NLP Assignment: Fake News Analysis using NLP Techniques

Section 0: Setup

- Import necessary libraries including nltk, sklearn, gensim, pandas, and matplotlib.
- Download NLTK data for tokenization, tagging, and chunking.

Section 1: Load Dataset

- Dataset used: Fake and True news articles from Kaggle.
- Concatenate and shuffle the dataset, combining title and text fields.

Section 2: Tokenization & Normalization

- Sentence Tokenization using sent_tokenize().
- Word Tokenization using word_tokenize().
- Normalize text: lowercase, remove punctuation, and filter stopwords.

Section 3: Stemming and Lemmatization

- Use PorterStemmer and WordNetLemmatizer.
- Compare results of stemmed vs lemmatized outputs.

Section 4: POS Tagging and Chunking

- Use pos_tag() to tag parts of speech.
- Apply RegexpParser for extracting noun phrases.

Section 5: Named Entity Recognition (NER)

- Use spaCy to extract named entities such as PERSON, ORG, DATE, etc.

Section 6: WordNet

- Explore synonyms and definitions using WordNet.
- Example: synonyms of "news".

Section 7: Feature Extraction

- Apply TF-IDF Vectorizer to extract top textual features from the dataset.

Section 8: Document Similarity

- Compute cosine similarity between TF-IDF vectors of sample news articles.

Section 9: Word Frequency Plot

- Visualize top 20 words in fake news articles using CountVectorizer and matplotlib.