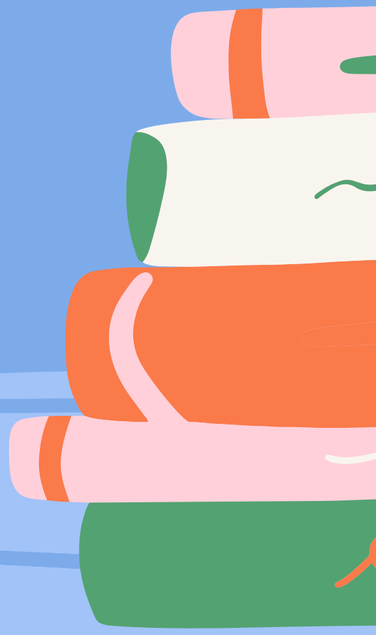
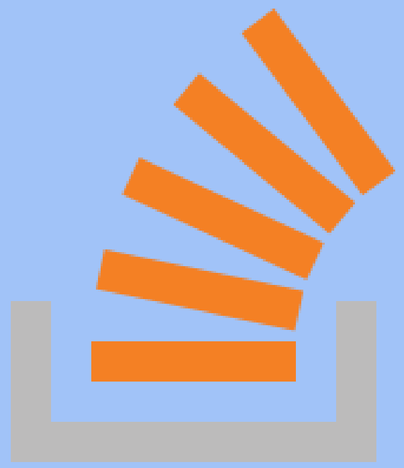
NATURAL LANGUAGE PROCESSING

Predict tags on StackOverflow with linear models

Pratik Chowdhury,
Elvis Dsouza,
Vedant Sahai



WHAT IS STACKOVERFLOW?



LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

OBTAINING DATASETS

We can easily find the datasets from the internet

We found a dataset online in TSV form, which we cleaned up by removing unneeded parameters with the help of Microsoft Excel to generate our current dataset
The dataset was loaded with the help of Pandas

LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

Perform Cleanup

Remove all the stopwords present. Perform cleanup

CONVERT THE SENTENCE TO IT'S SIMPLEST FORM

As our algorithm, which does not rely on order or context, removing common words or symbols will not cause any major issue

STOPWORD REMOVAL

Not every word will play a major role

Stopwords refer to the most common word which while provide context, play no role in a project like ours which does not rely on the context they can provide

REMOVE UNUSED SYMBOLS

Symbols will not be used by our dataset

{, }, [,] etc do not play any role in our project and hence can be removed safely

CHANGE CASE

CONVERT TO LOWER CASE

As upper and lower case would lead to different words, even if the meaning is the same, it is best to change to lower case so as to increase the size of the dataset instead of removing the words.

LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

Perform Cleanup

Remove all the stopwords present. Perform cleanup

Perform TF-IDF

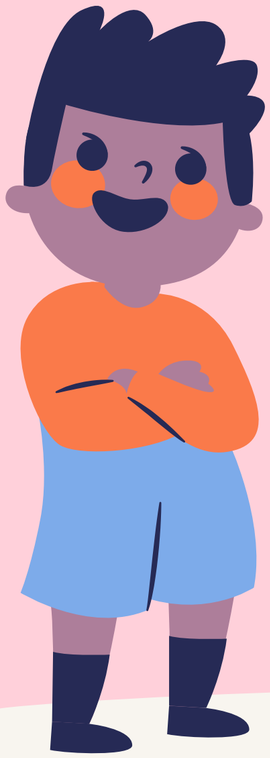
TF-IDF is used to find the relevance of a word within the document

TERM FREQUENCY

The frequency of a given word in a document

The weight of a word in a Document is simply proportional to it's Term Frequency

Term Frequency has a few pitfalls



INVERSE DOCUMENT FREQUENCY

It is a measure of how much information the word provides

As such, we check how rare or common it is within the document. It is the logarithmically scaled inverse fraction of the documents that contain the word (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient):

TF-IDF

Reflects how important a word is to a document in a collection or corpus

Multiply TF * IDF

CALCULATE TF-IDF

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

Perform Cleanup

Remove all the stopwords present. Perform cleanup

Perform TF-IDF

TF-IDF is used to find the relevance of a word within the document

Convert to Multi label Binarizer

Helps make it easier for Linear Model to comprehend the topic

MULTI-LABEL BINARIZER

Make it easier for the computer which does not understand labels like "git" but does understand 1 & 0

LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

Perform Cleanup

Remove all the stopwords present. Perform cleanup

Perform TF-IDF

TF-IDF is used to find the relevance of a word within the document

Convert to Multi label Binarizer

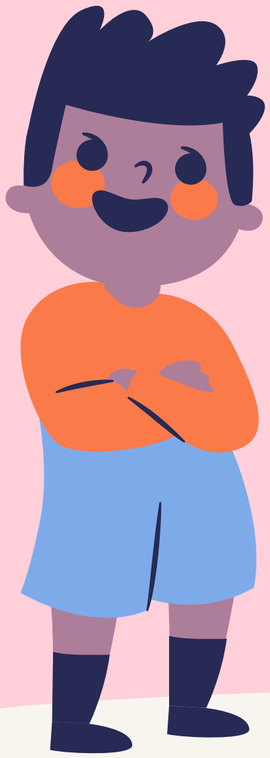
Helps make it easier for Linear Model to comprehend the topic

Train and Test our Linear Model

Check what went right and what went wrong

MULTINOMIAL LOGISTIC REGRESSION

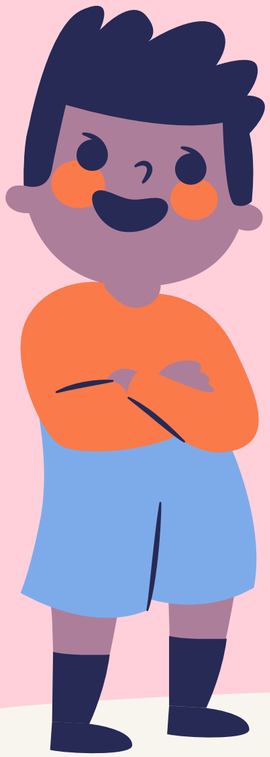
Easy to perform with the help of Scikit Learn



ONE vs REST

Helps perform Multiclass Classification

Compare one value (C++) with all of the other labels



LET US SOLVE THE PROBLEM

Obtain the Dataset

Obtain the dataset online. Train, test and validation

Perform Cleanup

Remove all the stopwords present. Perform cleanup

Perform TF-IDF

TF-IDF is used to find the relevance of a word within the document

Convert to Multi label Binarizer

Helps make it easier for Linear Model to comprehend the topic

Train and Test our Linear Model

Check what went right and what went wrong

Check the output

Based on the output, reach to a conclusion

Thank You