Tasks

1. Train Word2Vec on your selected corpus using Gensim in two different ways to obtain two different sets of word embeddings. The two ways could differ in terms of Skip Gram/CBOW, or how many epochs you use, or how much data you use for training, or using two different corpora, lowercased all words or not, lemmatized all words or not, or different vector lengths, or some other meaningful way. Save the two sets of learned word embedding vectors in plain text .txt files.

2. Come up with your own set of 20 words (each word should have its embedding) and use them in the visualization of the two sets of embeddings (you can use the posted A1_helper.py module).

3. Create your dataset of at least 10-word pairs and assign them similarity scores 0-1 based on your judgement (make sure you have a good range of scores; all words should have embeddings). Create your own tab-delimited input .txt file. Evaluate the two-word embeddings from Step 1 as well as the pre-trained Google News embeddings on this dataset using Pearson correlation coefficient.

4. Come up with any 5 words (they can be from Steps 2 or 3), and find their most similar five words using the three sets of word embeddings (two trained by you and the pre-trained Google News embeddings).