**Principles of Big Data Management**

**CSEE5540**

**Phase 1**

Team :

Apuroop Tummala (16296513)

Sameed shaik (16274073)

Tarun Teja Kasturi (16274556)

**Task 1**

Collect tweets using twitters streaming API(tweepy)

Python code has been written to collects tweets into a text file by using Twitter’s streaming API Tweepy. In order to connect to the Twitter’s Streaming API, we need to authenticate using credentials from our twitter developer account. The program which we have written filters data by "Sports" Keyword. The data is appended to a text file. We have collected around 600mb worth data of tweets.

**Task 2**

Hashtags and URLs extraction from the twitter data

We have written a python program to parse the text file from task1 into json format and to extract required Hashtags and urls and writes to two different files.

**Task 3**

WordCount Program Using Apache Hadoop.

Using mapreduce jar from hadoop we have saved the count of urls and hashtags on to a separate output folders.

WordCount Program Using Apache Spark.

We have created a rdd by using textFile function of the sparkcontext from the hashtags and urls textfile using scala. By using flatMap function, we generated words from each line by taking space as delimiter and appended one as value with the word as key. Using reduceByKey function we are adding values by the key and the output generated is the words and their sum is saved to a textfile.

All related outputfiles and logs are attached along with this file.