General Instructions

- The objective of this project is to practice some of the concepts introduced in the Spark lectures.
- We ask you to hand in an archive (zip or tar.gz) of your solution: code/scripts, README.txt file describing how to run your programs, a 5-10 page pdf report of your work (include your code in it).
- The report should list your answers (including code) and should also contain a short introduction and conclusion.
- The project is worth **10% of the total grade** for this module. The breakdown of marks for the project will be as follows:
  - Question 1 (code): 30%
  - Question 2 (code): 30%
  - Report: 40%
- **Due date: 11/03/2018**

1 CLI for Big Data

Download the following 6 books:

- `http://www.gutenberg.org/cache/epub/1524/pg1524.txt`
- `http://www.gutenberg.org/cache/epub/1112/pg1112.txt`
- `http://www.gutenberg.org/cache/epub/2267/pg2267.txt`
- `http://www.gutenberg.org/cache/epub/2253/pg2253.txt`
- `http://www.gutenberg.org/cache/epub/1513/pg1513.txt`
- `http://www.gutenberg.org/cache/epub/1120/pg1120.txt`

Using Bash commands and/or scripts in Bash, answer the following questions on the corpus:

1. what is the number of distinct words in the corpus? how many words start with the letter Z/z? how many words appear less than 4 times?
2. What are the most frequent words that end in *-ing*?
3. Which is more frequent: me/my/mine/I or us/our/ours/we?
4. Take one stopword (e.g., the, and) and compute the five words that appear the most after it. E.g. "the cat belongs to the old lady from the hamlet" → "cat ", "old" and "hamlet" would be candidates. The output should contains 5 lines with the words and their frequency.

2 Hadoop

Upload the 6 books in your HDFS. Modify the WordCount example (in Java) given in a previous practical to compute the following statistics on the corpus

1. (**One single job**) what is the number of distinct words in the corpus? how many words start with the letter Z/z? how many words appear less than 4 times? Think counters, and in particular user-defined counters. Starting off with the example given in the lecture, write your own Hadoop job.
2. (**One single job**) how many terms appear in only one single document? Such words may appear multiple times in one document, but they have to appear in only one document in the corpus.
3. (**One single job**) Take one stopword (e.g., the, and) and compute the five words that appear the most after it. E.g. "the cat belongs to the old lady from the hamlet" → "cat ", "old" and "hamlet" would be candidates. The output should contains 5 lines with the words and their frequency.