Application Development

Term Project Proposal

Team members

- Syed Najeeb Iqbal (21205)
- Azeem Ullah Hasan (09304)

Project Description

In this project, we will be working on an existing spam emails dataset that contains both spam and non-spam emails. This data set will be used to perform different kinds of tasks including:

- Mapping the most repeated words in spam and non-spam email by using word cloud
- Removing irregularities in data by using Python's built-in regular expressions
- Removing the stop words in an email, to find the keywords in it
- Sentiment analysis will be performed on the emails to classify them as having a positive, neutral, or negative sentiment.

Dataset/Web APIs

- Spam emails dataset from GitHub

Python Libraries to be Utilized

The following libraries will be mainly used for this project:

- Pandas for data manipulation and analysis
- Word Cloud for data representation in a cloud
- Matplotlib/Plotly for data visualization

The following libraries will be optional and used if the need arises, particularly on the NLP aspect of the project:

- Scikit, NLTK

Features

The following basic features will be added:

- Word Count, Most-Common Word(s), and other basic statistics
- Visualizations on the aforementioned statistics

Advanced features include:

- Emotions, tone, sarcasm, negations, bias, comparisons, etc.
- entity recognition, intent detection and other sentiment analysis aspects
- Application of tokenization, lemmatization, and stemming will be studied and applied wherever possible