Week 6 - Practice

Text vectorization

The goal is to practice with pre-trained word embeddings derived with

1. GloVe (https://nlp.stanford.edu/projects/glove/)
2. word2vec (https://code.google.com/archive/p/word2vec/)

The embeddings should be downloaded from the links above and can be loaded with the gensim library in Python. Choose the embeddings having the smallest dimensionality.

For each embedding type you will perform the following tasks:

Tasks:

I. Visualize a number of 20 words (at your choice) in a 3-dimensional space by
   1. choosing randomly the dimensions
   2. by using PCA and tSNE to extract 2- or 3-dimensional embedding

II. Compute the cosine similarity for 3 pairs of related words (synonyms/antonyms/same semantic field) and 2 pairs of unrelated words

III. Perform hierarchical clustering analysis for a set of ~25 words. Inspect the dendrogram.

Conclude on the ability of the 2 different embedding models to capture word similarity/relatedness.