First I started by understanding the problem

Next we cleaned removing links w regex then symbols and mentions and numbers w regex.

Main reason for this is to explain new stuff like lemmatizer and how it was used it in tfidf.

Basically the tokenizer returns individual words so basically not setting anything there is just like saying ok just follow the token pattern and get all words. So to be able to use lemmatizer in tfidf i had to make it split everything into an individual word and then apply lemmatizer to the words.

The token pattern '(?u)[\\b\\w+\'\w+\\b](file:///\\b\\w+\'\w+\\b)' this is the one I used since the stopwords from nltk recognizes didn’t as didn’t I couldn’t remove all apostrophes so I made the tfidf recognize a token(individual word) as first a boundary (maybe space) then any amount of word then an apostrophe, then any amount of word and then a boundary.

We use \\ or r basically the regex is\w but \ in python is an escape character so r basically tells python to recognize it as a normal symbol or we could use \\ to achieve the same thing.