Week 3 & 4 – Word2Vec

Lyubomira Dimitrova

February 2018

1 Tool

I used the gensim implementation of word2vec, and scikit-learn for the PCA transformation.

2 Plots

2.1 Tokenized corpus

- `all_words.png` Using the tokenization provided by the TreeTagger tool, I first plotted all words in the list given to us.
- `words.png` Secondly, I plotted only a number of words, namely: automobile, car, cord, food, fruit, furnace, midday, noon, stove, and string.

2.2 Tokenized & POS

- `all_words_pos.png` Plot of all given words, with '_POStag' added to every word. As mentioned in the previous exercise, the Bulgarian tagset is rather large, so again, only the first character of the POS-tag was used.
- `words_pos.png` Again, only the words automobile, car, cord, food, fruit, furnace, midday, noon, stove, and string, were plotted.

2.3 Tokenized & lemmata

- `all_words_lemmata.png` Plot of all given words, with '_lemma' added to every word. Words with the lemma '<unknown>' were not included in the sentences passed to gensim.
- `words_lemmata.png` Again, only the words automobile, car, cord, food, fruit, furnace, midday, noon, stove, and string, were plotted.

2.4 Only lemmata

Linguistically speaking, there is very little ambiguity in Bulgarian regarding the general POS-tag of the word (compared to English). Furthermore, the word forms are much more diverse (different verb forms for pretty much every tense-aspect-number-person combination; different noun forms since the article is part of the word), which makes using a simple tokenized corpus just as unsatisfactory. That's why I decided building sentences using only the lemmata might mean better word embeddings. I can't really tell whether the representation are better, but it was fun to try.

- `lemmata_only.png` All given words were plotted in a model consisting of sentences of lemmata. Again, the '<unknown>' lemmata were skipped in the sentence building.