Exercise Count Based Distributional Semantics

Natural Language Processing

Ulrich Heid, Christian Wartena, Johannes Schäfer

Summer Term 2022

Take the last model from the notebook `DistrSem1.ipynb` and download the small set of synonyms and non-synonym word pairs *STW German Synonyms* from `http://textmining.wp.hs-hannover.de/datasets.html`. Rank the words form this dataset according to their predicted similarity.

1. Evaluate the ranking using AUC
2. Investigate whether and how the result depends on the following parameters. Test at least 2 parameters, better 3 or all 4.
   a) Window size
   b) Corpus size (i.e. use only a part of the corpus)
   c) Lower and upper frequency bound of words to be included as context words.
   d) Number of dimensions

   You do not have to test all combinations. Just test each of the parameters independently while fixing the other parameters to a reasonable value. Note that you have to change frequency ranges for word inclusion if you reduce the corpus size!

Finally, if you are learning German, just for fun, you could try to rank the word pairs yourself (not looking at the labels) and compute the AUC of your personal ranking!