Домашнее задание 1. Программирование на R

Задание: Обработка и анализ текста на языке R

Описание задания

В этом задании вам предстоит написать скрипт на языке R для обработки текста. Вы будете работать с текстовым файлом, который необходимо загрузить в скрипт. Ваша задача — определить все слова в тексте, перевести их в верхний регистр, убрать знаки препинания, посчитать частоты слов и построить диаграммы для визуализации результатов.

Что важно: я думаю Вам хотелось бы знать какие слова на речи вы используете чаще всего[1], а какие реже. Поэтому те, у кого есть возможность, выгрузите тексты с распознанием речи (например, из мессенджера telegram). Студенты, у кого такой возможности нет, постарайтесь спарсить из того же telegram. Во всяком случае, задание будет засчитываться, если вы возьмете обычный текст, но респект и уважение тем, кто будет анализировать свои текстовые сообщения (и голосовые).

Сделаю небольшую ремарку: ничего страшного, если вы часто выражаетесь нецензурными словами и бранью, будет стимул стать лучше! Фильтровать базар – это искусство, а не обязательство.

Шаги выполнения задания

1. Загрузка текста:

Загрузите текстовый файл в скрипте. Вы можете использовать любой текстовый файл на ваш выбор (но желательно txt).

2. Обработка текста:

Определите все слова в тексте. Переведите все слова в верхний регистр. Удалите все знаки препинания. *Желательно*: приведите слова в единственное число, разнообразьте обработку, чтобы устранить повторения. Сделайте токенизацию и лемматизацию.

3. Анализ текста:

Посчитайте количество (частоты) каждого слова в тексте.

4. Визуализация данных:

[1] Сюда, сюдаааа, чиназес, скуф, SKU, skuf.

Постройте диаграммы для визуализации частот слов. Вы можете использовать гистограммы, столбчатые диаграммы или облака слов.

5. Требования к выполнению задания

Используйте функции и пакеты языка R для выполнения всех шагов. Код должен быть хорошо структурирован и прокомментирован. Результаты анализа должны быть представлены в виде диаграмм.

Отчетный файл высылать в формате .rmd (учил вас как с ним работать).