 REPORT ON

Spam Classfier

**Made by:-**

**Sagar Patidar** Reg No.-11804732 Section:-K18KKA08Roll no-08

**ABSTRACT**

In this project our main focus to make a project or algorithm to detect whether a email is spam or ham.”ham” is generally refers to good email or not spam. so its name is “Spam classfier”. In this project we are making a email clssfier which used naïve classifier to detect email as spam. But Nowdays, many emails contain harmful content link, viruses ,ads, phising pages and malware and that are dangerous for our system and privacy. It gives knowledge to the user about the fake e-mails and relevant e-mails.In this we used naïve bayes classfier to classify the mails.user mode dataset is provide by user and then an spam detection algorithm is work on it.we defined two categories spam or ham.and set their threshold frequency of words to detect mail is spam or not.

**RELATED WORK**

Many algorithms are made on this topic that help to detect spam in mails. Mostly people are using naïve bayes classfier to classify.But previously some more area of spam classification are known:-

**Content Based Filtering Technique:** Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification.

**Case Base Spam Filtering Method:** The data is then classified into two vector sets. Lastly, the machine learning algorithm is used to train datasets and test them to decide whether the incoming mails are spam or non-spam

**Heuristic or Rule Based Spam Filtering Technique:** This approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.

**Adaptive Spam Filtering Technique:** The method detects and filters spam by grouping them into different classes. It divides an email corpus into various groups, each group has an emblematic text.

**IMPLEMENTATION**

Spam is major problem in IT industry and a normal life .Implementaion of problem is divided in stages and then perform operation in stagewise. Firstly create dictionary and then save their keys values ,if the words are repeated then counter is increased to previous values. Then calculate their threshold frequencies. Then apply naïve bayes algorithm.

