

**Malignant-Comments Classifiers Project**

Submitted by:

Vandana Jain

Internship 10

**ACKNOWLEDGMENT**

The internship opportunity I had with FlipRobo was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this project period.

I would like to thank our SME for suggesting this project and for his whole hearted cooperation and constant encouragement throughout the project.

**INTRODUCTION**

* Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as inoffensive, but “u are an idiot” is clearly offensive.

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

* Motivation for the Problem Undertaken

The upsurge in the volume of unwanted comments called malignant comments has created an intense need for the development of more dependable and robust malignant comments filters. Machine learning methods of recent are being used to successfully detect and filter malignant comments. Build a model which can be used to predict in terms of a probability for comments to be malignant. In this case, Label ‘1’ indicates that the comment is malignant, while, Label ‘0’ indicates that the comment is not malignant.

* OBJECTIVE

The objective of identification of comments are :

* To give knowledge to the user about the bad comments and positive comments
* To classify that comments are malignant or not.
* SCOPE OF THE PROJECT:
* It provides sensitivity to the client and adapts well to the future malignant comments detection techniques.
* It considers a complete message instead of single words with respect to its organization.
* It increases Security and Control.
* It reduces IT Administration Costs.
* It also reduce Network Resource Costs.

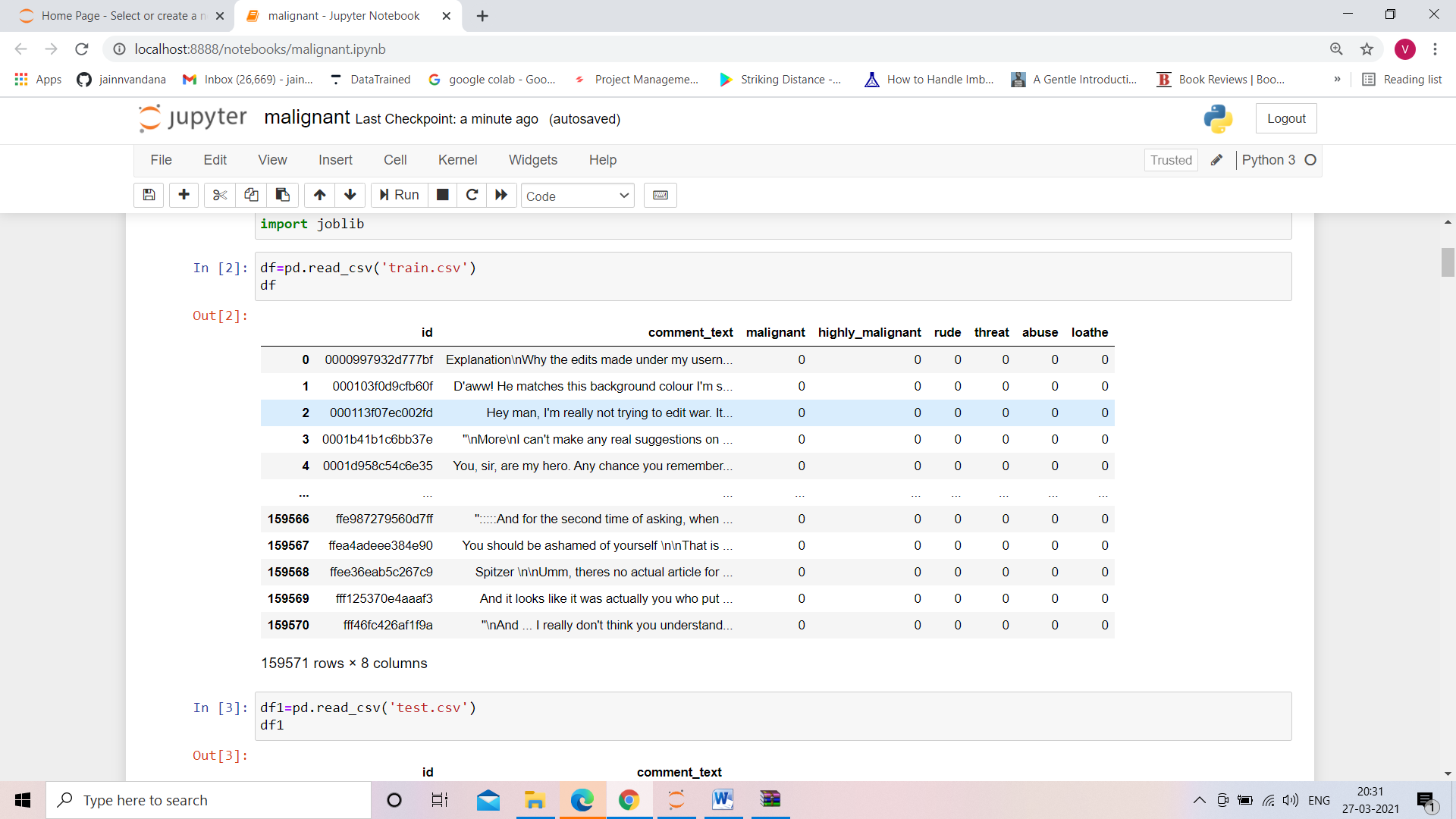
**Analytical Problem Framing**

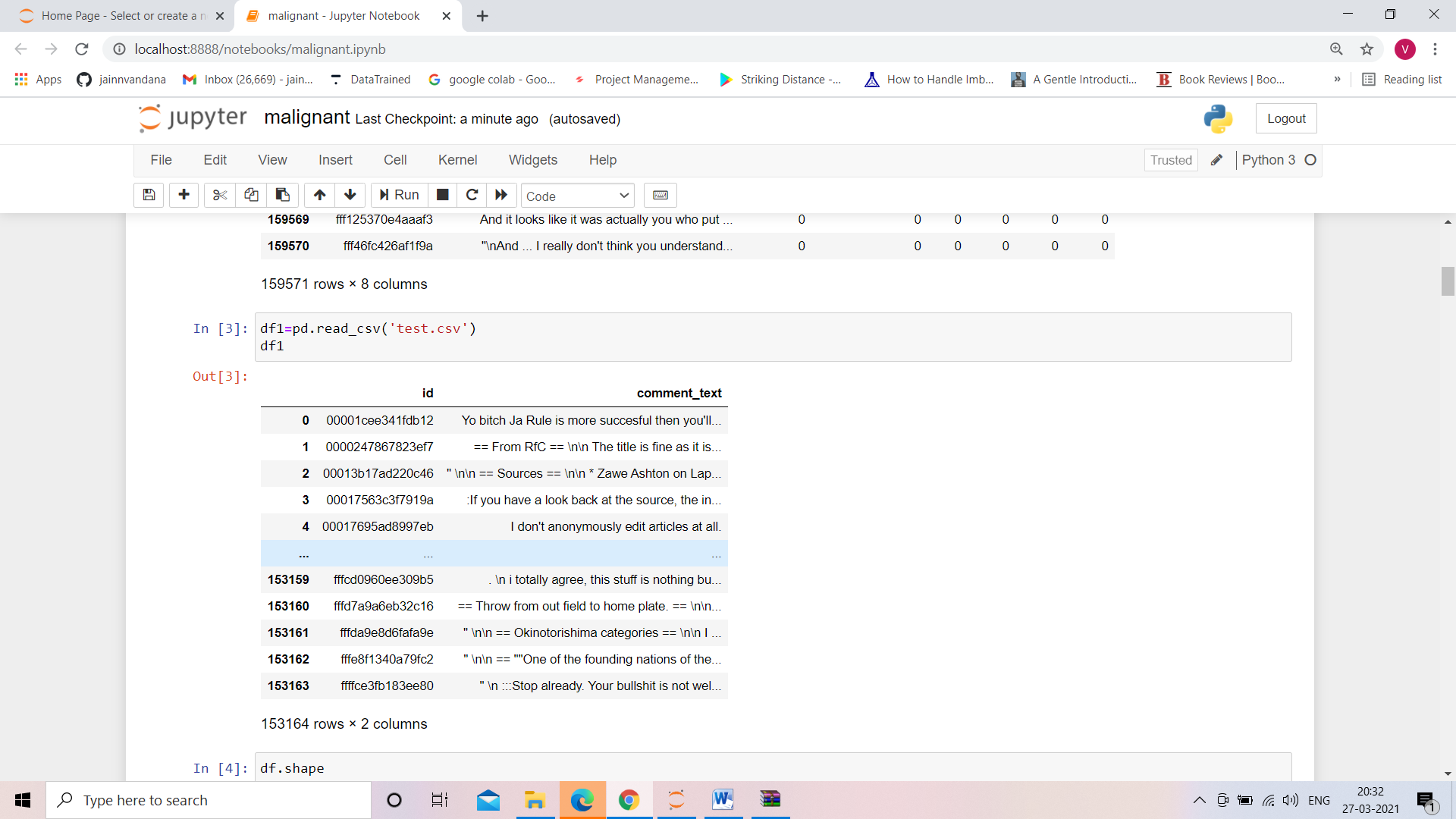
* Mathematical/ Analytical Modeling of the Problem

