PRINCIPLES OF BIG DATA MANAGEMENT

PROJECT – II

SPRING 2017

Submitted by: TEAM 7

Sri Sai Anusha Gandu (sgr43)

Sushma Mitta (smgp6)

Sadanand Kallakuri (sk789)

Abhilash Reddy Gaddam (aggg6)

CONTENTS

1. INTRODUCTION

1.1 About Twitter

Twitter is an online news and social networking service where users post and interact with messages, "tweets," restricted to 140 characters. Registered users can post tweets, but those who are unregistered can only read them. Twitter Inc. is based in San Francisco, California, United States, and has more than 25 offices around the world. Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and launched in July, whereby the service rapidly gained worldwide popularity. As of 2016, Twitter had more than 319 million monthly active users.

1.2 About the Project

Here, we have collected the tweets using twitter API through tweepy using the keywords Python, JavaScript and Ruby in JSON (JavaScript Object Notation) format. The tweets then collected have been analyzed and different SQL queries are written to obtain the result.

2. REQUIREMENTS

2.1 Languages

1. Python
2. Scala
3. SQL
4. Java

2.2 Software

1. IntelliJ IDEA 3.4 (IDE)
2. Python 3.6
3. JDK 1.8
4. Scala 2.12.1
5. Spark 2.1
6. Virtual Box (Cloudera)

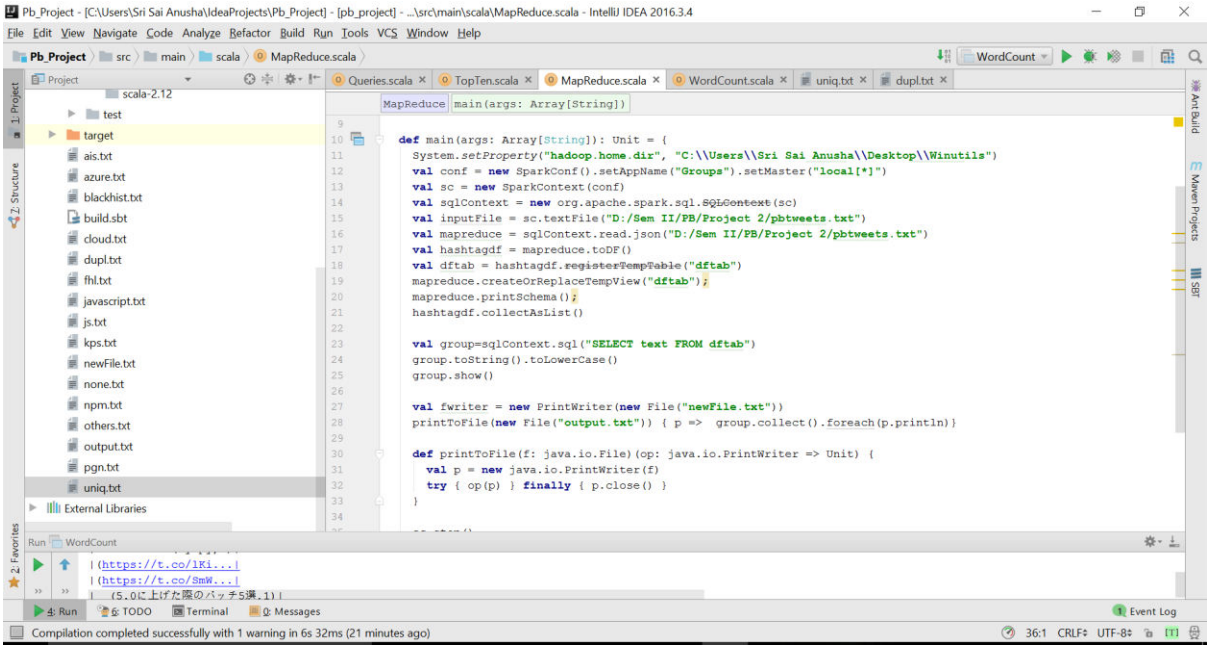
3. TASKS

1. To collect the tweets in JSON (JavaScript Object Notation) format.
2. Implement MapReduce to determine the uniqueness of the dataset.
3. To find the list of words in the tweets' text that are unique and duplicates.
4. Store the lists in two text files: uniq.txt and dupl.txt
5. To print the ratio of number of unique words to the number of duplicate words.

