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»

Выполнила: Аксенова 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**

Скачаем и распакуем все необходимые для выполнения лабораторной работы файлы. Затем напишем скрипт, собирающий список языков программирования на основе предоставленной страницы Википедии:

from bs4 import BeautifulSoup

import re

import pandas as pd

with open("List of programming languages - Wikipedia.html", "r", encoding='utf8') as f:

    contents = f.read().encode('utf-8').decode('ascii', 'ignore')

    soup = BeautifulSoup(contents, 'lxml')

    output\_rows = []

    for tag in soup.find\_all('div', {'class': "div-col columns column-width"}):

        s = re.sub(r'\([^()]\*\)', '', tag.text.strip())

        index = s.find('ISO')

        s = s[:index]

        index = s.find(',')

        s = s[:index]

        for i in s.splitlines():

            output\_rows.append(i)

df = pd.DataFrame(data = output\_rows, columns=["languages"])

df.reset\_index

print(df)

df.to\_csv('languages2.csv',  index=False)

Скрипт для разбора html-страницы пишем на языке Python. Файл с названиями языков программирования сохраним как csv.

Затем напишем программу на языке Scala для обработки xml-файла и определения 10 самых популярных языков программирования за каждый год начиная с 2010 на основе постов на stackoverflow.

Создадим список годов начиная с 2010 и заканчивая 2020 и считаем наш созданный csv-файл с названиями языков так же в виде списка. Затем разберем xml-файл, оставив только год публикации и тэг, который может состоять из названия языка программирования. Отфильтруем полученный RDD по годам и тэгам, оставив года с 2010 и тэги, содержашие в себе только какой-то из языков программирования. Подсчитаем количество постов для каждого языка, отсортируем по убыванию и выберем первые 10 за каждый год – получим 10 самых популярных языков программирования.

Выыедем результат на экран и сохраним итоговый датафрейм в формате parquet при помощи команды df.write.format("parquet").save("name.parquet").

Код на Scala: <https://github.com/darinaksena/2020.12.01_BD_LR_2.4> .

+----+------------+-----+

|Year| Language|Count|

+----+------------+-----+

|2010| java| 52|

|2010| javascript| 44|

|2010| objective-c| 23|

|2010| sql| 12|

|2010| ruby| 12|

|2010| linq| 10|

|2010| delphi| 8|

|2010| xslt| 3|

|2010| bash| 3|

|2010| applescript| 3|

|2011| java| 93|

|2011| javascript| 83|

|2011| sql| 35|

|2011| objective-c| 34|

|2011| ruby| 20|

|2011| delphi| 8|

|2011| bash| 7|

|2011| linq| 6|

|2011| matlab| 5|

|2011|actionscript| 4|

|2012| javascript| 132|

|2012| java| 124|

|2012| objective-c| 45|

|2012| sql| 36|

|2012| ruby| 27|

|2012| bash| 10|

|2012| r| 9|

|2012| scala| 6|

|2012| matlab| 6|

|2012| linq| 6|

|2013| javascript| 198|

|2013| java| 194|

|2013| sql| 70|

|2013| objective-c| 40|

|2013| ruby| 32|

|2013| r| 26|

|2013| bash| 12|

|2013| scala| 11|

|2013| linq| 10|

|2013| matlab| 8|

|2014| javascript| 239|

|2014| java| 228|

|2014| sql| 80|

|2014| objective-c| 52|

|2014| r| 28|

|2014| ruby| 24|

|2014| bash| 17|

|2014| matlab| 17|

|2014| linq| 8|

|2014| scala| 8|

|2015| javascript| 277|

|2015| java| 209|

|2015| sql| 52|

|2015| r| 43|

|2015| objective-c| 30|

|2015| ruby| 21|

|2015| matlab| 16|

|2015| scala| 15|

|2015| bash| 13|

|2015| linq| 10|

|2016| javascript| 278|

|2016| java| 184|

|2016| r| 52|

|2016| sql| 50|

|2016| ruby| 24|

|2016| bash| 18|

|2016| scala| 16|

|2016| matlab| 15|

|2016| objective-c| 12|

|2016| linq| 11|

|2017| javascript| 246|

|2017| java| 204|

|2017| r| 56|

|2017| sql| 53|

|2017| objective-c| 19|

|2017| ruby| 17|

|2017| bash| 14|

|2017| scala| 12|

|2017| linq| 10|

|2017| haskell| 7|

|2018| javascript| 198|

|2018| java| 146|

|2018| r| 66|

|2018| sql| 49|

|2018| scala| 23|

|2018| bash| 12|

|2018| go| 11|

|2018| ruby| 10|

|2018| objective-c| 10|

|2018| kotlin| 8|

|2019| javascript| 135|

|2019| java| 95|

|2019| r| 37|

|2019| sql| 26|

|2019| bash| 11|

|2019| go| 9|

|2019| dart| 9|

|2019| matlab| 9|

|2019| kotlin| 8|

|2019| ruby| 8|

+----+------------+-----+

**ЗАКЛЮЧЕНИЕ**

В результате лабораторной работы был перобразован файл набора данных stackoverflow в parquet формат c помощью Apache Spark и сформирован отчёт с информацией о частоте обсуждения 10 наиболее популярных языков программирования в каждом году с 2010 года по сегодняшний день.