Machine Learning is defined by Tom Mitchell in his book as “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. Supervised learning is when the output is known for the corresponding inputs, and is also provided for the machine to learn.

* EDA (Exploratory data analysis)
* Data Preprocessing
* Feature Extraction
* Scoring & Metrics
* Data Sources and their formats

The data is provided to us from our client database. It is hereby given to us for the exercise to improve the selection of comments for malignant or not malignant. It is given in the csv file format.





* Data Set Description

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

The data set includes:

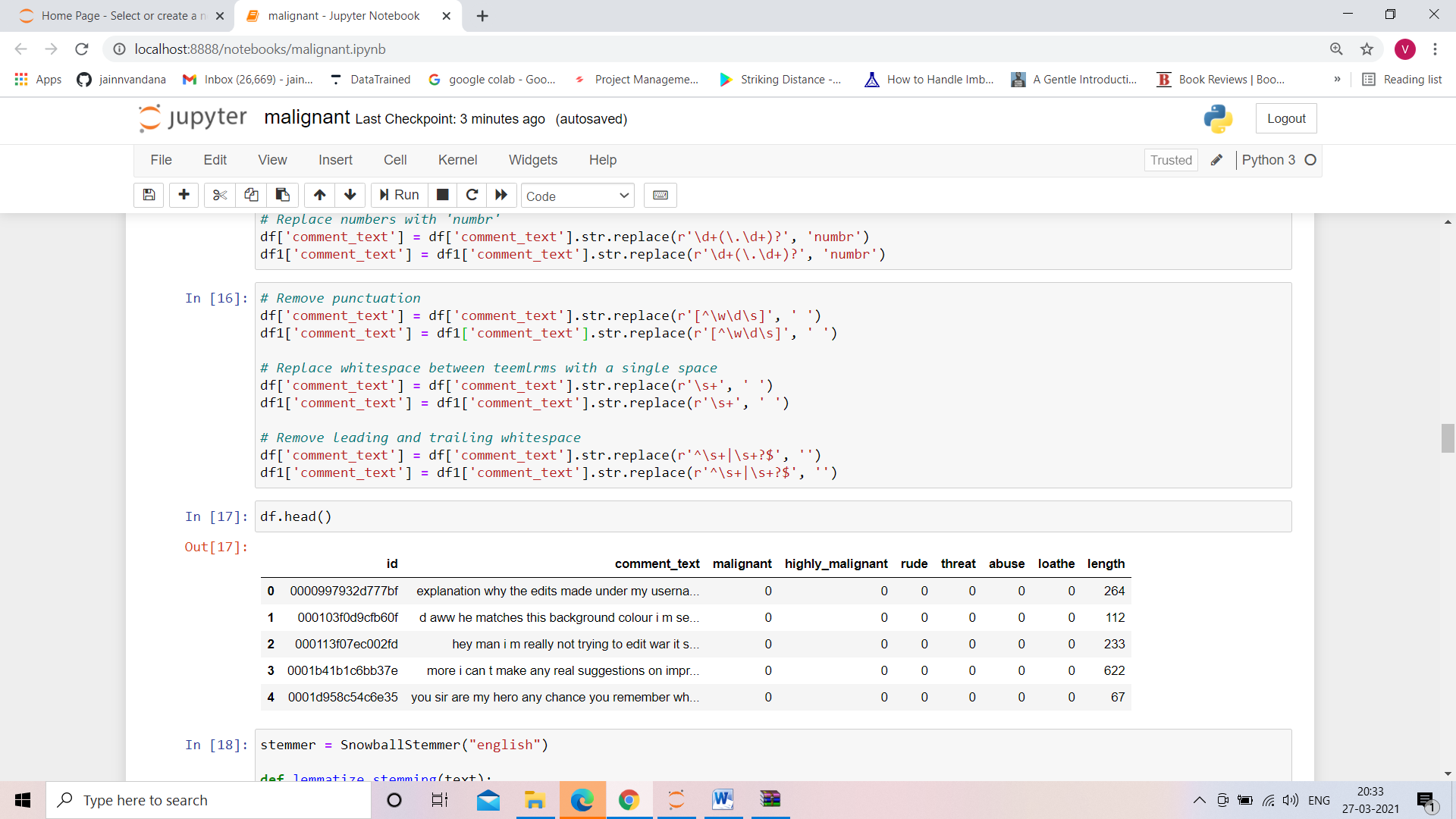
* + Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
  + Highly Malignant: It denotes comments that are highly malignant and hurtful.
  + Rude: It denotes comments that are very rude and offensive.
  + Threat: It contains indication of the comments that are giving any threat to someone.
  + Abuse: It is for comments that are abusive in nature.
  + Loathe: It describes the comments which are hateful and loathing in nature.
  + ID: It includes unique Ids associated with each comment text given.
  + Comment text: This column contains the comments extracted from various social media platforms.
* Data Preprocessing Done

