**Assignment 1** (20%)

CSE 5120 (Section 02) – Introduction to Artificial Intelligence – Spring 2022

*Submitted to*

Department of Computer Science and Engineering  
California State University, San Bernardino, California

*by*

Corbin Ulloa (006973376)

Date: *May 18, 2022*

*Email:*

*006973376@coyote.csusb.edu*

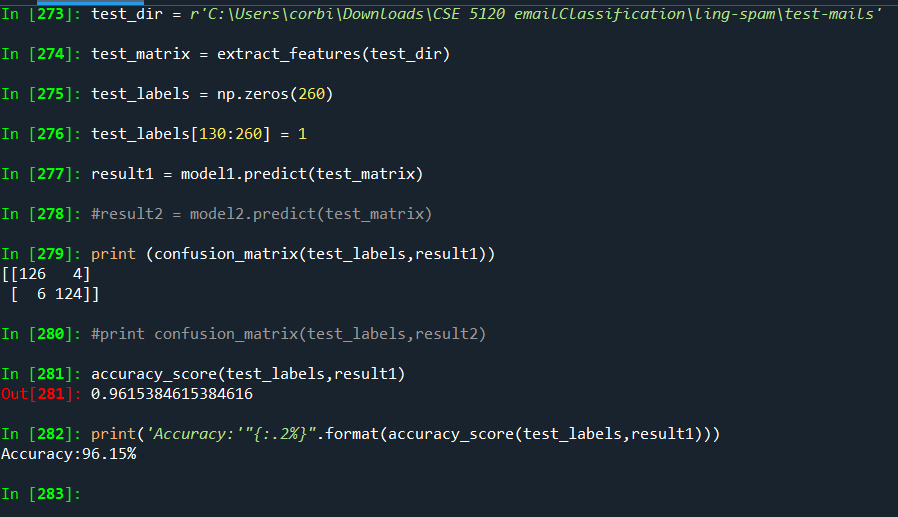
**Assignment Report**

Brief description of your work here acknowledging your collaboration with your class fellow (or a friend from other CSE 5120 section), and the capacity at which he/she collaborated with you, followed by the algorithms you implemented.

1. **Classifier\_lingSpam.py for Ling-spam dataset**

Your brief explanation of the dataset, your code solution, and any documentation with screenshots of your code Evaluation (results from classifier\_lingSpam.py)

The Ling-spam dataset is a collection of 260 emails (130 spam and 130 non-spam emails) that is used to develop an email classification model which classifies an email as spam or non-spam (i.e., ham) mail. In order to correctly identify emails, I implemented a function that returns a dictionary of words and their frequency along with extracting features of 3000 dimensions for each email of training set. After training the Support Vector Machine (SVM) classifier we print out the confusion matrix and an accuracy score of 96.15%.



1. **Classifier\_enron.py for Enron-spam dataset**

Your brief explanation of the dataset, your code solution, and any documentation with screenshots of your code Evaluation (results from classifier\_enron.py)

The Enron-spam dataset contains 33,716 emails in 6 directories containing ham and spam folders. Total number of non-spam emails and spam emails are 16,545 and 17,171. Since the directory path was in a different structure, I modified my helper functions by adding another os.path.join() to reach the directories with multiple paths. This allowed me to bypass the “permission denied” error when training model. Instead of using LinearSVC I used the SVC model where SVC uses the One-vs-One multiclass reduction. Hence, why it takes longer for SVC to converge larger number of samples as opposed to Linear SVC. However, SVC appears to output better results with a 97.97% accuracy.

Text

Description automatically generated