**Automated Specific Response System for Email**

**MSDS692 – Data Science Practicum 1**

**Progress Report for Week 3**

**Project Details:**  
The goal this week was to set up the project environment and begin preprocessing the data from the Enron email dataset for my automated email response system. The focus was on cleaning the data and preparing it for analysis and model development, which are important steps for my NLP-driven project.

**Project Timeline:**

* Week 1: Initial project submission (DONE)
* Week 2: Data sourcing and project redirection (DONE)
* Week 3: Environment setup and data preprocessing (DONE)

**Planned Work for the Week:**

* Complete setup of the Python environment and necessary libraries
* Start preprocessing the Enron dataset, focusing on cleaning and initial analysis.

**Progress for the Week:**

This week was dedicated to setting up the necessary environment for my project and beginning the main phase of data preprocessing. I successfully loaded the first chunk of the Enron email dataset, which involved reading 10,000 emails into my processing environment. Key steps in my preprocessing included cleaning the data by removing irrelevant information, calculating the length of each email body to measure content volume and performing initial text processing to shed away unnecessary characters and whitespace. Also,I initiated a more refined text analysis to prepare the email bodies for NLP tasks, focusing on tokenization and the removal of stop words to better make my dataset's quality for future modeling.

**Roadblocks/Issues:**

* Faced challenges with the large size of the dataset, requiring chunk processing to manage memory usage efficiently.
* The initial attempts to preprocess the data resulted in a blank CSV file, which was resolved by refining the data processing and saving process.

**Plan for Next Week:**

* Continue with data preprocessing, focusing on more in-depth text analysis and cleaning.
* Begin exploratory data analysis (EDA) to uncover patterns and insights within the email data.
* Start drafting the initial model framework for generating email responses.

**Resources for the Week:**

* Used Pandas documentation for data manipulation techniques.
* Explored NLTK and spaCy libraries for NLP preprocessing methods.

**Project Timeline:**

Week 4: Continue preprocessing and initial model training.

Week 5: Develop response generation logic and test models.

Week 6: Model evaluation, maybe begin report drafting.

Week 7: Final model refinements, complete report draft.

Week 8: Finalize report and prepare presentation.