Cloud Computing — Coursework

Dell Zhang
Birkbeck, University of London

2018/19

[Version: 2019-01-11]

The coursework consists of two MapReduce programming assignments which you need to complete *independently* on Amazon Web Services (AWS). The programs should be developed in Python 3.6+ with the module `mrjob`[1]. Although you may write and debug your program on a local machine, your final solution should run in the cloud using Amazon's Elastic MapReduce (EMR).

For each problem, please submit the following files in one zip package through Birkbeck's Moodle[2]:

- a Jupyter Notebook (`.ipynb`) which contains your main program and gives your answer to the question asked in the problem description,
- other Python source code files (`.py`) needed for the execution of your main program,
- the configuration file `mrjob.conf` with your AWS and SSH credentials removed,
- a JPEG format screen-shot image (`.jpg`) of your Amazon EMR clusters console that shows your program's "COMPLETED" *state* as well as the *elapsed time*, and also your AWS *account name* at the top-right corner, and
- a plain text document (`.txt`) that reports how much time your program took to run on EMR with how many map nodes & reduce nodes, and also roughly how much time you spent working on this problem [for statistical purpose only, not for assessment].

The coursework is an integral part of this module and contributes 20% to the overall mark. You should work independently to complete all stages/parts of it. The Department reserves the right to interview any student over the coursework if there is a reasonable suspicion that the student has not done the coursework by himself/herself.

[1] http://mrjob.readthedocs.org/en/stable/
[2] http://moodle.bbk.ac.uk/

1. (10 marks)
   Write a MapReduce program to calculate the conditional probability that a word w' occurs immediately after another word w, i.e.,

   $$\Pr[w'|w] = \text{count}(w, w')/\text{count}(w)$$

   for *each and every* two-word-sequence, i.e., bigram, (w, w') in the entire collection of over 200,000 short jokes (from Kaggle).
   https://www.kaggle.com/abhinavmoudgil95/short-jokes
   You program should ignore non-alphabetical characters and be case-insensitive when extracting bigrams from text.
   Which 10 words are most likely to be said immediately after the word "my", i.e., with the highest conditional probability $\Pr[w'|w = \text{my}]$?
   Please list them in descending order.

   (a) If you implement either the "pairs" pattern or the "stripes" pattern correctly, you can get up to 8 marks.

   (b) If you implement both the "pairs" pattern and the "stripes" pattern correctly, you can get up to 10 marks.

2. (10 marks)
   Write a MapReduce program to calculate the PageRank (with damping factor 0.85) score for *each and every* user in the Epinions who-trust-whom online social network (from the SNAP dataset collection).
   http://snap.stanford.edu/data/soc-Epinions1.html
   Which 10 users have the highest PageRank scores in this social network?
   Please list them in descending order.

   (a) If you implement the "simplified" PageRank algorithm correctly, you can get up to 8 marks.

   (b) If you implement the "complete" PageRank algorithm correctly, you can get up to 10 marks.