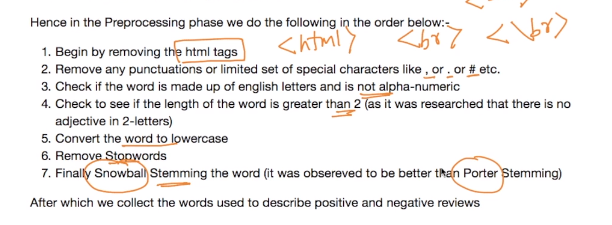
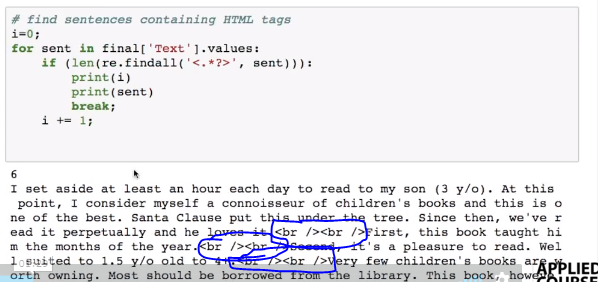
**After data cleaning we do text pre-processing which consist of multiple steps**





Above mentioned are the steps we will perform for text pre-processing.

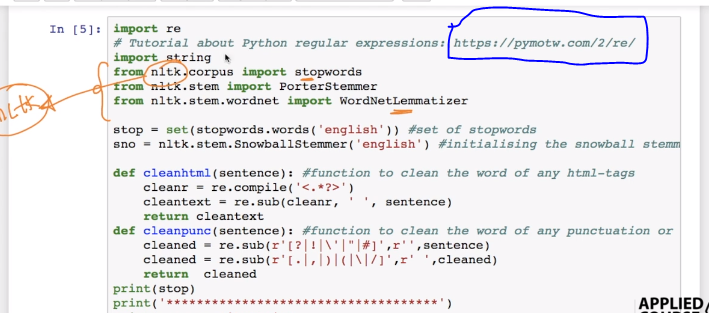
Firstly we need to remove all the html tags and we can do so by using below code.



Code to remove html tags and punctuations from any statements.

Function cleanhtml() is function which is removing html tags and cleanpunc() is function which is removing punctuations.

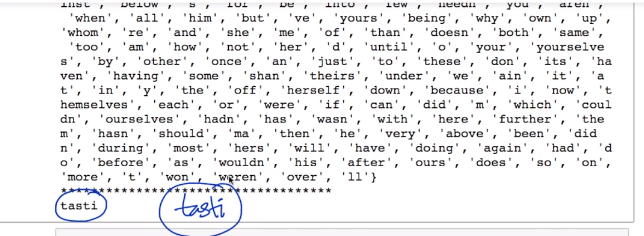
The blue block in below snapshot is link for tutorial of regular expression.



**In above code**



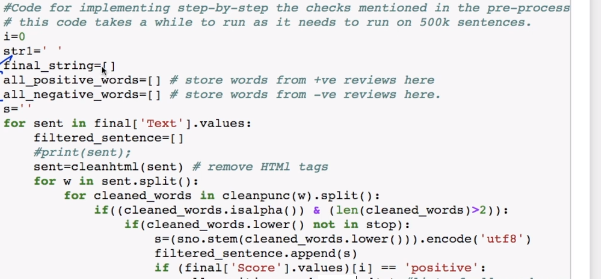
**The output of particular lines is as follows:**

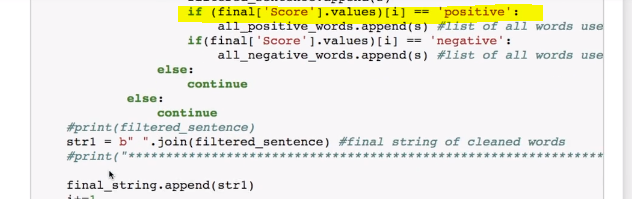


Basically printing stop gives us all he stop words scikit learn library has for English.

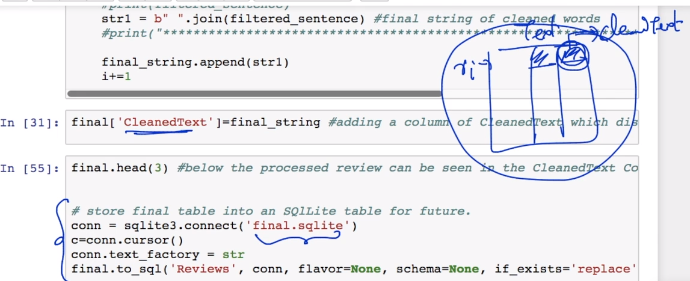
And stem word for tasty is **“tasti”**

Code for final pre-processing as per given steps at the start.





**Highlighted line is just to understand that this two images are appended from highlighted line.**



The Final string which we get after all the pre-processing is   
**“final\_string”** in above code and we are storing that sting in new column named **“CleanedText”** in final dataframe.

And than we are storing the intermediate dataframe to sqlite in last section of code.

As per updated notebook

Preprocessing Steps:

1. Remove URL’s: using regular expression
2. Remove html / extract text from html: using BeautifulSoup
3. Decontraction / expansion: using our function defined below
4. Remove words with digits (ex hello123): using regular expression
5. Remove Special characters: using regular expression
6. Remove stopwords: using stopwords provided by nltk

