COS 435, Spring 2016 - Problem Set 5

Due 11:59 pm Wednesday April 6, 2016 by DropBox submission
Due at 5:00 pm, Wednesday, April 6, 2016 if submitting handwritten work on paper.

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.

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 10am Thursday (4/7/16).
- Penalized 25% of the earned score if submitted by 4:30 pm Friday (4/8/16).
- Penalized 50% if submitted later than 4:30 pm Friday (4/816).

Submission

Submit your solutions as a PDF file using the Computer Science Department DropBox submission system for COS435 at https://dropbox.cs.princeton.edu/COS435_S2016/HW5 Name your file HW5.pdf. If you have not used this facility before, consult the instructions at

https://csguide.cs.princeton.edu/academic/csdropbox - student

Note that you are automatically enrolled in CS DropBox using the registrar's COS435 enrollment list.

You may hand write your solutions as long as they are legible. In this case, you may either scan your writing to produce a PDF file for submission through DropBox or turn in your document by 5:00 PM Wed. April 6 in the bin outside Prof. LaPaugh's office.

Problem 1 (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?

Problem 2:

The algorithm for hierarchical agglomerative clustering giving in Figure 17.8 of *Introduction to Information Retrieval* uses one priority queue for each cluster to efficiently find the most similar pair of clusters to merge. The priority queues are updated for each merge step by deleting the two clusters that have been merged and inserting the new combined cluster. Consider ***breaking ties*** when selecting the pair of clusters to merge by choosing the pair that results in the smallest combined cluster. What modifications would be needed in the algorithm and data structures of Figure 17.8? Be sure to address all the data structures, not just the priority queues. Would the running time be affected? Explain.

For Problem 1:

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 =

$$\begin{matrix} -0.40972 & 0.54966 & -0.19439 & -0.19354 & 0.67436 \\ -0.27231 & -0.59693 & -0.05582 & 0.58538 & 0.47301 \\ -0.53695 & 0.39476 & -0.03386 & 0.55283 & -0.49909 \\ -0.51397 & -0.40542 & -0.51446 & -0.48414 & -0.26907 \\ -0.45332 & -0.14614 & 0.83263 & -0.28260 & 0.00260 \end{matrix}$$

Σ=

$$\begin{matrix} 3.73682 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 2.20586 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 1.19304 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.70548 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.49931 & 0.00000 & 0.00000 \end{matrix}$$

V^T=

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

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