Comparative Analysis of Models with Manually Created Knowledge Graphs

Group 9: Mayank Tamakuwala, Nithin Bhat, Agasti Mhatre

1. Description

* Find the best POS algorithm for tagging news articles. We want to test this by generating a knowledge graph in the form of (Subject, Predicate, Object), and then using this graph to train/test each algorithm.

1. Dataset

* <https://www.globenewswire.com/newsroom>. We are going to manually go through 300 news articles and label them in the form (Subject, Predicate, Object). The dataset contains 207 different categories worth of articles, so we will likely pick out 300 news articles from different categories to diversify our samples.

1. Methodology and Expected Results:

* The major outside tools that we plan to use are Scikit-Learn, spaCy (for NLP), Transformers, Matplotlib, and Seaborn for visualization. We are deciding on the models that can generate the knowledge graphs for use. Those models will most probably include **Viterbi POS Algorithm** and BERT model variants like ALBERT or RoBERTa to do prediction and compare their training time as well as performance.
* The main results we want to have, are the labels to the words showing (Subject, Predicate, Object) pairs for an inputted data/dataset.

1. Timeline:

* Week One and Week Two: Finish manually labelling 300 news articles to create a ground truth and train/testing datasets for the model
* Week Three: Train and test different models and finetune/optimize their performance
* Week Four: Create and compare the performance visualizations for the different models

1. Responsibilities:

* Our team has three members: Mayank Tamakuwala, Nithin Bhat, and Agasti Mhatre. Each member is going to handle around 100 news articles to create the knowledge graph dataset manually. Then, all of us are going to handle the creation of at least one algorithm that generates a knowledge graph and then the evaluation of those models.