**FLOW CHART**

INPUT

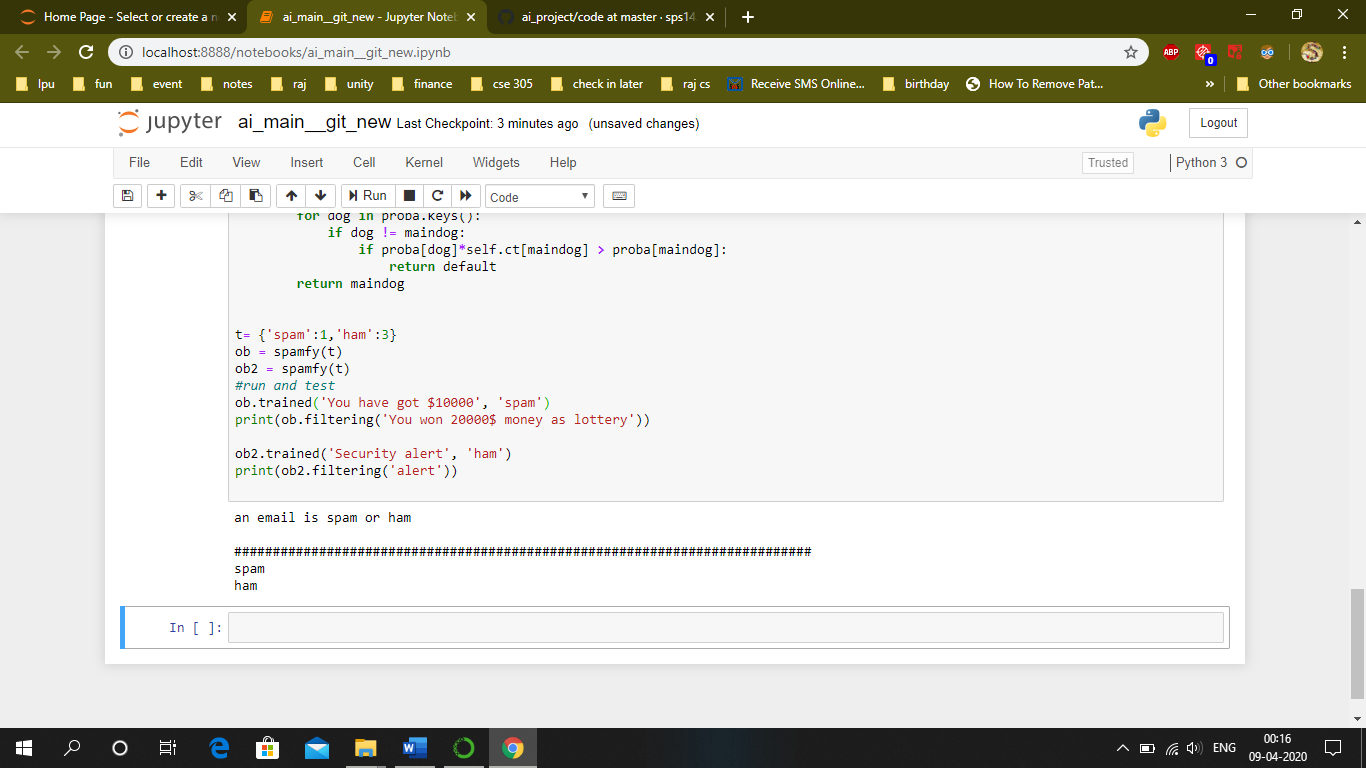
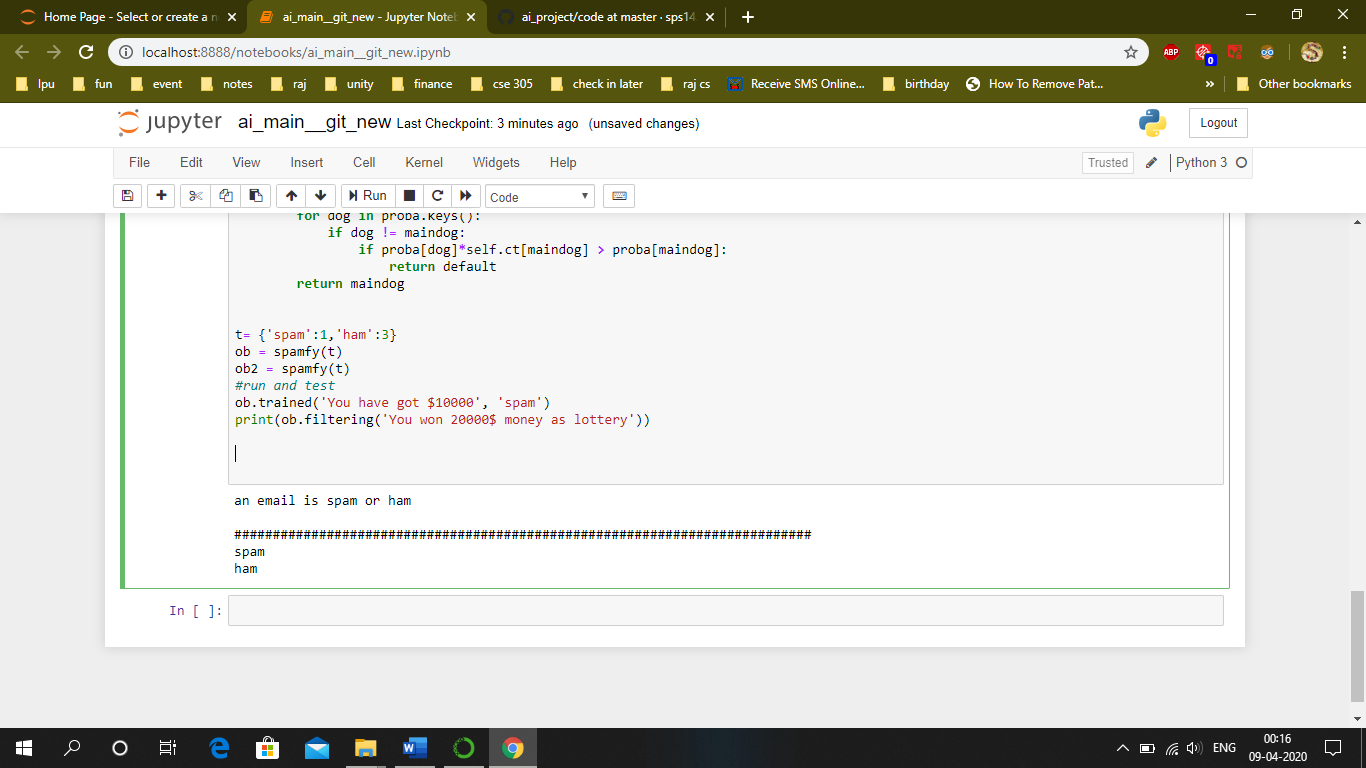
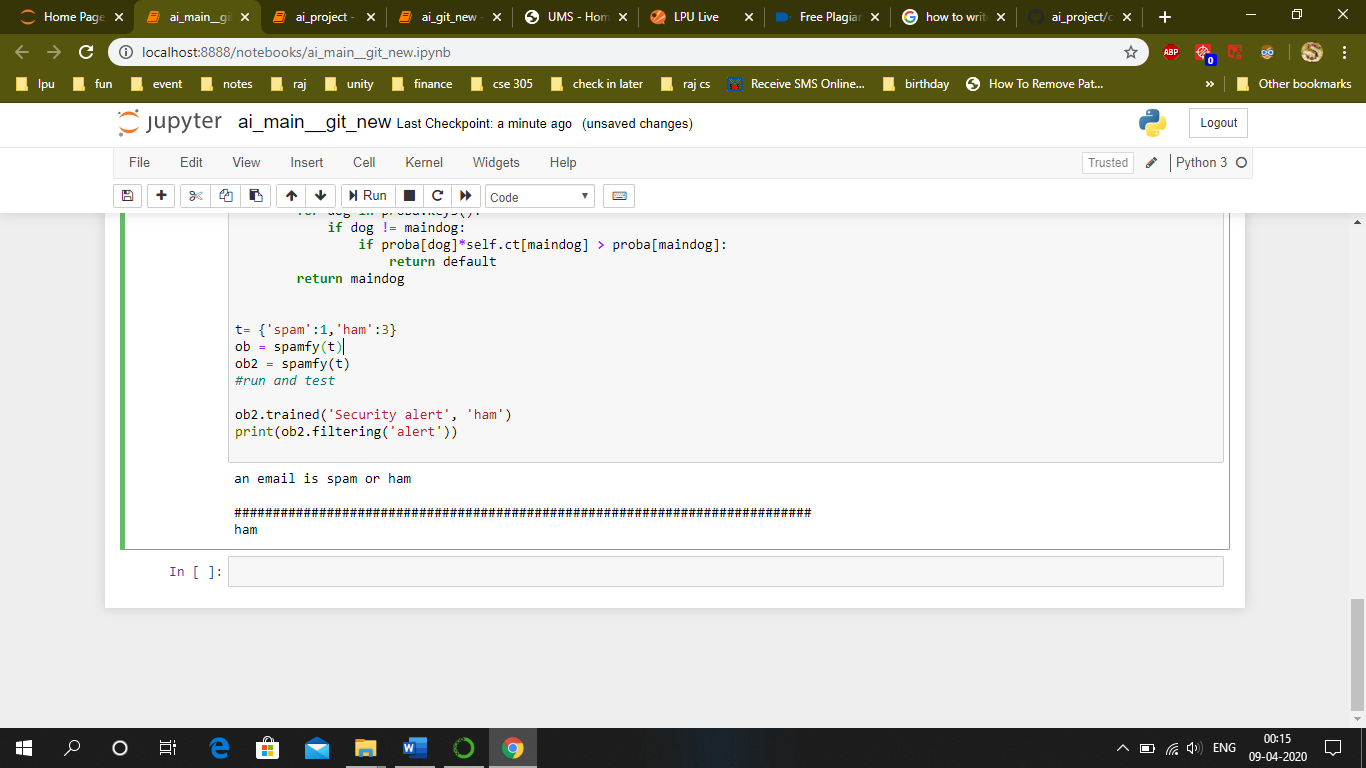
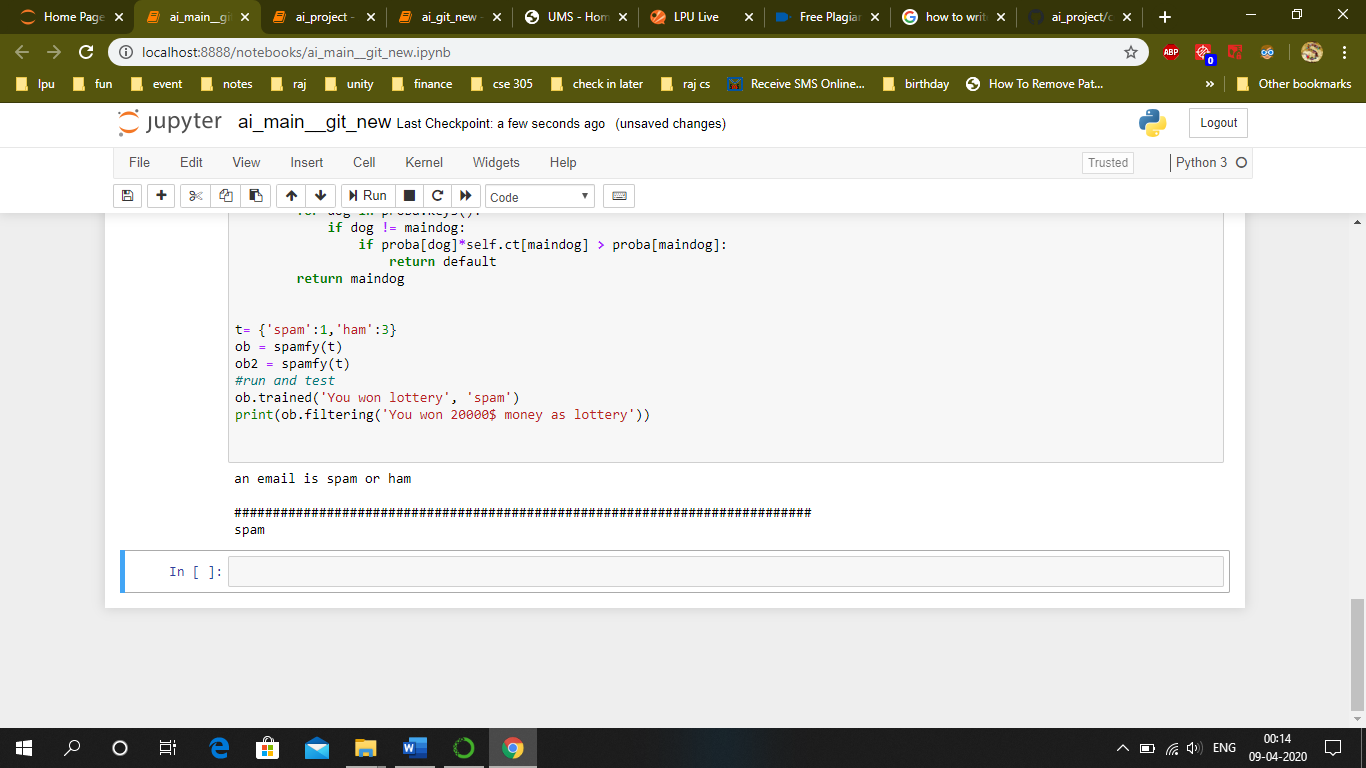
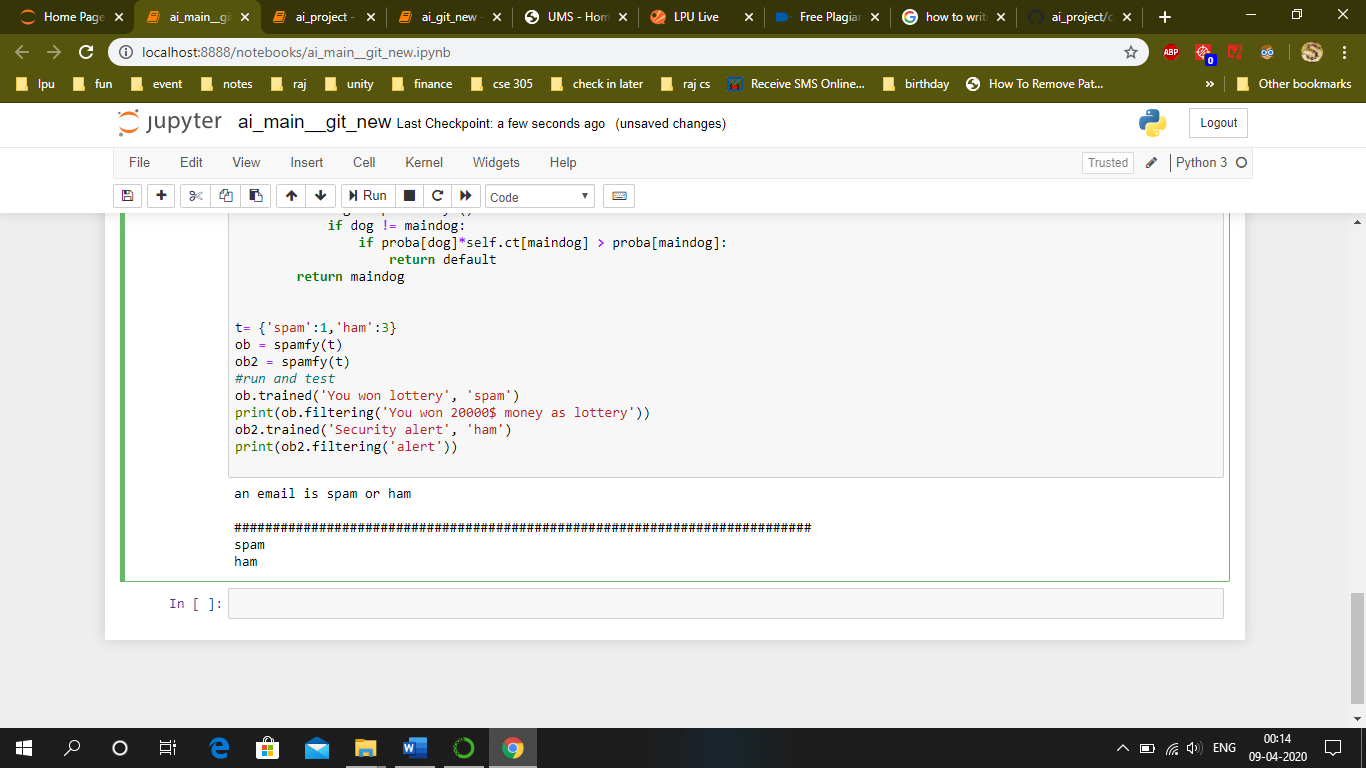
EMAILS

SPAM

Algorithm to check spam or not

HAM

**RESULT**

****

**LIBRARY USED:-**

|  |  |
| --- | --- |
| **Library name** | **Specification of library** |
| RE-Regular Expression | A regular expression is a special sequence of characters that helps you match or find other  strings or sets of strings, using a specialized syntax held in a pattern |
| OS- Operating System | The functions that the OS module provides allows you to interface with the underlying operating system that Python is running on – be that Windows, Mac or Linux |

**REFERENCES**

<https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/>

<https://www.guru99.com/nltk-tutorial.html>

<https://www.geeksforgeeks.org/naive-bayes-classifiers/>

[https://www.kaggle.com/benvozza/ classification](https://www.kaggle.com/benvozza/spam-classification)

<https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

<https://www.geeksforgeeks.org/naive-bayes-classifiers/>

<https://www.youtube.com/watch?v=CPqOCI0ahss>

<https://www.nltk.org/>

<https://www.guru99.com/nltk-tutorial.html>