Computational Semantics 2019-2020
First graded assignment

In this graded assignment, you will explore word similarity with distributional semantics. Submit a compressed file containing a pdf document with your answers as well as the Python scripts. **Important:** you need to submit your own answers to the exercises. The code can be the same for the same group; the answers to the questions must be written individually (of course you discuss the data with each other).

Exercise 1. In this exercise, you will implement a classifier by hand, and evaluate its performance on a test set. The classifier needs to decide when two words are highly similar, that is, when they are synonyms or near synonyms. Your classifier will decide that two words are highly similar if their cosine similarity is larger than 0.5. See folder python_basic for data and instructions.

Quantitative evaluation : What do you conclude on the basis of the Precision, Recall, and Accuracy the model obtained in terms of its ability to distinguish highly similar from dissimilar words? More precisely:

- Did the method achieve a high Precision? Why / why not? (Here and below, state your hypotheses for the reasons)
- Did the method achieve a high Recall? Why / why not?
- Did the method achieve a high Accuracy? Why / why not?

Qualitative evaluation (use examples in your answers!):

1. Examine the output for specific word pairs, i.e., whether the classifier was correct ("c" vs. "x"), the similarity estimate produced by the method (i.e., the cosine similarity), the target (gold label / reference / ground truth), and the part-of-speech. What do you observe?

- On which kinds of word pairs did the model fail or succeed, respectively?
- What patterns do you observe with respect to the semantic relationship between the words in correct vs. wrong word pairs?

2. Given the results and your observations, what do you conclude with respect to the effectiveness of a method based on the threshold of 0.5? How could this method be improved?

the use of word embeddings and the cosine similarity to distinguish close-synonyms from unsimilar words? Relate this to how the embeddings are obtained.

3. Compute the results you obtain for each PoS (part-of-speech, i.e., A(djectives), N(ouns), V(erbs)) separately. What differences do you observe? Also inspect the individual similarity estimations.

Exercise 2. In the previous exercise, we have modeled word similarity as a binary phenomenon (yes/no). However, similarity is a graded notion. In the second part of the assignment, you are going to work with graded similarity judgments. For these, recall/precision/accuracy do not work (those are for classification tasks). You will use correlation to evaluate the model. See folder python_advanced for data and instructions.

Quantitative evaluation :

- What do you conclude on the basis of the obtained correlation coefficient in terms of the ability of the model to account for human behaviour on judging word similarity?
- On the official SimLex-999 website (https://fh295.github.io/simlex.html) you can find a list of correlation coefficients for several models (see "state-of-the-art").
  - How does the skipgram-GoogleNews model compare against the state of the art listed there?
  - Address in particular the comparison to the first model listed there (Skipgram (Word2Vec) model trained on 1bn words of Wikipedia), as well as the human upper bound. How do you explain the different correlation coefficients?

Qualitative evaluation (use examples in your answers!):
Do either of the following two options:

- Examine the individual similarity estimations for the word pairs produced by the model (i.e., the cosine similarities) by comparing them to the human judgements. What do you observe? That is, on which word pairs did the model fail or succeed, respectively? Can you observe some pattern with respect to the category/specificity/any-other of the two words or the semantic relation between them?
- Compute the correlation between model predictions and human judgements for each PoS (part-of-speech, i.e., A(djectives), N(ouns), V(erbs)) separately. How do you explain your observation when comparing the results? Also inspect the individual similarity estimations.

Exercise 3. Based on the results of the two exercises above, what are your general conclusions about modeling word similarity with distributional semantics?