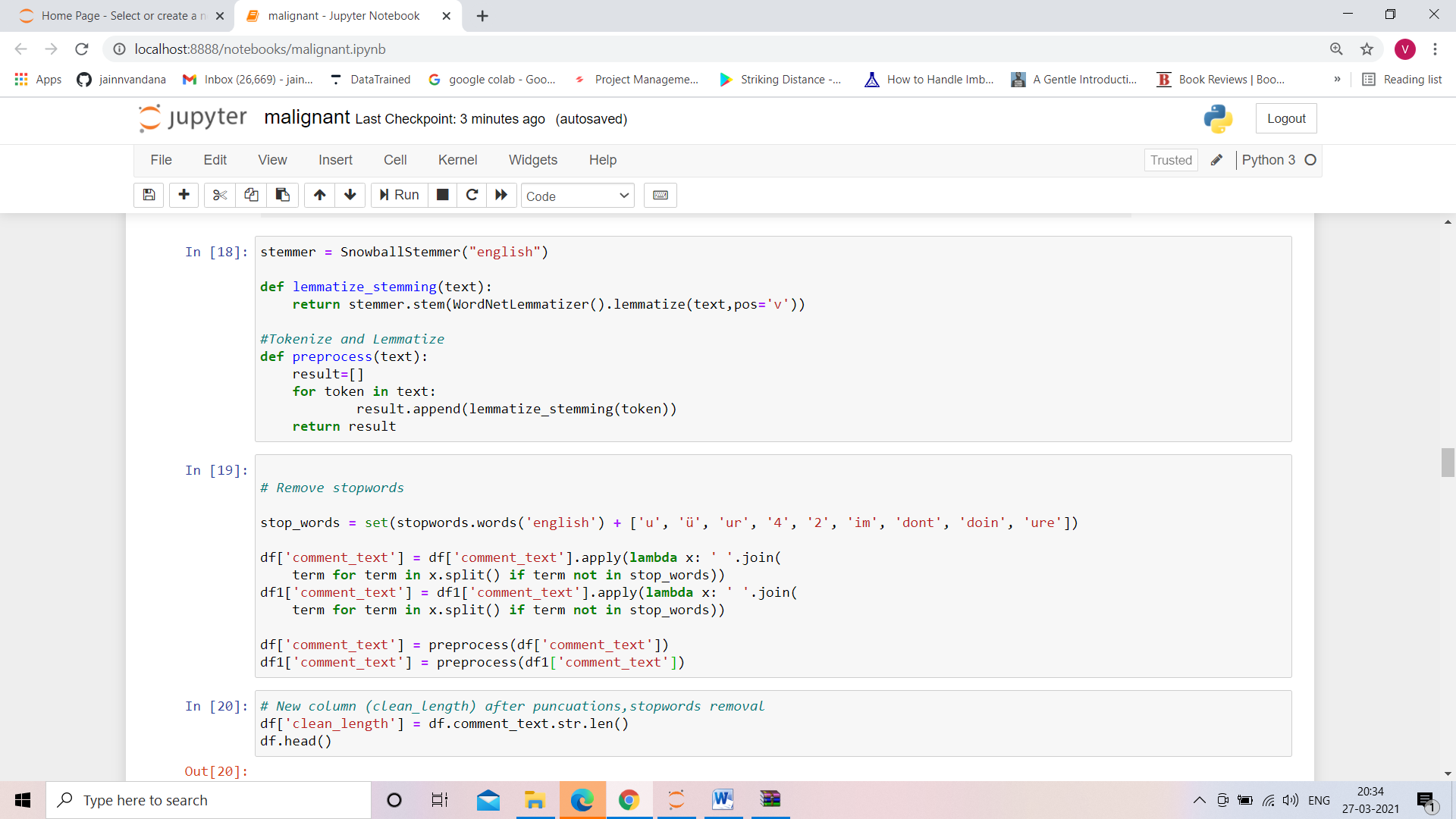
The dataset that will be used to train the model has some challenges. Text Cleaning is a very important step in machine learning because your data may contains a lot of noise and unwanted character such as punctuation, white space, numbers, hyperlink and etc.

Some standard procedures are:

* convert all letters to lower/upper case
* removing numbers
* removing punctuation
* removing white spaces
* removing hyperlink
* removing stop words such as a, about, above, down, doing and the list goes on… Sometimes, the extremely common word which would appear to be of very little value in helping select documents matching user need are excluded from the vocabulary entirely.
* **Word Stemming:** Stemming algorithms work by removing the end or the beginning of the words, using a list of common prefixes and suffixes that can be found in that language.
* **Word lemmatization**: Lemmatization is utilizing the dictionary of a particular language and tried to convert the words back to its base form. It will try to take into account of the meaning of the verbs and convert it back to the most suitable base form.





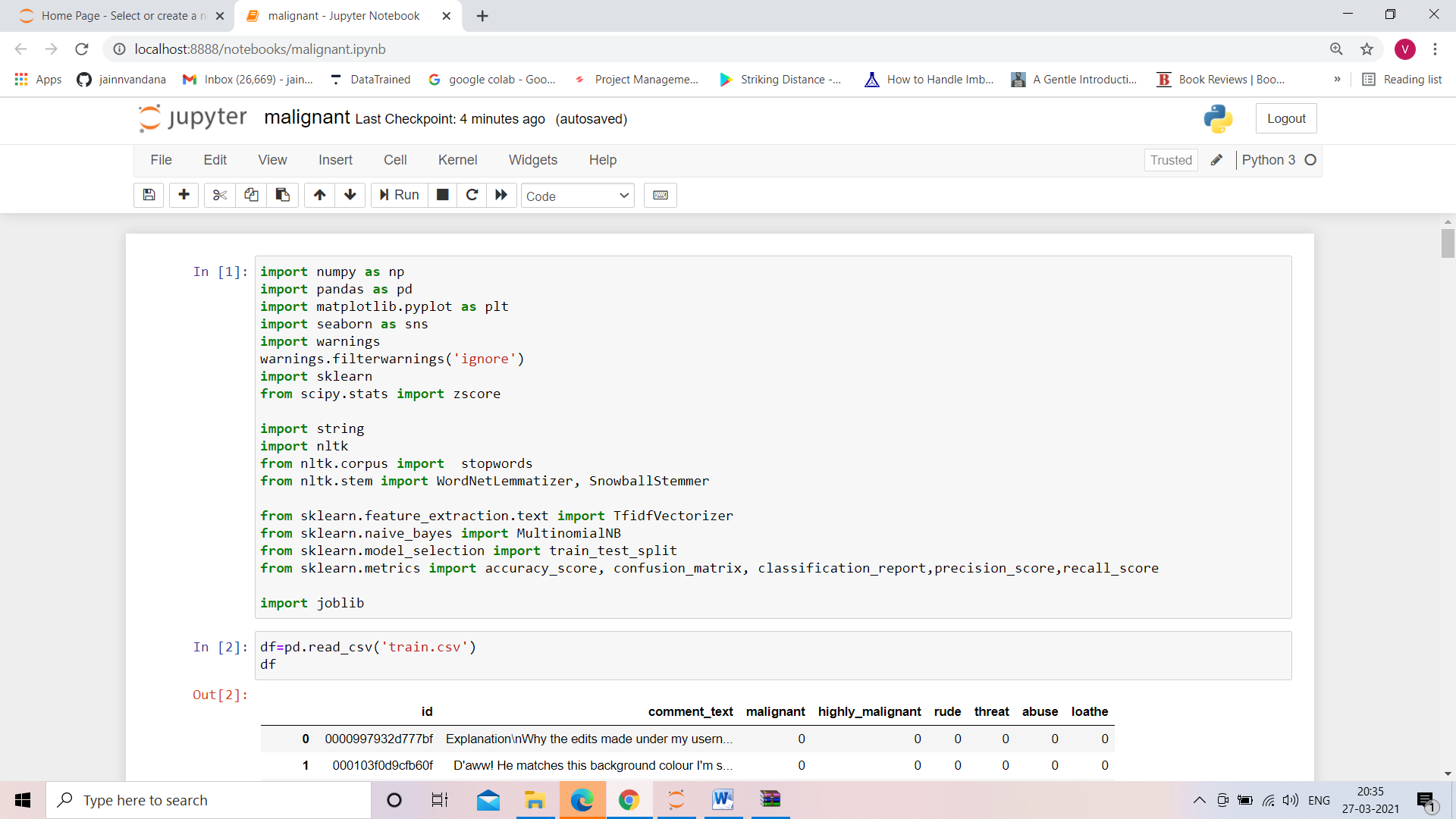


* Hardware and Software Requirements and Tools Used

Hardware : Since the computational aspect of the project is of importance to PANDA, it is important to know the hardware that was used in the evaluation process. The training and evaluation of the neural network model has been done on a Windows 10 computer using a quad-core CPU at i3.

