Maser's thesis outline

Jonas Folkvord Triki

February 13, 2020

1 Introduction

In this section, I am first going to introduce word embeddings and some history behind it (why it is used, where it started, etc.). Following, I will give some classical examples of its usage and mention where it is used as of today. I will also give a quick introduction to all the methods used in this thesis.

2 Word embeddings

In this section, I will first dig deeper into the word2vec method for computing word embeddings. I will explore its different architectures and will implement a simple model to create word embeddings of English texts.

I might also take a look at other methods other than word2vec for computing such word embeddings.

I will also use some pre-trained networks (already existing weights) to test on my set of texts, due to time and computational constraints. Here I will use texts from other languages, such as Norwegian.

3 Analysis of embeddings

In this section, I will analyze both my results and state of the art results from the previous section. I will use unsupervised methods such as dimensionality reduction (PCA, UMAP, TSNE, etc.) and clustering (K-means, Agglomerative, Spectral, HDBSCAN, etc.). I will also visualize my results and discuss my results (word embeddings of English vs. Norwegian).

4 Topological analysis

In this section, I will perform some hypothesis testing using topological methods (Vietoris-Rips/Čech complexes, persistence diagrams, etc.). For example, I would like to see if words like "north, south, west, east" form a circle. I will also look into unsupervised hypothesis testing and look for topological properties of word embeddings in general. I will also like to see if one could improve training of neural networks by using topological properties found from word embeddings.