Part-of-Speech (POS) Tagging Representation with Network Graphs

1. Introduction

1.1. Problem Statement

The goal of this project is to develop a pipeline that performs Part-of-Speech (POS) tagging on text articles and represents the results as a network graph. In this graph Word nodes represent individual words in the text, with attributes that capture the word itself.

POS tag nodes represent the grammatical categories (e.g., NOUN, VERB, ADJ) that classify the words, The relationships between words and their corresponding POS tags are represented as edges connecting word nodes to their respective POS nodes

1.2. Objectives

- Develop a POS tagging pipeline.
- Represent POS tagging results using a network graph.
- Optimize the pipeline for time efficiency when processing articles.
- Provide clear and reproducible documentation.

2. Data Processing

The goal is to preprocess the text data efficiently so that the article's words can be represented as nodes in a network graph, this process involves cleaning by removing punctuation, and extra spaces and normalize articles

3. POS Extraction

This step in volves experiment Traditional methods vs deep learning methods, Traditional methods are faster and suitable for real-time graph construction, Deep learning models provide better accuracy but at the cost of time and computational resources.

4. Tools Used

- spaCy: Tokenization and POS tagging.

- NetworkX: Graph construction and visualization.
- Python Libraries: Pandas, numpy, matplotlib

5. External Resources

- spaCy documentation
- NetworkX tutorials.
- Research papers on POS tagging and graph representation.

6. Project Learnings

6.1.Challenges

- Processing long articles within acceptable time constraints
- Balancing accuracy and efficiency when comparing traditional and deep learning methods