DETAIL PROJECT REPORT

SPAM-HAM CLASSIFIER

DETAIL-PROJECT REPORT

Table of Contents

DETAIL-PROJECT REPORT

1) Objective

Development of a predictive model which can classify the message type precisely.

2) Benefits

1) It will help customer to save time and be efficient in their work

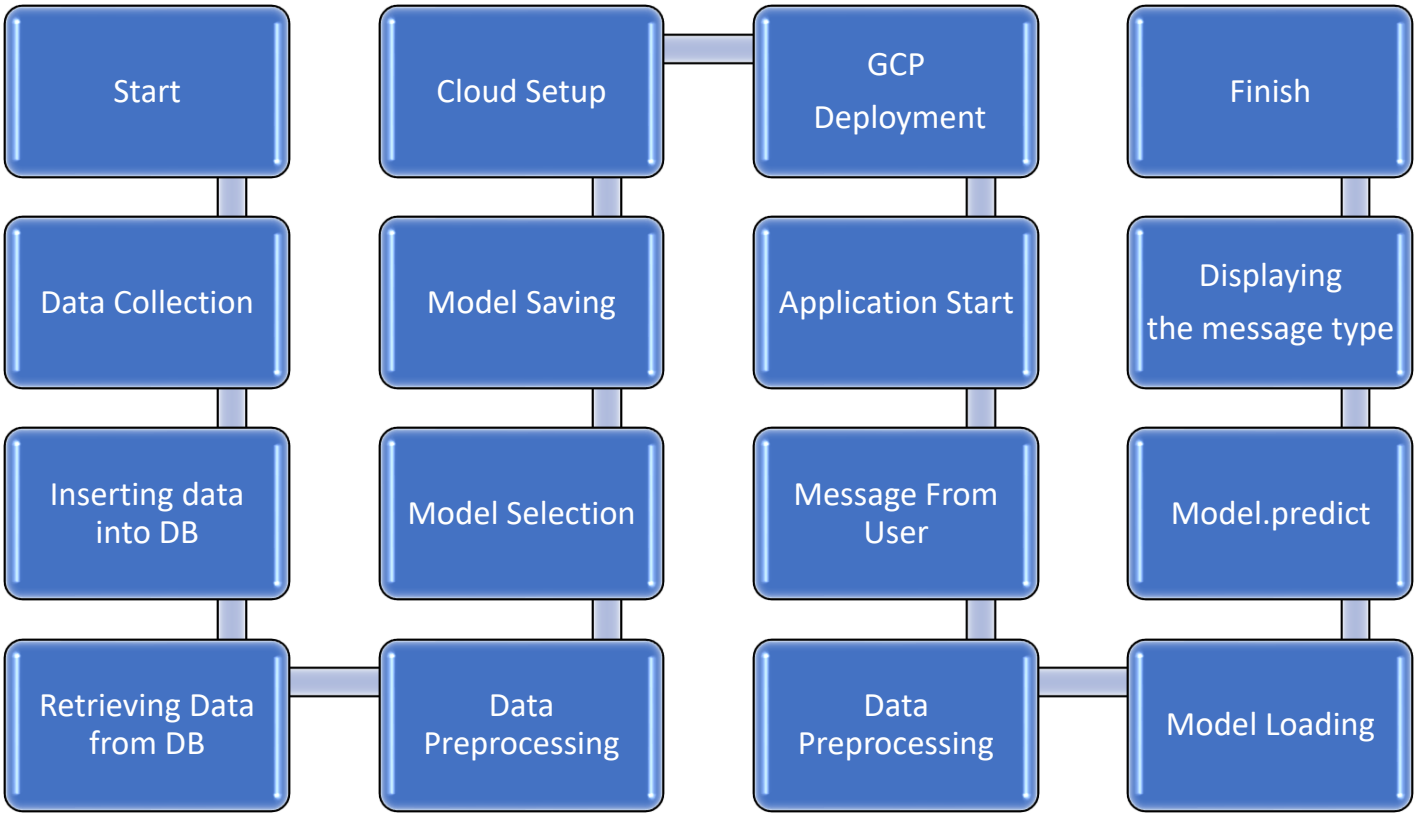
2) It will help to prevent from fraud, scam.

