



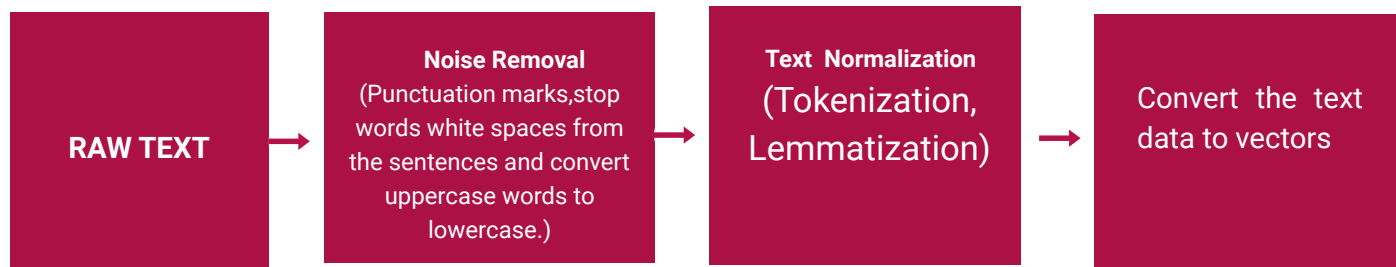
# SPAM VS HAM MODEL TRAINING AND DEPLOYMENT ON FLASK

# Data Preprocessing

- Data preprocessing is the important step in the Data Science project the model performance depends upon the how the data is processed.
- In this project data preprocessing involves in Text processing, feature selection and splitting the data to train and test dataset. Train set is to train the model and testset is to test the model.



# Text Processing



—————→ Text Processing Pipeline ←————

# Why Text Cleaning?

- Cleaning the data is removing the punctuations and special characters from the data that are not useful for the classifying it as spam or ham.
- If we use these characters for the training the model then the model tends to look for these characters in the SMS to classify it as spam or ham. So we remove these characters.
- Next we remove the stop words because there is no importance for the stop words in this problem if there is increases the model complexity.
- Next Tokenization of the SMS, the SMS message is converted into tokens and thereafter Lemmatization is performed.

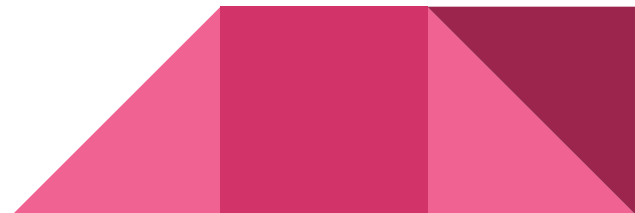


# Removing Punctuation And Stop Words

- The punctuation marks and all types of special characters from the sentences are removed and the characters are converted to lower letters and then the stop words are removed.

EX:- Before → “Sorry my roommates took forever, it ok if I come by now?”

After → “sorry roomates took forever ok come ”



# Tokenization

- Tokenization is convert a text into smaller pieces. Tokenization can be of various types like word tokenization,sentence tokenization,character tokenization.Here word tokenization is done
- Then Output from removing stop words and punctuations are taken and tokenization is performed.
- EX- Input :- “sorry roomates took forever ok come”  
Output:-[“sorry”,“roommates”, “took”,“forever”,“ok”,“come”]



# Why Lemmatization?

- Lemmatization is the process to get the tokens to the normal form, it is similar to stemming but it brings context to words.
- Ex: when performing stemming on a word like 'roommates' it gives result 'roommat' where it strips 'es' and 's' from the words do not care about the meaning of it.
- While in Lemmatization the word 'roommates' is normalized to 'roommate' where it keeps the context of the word. Output from tokenization is taken and lemmatization is performed.
- Input: ["sorry", "roommates", "took", "forever", "ok", "come"]

output:

sorry	roommate	took	forever	ok	come
-------	----------	------	---------	----	------



# Why CountVectorizer?

- To train the model direct SMS can not be sent to it for training. the model can not understand that, SMS need to be converted to vectors.
- Countvectorizer converts the words to vectors based on their count appearing on sentence.

Ex: The output of lemmatizer is taken and countvectrization is performed.

Input:-"sorry roommate took forever ok come"

output:-

Word	sorry	roommate	took	forever	ok	come
Count	1	1	1	1	1	1



# Train\_Test\_Split

- After the data that is transformed from words to vectors are splitted to test and train where train is send for model training, the test data is to test the trained and calculate the accuracy of the model.
- Sklearn provide the feature “`sklearn.model_selection.train_test_split`” to split the data to train and test (20% of data is taken from testing and 80% for training with `random_state=0`)



# Modelling

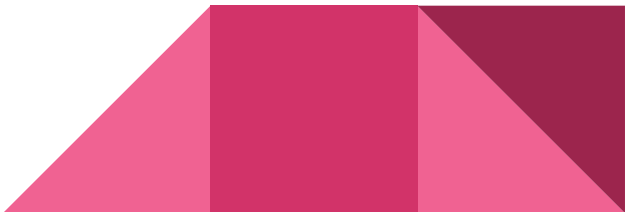
# Model Training

- The train set is sent for the training and training is done on different models and checked which model gives best accuracy.

Model	Multi Naive Bayes	Logistic Regression	Random Forest	MLP
Accuracy	0.98657	0.97757	0.97488	0.98475



# Model Testing

- After checking with some sentences the model is biased to certain words to detect it as Spam words like:- Free,Call,gift,text,reply and claim.
  - These words are occurred more in the Spam data in the data set when a spam message with different words were given then the model detected is not performing as expected.
  - EX: “Respond to this Number to process the gift” this is detected as HAM
  - The MultiNominalNavieBayes does well when the accuracy is concerned so the model is saved for the deployment. And the CountVectorizer was also taken and saved to process the input sentences.
  - Models are Saved by using Pickle.
- 

# Deployment on Flask

# What is Flask?

- Flask is a web framework where we can deploy the model and run the model
- Web page created here acts as a user interface,Where user can enter the Message there to classify it as SPAM or HAM.



# How is was Deployed?

- A folder named 'templates' is created and index.html file i.e the web page is stored in it.
- An app.py is created and the model,countvectorizer is loaded.
- Run app.py on command prompt(python app.py)
- Open the "<http://127.0.0.1:5000/>" on the browser.



# The Page Looks Like

---

## **SPAM SMS Predictor**

The Message can be entered in the text box and click on Predict button to get the message as SPAM or HAM.







Thank You