CS 6350 - ASSIGNMENT 1

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.

- The deadline for this assignment is Wednesday September 7 at 11:59 PM. No extensions are allowed.

- 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 1

Part 1 (50 points)

For the first part, you will need to write **Java** code to download the following books (the .txt version only) and then upload them to a HDFS directory called assignment1.

- The Outline of Science, Vol. 1 (of 4) by J. Arthur Thomson
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/20417.txt.bz2
- The Notebooks of Leonardo Da Vinci
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/5000-8.txt.bz2
- The Art of War by 6th cent. B.C. Sunzi
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/132.txt.bz2
- The Adventures of Sherlock Holmes by Sir Arthur Conan Doyle
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/1661-8.txt.bz2
- The Devil's Dictionary by Ambrose Bierce
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/972.txt.bz2
- Encyclopaedia Britannica, 11th Edition, Volume 4, Part 3
  http://www.utdallas.edu/~axn112530/cs6350/lab2/input/19699.txt.bz2

You should then decompress the files on the HDFS filesystem using any of examples shown in class and also delete the .bz2 (compressed) files when you are done.

Deliverable:

Only one zipped file containing your source code for the project. It should be turned in through eLearning only.

Part 2 (50 points)

For this assignment, you will use Twitter Search API to download tweets about anything that you care about. For example, you could search for "UTD" and get a list of tweets. Additionally, you have to search for the same topic for 6 different timelines to get 6 different files. For example, you can search for UTD on 6 different days to get 6 input files for MapReduce program to process. The MapReduce program will then look for Hashtags i.e. those words starting with "#" to find which Hashtags are trending for a topic.

Here are the requirements:

1. You have to use the API i.e. sign up for a developer account and use a programming language (preferably Java, but not required) to search for tweets. You might find some useful information here:
https://dev.twitter.com/rest/public/search

2. You have to search for the same topic on 6 different timelines. It is up to you to choose time intervals e.g. one day, one week, etc. But you need to get 6 medium sized input files containing tweets. Remember to search for a topic and not a hashtag.

What to submit:

1. Your code to get tweets and your code to upload those tweets on HDFS. It can be in the same Eclipse project.
Do not turn in your input files as this will consume a lot of space on eLearning.
4. A Readme file indicating what topic you searched for, which timelines you used, and what was the size of the input files.