Software : anaconda 3 , windows 10 ,Microsoft office.

Tools used : python , machine learning libraries, Nltk, Nlp libraries.

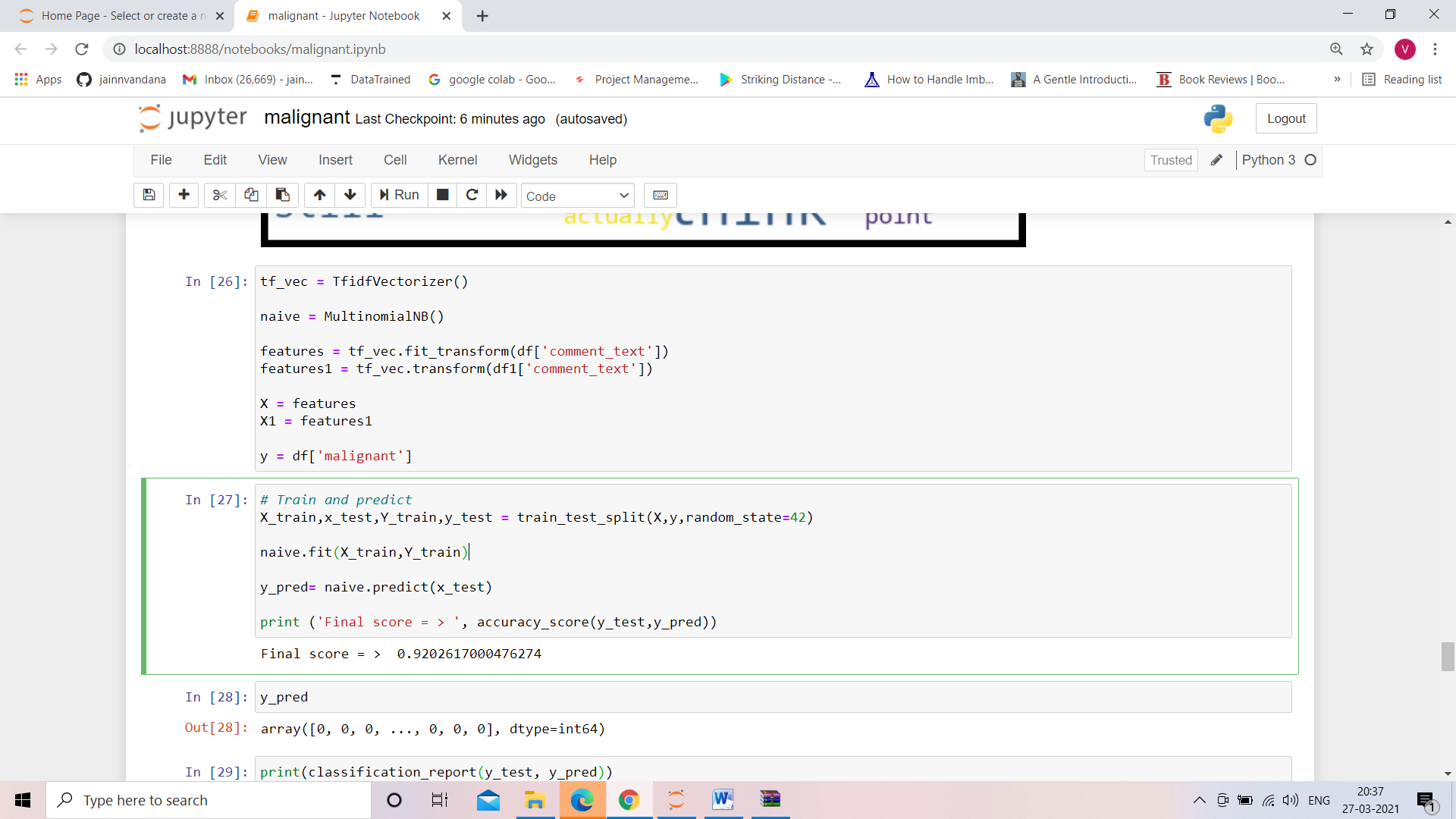


**Model/s Development and Evaluation**

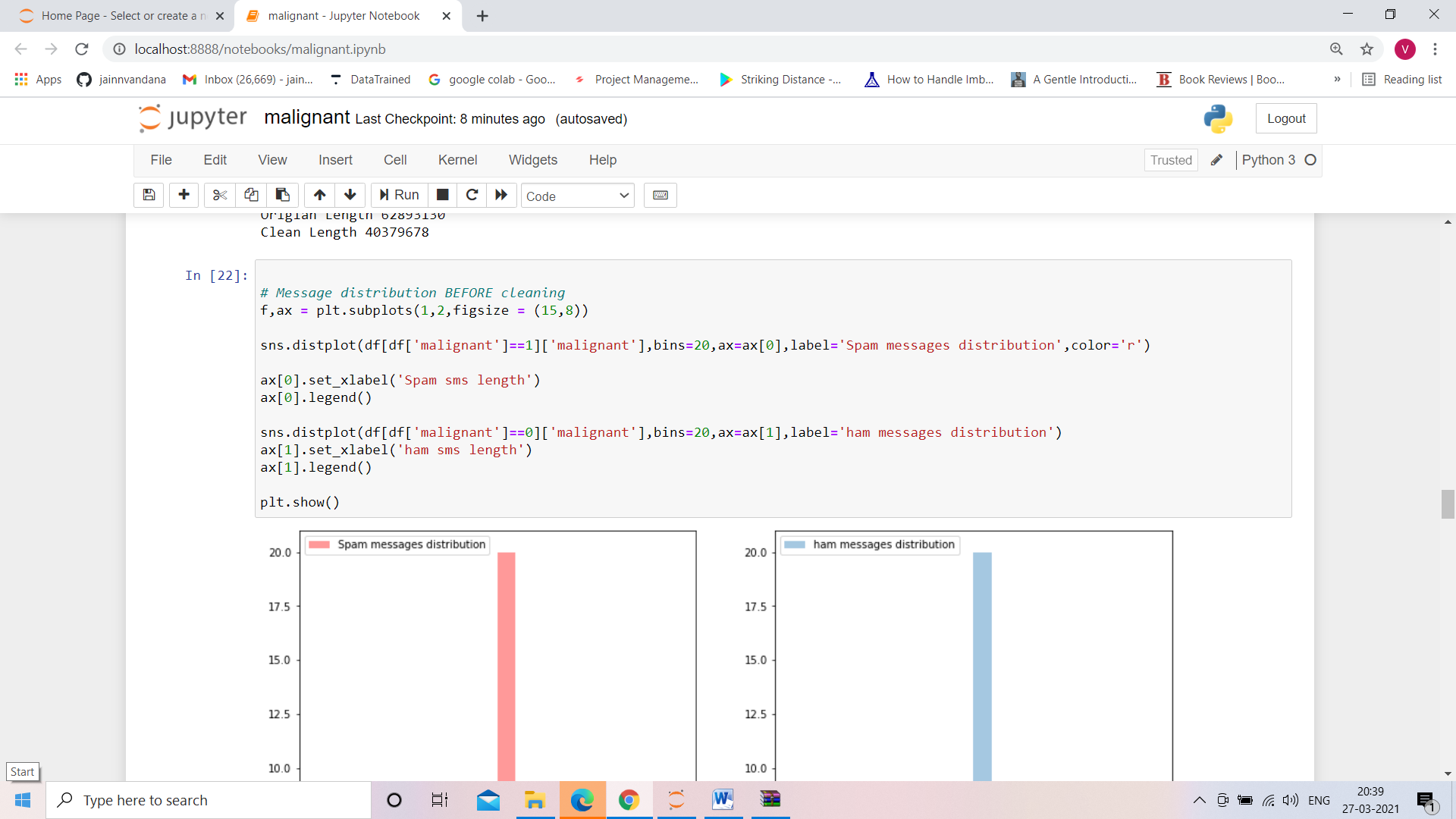
* Identification of possible problem-solving approaches (methods)

