Final exam questions

(50 points) Part 1: MapReduce

Count bigrams: Take the word count from the previous problem and extend it to count bigrams. bigrams are sequences of two consecutive words. (Recall our lecture on shingles.) Don't worry about doing anything fancy in terms of tokenization. You can use Java's StringTokenizer.

Deliverables (on Pride and Prejudice):

1. (10 points) How many unique bigrams are there?
    a. Give 3 examples of such bigrams.
2. (5 points) List the top ten most frequent bigrams and their counts.
3. (10 points) What fraction of all bigrams occurrences does the top ten bigrams account for? That is, what is the cumulative frequency of the top ten bigrams?
4. (5 points) How many bigrams appear only once?
    a. Give 3 examples of bigrams that appear only once.
5. (10 points) What are the five most frequent words following the word "light"? What is the frequency of observing each word?
    a. Describe in detail the procedure that you used to answer this question.
6. (10 points) Running in the cloud. Otherwise, 0 points.

Notes:

- This is an extension of your word counting project. You can reuse much of its implementation.
- For 1 – 6 give screenshots to show the running of your code and convincingly show that you implemented and ran your code.

(50 points) Part 2: Pandas

For this problem use the data crawled from your Project 2.

1. (5 points) Describe the that you use to solve this problem. (Hint: get it from your report.)
2. (5 points) Import the Pandas package.
3. (5 points) Load data into separate Data Frames. (All of you have collected multiple files. Work with at least 2 files for this problem.)
    a. Use csv functions to load data from csv files if your data is in csv.
    b. Use json functions to load data from json documents if your data is in json.
4. (5 points) Check the data-type of each of column by outputting the dttypes attribute of your DataFrame.
5. (5 points) Show an example of sorting one of your DataFrames by a column. Give the top-15 entries in descending order.
6. (10 point) Give an example of using filtering.

    a. Give an example for horizontal filtering/slicing where you select a subset of the columns. This corresponds to a projection in a SELECT statement.
    b. Give an example for vertical filtering/slicing where you select a subset of the rows in your DataFramer according to some criteria. This corresponds to WHERE in SELECT statement.
7. (10 points) Show an example where you merge two DataFrames.
8. (5 points) Export the merged DataFrame to
    a. A csv file if your input data is in json documents.
    b. A json file if your input data is in csv files.

Deliverables:

- A report that details your solutions.
- Include screen captures for every step of your solutions to convincingly show that
    - You have a working solution.
    - You are able to execute your implementations of the solutions.
- Upload the source code (exclude packages or libraries you might have used) in canvas.
- Give a ReadMe file where you describe how one can run your code.