3) Scam rate will be reduced.

3) Data Sharing Agreement

- File name can be anything
- Columns inside the file will be 4
- Columns are V1,V2,V3,V4
- Columns datatypes are [string,string,string,string]

4) Architecture



5) Data Validation and Data Transformation

- Name Validation: We don't have any problem with the name of the file which will be given by the client.
- Number of columns: There will be only 4 columns.
- Name of columns: V1,V2,V3,V4 are the name of the columns.
- Data type of columns: The data type of the columns is string.
- Null Values in the column: If null values is present in the table then that particular row will be deleted.

6) Data Insertion in Database

a) Table Creation: I am using Cassandra Database for my project. Inside keyspace training. I am creating a table named all_train_files, table will only be created if it doesn't exist

b) Data Insertion: The table will inserted into the table all_train_files.

7) Model Training

7.1) Data export from db

All the data from the database is retrieved and stored into a csv file called Training_file.csv.

7.2) Data Preprocessing

1) First I checked whether is there any null values present inside the data and I found that there is a single row only that's why I directly removed it.
2) After removing the null values. I handled the first column V1 which was containing the message type as spam or ham and renamed it as target, and after that I replaced spam with 1 and ham with 0.
3) I deleted the column V3 and V4 as it were not containing much more information and were mistakenly created I think.

4) After that My main task was to handle the message which was present in column V2 and I renamed it into message, and do some pre-processing in the language. All the pre-processing done is given below
    - Lowercasing each word present in the message.
    - Word tokenization Ex: My name is Suraj, After tokenization [My, name, Is, Suraj].
    - After word tokenization I am only taking the words which are not stop-word(is, are, the…) and punctuation(?,!..) because they don't contribute much to the message.

    - Now stemming will be done on each word. For example sleeping will be converted to sleep only.

5) Now my entire corpus is ready to be served to any word vectorizer like bag of word,1gram, tri-gram, tf-idf vectorizer. After experimenting on my data I got to know that tfi-idf is performing way better than other algorithms.

7.3) Model Training Part

Model training part took a lot of my entire project time and is the most crucial part of any project development.

- First as it is a classification based problem and is a spam-ham classifier so I took the algorithm gaussian naïve bayes. But it wasn't giving good accuracy so,
- I moved to another algorithm that is Multinomial Naïve Bayes and it was giving a best accuracy of 0.93.
- I thought I might get better accuracy than this also so I thought to experiment upon different algorithms like Decision Tree Classifier. Random Forest classifier, SVM , Logistic Regression, XGB and few others also. But also I wasn't able to cross the accuracy of 0.93.
- I even tried bagging, Stacking and many more things but nothing improved.
- So I decided to go with MNB because of its accuracy.

- I saved my model as model.pickle .

- Now I created an API for my model using Flask.
- Finally My project is created and I am going to deploy it to the AWS

7.4) Prediction

1) Now its time for prediction part, I will collect the message from the user

2) Data pre-processing is performed same as while training, I am

3) Now everything is just simple I am loading my pickle model and applying model.predict

4) The output will be the message type as spam or ham and I am going to display the output in the home page.

8) Q & A

1) What is the Source of Data?

Ans: UCI is the source of the data link:

https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

2) What was the type of Data?

Ans: Data type of every column is string.

3) How logs are managed?

Ans: I am managing the logs using logging module.

4) What techniques are you using for data pre-processing??

Ans: See Data Pre-processing above I have mentioned there in detail.

5) How training was done

Ans: After a lot of research I got to know that my data fits very good with MNB algorithm so I have selected MNB algorithm for my model training.

6) What are stages of Deployment?

Ans: Deployment has been done to Amazon Web Services and I have deployed the application in the production server. Link

http://ec2-3-140-241-66.us-east-2.compute.amazonaws.com:5000/