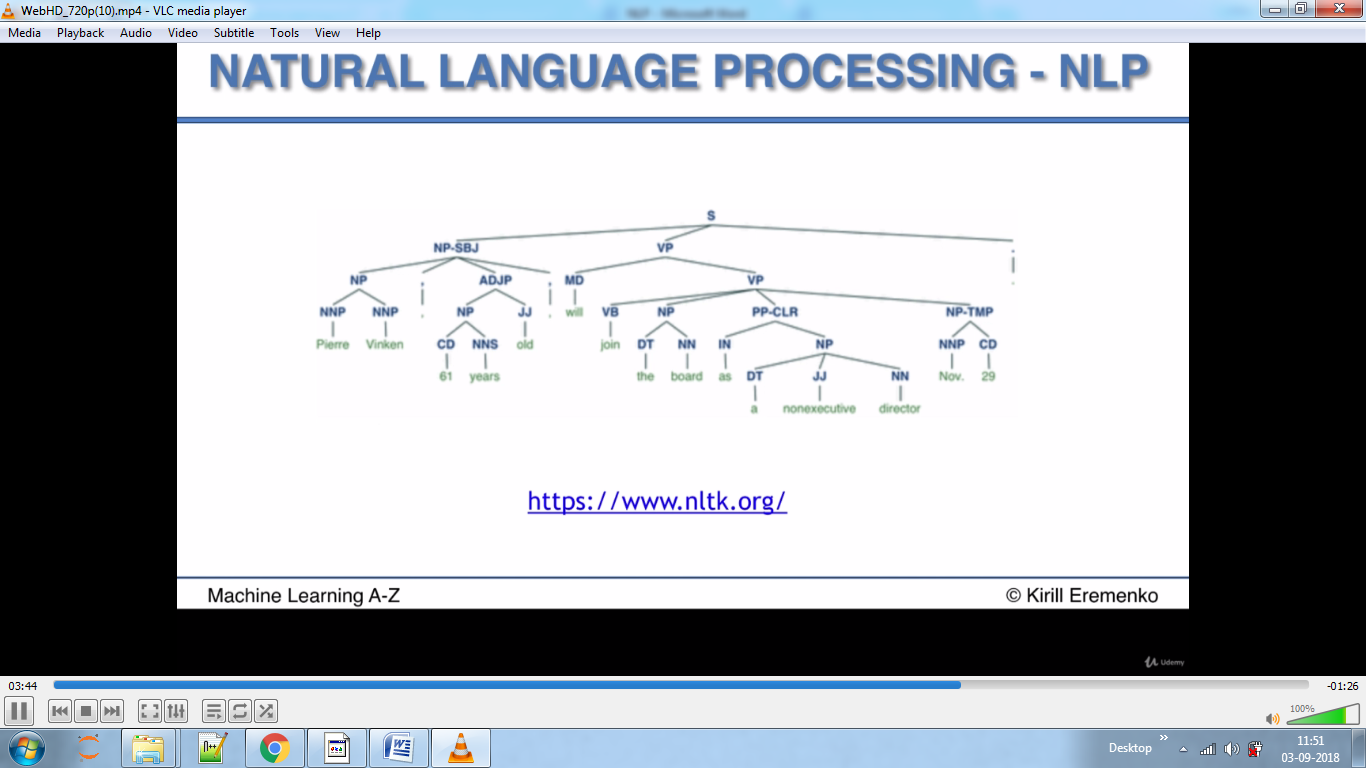
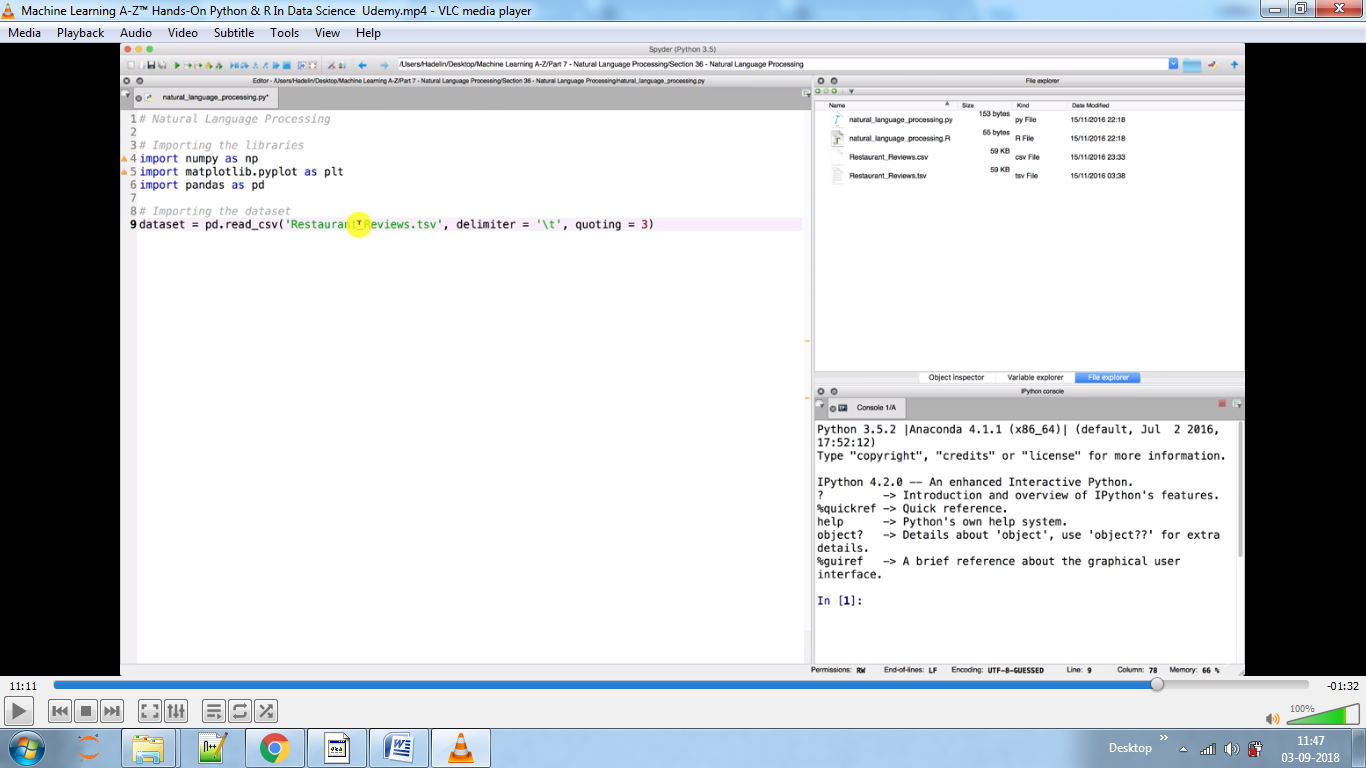


