IST 322: Introduction to Natural Language Processing

Final Project Description

Introduction

For your final project, you will work in teams to design and implement a **complete Natural Language Processing (NLP) research pipeline** using a **self-collected dataset** containing at least **10,000 textual records**.

This project challenges you to identify a **real-world problem** that can be addressed through text data, formulate a **research question**, collect and curate an appropriate dataset, and execute the **end-to-end NLP workflow**—from raw data acquisition and preprocessing to advanced analysis, modeling, and interpretation.

Through this project, you will demonstrate your ability to integrate technical, analytical, and interpretive skills in applied NLP research.

Each team should:

1. **Clearly define** the problem and research question.
2. **Collect, clean, and preprocess** the text data.
3. **Conduct exploratory and analytical tasks**, including sentiment analysis, topic modeling, and supervised learning.
4. **Document** all steps in a structured and reproducible manner.
5. **Present** both technical findings and interpretive insights in a comprehensive final report and presentation.

Your dataset and research topic may come from any real-world domain that interests you, such as:

- **Customer feedback:** Amazon, Yelp, IMDB, TripAdvisor, or other review platforms.
- **News and journalism:** NewsAPI, Reuters, BBC, or Google News.
- **Social media and forums:** Reddit, Twitter/X, or other publicly accessible APIs.
- **Academic or technical content:** PubMed abstracts, ArXiv papers, or Wikipedia text.
- **Government, policy, or legal sources:** open.gov datasets, UN, or WHO archives.

Your dataset must include at least **10,000 distinct text entries**—such as reviews, posts, sentences, or documents—to ensure sufficient scale and diversity for **robust statistical and machine-learning analysis**.

Task 1: Create Corpus

- Define your research question and identify the text source for your study.
- Create a text corpus using open-source data or data retrieved from web-based sources.
- Save your corpus in a structured format (e.g., CSV or Excel) with relevant metadata fields such as document ID, text content, category label (if any), and source information.
- Generate summary statistics of your corpus, including:
    - Total number of documents
    - Average and maximum document length (in tokens or words)
    - Number of unique documents or entities
    - Any relevant category distributions

Task 2: Text Preprocessing

Perform text preprocessing for all textual data. This step should include:

- Tokenization and normalization
- Optional steps such as stemming, lemmatization, part-of-speech tagging, named entity recognition, stopword removal, etc.
- Create summary statistics after preprocessing, including:
    - Average review/document length (tokens)
    - Vocabulary size (unique tokens)
    - Cleaned text length and lexical diversity
- Include both per-category and overall summaries in your report or Excel file.

Task 3: Data Understanding and Preparation

- Conduct exploratory data analysis to understand your text distribution, label balance, and content quality.
- Identify potential issues (missing data, class imbalance, noise, duplicates) and describe how you addressed them.
- Prepare the cleaned data for subsequent tasks (e.g., aggregation, transformation, or feature engineering).
- Document and justify your data preparation decisions.

Task 4: Sentiment Analysis

- Apply sentiment analysis to your corpus using appropriate tools or models (e.g., rule-based lexicons, pretrained transformers, or supervised classifiers).
- Visualize and interpret sentiment trends across different categories or time periods.
- Formulate and test a short hypothesis: *Does sentiment align with other indicators such as ratings, categories, or time?*
- Provide data-based justification and concise interpretation.

Task 5: Topic Modeling

- Perform **topic modeling** (e.g., LDA, NMF, BERTopic) to uncover latent themes in your corpus.
- Experiment with hyperparameters such as the number of topics or vectorization methods.
- Evaluate models using:
    - Top representative words per topic
    - Topic coherence metrics
- Label and interpret each topic in relation to your research question.
- Discuss how the discovered topics provide insight into your text data.

Task 6: Supervised Learning

- Construct textual features from your processed data using data-driven methods (e.g., Bag-of-Words, TF-IDF, embeddings) and optionally knowledge-driven features (e.g., language tags).
- Define an appropriate **target variable** based on your research design (e.g., category label, or review rating).
- Train at least **three classifiers** (e.g., Logistic Regression, SVM, Random Forest, or Transformer-based models) with hyperparameter tuning.
- Compare models using performance metrics such as accuracy, F1-score, and AUC.
- Select the best model and explain why it performs best for your dataset.

Task 7: Deployment Plan

- Design a **deployment plan** describing how your NLP workflow could be integrated into a real-world application.
- Consider aspects such as:

- Automation of data collection and preprocessing
- Model retraining and updating
- Monitoring performance and fairness over time
- Ethical and privacy considerations in data use

  Provide a brief but realistic plan that connects your technical work to a broader applied context.

Final Project Deliverables

Your final submission should include:

1. **Excel File:** Containing multiple worksheets with corpus summaries, preprocessing statistics, and model results.
2. **Final Report (Word or PDF):**
   Include the following sections:
   1. Introduction – describe your research problem, motivation, and goals.
   2. Corpus Creation – describe data sources, collection methods, and corpus statistics.
   3. Text Preprocessing – detail preprocessing steps and resulting summaries.
   4. Data Understanding and Preparation – describe exploratory and cleaning processes. ***(The draft of part 1, 2, 3, and 4 are due on Nov 23, 2025)***
   5. Sentiment Analysis – outline your approach, findings, and interpretation.
   6. Topic Modeling – summarize modeling process, evaluation, and insights.
   7. Supervised Learning – describe model setup, results, and comparison.
   8. Deployment Plan – describe how your system could be applied in practice.
   9. References – cite all datasets, tools, and external resources used.

Each team submits **one final project report**, but **individual contributions** must be documented in an appendix or separate reflection section. Individual grades will consider both team outcomes and individual contributions.