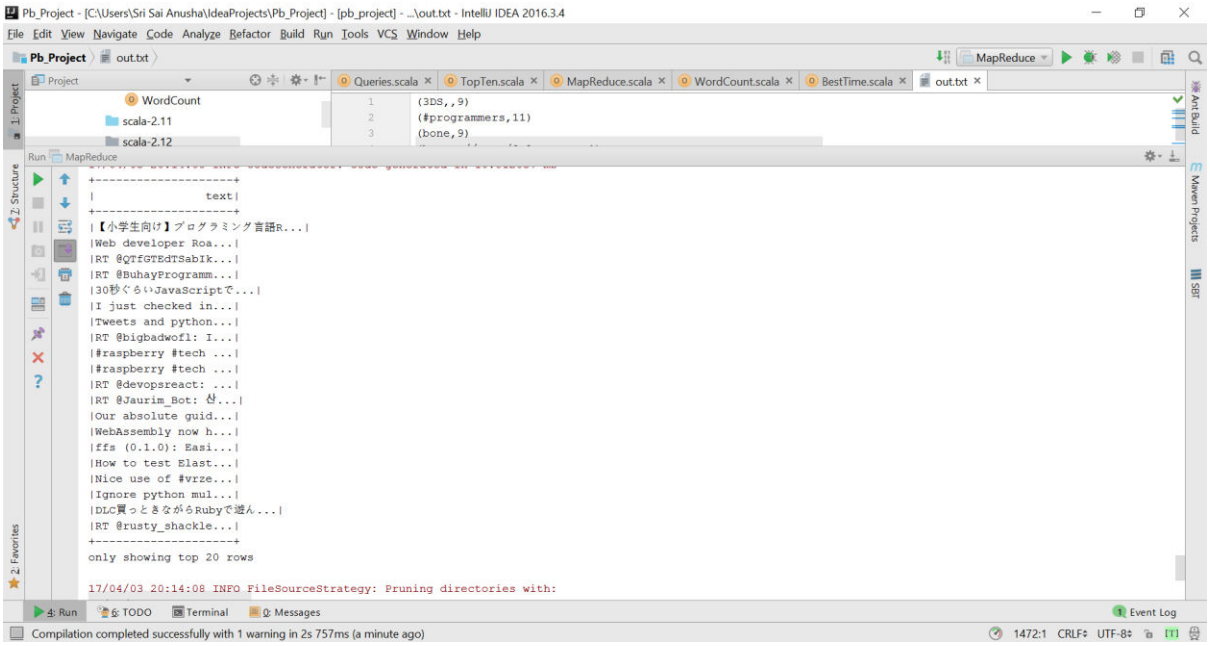
EXTRA REQUIREMENT

- To propose the best time to post a tweet on twitter.