NAÏVE BAYS CLASSIFIER

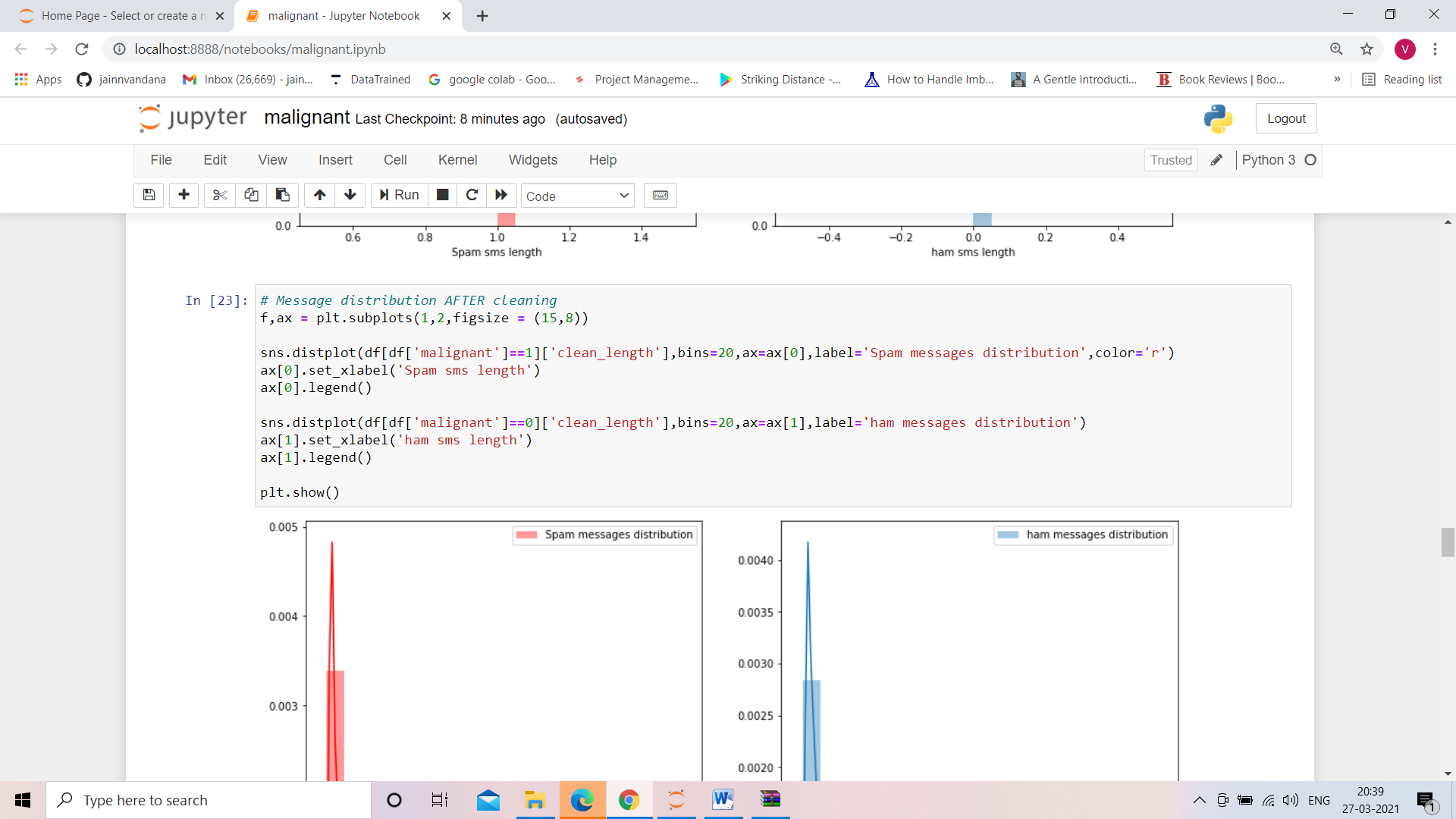
Simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset. Represent as a vector of feature values. It is very useful to classify the comments properly. The precision and recall of this method is known to be very effective.

