Cloud Computing — Coursework

Dell Zhang
Birkbeck, University of London

2014/15

[Version: October 19, 2014]

The coursework consists of two MapReduce programming assignments which you should complete independently on Amazon Web Services (AWS). Although you may debug in standalone mode on your local machine, your final solution should run on AWS using the Elastic MapReduce (EMR). The program should be written in Java or Python. If you would like to use any other language, please get the approval from the module tutor in advance. For each problem, please submit the following files in one zip package through Birkbeck's Moodle (http://moodle.bbk.ac.uk/):

- the source code of your program;
- a screen-shot image (in JPEG format) of your EMR Job Flows 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;
- a brief document (in plain text format) that gives the answer to the question asked in the problem description; and
- a brief document (in plain text format) that reports how much time your program took to run on AWS with how many mapper nodes & reducer nodes, and also roughly how much time you spent working on this problem [for statistical purpose only, not for assessment].

The coursework is worth 20 marks in total.

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 two-word-sequence (w, w') in the entire corpus of Jane Austen's works (from the Gutenberg project). You should ignore the two-word-sequences across paragraph boundaries.

http://www.gutenberg.org/cache/epub/158/pg158.txt
http://www.gutenberg.org/cache/epub/946/pg946.txt
http://www.gutenberg.org/cache/epub/1212/pg1212.txt
http://www.gutenberg.org/cache/epub/141/pg141.txt
http://www.gutenberg.org/cache/epub/121/pg121.txt
http://www.gutenberg.org/cache/epub/105/pg105.txt
http://www.gutenberg.org/cache/epub/1342/pg1342.txt
http://www.gutenberg.org/cache/epub/161/pg161.txt
Which 10 words are most likely to be said immediately after the word "for", i.e., with the highest conditional probability $\Pr[w'|w = \text{for}]$?
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 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.