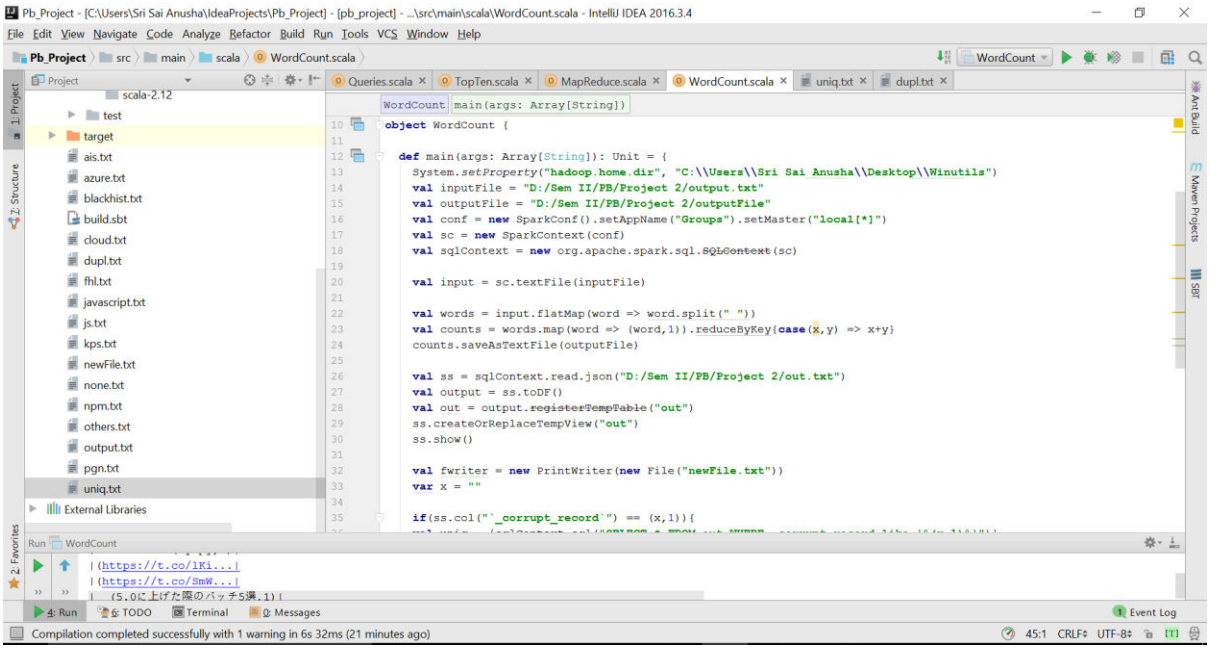
4. MAP REDUCE

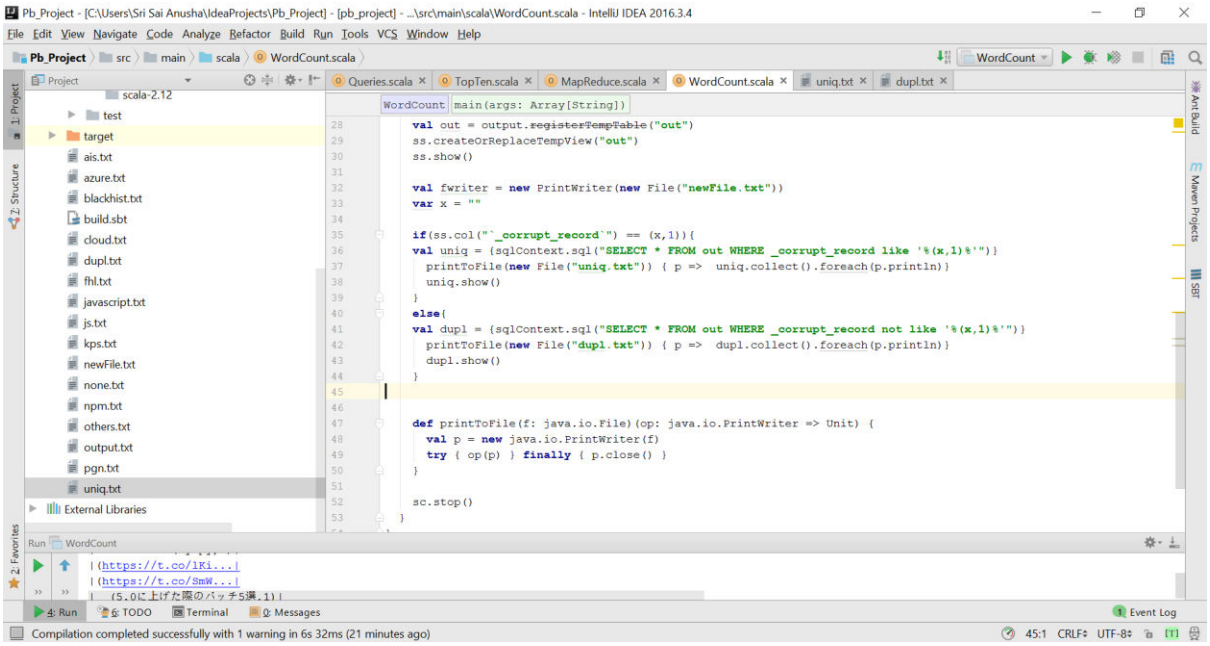


This bit of code extracts the text part from the tweets collected and saves it to a new text file named **output.txt**

