Exercise 11: Distributional Semantics

This exercise is fully optional. It will not be used to determine exam eligibility or course grades.

If you would like feedback, please email your solution to `claytong@coli.uni-saarland.de` *by* **July 21, 2016**.

Download some similarity judgements (between two words) here:
`http://alfonseca.org/eng/research/wordsim353.html`.

These scores are on a scale from 0 to 10. Here is a subset:

Word 1	**Word 2**	**Similarity**
tiger	tiger	10.00
tiger	jaguar	8.00
tiger	cat	7.35
tiger	carnivore	7.08
tiger	animal	7.00
tiger	mammal	6.85
tiger	fauna	5.62
tiger	organism	4.77
jaguar	cat	7.42
jaguar	car	7.27
jaguar	stock	0.92

Task 1

Use the data provided at
`http://www.coli.uni-saarland.de/~claytong/data/distributional_semantics.zip`
to build a word-context matrix. Use raw counts or tf-idf values for each context word. Vary the size of the context windows.

Compute the similarity between words by trying different metrics measuring the distance of the vectors representing the words. You can use e.g. a cosine or the Euclidian distance. Depending on what you pick, you might want to scale the result in order to make sure that the range is between 0 and 10.

How does your model compare to the human annotation? You can, for example, make a scatter plot to visualize the results. You can also calculate the Pearson product moment correlation coefficient or the Spearman's rank correlation coefficient.