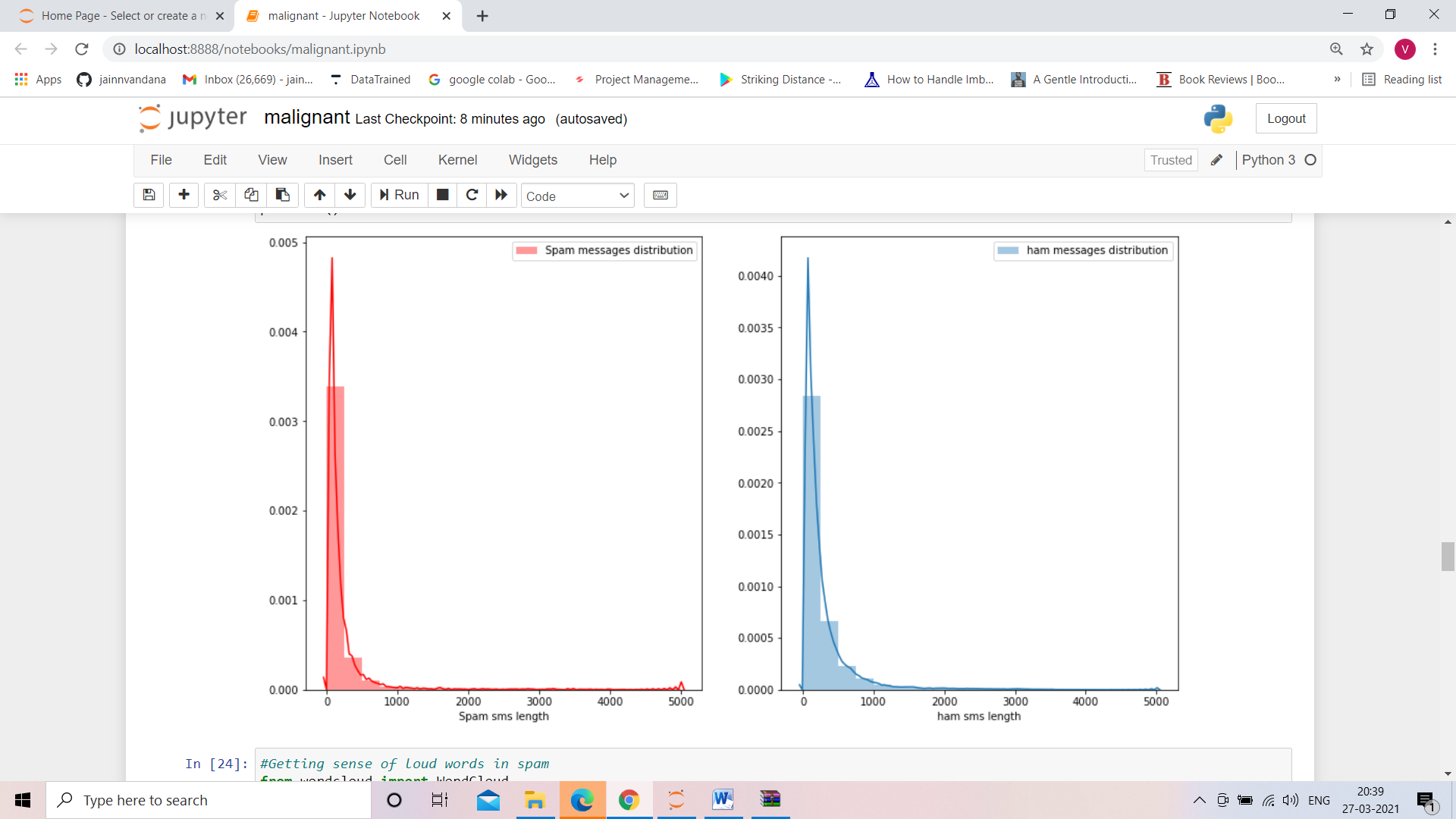


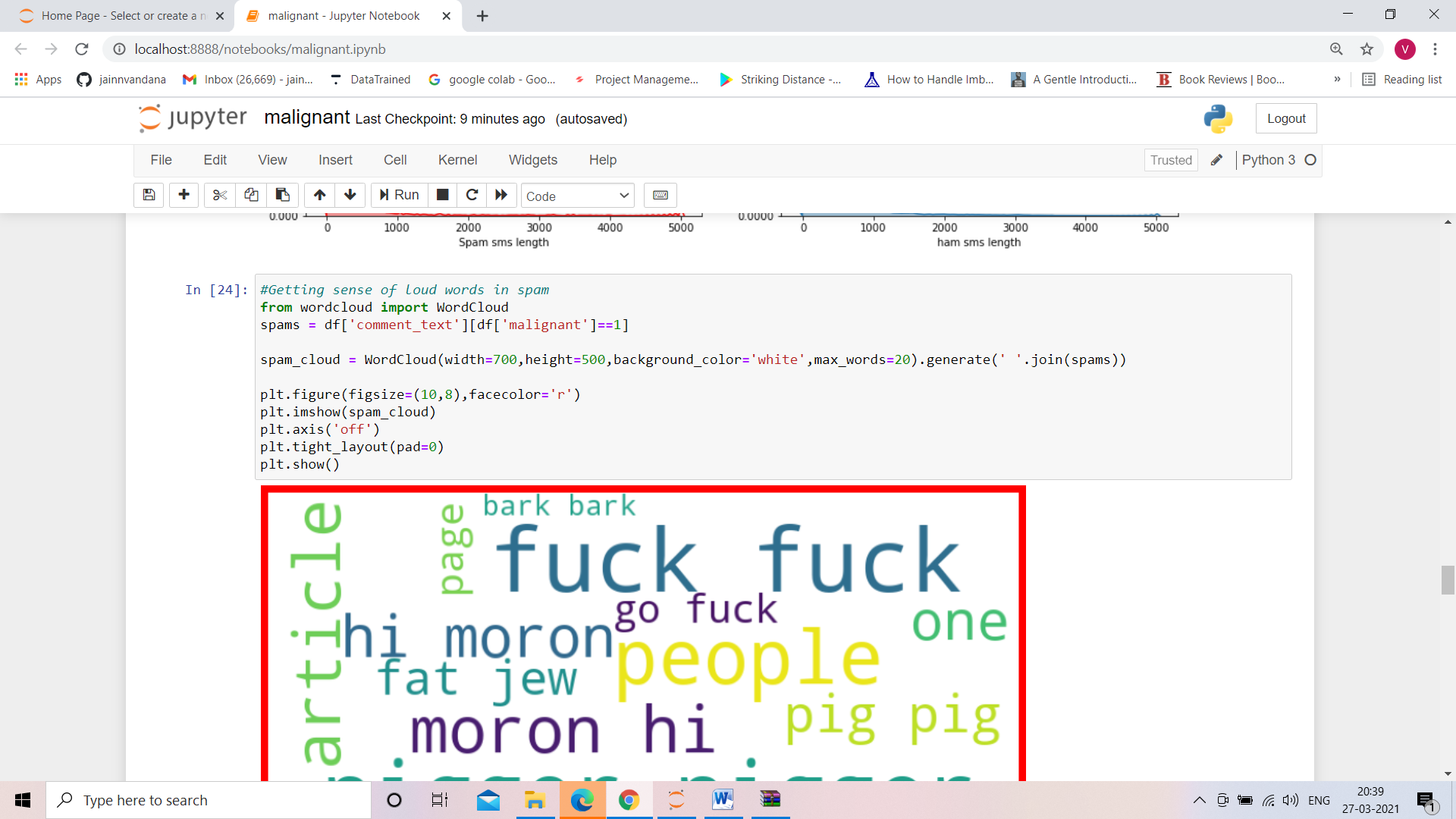
* Visualizations

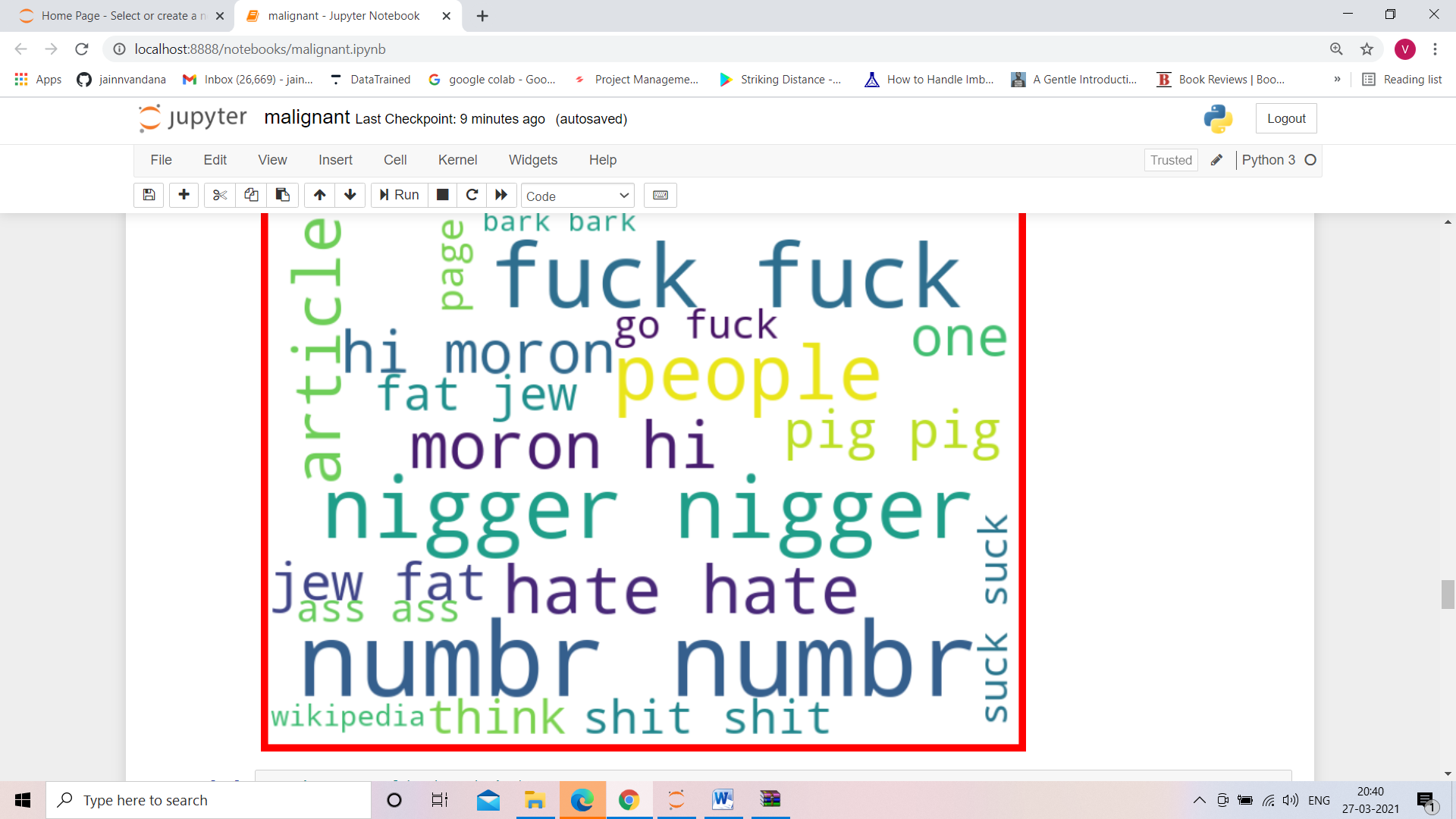


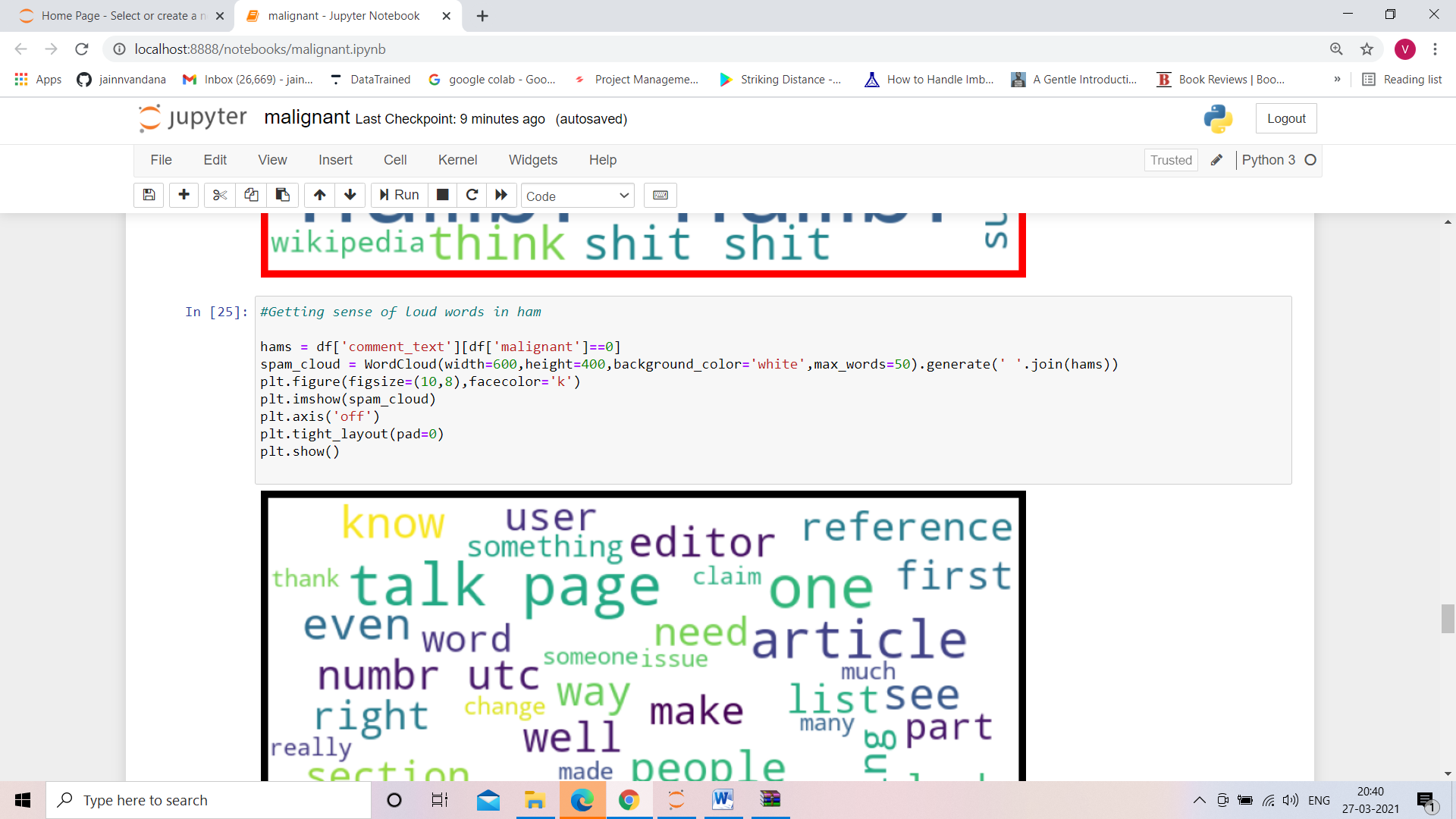


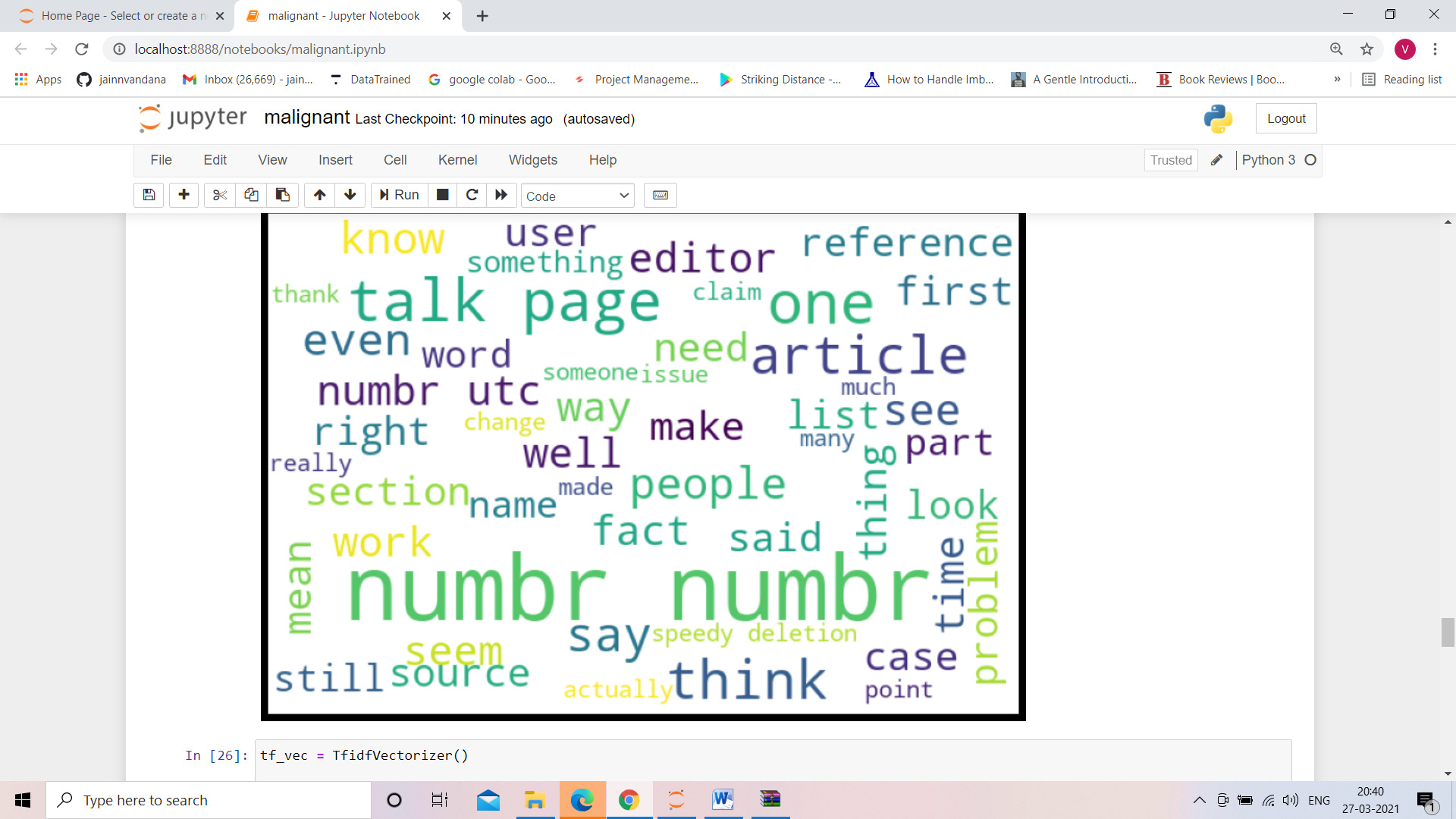












* Key Metrics for success in solving problem under consideration

When it comes to evaluation of a data science model’s performance, sometimes accuracy may not be the best indicator.

Some problems that we are solving in real life might have a very imbalanced class and using accuracy might not give us enough confidence to understand the algorithm’s performance.

In the malignant comments problem that we are trying to solve, the malignant comments data is approximately 10% of our data. If our