Homework I

Due date: Jan 25, 2023 (11:59pm)

Note: For written questions, you can either turn in a scanned copy of your handwritten answers or a PDF file of your answers. For programming questions, you need to submit your code. Please put your code and written answers in a zip file and submit it on Canvas.

Problem 1: Cosine and Dot Product Similarity (60 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 Canvas (Assignment/Homework I/hw1_data.zip). Upon unzipping the file, you can see two folders. One folder named docs contains all documents, with one file for each document. 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. In particular, you need to submit in this homework:

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

2. By looking at the content of the original documents, decide the relevance of the returned documents to the query, and compare the performance of the two similarity measures.

Problem 2: Singular Value Decomposition (40 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 compute

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

using U, Σ and V, where I_d is an identity matrix of size $d \times d$.