*Start a new Python project folder … do not reuse the previous workspace … As with any keyboard-driven console-like environment, developing muscle -memory for the common commands is also part of the learning curve.*

**Stemming** is the process of reducing inflected (or sometimes derived) words to their word stem or root. Crudely chopping off the end of the word to leave only the base. This is another step in the filtering process by reducing variations of the same root word.

**Stemming examples** - the words in each set had similar meanings

Let us reduce (chop) them to its semantic meaning ...

Stemming / stemmed → Stem

Electricity / electrical → Electr

Berries / berry → Berri

Connection / connected /connective → Connect

There are cases where stemming will be wrong because the words had different meaning …

Meanness / meaning → Mean

So stemming is not always right.

Why do we care? Because it saves memory: keeping many variations as one in memory.

* Reduces the corpus of words the model is exposed to
* Explicitly correlates words with similar meanings

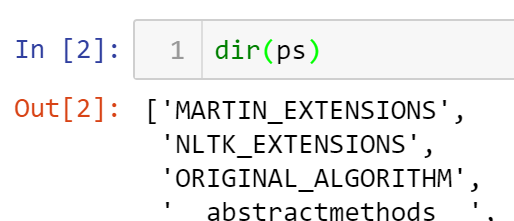
What are some stemmers?

* Porter Stemmer (the most popular)
* Snowball Stemmer
* Lancaster Stemmer
* Regex-Based Stemmer

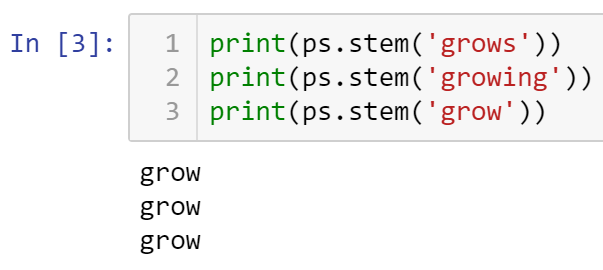
**The Porter Stemmer**



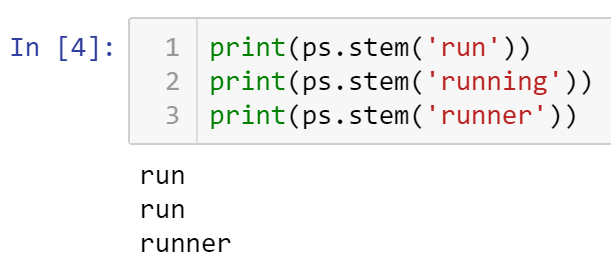
Checking what attributes and methods to use …

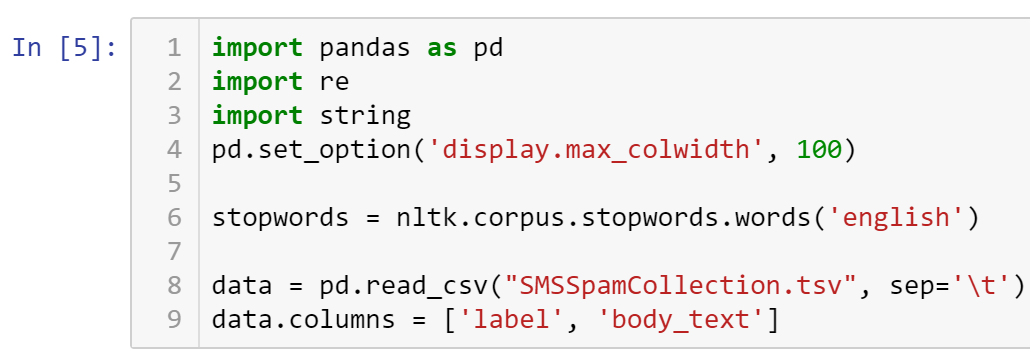


‘stem’ method is one of the member of ps.



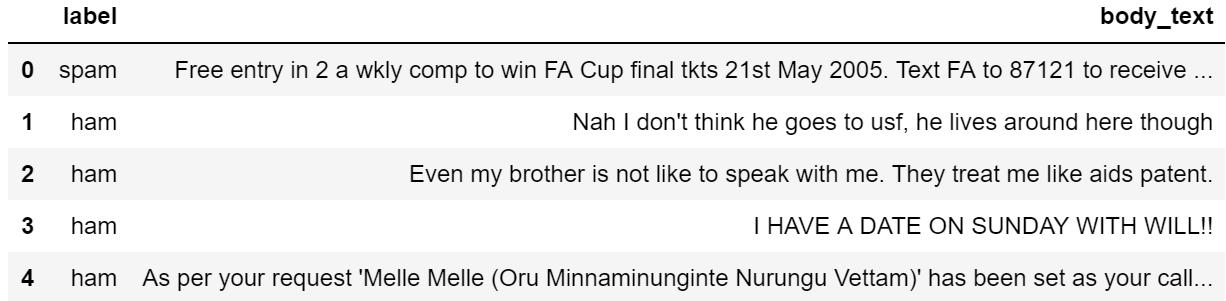
Although stemmer is not perfect, it knows that the last word below is different from the first two.





