COS 435, Spring 2015 - Problem Set 5

Due at 1:30PM, Wednesday, April 8, 2015.

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. For each problem, list the students with whom you discussed general methods of solving the problem (excluding very brief casual conversations).

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. (4/8/15).
- Penalized 25% of the earned score if submitted by 4:30pm Friday (4/10/15).
- Penalized 50% if submitted later than 4:30 pm Friday (4/10/15).

Problem 1 (2013 exam 2 problem)

The example of recommendation by content filtering given in class (slides 5, 6 and 7 of "recommender systems and search" under 3/25/15) uses a very simple method: Books are characterized by a weight for each of a set of characteristics, giving a vector characterizing the book. For each book characteristic, the weights of all books purchased by a user are averaged to produce a vector characterization for the user. A new book is scored by taking the dot product of the new book's vector of characteristics and the vector characterizing the user. If a user gives an explicit preference weight for a characteristic, this weight replaces the average for that characteristic in the vector characterization for the user.

Part A: Develop a recommendation method based on content filtering that combines for ***each*** characteristic both a user's expressed preferences and a user's ratings of books read. Now, instead of simply knowing whether the user purchased a book or not, all purchased books have a 1 (worst) to 5 (best) integer rating from the user (with a default of 3 in case a users does not bother to rate a book). Books are still characterized by weights for

different book characteristics, each weight a real number ranging between 0 and 1, inclusive, that indicates the degree to which a book has the characteristic. Users may still give integer-valued preference values for characteristics, ranging between -2 (strongly dislike) and 2 (strongly like). A user may give preference values for some but not all characteristics. Give precise descriptions of your algorithm for scoring new books for the user and of the criteria for recommending a new book. ***Your algorithm will be judged on the potential for effectiveness and efficiency.***

Part B: Apply your method to the example below. An "X" indicates a characteristic for which the user did not give a preference value:

	1st person	Romance	Mystery	Sci-fi	User rating
Book 1	0	1	1	0	5
Book 2	0	1	0	0	4
New book A	1	.5	0	0	
New book B	0	1	0	.2	
User pref.	0	1	X	-2	

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 =

-0.40972	0.54966	-0.19439	-0.19354	0.67436
-0.27231	-0.59693	-0.05582	0.58538	0.47301
-0.53695	0.39476	-0.03386	0.55283	-0.49909
-0.51397	-0.40542	-0.51446	-0.48414	-0.26907
-0.45332	-0.14614	0.83263	-0.28260	0.00260

Σ=

3.73682	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	2.20586	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	1.19304	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.70548	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.49931	0.00000	0.00000

V^T=

-0.37465	-0.33173	-0.51219	-0.25333	-0.21041	-0.47542	-0.39088
0.36189	-0.52065	0.17810	0.42814	-0.45440	-0.34169	0.24435
0.50659	0.21990	0.07536	-0.19132	-0.47801	0.19151	-0.62254
0.10871	-0.25707	-0.57755	0.50928	0.14350	0.52655	-0.17698
0.35622	0.41365	-0.18266	0.35102	0.40844	-0.58592	-0.18787
0.50226	-0.49771	-0.00455	-0.50226	0.49771	0.00000	0.00455
0.28473	0.29260	-0.57733	-0.28473	-0.29260	0.00000	0.57733

computed with *www.dotnumerics.com/MatrixCalculator*
verified using *http://comnuan.com/cmnn01004/*