Presentation & Demo Link:
https://drive.google.com/file/d/1yoOIe-9Z9sIoPqKi2Frb3H-8q1c7M_iY/view?usp=sharing

Email Author Classification using LinearSVC

This project is focused on classifying the authors of email messages based on their content. It uses data from the Enron email dataset and implements a machine learning pipeline involving text preprocessing, TF-IDF vectorization, and modeling via LinearSVC

📁 Project Structure

- **Text Preprocessing**: Emails are cleaned by removing URLs, punctuation, and stopwords, and then stemmed.
- **Vectorization**: Processed text is transformed into TF-IDF vectors.
- **Modeling**: We use a variety of methods, but the most accurate one is LinearSVC

🚀 Features

- Efficient text cleaning using NLTK.
- Scalable vectorization using `TfidfVectorizer`.
- sklearn linear models.
- pytorch neural networks
- Evaluation via accuracy and F1 score

🧪 Requirements

- Python 3.x
- Scikit-learn
- NLTK
- NumPy
- Pandas
- Matplotlib (for visualization)

Install dependencies with:

``` bash
pip install scikit-learn nltk numpy pandas matplotlib
```

🗂️ Data

We are using the Enron email dataset, found in this link https://www.kaggle.com/datasets/wcukierski/enron-email-dataset . Make sure to unzip the file completely as we had some problems doing that because of its large size. We created the following path to the emails that we were interested on for our project. Replace this on your side with a path to the Sent_Items_only directory in the Enron dataset sure you can access the Sent_Items_only

    /WAVE/projects/CSEN-140-Sp25/HHJ140Proj/Sent_Items_only

Ensure this directory structure exists and contains subfolders for each user, each with a `sent_items` folder of email files.

🛠️ Usage

You can run the notebook step-by-step to:

1. Clean and preprocess the email data.
2. Vectorize the text with TF-IDF.
3. Train traditional ML/Data Science Methods
4. Train Neural Network
5. Check how we are optimizing the values for each of these models.

📈 Results

Performance is evaluated using metrics such as precision, recall, and F1-score

📌 Notes

- The project is sensitive to the dataset structure and email formatting.