algorithm predicts all the comments as non-malignant, it will achieve an accuracy of 92%. And for some problem that has only 1% of positive data, predicting all the sample as negative will give them an accuracy of 99% but we all know this kind of model is useless in a real life scenario.

Precision & Recall is the common evaluation metrics that people use when they are evaluating class-imbalanced classification model.

Let’s try to understand what questions Precision & Recall is trying to answer,

* Precision: What proportion of positive identifications was actually correct ?
* Recall: What actual proportion of actual positives was identified correctly?

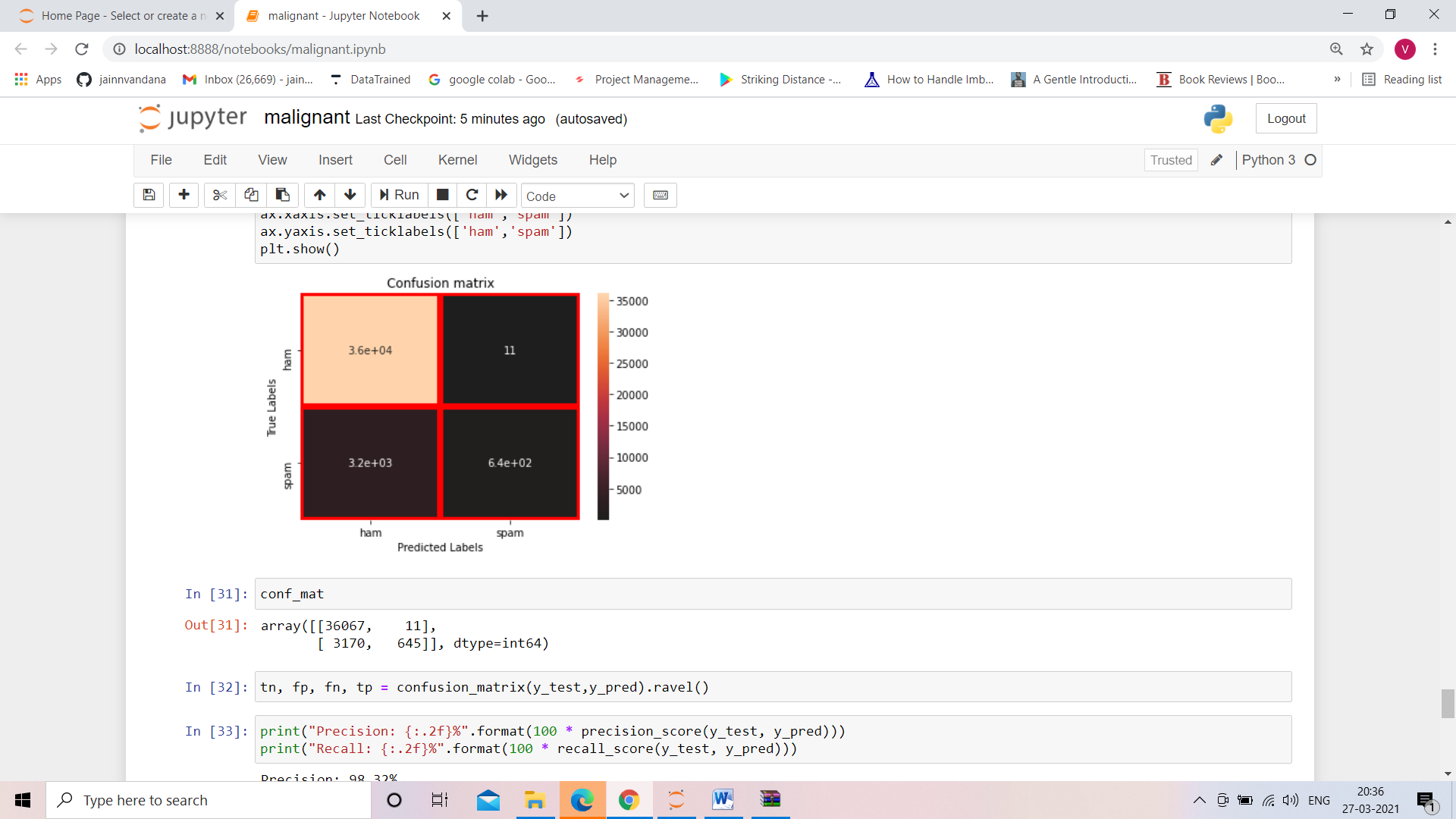
So, precision is evaluating, when a model predict something as positive, how accurate the model is. On the other hand, recall is evaluating how well a model in finding all the positive samples.

Confusion Matrix

Confusion Matrix is a very good way to understand results like true positive, false positive, true negative and so on.

Sklearn documentation has provided a sample code of how to plot nice looking confusion matrix to visualize your result..

Confusion Matrix of the result



**CONCLUSION**

Now that we have implemented the algorithm.

Label ‘1’ has approximately 10% records, while, label ‘0’ has approximately 90% records. Let’s have the results. We have implemented the dataset using Naive Bayes. We have the accuracy score of prediction of malignant comments is 92%.