LINMA2472 – Algorithms in Data Science

HW 2 – module "Embeddings"

Please prepare a written report with appropriate figures based on the results from the assignments.

Assignment 1: node embedding and clustering

Take the character co-occurrence network from the previous homework. If you haven't done it already, then add weights to the edges that represent how many times the co-occurrence happened. Thus the co-occurrence network is a sort of an embedding of the story into the metric space of a graph. The goal is to compare this embedding with the embedding based on text.

Use the partition into communities as computed by the Louvain method as a benchmark.

1) Train a Word2Vec model on the corpus of your book. Cluster the vectors of characters using k-means clustering algorithm using the same number of clusters as output by the Louvain algorithm. Compare the two clusterings using **Jaccard similarity**. Find the parameters of W2V that best resemble the Louvain clusters and those that produce a completely different clustering. Which clustering, in your opinion, better reflects the real communities from the story?

2) Visualise the embeddings and clusters in 2D using a dimensionality reduction algorithm of your choice.

Assignment 2: text embedding and classification

Attached you find two real datasets of comments from Reddit – one from the subreddit /r/JoeBiden of supporters of Joe Biden and one from the subreddit /r/The_Donald of supporters of Donald Trump. Both sets were collected in March, 2020. Imagine your goal is to learn if those two groups speak the same type of language.

1) Perform an embedding of the comments using BERT and Doc2Vec (Word2Vec for sentences). Remember to first do the appropriate text preprocessing.

2) Train a classifier of your choice to separate the comments of one subreddit from another. Perform training on the provided training set. Test your classifier on the

provided test set of comments. Achieve and report the best **accuracy** of your classifier on training and test sets. Best results will be announced on the lecture.

Report guidelines:

- We highly recommend to google about specific terms if they are not clear.
- Write in a concise and structured manner. No long sentences, only relevant information.
- You may present your data and the preprocessing steps, but remember that this isn't the main goal of the report
- Any numerical results that can be presented in a table should be presented so.
- Round numbers up to 3rd digit, unless it's really necessary. Don't copy-paste 10 digit floats.
- Plots must be easy-to-read. Must include labels on axes, legend if more than one curve is shown, title or a caption, explaining what the plot is about.