Семинарское задание №1
LSI

Мосиенко Константин Викторович

2018

Вам дана коллекция документов: txt.tar.gz. Воспользовавшись моделью LSI, необходимо каждый документ представить вектором в пространстве скрытых переменных: $d_{LSI} = d^T T S^{-1}$. Далее, для каждого документа необходимо вычислить его схожесть с остальными документами в соответствии с моделью. Полученный результат необходимо сохранить в виде матрицы в текстовый файл в формате tsv (убедитесь, что я смогу прочитать его с помощью pandas). В полученной матрице на позиции i, j находится схожесть документов i и j. Перед посылкой убедитесь: матрица симметричная, на диагонали стоят единицы, все числа по модулю не превосходят 1.

Также необходимо составить отчёт, в котором в удобочитаемом виде для каждого документа выписать 5 самых ближайших по выбранной метрике.