ASSIGNMENT 1 – Part 1

## Due Date: Friday January 29 before class on eLearning

This is an incremental assignment. 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 your HDFS directory.

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

You should then decompress the files on the server using any of examples shown in class.

You will use these in the next weeks to run WordCount algorithm using MapReduce computing paradigm. Try to make your code as efficient as possible and incorporate error checking e.g. what would happen if the file already exists.

**Deliverable:**

Only one zipped file containing your entire project. It should be turned in through eLearning only.