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 indicated on eLearning.

- 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 (20 points)

For the first part, you will need to write **Java** code to download the following books 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.

Hints:

You might want to look at the URL class:
https://docs.oracle.com/javase/7/docs/api/java/net/URL.html

Decompression can be understood by looking at the class FileDecompressor.java of Lab 2.

Deliverable:

- Only one zipped file containing your source code for the project. You can submit your zipped Eclipse project also.
- README file indicating how to run your code

Part 2 (20 points)

Download another text corpus i.e. dataset from this link:
https://www.corpusdata.org/formats.asp

You are free to choose any corpus, but you should download the “linear text” version of the data.
Like in part 1, you have to automate the process of getting this data into HDFS using Java code. You can look at code examples in Tom White’s book.

Hints:

In this case also, you might want to look at the URL class mentioned previously. For reading and downloading a zip file, you can also look at the ZipInputStream class:
https://docs.oracle.com/javase/7/docs/api/java/util/zip/ZipInputStream.html

Deliverable:

1. A Readme file indicating what corpus you chose.
2. Your code or Eclipse project in a zipped format.
3. README file indicating how to run your code