Technische Universität Dortmund
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2024/25
12/10/2024

Exercise sheet 9

Natural Language Processing

Hand-in (voluntarily): 12/16/2024 until 10:00 a.m. via Moodle
Please submit a `.py`, `.ipynb`, `.R` or `.rmd` file!

Task 1

In moodle you will find the file `trek.json` and `characters.csv`. The first file contains transcripts of 5 Star Trek tv shows, separated into the individual episodes. The second file contains the name of characters, the tv show they appear in and their respective rank or role in the show.
In this exercise, we will investigate, how well Word2Vec models the relationships between characters in the Star Trek franchise and how different window sizes can change the relationships that are being mapped by the model.
Please note: The names "obrien" and "tpol" originally contained an apostrophe. For Word2Vec to recognize the characters correctly, you have to remove each apostrophe with an **empty** string!

Task 2

Preprocess the texts so that they are fit for an analysis. Argue the use the preprocessing steps you take for the given analysis.

Task 3

Train a Word2Vec model on all transcripts with a window size of two (i.e. two words in each direction) and a vector dimension of 300. Train another model with the same parameters and only change the window size to ten.

Task 4

We will now use the characters from `characters.csv` and see, how well Word2Vec differentiates the different tv shows. Calculate the cosine similarities of all possible character pairs for both models. Then, calculate the average similarity between all character pairs within each tv show and the average pairwise similarity to all characters of a different tv show. In the end you should have a 5x5 matrix, containing average pairwise similarities between and within all 5 tv shows.
What do you notice? Which model does differentiate the characters of a tv show better from other tv shows?

Task 5

Repeat task for for the `role`-column, which contains information of the role the characters represent in the tv show. Again, compare the inner vs. outer similarities within these groups. Which model works better for this task?

Recommended packages & functions

R: `word2vec::word2vec()`, `word2vec::as.matrix.word2vec()`, `proxy::dist()`

Python: `gensim.corpora.Dictionary`, `gensim.corpora.Dictionary.doc2bow()`, `gensim.models.Word2Vec`, `scipy.spatial.distance.cdist()`