритм работы программы состоит из следующих шагов:

1. Загрузка и обзор данных
2. Очистка и предварительная обработка текста
3. Удаление стоп-слов
4. Перевод слов в основную форму
5. Подсчёт статистики встречаемости слов в тексте
6. Визуализация популярности слов в виде облака

Рассмотрим каждый шаг алгоритма подробнее:

1. *Загрузка данных.*

На этом этапе происходит загрузка файла с текстом в оперативную память компьютера с помощью встроенной функции, указывается режим чтения и кодировка.

2. *Предварительная обработка (препроцессинг) текста.*

Для проведения частотного анализа и определения тематики текста выполняется очистка текста от знаков пунктуации, лишних пробелов и цифр с помощью встроенных функций работы со строками. Все символы переводятся в нижний регистр (прописные буквы). Далее выявляются нестандартные знаки препинания, которые вносятся в готовую библиотеку знаков.

Для удаления символов используется поэлементная обработка строки – разделение исходной строки на символы, удаление «лишних» символов и объединение оставшихся символов в строку.

3. *Токенизация текста.*

Для последующей обработки очищенный текст разбивается на составные части – токены. В анализе текста на естественном языке применяется разбиение на текстовые единицы: символы, слова и предложения. Процесс разбиения называется токенизация. В данном алгоритме текст разбивается на слова.

4. *Подсчет статистики встречаемости слов в тексте.*

С помощью загруженной заранее библиотеки программа строит график, показывающий самые часто встречающиеся слова и количество их повторений. На данном этапе наибольшие частоты имеют союзы, предлоги и другие служебные части речи, не несущие смысловой нагрузки, а только выражающие семантико-синтаксические отношения между словами. Для того чтобы результаты частотного анализа отражали тематику текста, необходимо удалить эти слова из текста.

5. *Удаление стоп-слов*

К стоп-словам (или шумовым словам), как правило, относят предлоги, союзы, междометия, частицы и другие части речи, которые часто встречаются в тексте, являются служебными и не несут смысловой нагрузки – являются избыточными. Для их удаления используется библиотека стоп-слов, загружаемая в программу. После этой процедуры результаты частотного анализа становятся более информативными и точнее отражают основную тематику текста.

6. *Визуализация популярности слов в виде облака.*

В завершение работы результаты частотного анализа текста визуализируются в виде «облака слов». Облако тегов (облако слов) — это визуальное представление списка категорий или тегов. Обычно используется для описания тегов и данных на веб-сайтах или для представления неформатированного текста. Ключевые слова чаще всего представляют собой отдельные слова, и важность каждого ключевого слова обозначается размером шрифта или цветом (рис.1, рис. 2).

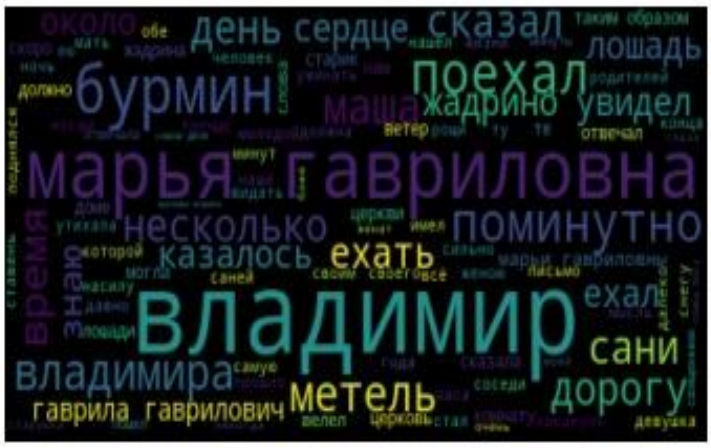

Рис.1 Результат обработки повести А.С. Пушкина «Метель»

Рис. 2 Результат обработки научно-популярного текста «Тайна черных дыр»

Список литературы:

1. Анализ текстов. Частотные характеристики текстовых сообщений. [Электронный ресурс] Сайт. URL: https://web.archive.org/web/20131213121450/http://www.statistica.ru/local-portals/data-mining/analiz-tekstov/

2. Криптоанализ. [Электронный ресурс] Онлайн-энциклопедия. URL: https://ru.wikipedia.org/wiki/%D0%9A%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7
3. Частотный анализ русского текста и облако слов на Python. [Электронный ресурс] Сайт. URL: https://habr.com/ru/post/517410/
4. Облако тегов. [Электронный ресурс] Онлайн-энциклопедия. URL: https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D0%BB%D0%B0%D0%BA%D0%BE_%D1%82%D0%B5%D0%B3%D0%BE%D0%B2