Web Retrieval and Mining Assignment #2

B0591025 Yu-Neng Wang

April 2, 2020

Problem

1. **IDF**

$$
\begin{aligned}
&Change\ base\ from\ b_1\ to\ b_2 \\
\Rightarrow\ & idf_{t,b_2} = \log_{b_2} \frac{N}{df_t}\ (from(6.7)) \\
\Rightarrow\ & idf_{t,b_2} = \frac{\log_{b_1} \frac{N}{df_t}}{log_{b_1} b_2} = \frac{idf_{t,b_1}}{log_{b_1} b_2} = log_{b_2} b_1 \times idf_{t,b_1} \\
\Rightarrow\ & Score(q,d)_{b_2} = \sum_{t \in q} tf\text{-}idf_{t,d,b_2} = \log_{b_2} b_1 \times Score(q,d)_{b_1}
\end{aligned}
$$

The base in (6.7) only affect the scale of score calculation in (6.9). Hence, the relative score would be closer/farther if we choose a larger/smaller base, but the original rank relationship among queries and documents keeps the same.

2. **SVD**

(a)

$$
CC^T = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \Rightarrow det(CC^T - \lambda I) = -\lambda^3 + 4\lambda^2 - 3\lambda = -\lambda(\lambda - 1)(\lambda - 3)
$$

$$
\Rightarrow two\ largest\ eigenvalues\ of\ CC^T\ are\ 3\ and\ 1.
$$

$$
C^TC = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \Rightarrow det(C^TC - \lambda I) = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3)
$$

$$
\Rightarrow two\ largest\ eigenvalues\ of\ C^TC\ are\ 3\ and\ 1.
$$

Hence, their two largest eigenvalues are the same.

(b)

$$
\begin{aligned}
C_1 = U\Sigma V^T &= \begin{bmatrix} -0.816 & 0.000 \\ -0.408 & 0.707 \\ -0.408 & 0.707 \end{bmatrix} \begin{bmatrix} 1.732 & 0.000 \\ 0.000 & 0.000 \end{bmatrix} \begin{bmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{bmatrix} \\
&= \begin{bmatrix} 1.0 & 1.0 \\ -0.5 & 0.5 \\ -0.5 & 0.5 \end{bmatrix} \\
&\Rightarrow Frobenious\ norm\ = \sqrt{4 * (0.5)^2} = 1 \\
&= \sigma_2
\end{aligned}
$$