Printing the first 5 rows …

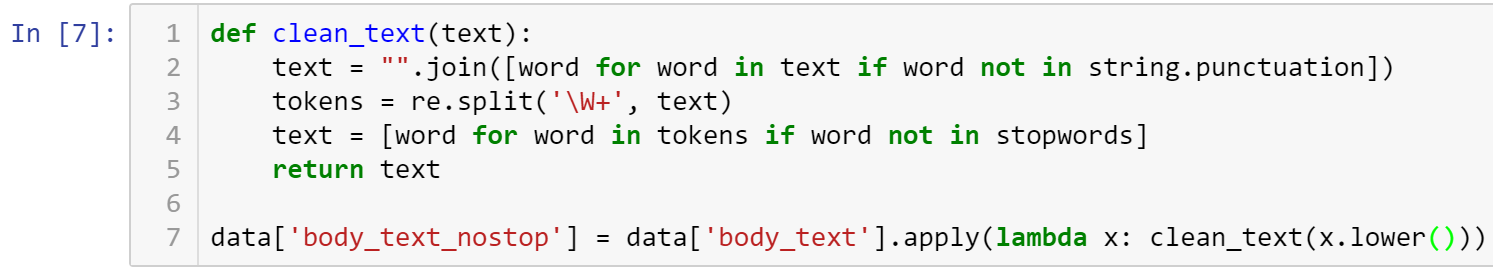




Time to clean up this text …

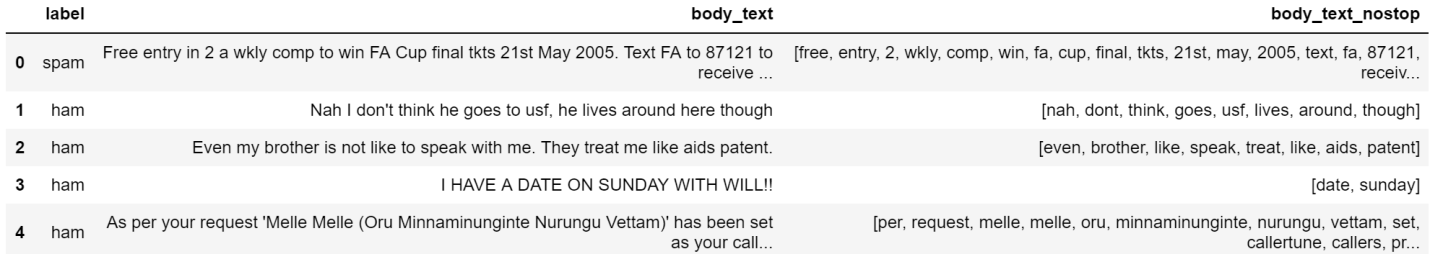
Create a function that removes punctuation, tokenize it, and remove stop words.

Then apply the function to the data column using lambda function.



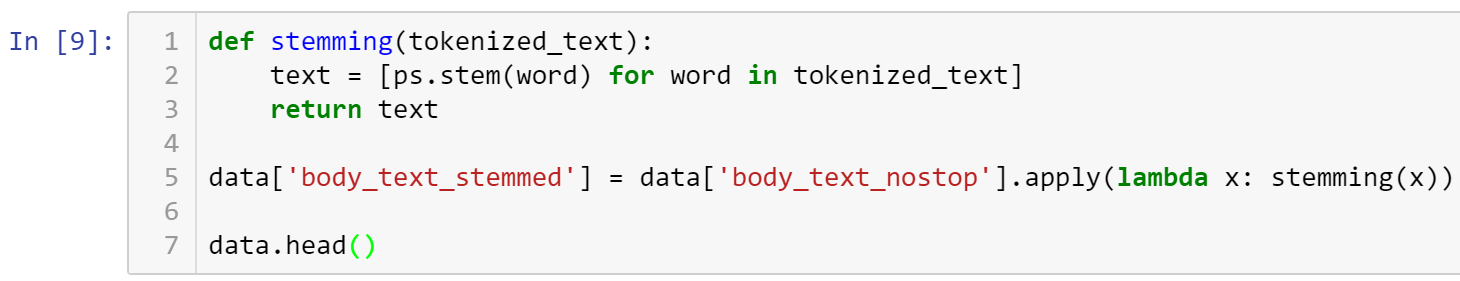
Verify that the clean\_text() function is working by printing the first 5 rows …

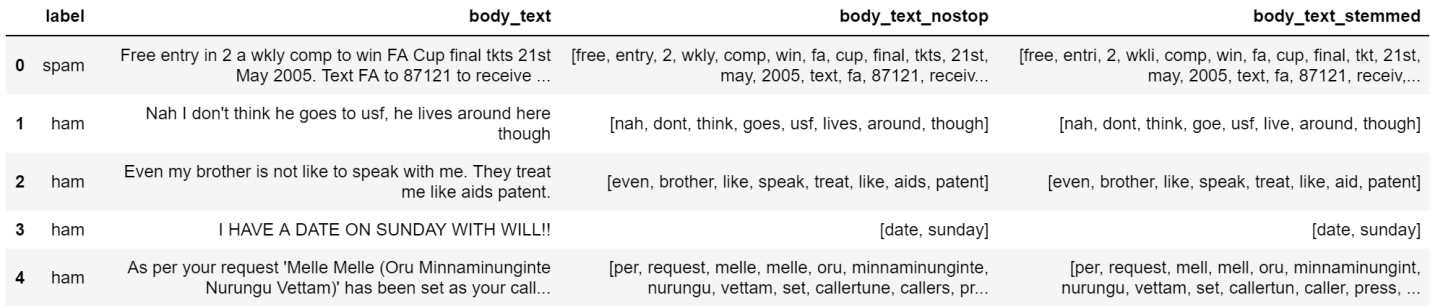




Now, let’s stem the sanitized data …

Create a function that will return the stem word for each word in the tokenized text.





Study the stemmed words and make observations. For example, lives is stemmed as live.

* All submissions should be separate from other exercises and quests. Please do not lump all your answers into one document and re-using that same workspace to gain multiple points. Thanks.
* Place your name at the bottom of your code, download your Python program in html format, and submit your work in Canvas.