I want us to take a moment to peak into the corpora we all downloaded! The NLTK corpus is a massive dump of all kinds of natural language data sets that are definitely worth taking a look at.

Almost all of the files in the NLTK corpus follow the same rules for accessing them by using the NLTK module, but nothing is magical about them. These files are plain text files for the most part, some are XML and some are other formats, but they are all accessible by you manually, or via the module and Python. Let's talk about viewing them manually.

Depending on your installation, your nltk\_data directory might be hiding in a multitude of locations. To figure out where it is, head to your Python directory, where the NLTK module is. If you do not know where that is, use the following code:

import nltk

print(nltk.\_\_file\_\_)

Run that, and the output will be the location of the NLTK module's \_\_init\_\_.py. Head into the NLTK directory, and then look for the data.py file.

The important blurb of code is:

if sys.platform.startswith('win'):

# Common locations on Windows:

path += [

str(r'C:\nltk\_data'), str(r'D:\nltk\_data'), str(r'E:\nltk\_data'),

os.path.join(sys.prefix, str('nltk\_data')),

os.path.join(sys.prefix, str('lib'), str('nltk\_data')),

os.path.join(os.environ.get(str('APPDATA'), str('C:\\')), str('nltk\_data'))

]

else:

# Common locations on UNIX & OS X:

path += [

str('/usr/share/nltk\_data'),

str('/usr/local/share/nltk\_data'),

str('/usr/lib/nltk\_data'),

str('/usr/local/lib/nltk\_data')

]

There, you can see the various possible directories for the nltk\_data. If you're on Windows, chances are it is in your appdata, in the local directory. To get there, you will want to open your file browser, go to the top, and type in %appdata%

Next click on roaming, and then find the nltk\_data directory. In there, you will have your corpora file. The full path is something like:   
C:\Users\yourname\AppData\Roaming\nltk\_data\corpora

Within here, you have all of the available corpora, including things like books, chat logs, movie reviews, and a whole lot more.

Now, we're going to talk about accessing these documents via NLTK. As you can see, these are mostly text documents, so you could just use normal Python code to open and read documents. That said, the NLTK module has a few nice methods for handling the corpus, so you may find it useful to use their methology. Here's an example of us opening the Gutenberg Bible, and reading the first few lines:

from nltk.tokenize import sent\_tokenize, PunktSentenceTokenizer

from nltk.corpus import gutenberg

# sample text

sample = gutenberg.raw("bible-kjv.txt")

tok = sent\_tokenize(sample)

for x in range(5):

print(tok[x])

One of the more advanced data sets in here is "wordnet." Wordnet is a collection of words, definitions, examples of their use, synonyms, antonyms, and more. We'll dive into using wordnet later.