Семинарское задание №4
Latent Semantic Clustering

Мосиенко Константин Викторович

2017

Необходимо выполнить следующие шаги:

- Построить матрицу *терм-документ* $A = \{a_{ij} = tf_{ij}\}$ как указано в конспекте лекции.
- Отнормировать её следующим образом:

$$\hat{a}_{ij} = g_i \log(a_{ij} + 1), \quad g_i = \sum_{j=1}^{|D|} \frac{p_{ij} \log p_{ij}}{\log |D|}, \quad p_{ij} = \frac{tf_{ij}}{\sum_{k=1}^{|D|} tf_{ik}}$$

  Обратите внимание, что нулевые элементы стоит оставить нулевыми без попыток нормировки.
- Выполнить SVD над полученной матрицей: $\hat{A} = TSD^T$.
- Перебрать r от 50 до 1000 и для каждого r:
  - Вычислить норму Фробениуса $||\hat{A} - \hat{A}_r||_F$, где $\hat{A}_r = T_r S_r D_r^T$. Такое приближение оставляет только векторы, соответствующие r максимальным сингулярным числам (либо равнозначно занулению остальных сингулярных чисел). Подумайте, как вычислить такую норму, не прибегая к умножению матриц, воспользовавшись свойством $||B||_F^2 = \sum_{i=1}^{n} s_i^2$, где s_i - сингулярные числа матрицы B.
- Постройте график зависимости $||\hat{A} - \hat{A}_r||_F$ от r. Обратите внимание на то, что должно быть понятно, что происходит на графике: если зависимость слишком «экспоненциальная», то необходимо прологарифмировать одну или обе оси.
- Зафиксируйте на выбор некоторое значение r, обоснуйте выбор. Переведите все документы в пространство размерности r по формуле:

$$d_{LSI} = d^T T_r S_r^{-1}$$

- Кластеризуйте полученные векторы на k кластеров алгоритмом k-means, выбрав в качестве функции расстояния косинусную меру[1]. Конкретное значение k необходимо подобрать эмпирически «на глаз», попробовав различные значения и обосновав выбор.

Необходимо прислать мне код и отчёт, причём, код я буду только читать, но не запускать, поэтому основной упор надо сделать на отчёт. В нём должен быть график, обоснование выбора r, обоснование выбора k и для каждого кластера имена файлов, попавших в кластер.

Коллекцию документов можно взять со страницы курса: txt.tar.gz.

[1] http://stackoverflow.com/questions/5529625/is-it-possible-to-specify-your-own-distance-function-using-scikit-learn-k-means