CS5344 Lab 2

AY2018/2019 Semester 2

This purpose of this lab is to familiarize you with using RDD in Apache Spark. You will work with multiple text files and process words as key-value pairs. **This is an individual lab assignment.**

Task: Write a Spark program to find the common words in a set of documents.

Algorithm:

Step 1. Remove stopwords from the input documents.
Use the list of stopwords in the stopwords.txt file provided.

Step 2. Compute the count of every word in the resulting documents.
This is similar to the Word Count program in Lab 1.

Step 3. Identify the words common to all the documents and keep the smallest count of these common words. For example, if you have 3 documents and the word "dog" occurs 10, 5 and 8 times respectively, then the count for "dog" is 5.

Step 4. Sort the list of common words in descending order.

Input: (a) Set of documents (in "datafiles" folder),
(b) stopwords to remove (in *stopwords.txt*).

Output: One line per word in the following format: <*word*> <*count*>
You should sort the common words in descending order of occurrence.

Deliverables: Upload your executable Spark program (with proper documentation for important steps in the code) to the Lab2 folder in IVLE.

Important Notes:

- Specify the configuration of your program clearly.
  - For Python program, specify the python version along with the supporting packages used.
  - For Java program, provide the pom.xml configuration file and specify the directory structure of your source code files.
- Your code should be executable either on the virtual machine configuration given in Lab 1 or on stand-alone Spark configuration.

References:

- https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#transformations
- https://spark.apache.org/examples.html