Exercises with word embeddings

We will experiment with word embeddings produced by

- word2vec https://code.google.com/archive/p/word2vec/
- GloVe http://nlp.stanford.edu/projects/glove/

you are encouraged to experiment with both, but one of them is sufficient for this set of exercises

1. Download the code, compile, make sure the application runs
2. Experiment with pre-trained vectors, checking whether the distances between words that you think as similar and different are reflected in the word vectors.

   Web pages of both applications provide some examples that you can use for inspiration.

   You can download pre-trained vectors from the software web pages, or other sources on the internet. A number of samall pre-trained models are also provided here.

   For vectors in text format, you can use any programming environment. For word2vec binary format, either you can use `distance` command that comes with word2vec, or packages like `gensim` may be helpful here.

   You can also make use of the data set from http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/ which includes human judgments. Check whether the human judgments correlate with the distances from the word vectors.

   Other data for measuring similarity of words can be found in word2vec source code as `questions-words.txt`, and at http://research.microsoft.com/en-us/um/people/gzweig/Pubs/myz_naacl13_test_set.tgz

   In general, you may want to compare synonyms/antonyms etc. It is also interesting to try to discover how they handle homonyms.
3. Perform PCA on a subset of word vectors that you expect to find interesting differences/similarities, and plot the first two components.
4. Train word vectors on a small corpus you build, either from a small text collection, or even sentences phrases you typed-in for testing.

   You are encouraged to try this on larger data sets later, also from different languages. But to save time (and your computer's battery) a toy setup is enough for getting familiar with the input/output formats and parameters of these models.