**Natural Language Processing – Project 3**

Comprehensive NLP Analysis of Semantic Scholar Articles

This project aimed to conduct a multi-faceted Natural Language Processing (NLP) analysis on a corpus of articles related to NLP, collected from Semantic Scholar. The goal was to extract key insights, identify prevalent topics, and compare the effectiveness of various NLP techniques in understanding the content and trends in NLP research.

Data:

The source was taken from Semantic Scholar (<https://www.semanticscholar.org/>)

We focus on articles related to Natural Language Processing

Methodology

The project employed a diverse range of NLP techniques:

1. Preprocessing:

- Tokenization

- Lemmatization

- Stop words removal

2. Feature Extraction and Representation:

- TF-IDF (Term Frequency-Inverse Document Frequency)

- Word2Vec

- Autoencoder

3. Named Entity Recognition (NER)

4. Exploratory Data Analysis (EDA)

5. Text Summarization:

- Applied to each article's abstract

6. Advanced Language Model Application:

- Utilized GPT model for question answering on the most common subjects

Key Analyses

1. Identification of key terms and concepts in NLP research

2. Exploration of semantic relationships between NLP-related terms

3. Discovery of latent topics and themes in the corpus

4. Recognition of named entities specific to NLP domain

5. Summarization of research abstracts for quick insights

6. In-depth exploration of common subjects using GPT-based question answering

The results:

\*TF-IDF:

The TF-IDF analysis reveals that the corpus is strongly focused on core NLP concepts and methodologies. It captures both the fundamental aspects of language processing and the diverse applications of NLP techniques. The results suggest a field that is deeply rooted in linguistic and computational methods while continuously expanding into new domains and applications.

\*word2vec:  
The image shows a 2D projection of Word2Vec embeddings for the top 20 common words in the NLP corpus. The spatial relationships between words indicate semantic similarity or relatedness in the vector space.

\* Autoencoder:

The bar chart shows the top 20 most important words as determined by the Autoencoder model, with importance measured by reconstruction error.

\*NER:  
 The bar chart displays the frequency of various entity types identified in the NLP articles corpus.  
 The results align well with previous findings from TF-IDF, Word2Vec, and Autoencoder analyses, particularly in emphasizing the data-driven and quantitative aspects of NLP research.

\*EDA:  
 The bar chart displays the number of NLP articles published each year from 2005 to 2024. our EDA provides valuable insights into the evolution of NLP research over time. It clearly shows the field's growth from a relatively niche area to a major focus of computer science and linguistics research. The data reflects the impact of technological advancements, particularly in machine learning and deep learning, on NLP research output.

**Issue with the Existing NER Algorithm**: Common Named Entity Recognition (NER) algorithms, such as CRF, LSTM, or even BERT, primarily rely on the immediate context within the text to identify entities. In some cases, entity recognition can be challenging when the immediate context is not clear enough, especially in complex or lengthy texts. For example, identifying an entity might be difficult if it first appears with little context but becomes clearer later in the text.

**Proposed Change - Incorporation of Contextual Features**: The proposed change to the NER algorithm is to incorporate broader contextual features into the model by utilizing information from previous or additional text sections. Instead of focusing only on the immediate context, a broader context layer can be added to the model. This can be achieved using an attention mechanism that can focus on parts of the text where an entity is more likely to appear with a clear context, such as in previous sentences.

**Rationale**: Incorporating broader contextual features can improve the accuracy of NER by enabling the model to better understand the full meaning of entities in texts. For example, if a particular entity is briefly mentioned at the beginning of a document and then discussed in more detail later, the model can use the broader context to identify that entity more accurately.