**Final Report: Natural Language Processing Coursework**

**Table of Contents**

1. **Summary**
2. **Introduction**
3. **Methodology**
   * 3.1 Preprocessing
   * 3.2 Task 1: Introduction to NLP and Data Collection
   * 3.3 Task 2: Text Preprocessing and Tokenization
   * 3.4 Task 3: POS Tagging and Named Entity Recognition
   * 3.5 Task 4: Sentiment Analysis
   * 3.6 Task 5: Topic Modeling
   * 3.7 Task 6: Stylometric Analysis
   * 3.8 Task 7: Document Clustering
   * 3.9 Task 8: Dependency Parsing
4. **Results and Discussion**
   * 4.1 Key Insights
   * 4.2 Challenges and Limitations
5. **Conclusion**
6. **References**
7. **Appendix**

**1. Summary**

This report details the execution of an extensive Natural Language Processing (NLP) project. Spanning from foundational tasks like data collection and preprocessing to advanced analyses such as document clustering and dependency parsing, each task is addressed systematically. Using real-world datasets of Reddit posts and comments during the COVID-19 pandemic, the study applies techniques such as POS tagging, Named Entity Recognition (NER), sentiment analysis, and stylometric evaluation. The results highlight key thematic and emotional patterns, emphasizing the relevance of NLP in understanding large-scale textual data.

**2. Introduction**

Natural Language Processing (NLP) bridges the gap between human communication and computational analysis. This coursework employs NLP to analyze text data from Reddit, a platform known for diverse discussions. The objective is to extract insights from textual content during the COVID-19 pandemic, focusing on sentiment, topic modeling, and more. This report provides a structured narrative, detailing the methodologies, results, and interpretations.

**3. Methodology**

**3.1 Preprocessing**

Preprocessing steps included:

* **Lowercasing:** Ensures uniformity.
* **Special character removal:** Reduces noise.
* **Stopword removal:** Focuses on meaningful words.
* **Tokenization and Lemmatization:** Standardizes text.
* **N-grams:** Captures contextual patterns.

**3.2 Task 1: Introduction to NLP and Data Collection**

* **Introduction to NLP:** Explains NLP's importance and applications, including chatbots, sentiment analysis, and translation systems.
* **Data Collection:** Utilized Reddit datasets containing posts and comments. Files processed included:
  + coronavirus\_reddit\_raw\_comments.csv
  + coronavirus\_reddit\_posts.csv

**3.3 Task 2: Text Preprocessing and Tokenization**

* Applied preprocessing to clean the text.
* Tokenized text data and generated n-grams for deeper context analysis.
* Outputs included processed datasets saved for subsequent tasks.

**3.4 Task 3: POS Tagging and Named Entity Recognition (NER)**

* **POS Tagging:** Used SpaCy for batch processing, identifying frequent parts of speech like nouns and verbs.
* **NER:** Extracted entities such as PERSON, DATE, and ORGANIZATION, visualizing entity relationships with heatmaps.
* **Key Visuals:**
  + Bar charts of common POS tags.
  + Heatmap of entity relationships.

**3.5 Task 4: Sentiment Analysis**

* Applied VADER and TextBlob for sentiment classification.
* Analyzed sentiment distributions across posts and comments.
* **Key Visuals:**
  + Sentiment distribution bar charts.
  + Correlation heatmaps linking sentiment with stylometric features.

**3.6 Task 5: Topic Modeling**

* Utilized LDA to extract dominant topics from posts and comments.
* Topics included themes like "Health Policy" and "Vaccination Progress."
* **Key Visuals:**
  + Word clouds representing key terms per topic.

**3.7 Task 6: Stylometric Analysis**

* Extracted features like vocabulary richness and readability.
* Conducted PCA and K-Means clustering for stylistic grouping.
* **Key Visuals:**
  + Histograms of stylometric features.
  + PCA scatterplots of clustered texts.

**3.8 Task 7: Document Clustering**

* Used Doc2Vec for document embeddings.
* Applied K-Means to cluster semantically similar documents.
* **Key Visuals:**
  + PCA scatterplots of clustered document vectors.

**3.9 Task 8: Dependency Parsing**

* Analyzed sentence structures using SpaCy.
* Extracted advanced structures like noun chunks and clauses.
* **Key Visuals:**
  + Dependency tree diagrams.

**4. Results and Discussion**

**4.1 Key Insights**

1. **Sentiment Analysis:**
   * Comments exhibited more emotional variability than posts.
   * Negative sentiments correlated with longer texts, reflecting detailed complaints.
2. **Topic Modeling:**
   * Captured pandemic-related themes, aiding policymakers in understanding public concerns.
3. **Stylometric Analysis:**
   * Vocabulary richness distinguished emotional texts from factual posts.
4. **Clustering:**
   * Segregated documents into distinct themes, offering actionable insights.

**4.2 Challenges and Limitations**

* Processing time for large datasets was a constraint despite optimizations.
* Some models, like LDA, struggled with thematic overlap.

**5. Conclusion**

This NLP coursework demonstrates the potential of automated textual analysis in deriving insights from large datasets. By systematically applying a range of techniques, the project highlights thematic, stylistic, and emotional patterns in pandemic-related discussions. These findings contribute to understanding public sentiment, aiding informed decision-making.

**6. References**

1. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing*. Pearson.
2. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
3. SpaCy Documentation. Retrieved from <https://spacy.io/>
4. NLTK Documentation. Retrieved from <https://www.nltk.org/>

**7. Appendix**

**Figures and Tables**

* Sentiment distribution bar charts.
* PCA scatterplots for clusters.
* Word clouds for topics.
* Heatmaps for entity relationships.

**Code Snippets**

* Provided in the Jupyter notebook accompanying this report.