Школа лингвистики, 2023-24 уч. год
Линейная алгебра и математический анализ
Мера tf-idf. Латентный семантический анализ. (12.12.2023)
Д. А. Филимонов

1 Латентный семантический анализ

Задача 1. Для следующих наборов предложений постройте терм-документные матрицы в метрике tf-idf.

(a)
$$\begin{cases} \text{Я сел на стул.} \\ \text{Я сел за стол.} \\ \text{Он сел на стол.} \end{cases}$$

(b)
$$\begin{cases} \text{Быть, или не быть?} \\ \text{Может, не надо?} \\ \text{Быть того не может!} \end{cases}$$

Задача 2. Пусть дана терм-документная матрица X и известно её сингулярное разложение

$$X = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{4}{\sqrt{18}} & \frac{1}{\sqrt{18}} & \frac{1}{\sqrt{18}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{pmatrix}$$

(a) Оставьте двумерное семантическое пространство. Запишите координаты терминов и координаты документов в новом пространстве.
(b) Найдите координаты нового документа (запроса) $q = (1; 3)$ в 2-мерном семантическом пространстве.
(c) Если считать, что вектор q — это поисковый запрос, то в каком порядке стоит ранжировать документы в поисковой выдаче?

Задача 3. Пусть дана терм-документная матрица X и известно её сингулярное разложение

$$X = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 2 & 0 \end{pmatrix} = \begin{pmatrix} \frac{9}{15} & 0 & 0 & -\frac{4}{5} \\ \frac{8}{15} & -\frac{1}{3} & -\frac{2}{3} & \frac{2}{5} \\ \frac{8}{15} & \frac{2}{3} & \frac{1}{3} & \frac{2}{5} \\ \frac{4}{15} & -\frac{2}{3} & \frac{2}{3} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

(a) Спроецируйте документы и термины на 2-мерное подпространство главных компонент. Какие координаты будут у терминов и документов в этом семантическом пространстве?
(b) Найдите проекцию вектора $q = (0, 0, 1, 1)$ на найденное 2-мерное семантическое подпространство.
(c) Если считать, что вектор q — это поисковый запрос, то в каком порядке стоит ранжировать документы в поисковой выдаче?