COMP-SCI 5540 Principles of Big Data Management

**GitHub URL:** <https://github.com/NBALAJI95/MapReduce_Project>

# Project 2 Report (04/08/2017)

# Project Group - 6

**By**

Balaji Natarajan (bnqm6 - 53)

Mohammed Belhadj Larbi

Mustafa A. Muswadi

Nancy Yang (80)

# INTRODUCTION:

This project is to develop a system to analyze Twitter’s data using MapReduce based on Hadoop framework.

# PROJECT GOAL AND OBJECTIVES:

Using Python program collecting tweets and implement MapReduce programs to determine the vocabulary uniqueness of our dataset. The MapReduce program will find the word list of duplicates and unique word in the tweets’ text. We also implemented a MapReduce program to determine the best time to tweet.

# PLATFORM/ALGORITHM USED

# Cloudera VM

# Hadoop Streaming API for writing MapReduce programs in Python

# WORK ACCOMPLISHED

## MAIN REQUIREMENTS

# For finding the duplicates and unique work in the tweets’ text.

## SCREENSHOTS

# *Collected tweets*

# https://lh3.googleusercontent.com/wq0dyGSZEKZcp8ZXO6fvv_fhQ-w668flu7Xm9I5lBxMxtA4-odW9h99-aj45RiiTCdCNEXivPWIGsjBIXgj0obOcMQlqWNJ1burGwxht-mmfj3Oy3Q4Vr8190pjasLlQe9tvE4Nm

# *Text extracted from tweets*

# 

# *Unique words in text of tweets (Result of unique words map-reduce program)*

# 

# *Duplicate words in text of tweets (Result of duplicate words map-reduce program)*

# 

# *Main requirements output (ratio of unique words to duplicates)*

# 

## EXTRA REQUIREMENT

# For finding best time to tweet.

## METRIC

# The time when most tweets are found (most people tweet) is the best time to tweet.

# Reason: The time most people tweet is when most users will be online creating potential for viewing, favoriting and retweeting.

## SCREENSHOTS

# *Map reduce output for finding best time (UTC) to tweet*

# 

# *Top 10 best time to tweet*

# 

# PROJECT SUMMARY

* Found the list of word of duplicates and unique word in the tweets’ text.
* Stored them into two files: dups.txt and uniqs.txt.
* Printed the ratio of the number of unique words to the number of words with duplicates.
* Implemented a MapReduce program to determine the best time to post a tweet.
* Proposed the metric/criterion of your choice based on the tweet JSON format.
* Got the result of top ten best times to post a tweet on twitter.

# REFERENCES

# <https://www.coursera.org/learn/big-data-introduction>

# <https://www.youtube.com/watch?v=QNB1SZm2jS4&t=345s>