Homework 1
Ada Huang
CSCE 633 - 600
January 24, 2023

1 Cosine and Dot Similarity

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

Ans:

Below are the top ten documents with similarity scores ranking from high to low using dot product similarity and cosine similarity compared to each query. The format is

```
(score, document name)
```

Dot Product Similarity:

```
[[(170, '51060'),
  (21, '49960'),
  (11, '51153'),
  (10, '51165'),
  (9, '51164'),
  (7, '51144'),
  (7, '51120'),
  (6, '51135'),
  (6, '51130'),
  (5, '51161')],
 [(64, '59905'),
  (43, '60103'),
  (30, '59850'),
  (19, '59874'),
  (17, '59873'),
  (15, '59909'),
  (12, '59871'),
  (11, '59870'),
  (7, '60210'),
  (6, '59913')],
 [(9, '10011'),
  (7, '59913'),
  (5, '59850'),
  (5, '10083'),
  (4, '10089'),
  (3, '59874'),
  (3, '59873'),
  (3, '10074'),
  (3, '10066'),
  (2, '59908')],
 [(10, '59849'),
  (6, '59850'),
  (6, '51151'),
  (4, '51211'),
  (3, '60213'),
  (3, '60152'),
  (3, '59905'),
  (3, '59874'),
  (2, '60187'),
  (2, '60103')],
 [(23, '102610'),
  (17, '101666'),
  (7, '102591'),
  (6, '102647'),
  (6, '102604'),
  (5, '102627'),
  (5, '102609'),
  (4, '102648'),
  (4, '102608'),
  (4, '102598')]]
```

Cosine Similarity:

```
[[(0.561, '51060'),
  (0.275, '51144'),
  (0.24, ’51120’),
  (0.231, ’51164’),
  (0.217, ’51135’),
  (0.21, ’51158’),
  (0.201, ’51184’),
  (0.194, ’51171’),
  (0.178, ’51130’),
  (0.173, ’51161’)],
 [(0.266, ’59905’),
  (0.251, ’60103’),
  (0.225, ’60170’),
  (0.22, ’60210’),
  (0.21, ’59850’),
  (0.151, ’60195’),
  (0.15, ’59909’),
  (0.131, ’60198’),
  (0.129, ’59870’),
  (0.128, ’60200’)],
 [(0.224, ’10064’),
  (0.173, ’10083’),
  (0.171, ’10063’),
  (0.135, ’10089’),
  (0.132, ’10052’),
  (0.129, ’10013’),
  (0.113, ’10066’),
  (0.091, ’59913’),
  (0.081, ’10067’),
  (0.076, ’101639’)],
 [(0.112, ’102656’),
  (0.105, ’102660’),
  (0.098, ’51207’),
  (0.093, ’51151’),
  (0.093, ’10027’),
  (0.087, ’59849’),
  (0.083, ’60213’),
  (0.078, ’102675’),
  (0.066, ’51206’),
  (0.066, ’102626’)],
 [(0.392, ’102598’),
  (0.374, ’102610’),
  (0.277, ’102647’),
  (0.231, ’101666’),
  (0.22, ’102609’),
  (0.205, ’100521’),
  (0.192, ’102617’),
  (0.173, ’102608’),
  (0.164, ’102634’),
  (0.161, ’102622’)]]
```

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.

Ans: By looking at the content of the original documents, I found that dot product similarity is highly affected by the length of text, as it is based on the magnitude of the vectors, while cosine product similarity is less affected by the length of text, as it is based on the angle between vectors. So, when using dot product similarity, a high score does not necessarily indicate that the texts are similar in content. It could also be affected by the length of the texts.

Both dot product similarity and cosine similarity are insensitive to synonymy, meaning they cannot distinguish between words or phrases that have the same meaning but different wording.

2 Singular Value Decomposition

Let $X \in \mathbb{R}^{n\times d} (n \geqslant d)$ denote a matrix with the singular value decomposition given by $X = U\Sigma V^\intercal$, where $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times d}$ are orthonormal matrices satisfying $U^\intercal U = I_d$ and $V^\intercal V = I_d$, and $\Sigma = diag(\sigma_1, \ldots, \sigma_d)$ is a diagonal matrix with $\sigma_i \geqslant 0, i = 1, \ldots, d$. You are asked to compute

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

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

Ans:

$$
\begin{aligned}
&(\lambda I_d + X^\top X)^{-1} X^\top \\
&= (\lambda I_d + (U\Sigma V^\top)^\top U\Sigma V^\top)^{-1} (U\Sigma V^\top)^\top \\
&= (\lambda I_d + V\Sigma U^\top \cdot U\Sigma V^\top)^{-1} (V\Sigma U^\top) \quad \text{since } \Sigma = \Sigma^\top \\
&= (\lambda I_d + V\Sigma^2 V^\top)^{-1} (V\Sigma U^\top) \quad \text{since } U^\top U = I_d \\
&= (V(\lambda I_d + \Sigma^2) V^\top)^{-1} (V\Sigma U^\top) \quad \text{since } V^\top V = I_d, V \text{ is a square matrix, } VV^\top = I_d \\
&= (V^\top)^{-1} (\lambda I_d + \Sigma^2)^{-1} \cdot V^{-1} (V\Sigma U^\top) \quad \text{if } A, B \text{ are invertible, then } (AB)^{-1} = B^{-1}A^{-1} \text{ since } AB \cdot B^{-1}A^{-1} = I_d \\
&= V(\lambda I_d + \Sigma^2)^{-1} V^\top V\Sigma U^\top \quad \text{here } V^\top V = I_d \\
&= V \begin{bmatrix} \frac{1}{\lambda+\sigma_1^2} & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \frac{1}{\lambda+\sigma_d^2} \end{bmatrix} \Sigma U^\top \\
&= V \begin{bmatrix} \frac{\sigma_1}{\lambda+\sigma_1^2} & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \frac{\sigma_d}{\lambda+\sigma_d^2} \end{bmatrix} U^\top
\end{aligned}
$$

In the last two matrix representations, all the non-diagonal entries are zeros.