**Group HW 2**

**Required: Must do only using Spark framework**

**General Overview**:

1. Use the file which have been uploaded on canvas as the input.
2. Get rid of all the punctuations, i.e., any character other than [a to z] and [0

to 9] and [spaces]. It will be helpful to use a regular expression to replace

them with empty strings.

1. Also, get rid of “a”, “as”, “and”, “the”, etc. stopwords and do not count them.
2. Develop the Java/python Mapper/Reducer code to implement
3. You should submit

- Your code file(s).

- Sample a screenshot of your output.

- Commands to run the code (E.g., Shell commands, Spark

commands)

- A description.txt file with a short description about your map/reduce

procedure.

Get rid of all punctuations (Group work)

Get rid of stop-words (Group work)

**Individual Part 1**

Consider the words are NOT case sensitive (which means

"Jingle" and "jingle" are the same word). Calculate word count

(occurrence of each word) in the file alice.txt and show the top 200 words with their count in descending order.

Sample output:

Word Count

jingle 7

hello 4

**Individual Part 2**

Consider the words are NOT case sensitive (which means

"Jingle" and "jingle" are the same word). Show the average length

of words starting with each letter.

Sample output:

a 2.5

b 5

c 3.2

**Individual Part 3**

Consider the words are NOT case sensitive (which means

"Jingle" and "jingle" are the same word). Show top 10 starting letters having highest average length words. Also show the total word count for these starting letters.

Sample output:

<starting letter> <word count> <avg word length>

b                                  17                   5

c                                  11                   3.2

a                                  20                   2.5

[Note:

\*\*\* It is a good idea to first take a small subset of the provided input file or create

a small dataset by yourself and then complete the programming. So that you can

test the correctness of your code. Then you can run your program for the whole

file and submit the output.

\*\*\* Each student in a group needs to do one separate part. Rest are collaborative

efforts. However, submit the code as one whole project].