ENGG4030/ ESTR4300 Spring 2018 Homework 2

Release date: Mar 1, 2018

Due date: Mar 15, 2018 (Thursday) 11:59am (i.e. noon-time)

The solution will be posted right after the deadline, so NO late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student_________________________) Date:______________________________

Name_________________________________ SID_______________________________

Submission notice:

- Submit your report in a single PDF document on Elearning

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [70 marks + 20 Bonus marks]: Find frequent item sets

In this problem, we use the Shakespeare data set. The original data set has been pre-processed as follows:

- Apply a sliding window of 40 words on each work. All the 40 words in one window make up a basket.
- Delete duplicate words in one basket, then filter out some common words
- You can download the pre-processed data from the following link:

http://mobitec.ie.cuhk.edu.hk/engg4030Spring2018/homework/shakespeare_basket.zip

- Each line of this input is a tab separated list of words which corresponds to one basket.

The threshold for a frequent pair is defined as s=0.005. The frequency of a pair = *Occurrence of pair (i, j)* / *Total number of baskets*. For Q1(a), (b), (c) and (d), if the number of frequent pairs is larger than 40, please only submit the **Top** 40 pairs (if any). Your result should consist of a frequent pair and its corresponding count.

- You are allowed to use linux command *sort* to post-process your result.

(a) [20 marks] Implement the A-Priori algorithm to find frequent pairs on a single machine

Refer to Lecture 3, Page 30 to implement A-Priori algorithm. You do not need to use MapReduce Framework for this sub-question. You can run this job on one single AWS machine. Note that dic10.ie.cuhk.edu.hk is only a client for our DIC cluster. Please do NOT run the job in this machine.

(b) [30 marks] Implement the SON algorithm on MapReduce to find frequent pairs

Implement the SON algorithm under the MapReduce framework to find the frequent pairs. Note that your code should be scalable. In other words, your code should allow multiple mappers or reducers in both jobs. You need to implement two MapReduce jobs:

- First MapReduce job should use **A-priori** algorithm to find the candidate pairs, which are frequent in at least one input file.
- Second MapReduce job counts only the candidate frequent pairs.

Tips:

- In the second MapReduce job, each map will load all the candidate pairs. You can pass them as a supplementary file.

Streamline and performance comparison:

- Wrap the two MapReduce rounds as single executable by putting those commands you type before in a shell script.
- Compare the overall execution time of (a) and (b) .
- Output the command you use to submit Hadoop job.

You can use the IE Data Intensive Cluster (DIC) or any other Hadoop cluster (e.g., the AWS cluster built in HW#0) in various cloud computing platforms of your choice to do this problem.

(c) [30 marks] SON on MapReduce to find frequent triplets

The threshold for a frequent triplets is defined as s=0.0025. The frequency of a triplet (i, j, k) = *Occurrence of triplet (i, j, k)* / *Total number of baskets*. If the number of frequent triplets is larger than 20, please only submit the top 20 triplets (if any).

Tip:

- In case of memory error, you may need to use multiple mappers/ reducers (*eg*. 20+).

(d) [20 Bonus marks] Use the PCY algorithm to filter the candidate pairs in the SON algorithm

Implement the SON algorithm under the MapReduce framework. And use the **PCY** algorithm to filter the candidate pairs in the first MapReduce job. You can use the following Python hash function.

HashFunction=hash(word_1 + word_2) mod 100000

For example, the result of the word pair (‘Monday’, ‘Tuesday’) can be implemented as follows:

HashFunction=hash(‘Monday’ + ‘Tuesday’) % 100000

Streamline and performance comparison:

- Wrap the two MapReduce rounds as single executable by putting those commands you type before in a shell script.
- Compare the overall execution time of (a), (b) and (d).
- Output the command you use to submit Hadoop job.

Part (d) is an optional (bonus) part for ENGG4030, but mandatory for ESTR4300.

Q2 [20 marks]: Parameter Design for Minhash/ Locality-Sensitive Hashing (LSH)

Let r be the number of rows within each band and B be the total number of bands within the Minhash signature matrix M. We want to design the system so that:

- For any pair of items with similarity greater than or equal to T1, the probability that they will be correctly identified as a similar-pair candidate should be at least P1.
- For any pair of items with similarity below T2, the probability that they will be mistakenly identified as a similar-pair candidate should be no more than P2.

(a) [10 marks] Derive the set of inequalities to govern the relationship between T1, T2, P1, P2, r and B so that the aforementioned accuracy/error requirements would be satisfied.

(b) [10marks] For T1=0.9, T2=0.7, P1=0.99 and P2=0.05, use your results in part (a) to derive a single pair of values for (r , B) so that the aforementioned accuracy/error requirements would be satisfied. In general, there can be multiple feasible solutions. You only need to produce one pair of solution. Show your steps.

Submission requirement:

- You need to submit BOTH your code and your result. Please place the relevant code and the result in **a SINGLE PDF** file.