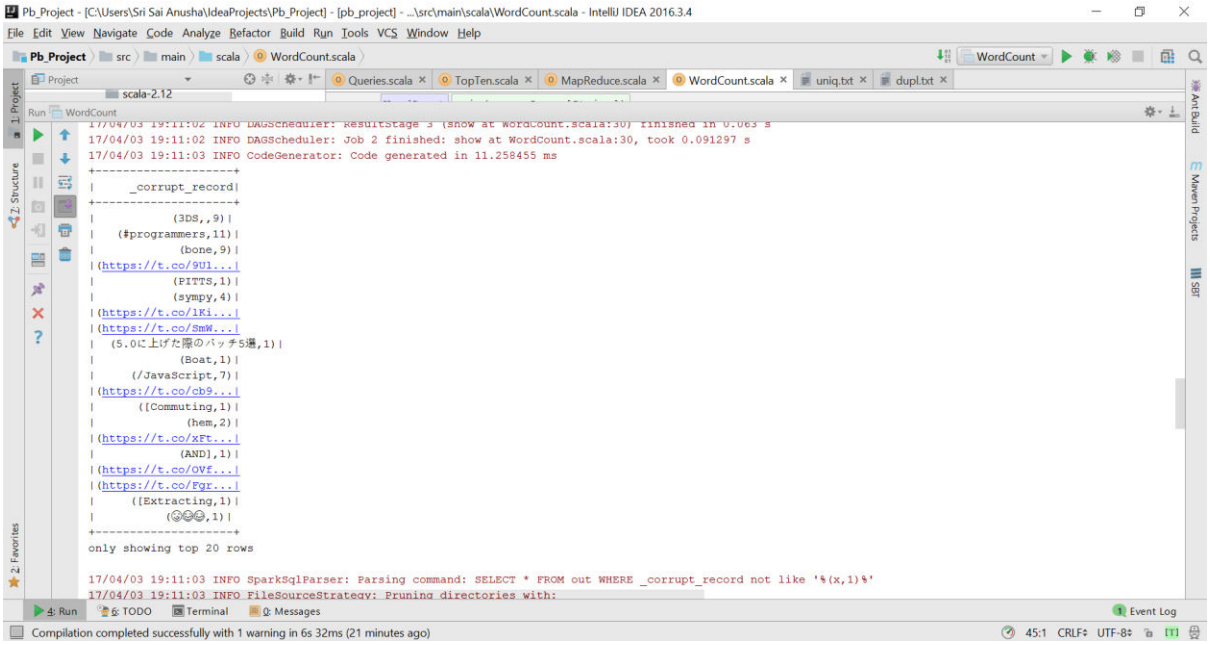


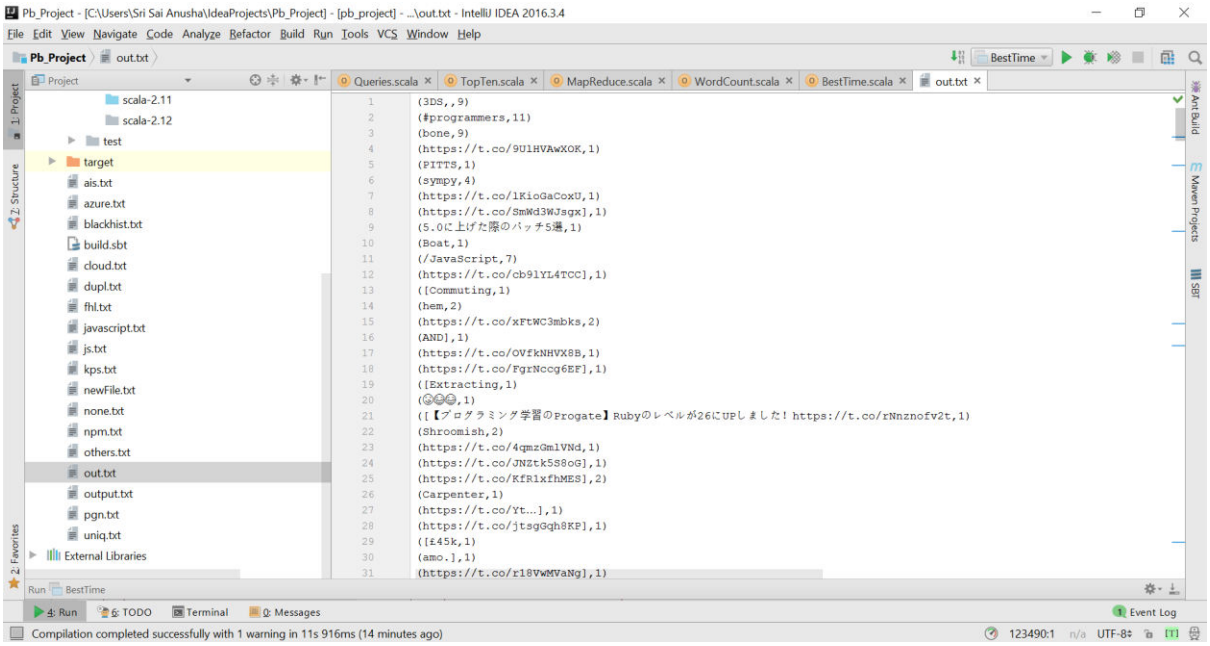
After performing extraction of text from tweets, map reduce function is performed on the text to obtain the output in the form of **(word, count)**



