Ludwig-Maximilians-Universität München Munich, 12.01.2017
Institut für Informatik
Prof. Dr. Matthias Schubert
Daniyal Kazempour

Big Data Management and Analytics

WS 2016/17

Tutorial 10: Text Processing & High Dimensionality Data

Assignment 10-1 *Finding similar items*

Suppose that the universal set is given by $\{1, \ldots, 10\}$. Construct minhash signatures for the following sets:

(a) $S_1 = \{3, 6, 9\}$

(b) $S_2 = \{2, 4, 6, 8\}$

(c) $S_3 = \{2, 3, 4\}$

1. Construct the signatures for the sets using the following list of permutations:
   - (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
   - (10, 8, 6, 4, 2, 9, 7, 5, 3, 1)
   - (4, 7, 2, 9, 1, 5, 3, 10, 6, 8)
2. Suppose that instead of using particular permutations to construct signatures for the threse sets, we use hash functions. The three hash functions we use are:
   - $h_1(x) = x \mod 10$
   - $h_2(x) = (2x + 1) \mod 10$
   - $h_3(x) = (3x + 2) \mod 10$
3. How does the estimated Jaccard similarity, derived from $(1.)$ and $(2.)$ compare with the true Jaccard similarity of the original data? How to reduce deviations in the approximated Jaccard similarities?

Assignment 10-2 *PCA - General Questions*

a) Please describe what a PCA aims for and under what circumstances it is most helpful.

b) Which possibly negativ consequences might arise when applying PCA to a dataset of unknown structure?

Assignment 10-3 *PCA*

Consider the $X \in \mathbb{R}^{M \times N}$ matrix containing six data points $X_i \in \mathbb{R}^2$.

$$X = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 5 & 6 \\ 6 & 6 \\ 7 & 6 \end{pmatrix}$$

Conduct a PCA on the given data,i.e. project the data onto a one-dimensional space. Please state the eigenvectors, eigenvalues, covariance matrix and visualize the data before and after the PCA.