CS 6350- ASSIGNMENT 1b

Please read the instructions below before starting the assignment.

- This assignment consists of two parts. Please create separate folders named **parti** and **partii** and zip them together for submission.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions on Piazza, and not through email to the instructor or TA.

ASSIGNMENT 1b

Part 1 (20 points)

In the first part of assignment 1, you uploaded 6 large files on HDFS. In this part, you will run a modified version of the WordCount algorithm on all of the files taken together using MapReduce.

Below are the requirements:

1. You will only consider words that have either a positive or negative emotion. A list of such words can be downloaded from http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

2. From the map method, you have to output (x, 1) for each term that matches the list, where x could be "positive" or "negative".

3. At the end of the program, you have to output the total count of positive and negative words i.e. something like:

Total count of positive words: ______
Total count of negative words: ______

Remember, you have to output a global count of positive and negative words for all the files combined. You have to accomplish this using the code within MapReduce solution.

Hint: The class Path in org.apache.hadoop.fs can represent a directory or path in HDFS
Note: UTD cluster does not have DistributedCache. Please do not attempt to put files there, it may cause the cluster to crash.

Deliverable:

- Only one zipped file containing your source code for the project.
- Readme file indicating how to run your code and the output of your program.

Part 2 (20 points)

In the second part of assignment 1, you downloaded a second text corpus (e.g. Wikipedia). In this assignment, you will convert each word from this corpus that has a minimum length of 5 characters into its corresponding Part-of-Speech (POS).
You can learn about POS here: http://partofspeech.org/ or here: https://en.wikipedia.org/wiki/Part_of_speech#English

To convert words to POS, you can use one of the following methods:
- Use a POS list. For example, http://icon.shef.ac.uk/Moby/mpos.html
- Use an API. For example, Oxford Dictionary API, https://developer.oxforddictionaries.com/
- Any other technique that you can thing of

Below are the requirements:

1. From the mapper, for each word having length greater than or equal to 5, you have to output the word length as the key and part of speech as the value.
For example, if you encounter the word "birthday", you will output (8, "noun") from the mapper.
Similarly, if you encounter the word "believe", you will output (7, "verb") from the mapper.

2. From the reducer, for each key you will output following summary statistics:
Example:

Length: 7
Count of Words: 300
Distribution of POS: {noun: 60; pronoun: 30; verb: 40.; ….}
Number of palindromes: 10

Length: 8
Count of Words: 380
Distribution of POS: {noun: 90; pronoun: 30; verb: 70.; ….}
Number of palindromes: 0

Note: UTD cluster does not have DistributedCache. Please do not attempt to put files there, it may cause the cluster to crash.

What to submit:
- Only one zipped file containing your source code for the project.
- Readme file indicating how to run your code and the output of your program.