ANALYSIS REPORT

ASSIGNMENT 1

PROBLEM 1.1

WordCount problem using MapReduce and Hadoop

- The WordCount problem using MapReduce was run on two different data sets having varying lengths of data size.
- This was done to observe the runtime values of each dataset using the same program.
- As mentioned in the excel document, the dataset with larger number of words took longer to run as compared to the dataset with comparatively lesser number of words by a slight margin.
- Looking at this output, we can conclude that normal sorting and calculating the number of words would take huge amounts of processing time as compared to running it using Hadoop, which efficiently maps similar data and reduces it to return accurate output in much lesser time.

TopMovie problem using MapReduce and Hadoop

- The TopMovie problem uses MapReduce to return the top 10 most viewed movies given by the dataset.
- The dataset present gives the list of movies and the corresponding views for each movie.
- The MapReduce problems maps each movie with its count and using a TreeMap the top 10 movies are sorted.
- This is passed to the reducer class to consolidate top 10 most viewed movies and return it as output.
- Looking at this output, the MapReduce program can be used for faster shortlisting and exploiting data structures to return results quicker than brute force sorting/searching algorithms.