Module 2 (Python 3)

June 8, 2019

1 Module 2 (Python 3)

1.1 Basic NLP Tasks with NLTK

```
In [ ]: import nltk
        nltk.download('book')
```

```
[nltk_data] Downloading collection 'book'
[nltk_data]    |
[nltk_data]    | Downloading package abc to /home/jovyan/nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package brown to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/conll2000.zip.
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/conll2002.zip.
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/dependency_treebank.zip.
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/genesis.zip.
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/gutenberg.zip.
[nltk_data]    | Downloading package ieer to /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/ieer.zip.
```

```
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/inaugural.zip.
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    |   Unzipping corpora/movie_reviews.zip.
[nltk_data]    | Downloading package nps_chat to
[nltk_data]    |     /home/jovyan/nltk_data...
[nltk_data]    | Error downloading 'nps_chat' from
[nltk_data]    |     <https://raw.githubusercontent.com/nltk/nltk_data
[nltk_data]    |     /gh-pages/packages/corpora/nps_chat.zip>:
[nltk_data]    |     <urlopen error [Errno 110] Connection timed out>
```

Out[ ]: False

In [ ]:
```
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)

/opt/conda/lib/python3.6/site-packages/nltk/corpus/util.py in __load(self)
     79             except LookupError as e:
---> 80                 try: root = nltk.data.find('{}/{}'.format(self.subdir, zip_name))
     81                 except LookupError: raise e


/opt/conda/lib/python3.6/site-packages/nltk/data.py in find(resource_name, paths)
    674     resource_not_found = '\n%s\n%s\n%s\n' % (sep, msg, sep)
--> 675     raise LookupError(resource_not_found)
    676


LookupError:
**********************************************************************
  Resource nps_chat not found.
  Please use the NLTK Downloader to obtain the resource:
```

```
  >>> import nltk
  >>> nltk.download('nps_chat')

  Searched in:
    - '/home/jovyan/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/local/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/local/lib/nltk_data'
    - '/opt/conda/nltk_data'
    - '/opt/conda/share/nltk_data'
    - '/opt/conda/lib/nltk_data'
**********************************************************************



During handling of the above exception, another exception occurred:


    LookupError                               Traceback (most recent call last)

    <ipython-input-5-2a5ed6d8b2c3> in <module>()
----> 1 from nltk.book import *


    /opt/conda/lib/python3.6/site-packages/nltk/book.py in <module>()
     31 print("text4:", text4.name)
     32
---> 33 text5 = Text(nps_chat.words(), name="Chat Corpus")
     34 print("text5:", text5.name)
     35


    /opt/conda/lib/python3.6/site-packages/nltk/corpus/util.py in __getattr__(self, attr)
    114             raise AttributeError("LazyCorpusLoader object has no attribute '__bases_
    115
--> 116         self.__load()
    117         # This looks circular, but its not, since __load() changes our
    118         # __class__ to something new:


    /opt/conda/lib/python3.6/site-packages/nltk/corpus/util.py in __load(self)
     79             except LookupError as e:
     80                 try: root = nltk.data.find('{}/{}'.format(self.subdir, zip_name))
---> 81                 except LookupError: raise e
     82
     83         # Load the corpus.
```

```
    /opt/conda/lib/python3.6/site-packages/nltk/corpus/util.py in __load(self)
     76         else:
     77             try:
---> 78                 root = nltk.data.find('{}/{}'.format(self.subdir, self.__name))
     79             except LookupError as e:
     80                 try: root = nltk.data.find('{}/{}'.format(self.subdir, zip_name))


    /opt/conda/lib/python3.6/site-packages/nltk/data.py in find(resource_name, paths)
    673     sep = '*' * 70
    674     resource_not_found = '\n%s\n%s\n%s\n' % (sep, msg, sep)
--> 675     raise LookupError(resource_not_found)
    676
    677


    LookupError:
**********************************************************************
  Resource nps_chat not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('nps_chat')

  Searched in:
    - '/home/jovyan/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/local/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/local/lib/nltk_data'
    - '/opt/conda/nltk_data'
    - '/opt/conda/share/nltk_data'
    - '/opt/conda/lib/nltk_data'
**********************************************************************
```

1.1.1 Counting vocabulary of words

```
In [ ]: text7
```

```
In [ ]: sent7
```

```
In [ ]: len(sent7)
```

```
In [ ]: len(text7)
```

```
In [ ]: len(set(text7))

In [ ]: list(set(text7))[:10]
```

1.1.2 Frequency of words

```
In [ ]: dist = FreqDist(text7)
        len(dist)

In [ ]: vocab1 = dist.keys()
        #vocab1[:10]
        # In Python 3 dict.keys() returns an iterable view instead of a list
        list(vocab1)[:10]

In [ ]: dist['four']

In [ ]: freqwords = [w for w in vocab1 if len(w) > 5 and dist[w] > 100]
        freqwords
```

1.1.3 Normalization and stemming

```
In [ ]: input1 = "List listed lists listing listings"
        words1 = input1.lower().split(' ')
        words1

In [ ]: porter = nltk.PorterStemmer()
        [porter.stem(t) for t in words1]
```

1.1.4 Lemmatization

```
In [ ]: udhr = nltk.corpus.udhr.words('English-Latin1')
        udhr[:20]

In [ ]: [porter.stem(t) for t in udhr[:20]] # Still Lemmatization

In [ ]: WNlemma = nltk.WordNetLemmatizer()
        [WNlemma.lemmatize(t) for t in udhr[:20]]
```

1.1.5 Tokenization

```
In [ ]: text11 = "Children shouldn't drink a sugary drink before bed."
        text11.split(' ')

In [ ]: nltk.word_tokenize(text11)

In [ ]: text12 = "This is the first sentence. A gallon of milk in the U.S. costs $2.99. Is this
        sentences = nltk.sent_tokenize(text12)
        len(sentences)

In [ ]: sentences
```

1.2 Advanced NLP Tasks with NLTK

1.2.1 POS tagging

```
In [ ]: nltk.help.upenn_tagset('MD')

In [ ]: text13 = nltk.word_tokenize(text11)
        nltk.pos_tag(text13)

In [ ]: text14 = nltk.word_tokenize("Visiting aunts can be a nuisance")
        nltk.pos_tag(text14)

In [ ]: # Parsing sentence structure
        text15 = nltk.word_tokenize("Alice loves Bob")
        grammar = nltk.CFG.fromstring("""
        S -> NP VP
        VP -> V NP
        NP -> 'Alice' | 'Bob'
        V -> 'loves'
        """)

        parser = nltk.ChartParser(grammar)
        trees = parser.parse_all(text15)
        for tree in trees:
            print(tree)

In [ ]: text16 = nltk.word_tokenize("I saw the man with a telescope")
        grammar1 = nltk.data.load('mygrammar.cfg')
        grammar1

In [ ]: parser = nltk.ChartParser(grammar1)
        trees = parser.parse_all(text16)
        for tree in trees:
            print(tree)

In [ ]: from nltk.corpus import treebank
        text17 = treebank.parsed_sents('wsj_0001.mrg')[0]
        print(text17)
```

1.2.2 POS tagging and parsing ambiguity

```
In [ ]: text18 = nltk.word_tokenize("The old man the boat")
        nltk.pos_tag(text18)

In [ ]: text19 = nltk.word_tokenize("Colorless green ideas sleep furiously")
        nltk.pos_tag(text19)
```