School of Computer Science

COMP47470

Project 2
Hadoop

Teaching Assistant:	Julia Boes & Ersi Ni
Coordinator:	Dr Anthony Ventresque
Date:	Thursday 28th March, 2019
Total Number of Pages:	2

General Instructions

- This project will consist in adding datasets to your Hadoop/HDFS platform (single node) and to run some Hadoop jobs.
- You are encouraged to collaborate with your peers on this project, but all written work must be your own. In particular we expect you to be able to explain every aspect of your solution if asked.
- We ask you to hand in an archive (zip or tar.gz) of your solution: code/scripts and a 5-10 page pdf report of your work.
- The report should include the following sections:
  1. a short introduction
  2. a requirement section that answers the question *what* is the system supposed to do
  3. an architecture/design section that answers the question *how* your solution has been designed to address the requirements described in the previous section
  4. a series of sections that describe the different challenges you faced and your solutions. For instance, take one of the script, describe the difficulty you faced and your solution. These sections can be short – the objective here is to show how you crafted the solutions with the tools you have learnt so far.
  5. a short conclusion
- **Due date: 14/04/2019**

1 Big Data job

Download an archive of debates from the European Parliament here:

- `http://csserver.ucd.ie/~aventresque/COMP47470/2019/Project2/`

Note that this is part of a bigger data collection exercise that you can find here: `http://www.statmt.org/europarl/`

Create a repository in your HDFS and Upload the text files there. Modify the WordCount example (in Java) given in one of your previous practical to answer some of the following questions:

1. run the WordCount example. How many unique words does the corpus contain?
2. Make five changes to the map function to reduce the number of unique terms found (e.g. remove non-alphanumeric terms) and run the job again. Which changes did you make and how do they influence the number of unique terms found?
3. How many words appear less than 4 times?
4. Which is more frequent: me/my/mine/I or us/our/ours/we? (use counters in the reduce function?)
5. 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 *(*COMP47470_2019_LECT8.PDF *p25 and following ones should give you an intuition on what the solution could be).*
6. Take a list of stopwords - e.g., this one `https://github.com/igorbrigadir/stopwords/tree/master/en` and run the WordCount job while filtering these. What is the impact on the output of the WordCount job? (number of words etc.) *(*COMP47470_2019_LECT9.PDF *p15 and following ones should give you an intuition on what the solution could be)*
7. Take one "political concept" (e.g., "justice", "citizen", "war", "hegemony", "nationality") and compute the five words that appear the most after it.