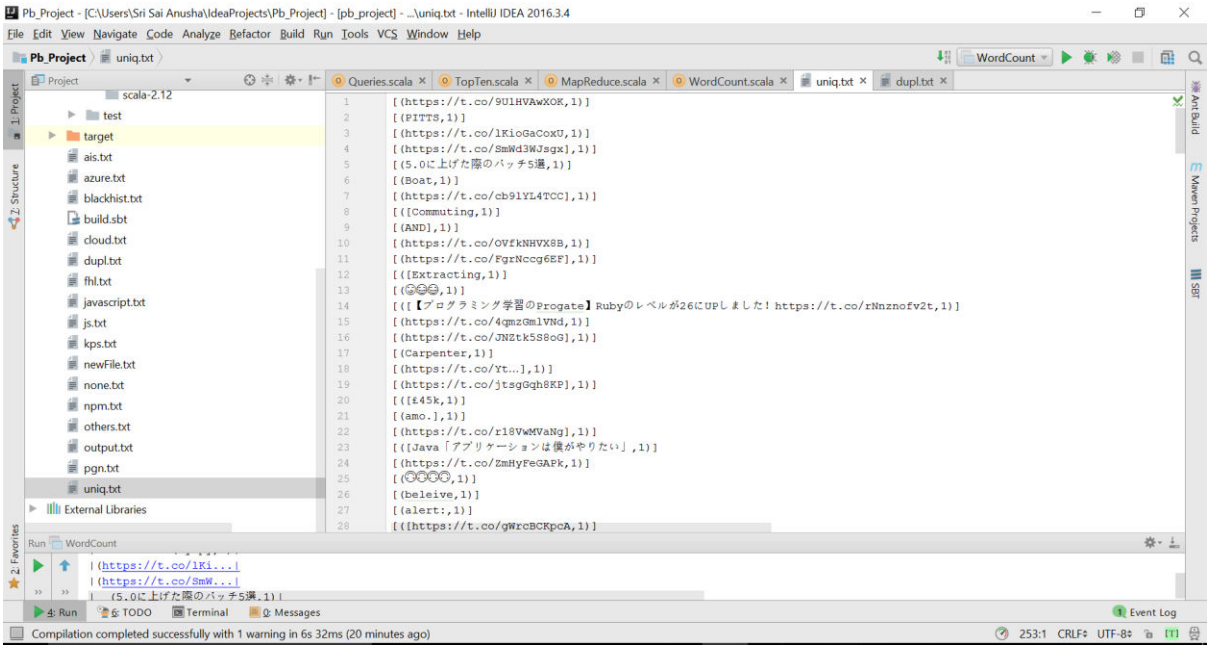


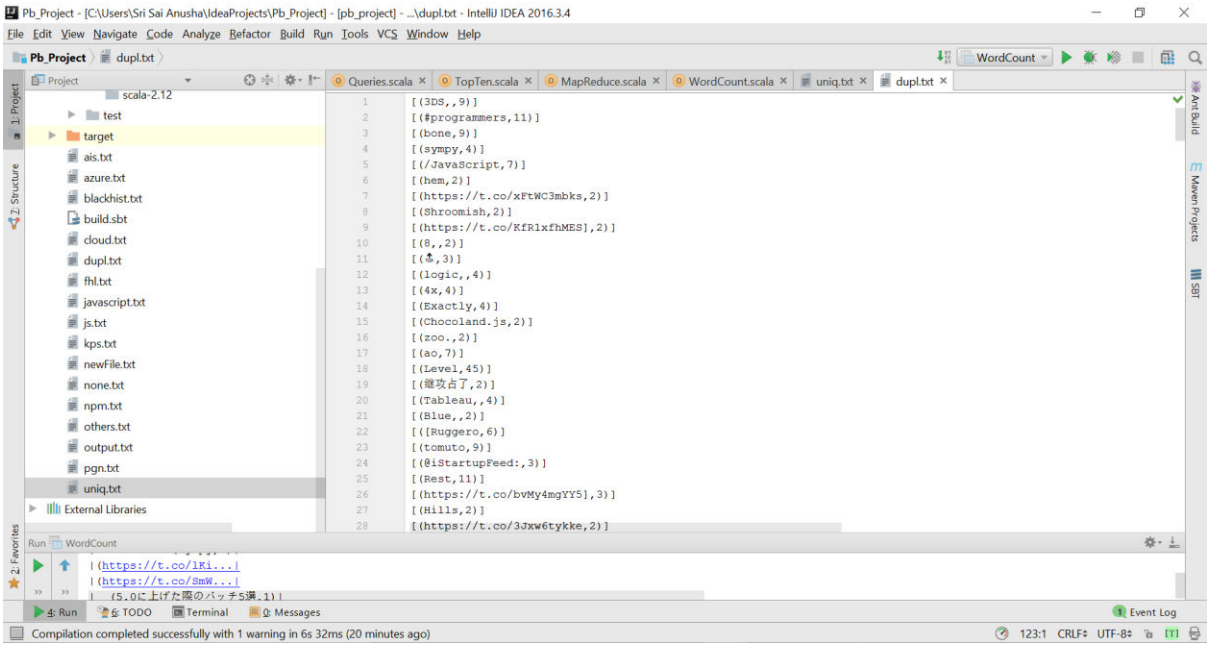
The output obtained through map reduce is shown below as a table and text file.



