Semantic Computing

Tutorial 5

Summer Semester 2018

Exercise 1

Data cleaning and analysis: Today we will work with the real-world data set of company e-mails called Enron (`http://www.cs.cmu.edu/~enron/`). The Enron Corporation, an energy, commodity and service corporation, went bankrupt in 2001 as a result of fraudulent business practices. In total 0.5 million e-mails send by Enron executives between 2000 and 2002 were recorded and published along with financial information. The goal of processing this dataset with machine learning algorithms is to detect emails of interest which might be involved in fraudulent activities.

- Download the code template from `https://github.com/dgromann/SemanticComputing`
- Iteratively load all the .txt files of the provided ecron folder, read their lines, and extract the following information from the contained e-mails: from (e-mail address of sender), to (e-mail address of receiver), and body (main text body of the e-mail); remove digits in the process
- Convert the raw text (body) to a matrix of TF-IDF features (using TfidfVectorizer) while removing stop-words
- Plot the result using the provided code
- Use the TF-IDF matrix to obtain the 20 top keywords

Exercise 2

Clustering:

- Apply k-means to the produced TF-IDF matrix using a cluster size of 3
- Plot the resulting clusters (template provided)
- Print the top 25 words per cluster and try to guess what the topic of individual clusters might be
- Try again with two larger cluster sizes and compare the result based on the top words for each cluster