Lab 2 Report - NLTK

Text Preprocessing and NLP Analysis

1. Objective

To apply fundamental Natural Language Processing (NLP) techniques on the constructed news corpus using the NLTK library and perform basic sentiment analysis and classification.

---

2. Dataset

The dataset used was `news_corpus.jsonl`, created in Lab 1 from the Fake and Real News dataset.
The corpus contains labeled news articles (0 = Fake, 1 = Real).

---

3. Methodology

The following NLP techniques were implemented:
a) Tokenization

- Word-level and sentence-level tokenization using NLTK.
  b) Stopword Removal
- Removed common English stopwords to reduce noise.
  c) Stemming
- Applied Porter Stemmer to reduce words to root form.
  d) Lemmatization
- Used WordNet Lemmatizer to obtain meaningful base forms.
  e) POS Tagging
- Assigned grammatical tags (noun, verb, adjective, etc.) to tokens.
  f) Named Entity Recognition (NER)
- Identified entities such as persons, locations, and organizations.
  g) Sentiment Analysis (VADER)
- Computed compound sentiment scores for each article.
  h) Naive Bayes Classification
- Built a probabilistic classifier using word frequency features.

- Evaluated model performance using accuracy.

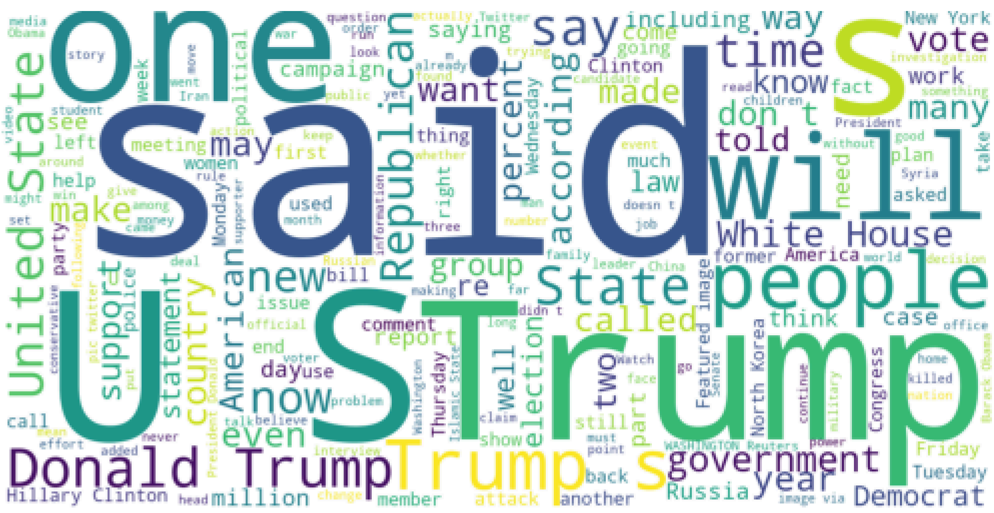
  i) Word Cloud Visualization

- Generated word clouds to visualize frequent terms in the corpus.

---

4. Results

- Successfully implemented a complete NLP preprocessing pipeline.
- Sentiment polarity scores were generated for all articles.
- Naive Bayes classifier achieved reasonable performance.
- Word cloud visualizations highlighted common vocabulary patterns.



---

5. Conclusion

This lab demonstrated practical implementation of core NLP preprocessing techniques and basic text classification using NLTK. The processed data and extracted features provide a strong foundation for advanced machine learning models in the next lab.

6. Git Link

(Lab 2 Git Link:-)[https://github.com/MYTH-il/NLP/blob/main/lab2_nltk.ipynb]