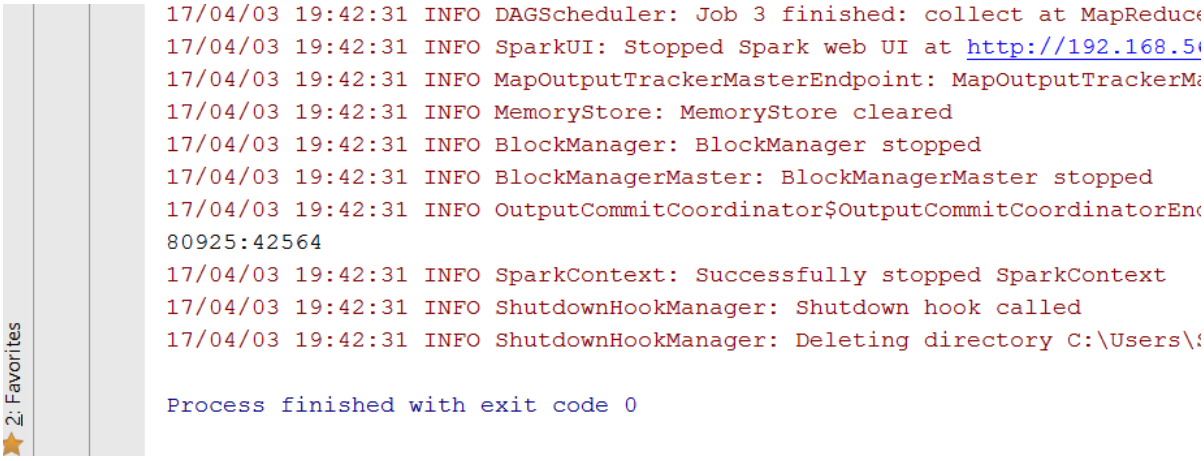


This output file is an aggregation of unique words and duplicate words in the text. These files are now segregated into uniq.txt and dupl.txt



