Homework 5
Semantic Similarity between Words

Compare methods for similarity measurement using NLTK (http://nltk.org).
Data: WordSimilarity-353 corpus http://alfonseca.org/eng/research/wordsim353.html
It is splitted into two datasets: similarity and relatedness.

1. Write program in Python to measure similarity for each pair of concepts with the following methods:

- Path
- Wu & Palmer
- Resnik
- Jiang & Conrath
- Lin

The methods take concepts as input, so you need to point POS and word sense number. All words in the corpus are nouns. For word sense disambiguation you can take sense numbers from WordNet, measure similarity for each pair of senses and take maximum.

2. Measure how good results are through correlation between gold standard and results of the method. Hint: Excel can calculate correlation.

3. Measure correlations between different methods.

4. Study results and write a half-page discussion of the results:

- which method shows better results,
- how the correlation with gold standard changes for different dataset,
- what are the tricky examples for different methods, why they are tricky
- what is the correlation between methods, is there anything unexpected

Submission:
1. You code and ReadMe file, explaining how
2. Tables of results for two data sets:

Pair of Concepts	Gold Standard	Path	…	Lin
bank river	0.4	0.5		0.4
…				
Correlation	1	0.6		0.9

3. Tables of correlation between methods for two datasets:

	Path	Wu & Palmer	…	Jiang & Conrath
Wu & Palmer	#	#	#	#
Resnik		#	#	#
…			#	#
Lin				#

4. Discussion

Some basics of python: http://code.google.com/edu/languages/google-python-class/