Natural Language Processing

Assignment 1: Basic-Level - Text Preprocessing for AI Solution Platform

Overview: One Hour AI Solutions needs to process client problem descriptions to make them ready for analysis. Your task is to implement basic text preprocessing techniques covered in the educational material.

Requirements:

1. Use the provided "client_problems.csv" dataset containing AI problem descriptions from clients
2. Implement a complete text preprocessing pipeline that includes:
    - Tokenization (splitting text into words)
    - Converting text to lowercase
    - Removing punctuation and special characters
    - Removing stop words
    - Performing stemming or lemmatization
3. Compare the results of stemming vs. lemmatization on 5 example sentences
4. Calculate and report basic statistics about the processed text (average word count before and after, most common words)

Expected Output:

- Processed text data with all preprocessing steps applied
- Comparison table showing original text, text after stemming, and text after lemmatization
- Summary statistics about the text corpus before and after preprocessing

Dataset Description: The "client_problems.csv" will contain columns: problem_id, problem_description, difficulty_level

Assignment 2: Intermediate-Level - Bag of Words and TF-IDF for AI Problem Classification

Overview: One Hour AI Solutions wants to categorize incoming client problems to route them to appropriate engineers. Your task is to implement text representation techniques (Bag of Words and TF-IDF) and use them for simple classification.

Requirements:

1. Use the provided "categorized_problems.csv" dataset containing problem descriptions and their categories
2. Implement:
    - A Bag of Words representation for the problem descriptions
    - A TF-IDF representation for the same descriptions
3. Use a simple Naive Bayes classifier to:
    - Train a model using both representations
    - Predict the category of new problem descriptions
4. Compare the performance of BoW vs. TF-IDF approaches
5. Identify the most influential words for each category using TF-IDF scores

Expected Output:

- Comparison of classification accuracy between BoW and TF-IDF representations
- List of top 5 most important words for each problem category
- Analysis of when each representation performs better or worse

Dataset Description: The "categorized_problems.csv" will contain columns: problem_id, problem_description, problem_category

Assignment 3: Moderately Advanced - Rule-Based Sentiment Analysis for Client Feedback

Overview: One Hour AI Solutions collects feedback after each session with an AI engineer. Your task is to build a rule-based sentiment analysis system to analyze these feedback comments.

Requirements:

1. Use the provided "session_feedback.csv" dataset containing client feedback after AI solution sessions
2. Implement a rule-based sentiment analyzer that:
    - Uses dictionaries of positive and negative words
    - Handles negation (e.g., "not bad" should be interpreted as positive)
    - Assigns sentiment scores to feedback text
    - Classifies feedback as positive, negative, or neutral
3. Create a function to identify specific aspects being mentioned (e.g., engineer knowledge, solution quality, time efficiency)
4. Evaluate your analyzer by comparing its predictions to human-labeled sentiments provided in the dataset

Expected Output:

- Overall sentiment distribution across all feedback
- Aspect-specific sentiment analysis (which aspects receive more positive/negative feedback)
- Evaluation metrics (accuracy, precision, recall) for your sentiment analyzer
- Discussion of limitations of the rule-based approach

Dataset Description: The "session_feedback.csv" will contain columns: feedback_id, feedback_text, human_labeled_sentiment