Cloud Computing

CS-351 **Spring 2019**

Homework 3 **Due 11:59:59 p.m., Friday, 17 May 2019 (7th Friday)**

Assignment Weight: 2.0

MapReduce

This homework is a MapReduce programming assignment which need to complete independently on Amazon Web Services (AWS). The program 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).

Please submit the following files in one zip package through Blackboard, Homework 3 by 11:59:59 p.m., Friday, 17 May 2019 (7th Friday):

- 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].

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

$$Pr[w'|w] = count(w, w')/count(w)$$

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.

[1] http://mrjob.readthedocs.org/en/stable/

Which 10 words are most likely to be said immediately after the word “my”, i.e., with the highest conditional probability $Pr[w'|w = my]$?

Please list them in descending order.

1. If you implement either the “pairs” pattern or the “stripes” pattern correctly, you can get up to 80% of your grade.
2. If you implement both the “pairs” pattern and the “stripes” pattern correctly, you can get up to 100% of your grade..