Assignment 2, CMPT 741. Due date: Nov 17 2017

Finding all pairs of objects (o1,o2) such that o1 and o2 are similar has many applications. For example, to recommend movies to a user Alice, the user-based collaborative filtering will identify a set of users that are similar to Alice’s taste in the ratings of items, where each user can be represented by a set of interested movies. In this assignment, you are asked to find all pairs of objects that have a similarity above a given threshold value t. Each object is represented by a subset of items from an item universe of size n. The m objects are represented by a n\*m matrix M, where each row corresponds to an item and each column corresponds to an object. The cell M(i,j)=1 if the j-th object contains the i-th item, otherwise, M(i,j)=0. The similarity between two sets (i.e., two columns) is computed by the Jaccard coefficient. We want to find all pairs of objects whose similarity is no less than a given threshold t.

Here are the details of the assignment:

* M will be given to you as a text file, with n=6000 and m=1000.
* Generate the signature matrix using p=100 random permutations through minhash
* Determine the proper b and r for t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b\*r=p. Show the S curve to justify your choice.
* Find candidate pairs of signatures using LHS, by choosing a hash function h with k=10,000 buckets.
* Determine FP (false positive) and FN (false negative) of the result in 4, i.e., the number of dissimilar signature pairs that are candidate pairs, and the number of similar signature pairs that are not candidate pairs.
* Find similar pairs of signatures by removing FP.
* Find similar pairs of objects from the remaining candidate pairs in 6 and determine the FP and FN of this result.

Repeat 2-7 using p=500 permutations.

Question 1: report the result in the table below (# refers to the number in the above list)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 3 (b,r) | 4(number of pairs) | 5(FP,FN) | 6(number of pairs) | 7 (FP,FN) |
| P=100 | (25,4) | 35659 | 33195, 994 | 1464 | 293, 6383 |
| P=500 | (125,4) | 118081 | 116891, 24 | 1190 | 10, 6374 |

Chosing best b and r for for 100 permut and t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b\*r=p. Show the S curve.

Best will be b = 25 and r = 4.



Chosing best b and r for for 500 permut and t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b\*r=p. Show the S curve.

Best will be b = 125 and r = 4.



For p=100 permutations:

Question 2: Draw the figure 1 for the FP and FN in 5 for different choices of b



Question 3: Draw the figure 2 similar to figure 1, but for the FP and FN in 7



A short discussion on how the value of b affects FP and FN in 5 and 7.

A proper b and r was calculated using a formula Pr(s) = 1- (1- s^r)^b, where Pr(s) is a probability of sharing the same bucket and s is a similarity of two sets.

A hash function used in 4 is

For all questions, explain your methods or formulas for computing b, r, FP, and FN. Also include the hash functions used in 4.

**For submission:** (1) one file contains your answers to the above questions, (2) one file contains the source code of any program you used (you can choose your own programming language). You should upload a single zipped file containing both (1) and (2) using the course system.

**Warning:** This assignment is designed to practice important concepts. You should do it individually without sharing your answers with others. Your understanding will be tested again in the final.

Marking: structure and clarity of reports, and correctness of results.

Figure 1