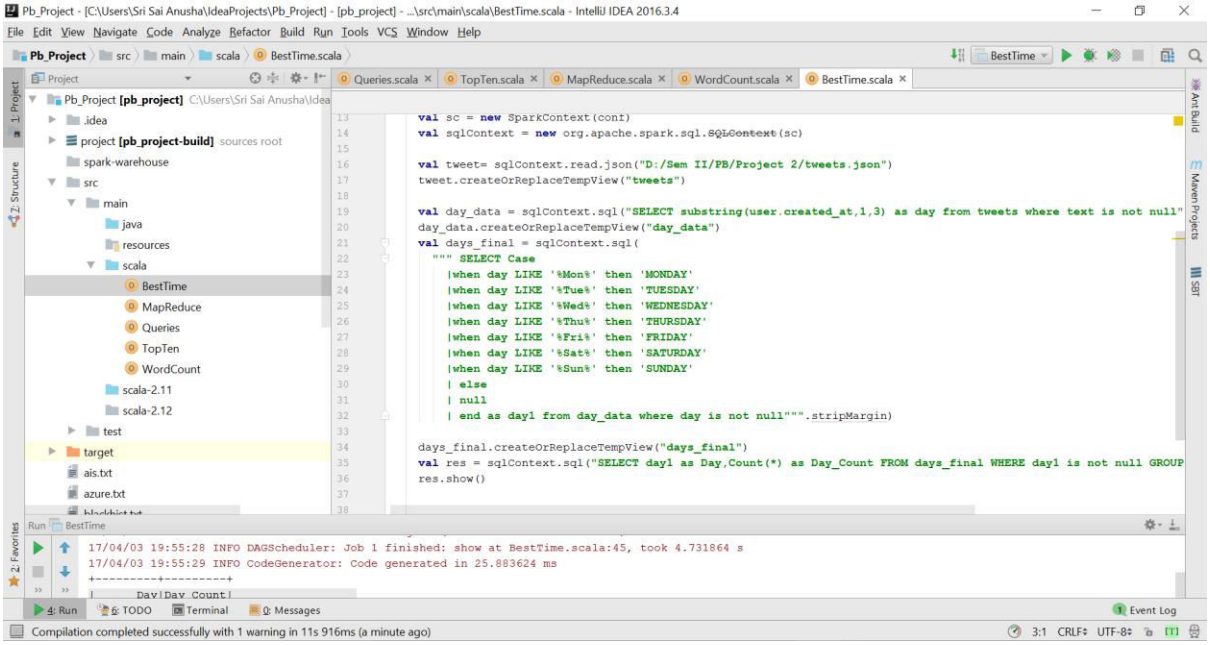


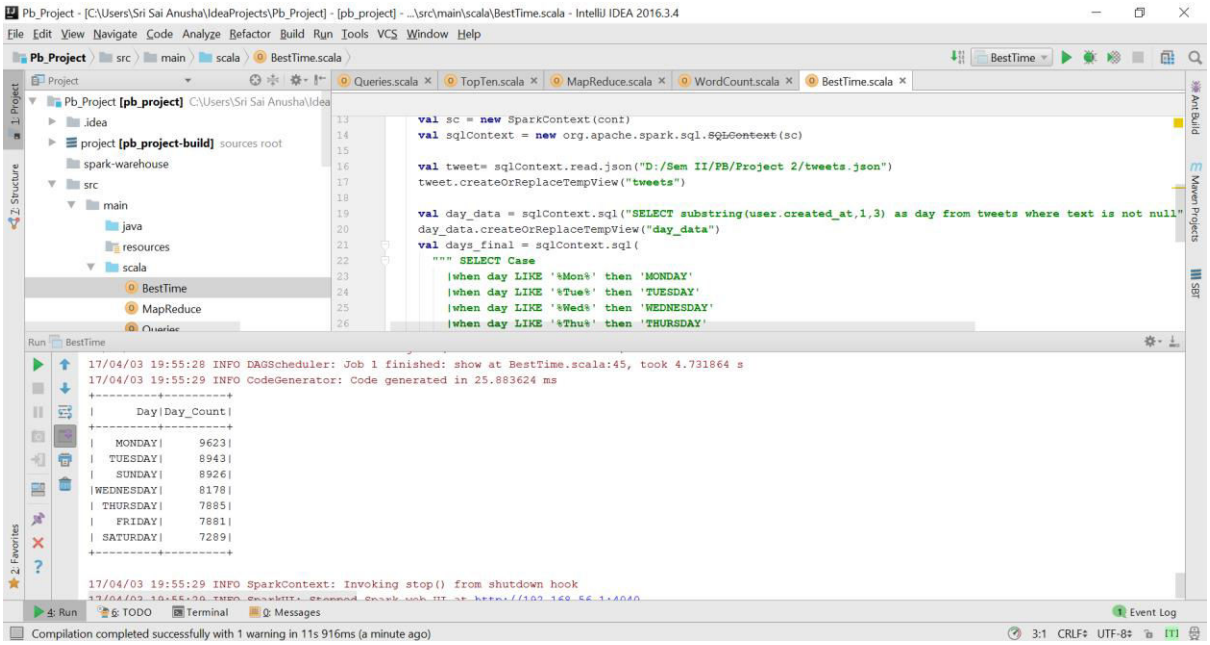
The ratio of the number of unique words to the number of duplicate words is shown below.



5. EXTRA REQUIREMENT

Based on the tweets JSON format, the metric to calculate the best time is the day of the week. According to the proposed metric, the output is as shown below.





According to the output thus obtained, the best day of the week to tweet is MONDAY and the least number of tweets are observed on SATURDAY

6. REFERENCES

- https://twitter.com/
- http://stackoverflow.com/questions/
- https://www.jetbrains.com/idea/
- docs.scala-lang.org
- alvinalexander.com
- www.tutorialspoint.com