Homework I

Due date: January 24, 2019 (Before the Class)

Note: For written questions, you need to turn in handwritten answers or a printed copy of your answers in class. If required, you need to submit your code on ICON.

Problem 1: Cosine and Dot Product Similarity (30 points)

In this homework assignment, you are required to compare the retrieval performance of two different similarity measures, i.e., dot product and cosine similarity. The document collection has already been preprocessed, with one file for each document. The collection of cleaned up documents and queries can be downloaded from ICON (go to your main course page on ICON and click Assignments and download the data sets in homework 1. Upon unzipping the file, you can see two folders. One folder named docs contains all documents, with one file for each document. Note that each file is a text file. You can open it using TextEdit in Mac, using NotePad in Windows and using Vim or Emacs in linux style system. Similarly, in the folder named queries one file is for each query.

You need first to extract the vocabulary out of the document collection and create a vector representation for each document and query. Let n be the number of unique worlds extracted from the document collection. Let $d = (d_1, \ldots, d_n)^\top \in \mathbb{R}^n$ denote a vector representation for a document where d_i is the term frequency of ith term in the vocabulary. Similarly, you can denote a query by $q = (q_1, \ldots, q_n)^\top \in \mathbb{R}^n$. Two similarity measures will be computed and compared. For dot product similarity, the document-query similarity is computed as

$$S_{\text{dot}}(d, q) = d^\top q = \sum_{i=1}^{n} d_i q_i = d_1 q_1 + d_2 q_2 + \ldots + d_n q_n$$

For cosine similarity, the document-query similarity can be computed by

$$S_{\cos}(d, q) = \frac{d^\top q}{\|d\|_2 \|q\|_2} = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

For each query, you are asked to compute the similarities between the query and all documents using both similarity measures, and return the first 10 documents with the largest scores (you can randomly break the tie when documents have identical scores). You will then compare the returned documents using different similarity measures, and discuss your observation.

1. (5’) Submit your code. Include a README file to describe how to run your code (e.g, what is the input, what is the output).

2. (15’) For each of the five queries and for each similarity measure, report the list of 10 most similar documents with their filenames (i.e., documents with the largest similarity scores).

3. (10') By looking at the content of the original documents, discuss the relevance of the returned documents to the query, and compare the performance of the two similarity measures.

4. (Optional 20'): Conduct the latent semantic indexing (LSI) on the term-document matrix for files in the docs directory and get top 5 singular values and their corresponding singular vectors. The left singular vectors are representations of concepts in the word space and the right singular vectors are representations of concepts in the document space. For each concept, you can get the top 10 words and top 10 documents by sorting the left and right singular vectors in descending order by magnitude. From the top 10 words for each concept, you can get a sense about the concept. You need to discuss the top 5 concepts, report the top 5 singular values and for each concept report the top 10 words and top 10 documents related to the concept. Now, If you need to retrieve similar documents for each query based on the top 5 concepts, how would you do it (describe your approach).

Problem 2: Singular Value Decomposition (20 points)

Let $X \in \mathbb{R}^{n \times d}$ $(n \geq d)$ denote a matrix with the singular value decomposition given by $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{d \times d}$ are orthonormal matrices satisfying $U^\top U = I_d$ and $V^\top V = I_d$, and $\Sigma = diag(\sigma_1, \cdots, \sigma_d)$ is a diagonal matrix with $\sigma_i \geq 0, i = 1, \ldots, d$. You are asked to simplify the following expression into products of three matrices, i.e, U, V and another diagonal matrix depending on σ_i and λ.

$$(\lambda I_d + X^\top X)^{-1} X^\top,$$

where I_d is an identity matrix of size $d \times d$.

Problem 3 (30 points)

There is a rare disease that only happens to 1 out of 100,000 people. A test shows positive 99% of times when applied to an ill patient and, 1% of times when applied to a healthy patient. Please answer the following questions.

1. What is the probability for you to have the disease given that your test result is positive?

2. What is the probability for you to have the disease when you did two tests and both of them show positive? Assume that two tests are conducted independent.

3. Assume that the patient keeps on trying the test, what is the minimum number of tests that the patient has to try to be 99% percent sure that he is actually ill? Assume that all tests are conducted independently.

Problem 4 (20 points)

Given a vector $\mathbf{y} \in \mathbb{R}^d$ and a scalar $\lambda > 0$, consider the following optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1.$$

Please derive the optimal solution of $\mathbf{x}$ expressed in terms of $\mathbf{y}$ and λ (you need to show detailed steps of derivation).