Below is POS –park of speech tag where the sentence gets broken down into parts of speech like adjective ,noun and Cardinal number.



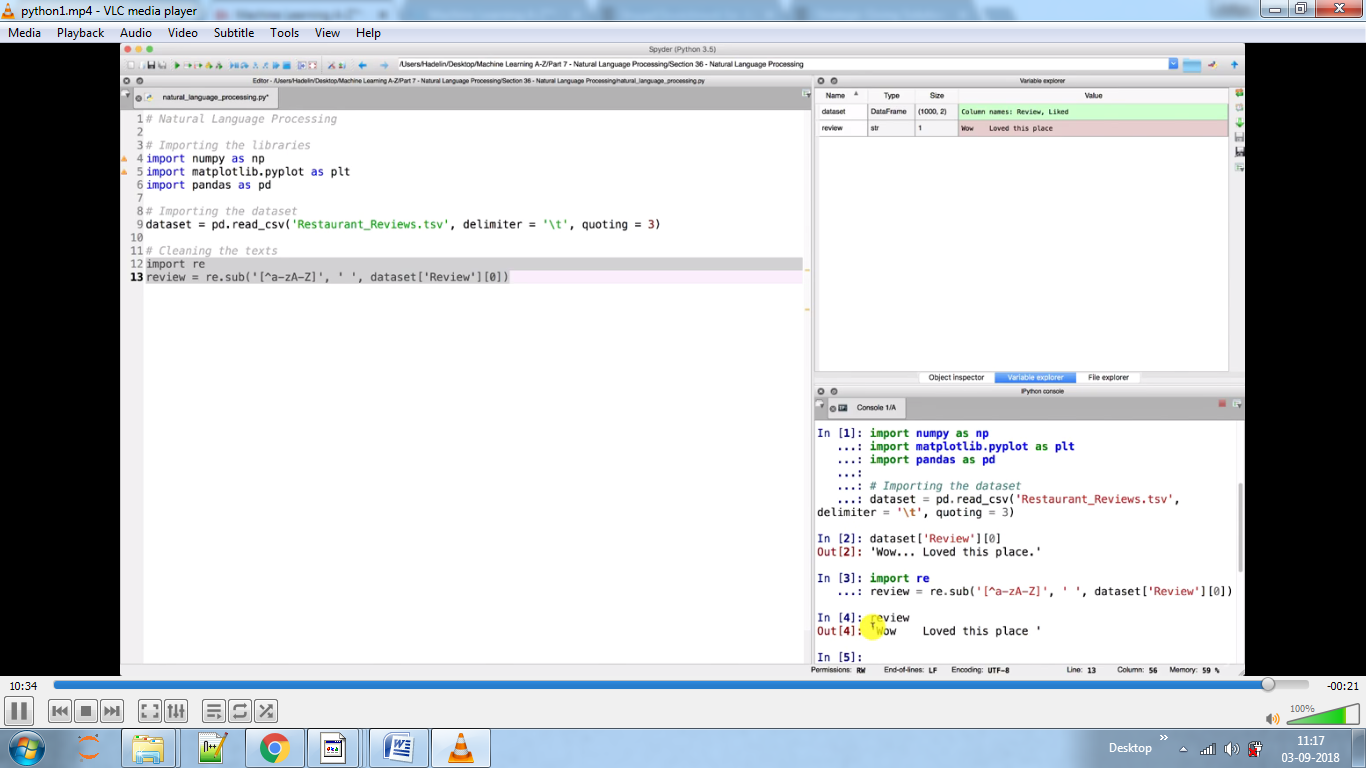
We use tsv file instead of csv ,as tab separated file will not be contaminated like the csv files get as reviews may contain commas so pandas will attach comma with the words in different columns which we don’t want,hence we use tsv as no person will ever put tab but they may put a comma in their review.

