МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное   
образовательное учреждение высшего образования  
«Самарский национальный исследовательский университет   
имени академика С.П. Королева»

(Самарский университет)

Институт информатики, математики и электроники

Факультет информатики  
Кафедра технической кибернетики

**Отчет по лабораторной работе №2**

Дисциплина: «Большие Данные»

## Тема: «Формирование отчётов в Apache Spark»

Выполнил: Бабонин Д.А.

Группа: 6133-010402D

Самара 2020

**ЗАДАНИЕ**

1. Преобразовать любой файл набора данных stackoverflow в parquet формат c помощью Apache Spark.
2. Сформировать отчёт с информацией о частоте обсуждения 10 наиболее популярных языков программирования в каждом году с 2010 года по сегодняшний день. Используйте теги входящие в список языков перечисленных в википедии https://en.wikipedia.org/wiki/List\_of\_programming\_languages.

**ХОД РАБОТЫ**

Подготовим необходимые файлы для лабораторной работы. Загрузим файлы programming-languages.csv и posts\_sample.xml. Напишем программу, которая формирует отчет о частоте обсуждения популярных языков программирования с 2010 по 2020.

Используемый язык для программы Scala

Сформируем список годов с 2010 по 2020 и список языков программирования. В файле post\_sample оставляем год публикации и тэг, состоящий из названия языка программирования. Произведем фильтрацию RDD по году и тэгу. Вычислим количество постов каждого языка программирования, отсортируем по времени, выберем первые 10 за каждый год и получим топ 10 языков программирования.

Выведем результат на экран, сохранив датафрейм в формате parquet при помощи команды df.write.format("parquet").save("name.parquet").

Ссылка на репозиторий: <https://github.com/DaniilBabonin/Big_Data_Lab_2.git>

Внутри репозитория в файле README.md содержится таблица по популярности языков за каждый год

**ЗАКЛЮЧЕНИЕ**

В результате лабораторной работы c помощью Apache Spark был преобразован файл набора данных stackoverflow в parquet формат. Сформирован отчёт с информацией о частоте обсуждения 10 популярных языков программирования в каждом году с 2010 года по 2020.