Project Documentation

Tejaswi Choppa
Tanishq Quraishi
Sarah Razavi Sharifian

July 31, 2023

1 Our Story

The goal of this project is to optimize the word sense disambiguation task. Word senses are easier to capture when visualised as a graph. We obtained the SemEval 2015 data online which is in a XML format. Corresponding to each lemma, we have the context in which the word occurs, its WordNet id and the POS tag. We wrote a DTD file to validate our XML data file along with XSD schema. We extracted the data into CSV format. Further, using the Neo4j we prepared the data into a graph database. We further queried our graph database and extracted sub-graphs.

2 Technology

The following technologies have been used for our project:

- XML, XSD, DTD
- Neo4j, Cypher, Bloom
- Python

2.1 Neo4j Guide

1. **Nodes:** [**Lemma** (id, text)] [**Sentence** (id, text)] [**Wordnet_id** (id)]
2. **Relationship Types**
   - FOUND_IN
   - HAS_SENSE (POS, index)

3 Extension

Our project extension involves a detailed graph visualization using Neo4j Bloom. Bloom is accessible through the Neo4j desktop app, and enables users to explore the graph database through a well-designed user interface. The motivation behind Bloom is to make graph databases more accessible, circumventing the use of Cyper, the querying language.