Data Science Certificate Diploma

DS650 Data Analytics

**Homework 3 (20 points)**

**You are expected to deliver a PDF document (this document) and one Jupyter notebook with your Python code.**

For the following questions, we will be working with an SMS Spam dataset provided as a .csv file. The dataset contains SMS messages labeled as either "spam" or "ham" (not spam). Our goal is to analyze, preprocess, and build a machine learning model to classify messages as spam or not. All questions must be answered using Google Colab.

To obtain reproducible results, call the following command in the beginning of your code:

np.random.seed(123)

a. Read the data into a data frame. Report in the table below the total number of spam and ham messages in the dataset. What do you observe? Which evaluation metric might be more useful (e.g. accuracy)? (2 points)

|  |  |
| --- | --- |
| **Label** | **Count** |
| Ham |  |
| Spam |  |

|  |
| --- |
| Answer: |

b.

We will convert the messages into tokens with different ways. Firstly, use the LemmaTokenizer class (that tokenizes the text, lemmatizes each word, and returns the cleaned tokens), the StemTokenizer class and the get\_wordnet\_pos function(maps POS tags to WordNet tags for lemmatization) from the count vectorizer.ipynb file from Activity Week 5 Folder. Secondly, test these two tokenizers on the sample message (displaying in the box below the tokens created for each method):

*sample\_msg = [“Congratulations you have won a 1000 dollars gift card Click here to claim your prize”]*

(2 points)

Answer:

c. Let’s return to on our data! Create a binary label column called b\_labels where 'ham' is mapped to 0 and 'spam' is mapped to 1. Split the dataset into training and testing sets (80% train, 20% test). Then convert the text data in the train and test set into numerical vectors using 5 different techniques:

* CountVectorizer with default arguments
* CountVectorizer with lemmatization
* CountVectorizer with stemming
* TF-IDF vectorizer with default arguments
* TF-IDF vectorizer with a maximum of 2000 features

Report and compare the shape of the resulting matrices for the training set.

(# of rows, # of columns)

(3 points)

|  |  |
| --- | --- |
| **Vector Model** | **Matrix Shape** |
| TFIDF | ( , ) |
| TFIDF with 2000 features | ( , ) |
| Count Vectorizer | ( , ) |
| Count Vectorizer with Lemmatization | ( , ) |
| Count Vectorizer with Stemming | ( , ) |

|  |
| --- |
| Answer: |

c. Train a Logistic Regression classifier using with all the above vectorizers as input features. Instead of simply training your models on the entire training set and evaluating on the test set, we will use 5-fold cross validation which is a methodology introduced in Lecture 2 for model evaluation and selection. For each vectorizer use 5-fold cross validation with grid search for selecting the hyperparameter C with options [1, 0.1, 0.01] and F1-score as the metric. Report the cross validation F1-score of these models.

(5 points)

|  |  |
| --- | --- |
| **Vector Model** | **CV F1-score Logistic Regression** |
| TFIDF |  |
| TFIDF with 2000 features |  |
| Count Vectorizer |  |
| Count Vectorizer with Lemmatization |  |
| Count Vectorizer with Stemming |  |

d. Which preprocessing technique works best? What is the optimal value for C? (1 point)

Answer:

e. Choosing the TF-IDF with 2000 features and Count Vectorizer with Stemming from above use 5-fold cross validation with a random forest classifier using F1 score as a metric with the following hyperparameters for tunning:

'n\_estimators': [50, 100, 200],

 'max\_depth': [None, 10, 20],

'min\_samples\_split': [2, 5, 10]

Report the cross validation F1 score and the best hyperparameters.

(2 points)

|  |  |
| --- | --- |
| **Vector Model** | **CV F1-score Random Forest** |
| TFIDF with 2000 features |  |
| Count Vectorizer with Stemming |  |

Answer:

f. Select the best model among random forest and logistic regression and evaluate the model on the test set and report the accuracy, precision, recall, and F1-score. Plot in the following box the normalized confusion matrix based on the test data.

(4 points)

|  |  |
| --- | --- |
| **Metric** | **Best Model Test Set** |
| Accuracy |  |
| Precision |  |
| Recall |  |
| F1-score |  |

Plot here

g. Examine false positives (ham classified as spam) and false negatives (spam classified as ham) from the test set. Show those misclassified cases in the box below and elaborate on what might have confused the algorithm.

(3 points)

Answer: