Assignment 03: Sentiment Analysis using Hadoop

Sentiment analysis is a process to identify emotions behind any text. In this assignment, you have to build a simple sentiment analysis MapReduce job. Following are the important files to perform the task:

1. twitter.json.tar.gz: A compressed file containing around 1 million tweets in a JSON format
2. positive_words.txt: A file containing list of positive words
3. negative_words.txt: A file containing list of negative words
4. stop_words.txt: A file containing list of stop words

Your MapReduce job should label each tweet either positive, negative, or neutral. For each tweet i, you need to ignore stop words, calculate positive words count p_i, and negative words count n_i. Then use the following formula to calculate the sentiment score s_i:

$$s_i = \frac{p_i - n_i}{p_i + n_i + 1} \tag{1}$$

A tweet i will be positive if s_i is greater than 0, negative if s_i is less than 0, and neutral if s_i is equal to 0. You must keep the original JSON document for each tweet and append additional attributed named *sentiment*. You need to produce three separate files for each category in JSON format.

Submission Guidelines

You need to prepare a small report contain screenshot of running your code, attach code in a zip file, and output file before **Tuesday 01 June, 11:59pm** and submit over Google Classroom.