COS 435, Spring 2014 - Problem Set 5

Due at 1:30PM, Wednesday, April 9, 2014.

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently.

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- Penalized 10% of the earned score if submitted by 11:59 pm Wed. (4/9/14).
- Penalized 25% of the earned score if submitted by 4:30pm Friday (4/11/14).
- Penalized 50% if submitted later than 4:30 pm Friday (4/11/14).

Problem 1 (from 2012 exam problem)

This problem concerns a recommender system that uses collaborative filtering with item-item similarities (nearest-neighbor item similarities method). The system designers have decided that all items rated by a user should have some amount of influence on a recommendation for the user. However, the system has a very large number of items. To cut down on computation time for each recommendation, the system designers have decided to partition items into clusters based on item similarity. Item similarity is defined using the cosine measure on the vector of ratings by users (see slide #25 of the lecture slides "recommender systems and search"):

$$\text{sim}(i,j) = \frac{\sum_{u \text{ in } U_{i,j}} (r(u,i) - r_u^{avg})\,(r(u,j) - r_u^{avg})}{\left(\sum_{u \text{ in } U_{i,j}} (r(u,i) - r_u^{avg})^2 \sum_{u \text{ in } U_{i,j}} (r(u,j) - r_u^{avg})^2\right)^{1/2}}$$

where sim(i,j) is the similarity between items i and j, r(u,i) is the rating of item i given by user u, r_u^{avg} is the average rating by user u, and $U_{i,j}$ is the set of users who rated both items i and j.

Part A Design a recommendation algorithm that computes a prediction $r^{pred}(u,i)$ of the rating user u would give to item i, based on the similarity between items. The algorithm must use the pre-computed clusters of items to improve the running time and must include some direct or indirect influence by all items rated by the user u. Your description should be precise and complete.

Part B Assuming clusters of items have been computed in advance, what is the computational cost of your algorithm of Part A? Break up the cost into pre-processing time that is not done every time a prediction is made and time to actually make the prediction. Your times should be in terms of some or all of these quantities: the number of items, the number of clusters of items, the average number of items in a cluster and the number of users. Be precise!

Part C Now suppose user u rates item i for the first time. What updates need to be done before the next recommendation? What is the computational cost of these updates?

Problem 2 (similar to an old exam problem)

On the next page is the 5x7 term-document matrix C for a set of documents under the set of terms model and the matrices U, Σ, and V^T that make up the singular value decomposition of C.

Part a. Give the matrices of the rank-three approximation of C. That is, give U'_3, Σ'_3, and V'^T_3.

Part b. What is the 3-dimensional representation of Doc 5 for the rank-three approximation?

Part c. What is the 3-dimensional representation of term "cat" for the rank-three approximation?

Part d. In the 3-dimensional representation, what is the similarity of "cat" and "cow"? of "cat" and "dog"? What are the dot product similarities of the original representations of these terms as given in matrix C?

C =

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7
cat	1	0	1	1	0	0	1
cow	0	1	0	0	1	1	0
dog	1	0	1	1	0	1	1
pig	0	1	1	0	1	1	1
rabbit	1	1	1	0	0	1	0

U =

```
-0.390  0.359 -0.383  0.062  0.754
-0.310 -0.616 -0.167 -0.697  0.106
-0.499  0.512  0.546 -0.391 -0.193
-0.518 -0.472  0.418  0.566  0.122
-0.484  0.081 -0.594  0.194 -0.607
```

Σ=

```
3.741 0.000 0.000 0.000 0.000 0.000 0.000
0.000 1.902 0.000 0.000 0.000 0.000 0.000
0.000 0.000 1.275 0.000 0.000 0.000 0.000
0.000 0.000 0.000 0.692 0.000 0.000 0.000
0.000 0.000 0.000 0.000 0.528 0.000 0.000
```

V^T=

```
 -0.367  -0.351  -0.505  -0.133  -0.221  -0.588  -0.272
 0.501  -0.529   0.253   0.269  -0.572  -0.071   0.021
-0.337  -0.269  -0.010   0.428   0.197  -0.141   0.756
-0.195   0.091   0.622  -0.566  -0.189  -0.384   0.251
-0.088  -0.719   0.143  -0.366   0.431   0.343  -0.134
 3.724   0.801   0.196  -0.001  -0.023   0.672   0.098
-10.904  0.538  -0.311  -0.320   0.689   0.373   0.000
```

using Bluebit Software Online Matrix Calculator: www.bluebit.gr/matrix-calculator/