Quoting parameter says to ignore double quotes in the file.



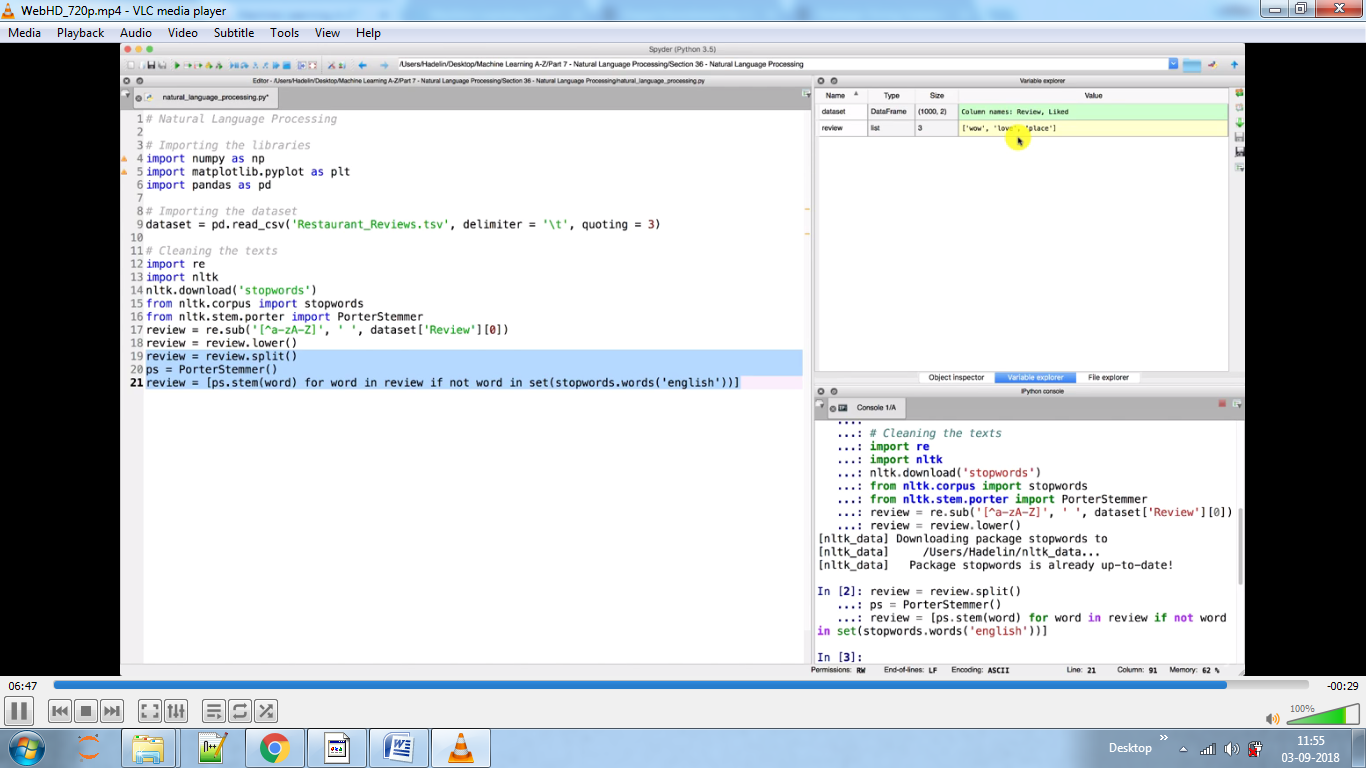
Here we will get rid of unnecessary words like the, a punctuation, numbers not relevant and taking words such as love, loved as love only

The sub removes numbers, punctuation which accept arguments which we want.



Here we want to separate words, hence we used 2nd argument, we don’t want <all the> to be<allthe>

PorterStemmer is used to convert loved into love



Creating bag of words model using tokenisation, here we have created table where each column represents words. If column has a value of 1 then that entry has that word in it.

We used max\_feature because we wanted to eliminate names, places which are irrelevant and hence randomly we took top 1500 words at a time.

Most times we use Naive bayes, Tree models classification for training NLP.Other models we can use are CART,C5.0 ,maximum entropy.