P13: Word Similarity

Project owner
Arvi Tavast

· Andreas Pung, Alfred Saidlo, Karl Vaba, Merlin Kasesalu

16.12.2020
Problem

• Analyze a huge amount of written Estonian text
• For the most common words, find the most similar words
• Provide this data to Arvi Tavast
Why?

- This data can be used for inclusion in dictionaries
- Coming up with similar words manually can be tedious
How?

• Imagine two points

(1,2)  (5,5)

• How close are they?
• Imagine two points

\[(1,2)\] \quad \text{(1,2)}

\[(5,5)\] \quad \text{(5,5)}

Obviously, we find the euclidean distance

\[\sqrt{(5-1)^2 + (5-2)^2} = \sqrt{16+9} = 5\]
How?

- If we have a method for transforming words into tuples of numbers (vectors), we can find their similarity by distance.
- Word2vec does exactly that.

Source: https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f
The results

• We ran word2vec models with different parameters (dimensionality, epochs, cbow vs skip-gram)

• For each model we put together a .csv file where we publish the 50 most similar words for 200000 most common words.
The results

• For example, one of our models says that the most similar words for “tark” are “arukas” and “intelligentne”

• Then again for the word “abitu” (helpless) one model thinks a similar word is “{politician name here}“
Evaluation

- Obviously with multiple models we need a way to compare them
- This is not straightforward for unsupervised learning
Evaluation

- Arvi provided us with a list of words with similar words attached
- We compared how “similar” each model is to this standard
Evaluation

- For that let's look at the standard:
  It says the most similar word to “lahe” is “tore”

- Now we can check if our model also thinks the word “tore” is similar to “lahe” (if it appears in the top N most similar words)

- We can repeat this for all words in the standard and get a success rate
As we can see, models trained using CBOW is more similar to the standard since more searches are successful.
Shortcomings

• We didn't manage to parse the largest files so our models had to be trained on a subset of the data
• While we were initially optimistic about training lots of different models, we limited ourselves to word2vec due to time constraints
•Thanks for listening!