Language Technology I, Winter 2016-2017

Exercise 7: Word Vectors

You can earn up to 10 points on this exercise.
You may work as a group of up to 3 people, but please submit your own version.
You may use any programming language you wish.

Please email your solution as a single PDF file to `langtech1saarlandws1617@gmail.com` *by* **3:00 PM GMT+1, January 18, 2017**.

We will soon send you a private link to a large text data set via the `langtech1saarlandws1617@gmail.com` email address using the email addresses we've collected. Download the data and do the following:

1. The data consists of thousands of documents. Randomly select and extract enough documents to generate at least 2 million words. This is our corpus. Lowercase and tokenize them (you can use the script from Exercise 4 or any other method). Report the unigram frequencies of the top 50 words in a table. (2 points)

2. Arbitrarily choose 15-20 words for fruit (e.g. "apple", "pear", etc.) in the corpus. Give the unigram frequency table. Then arbitrarily choose 15-20 words for junk food (single word, avoid brand names if possible; e.g. "cookie", "popcorn", etc.) in the corpus. Give the unigram frequency table. (1 point)

3. Construct word vectors by any algorithm you like using any programming environment you prefer, so long as they're at least 100-dimensional. Present, aligned, 100 dimensions of one of your fruit words and one of your junk food words. (1 point)

4. Project the vector space down to two dimensions using any non-trivial projection you like (SVD, PCA, t-SNE, etc.). Then present the following, stating which projection you use:

   (a) A plot of the vectors representing the fruit words. (2 points)

   (b) A plot of the vectors representing the junk food words. (2 points)

   Make sure the point labels (i.e., the words) are easy to read.

5. Use K-means clustering (there are packages in Python and other languages that will do this for you) to cluster the fruit and junk food vectors *together*. Force it to make 5 clusters, with whichever other parameters you like. For each cluster, list the members. (2 points)