Exercise Week 11

Sprachverarbeitung (VL + Ü)

Nils Reiter, nils.reiter@uni-koeln.de

June 20, 2023 (Summer term 2023)

This exercise is to be done on the server.

- The directory `/teaching/summer-2023/sprachverarbeitung/data` contains three corpora of different sources in three large plain text files. 'bt.txt' contains text from parliamentary debates from the German Bundestag, 'wp.txt' contains text from the German Wikipedia and 'st.txt' a corpus of steam reviews. Choose one of these as you like and find interesting. They have very different sizes.
- Train a word2vec model using the script `train-word2vec.py` in `/teaching/summer-2023/sprachverarbeitung/word2vec`. You'll need to change the filename (which is hardcoded in the script).
- Load and explore the module using the script 'use-word2vec.py' (again, change the filename within the script). Try to find a few good and a few bad examples. Run the script with 'python3 -i use-word2vec.py' to enter the Python interpreter. The function `model.wv.most_similar("universität")` will give you the ten most similar words to the word 'universität'. Beware: All tokens are lower-cased.
- Optional: Do all the above for a second corpus and compare the resulting similarities between word pairs.