Assignment #2

B05902120 / Yu-Ting, TSENG

Apr 19, 2019

Language Model

Language modeling has been well applied to IR. Given a corpus consisting of the following two documents, please calculate probabilities based on the corpus as a whole.
Document 1: the martian has landed.
Document 2: the latin pop sensation ricky martin.

What are P("sensation" | "pop"), P("pop" | "the"), and P("sensation" | "ricky") under a MLE-estimated bigram model, where MLE means "maximum likelihood estimation"?

P("sensation" | "pop") $= \frac{1}{1} = 1$; P("pop" | "the") $= 0$; P("sensation" | "ricky") $= 0$.

Consider P("pop martian") and P("pop martin"). Which should be higher? Does a MLE-estimated unigram model agree with this judgment? What about a MLE-estimated bigram model? If neither one agrees, please suggest another probabilistic model that might work well.

Unigram model:
P("pop martian") = P("pop") $*$ P("martian")$= \frac{1}{10} * \frac{1}{10} = 0.01$.
P("pop martin") = P("pop") $*$ P("martian")$= \frac{1}{10} * \frac{1}{10} = 0.01$.
Bigram model:
P("pop martian") $= 0$.
P("pop martin") $= 0$.

SVD

Let

$$C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{1}$$

be the term-document incidence matrix for a collection. Given that the SVD of the matrix C is

$$U = \begin{bmatrix} -0.816 & 0.000 \\ -0.408 & -0.707 \\ -0.408 & 0.707 \end{bmatrix}, \Sigma = \begin{bmatrix} 1.732 & 0.000 \\ 0.000 & 1.000 \end{bmatrix}, V^T = \begin{bmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{bmatrix} \tag{2}$$

please answer the following questions.

Show the first two largest eigenvalues of CC^T are the same as those of C^TC.

$$CC^T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{3}$$

$\det(CC^T - \lambda I) = \lambda * (\lambda^2 * 4\lambda + 3) = 0$; $\lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 0$.

$$C^TC = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{4}$$

$\det(CC^T - \lambda I) = \lambda^2 * 4\lambda + 3 = 0$; $\lambda_1 = 3, \lambda_2 = 1$.
The first two largest eigenvalues of CC^T are the same as those of C^TC.

Compute a rank 1 approximation C_1 to the matrix C. What is the Frobenius norm of the error of this approximation?

$$\Sigma_1 = \begin{bmatrix} 1.732 & 0.000 \\ 0.000 & 0.000 \end{bmatrix}, C_1 = U\Sigma_1V^T = \begin{bmatrix} 0.992 & 0.992 \\ 0.4996 & 0.4996 \\ 0.4996 & 0.4996 \end{bmatrix} \tag{5}$$

$$X = C - C_1 = U\Sigma_1V^T = U * \begin{bmatrix} 0 & 0 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix} \tag{6}$$

Frobenius norm $= (-0.5)^2 + (0.5)^2 + (0.5)^2 + (-0.5)^2 = 1$