

Lab 2: NLTK, SpaCy, and the UMLS

Michael Hogarth, MD, FACP, FACMI

Professor, Biomedical Informatics, Dept of Medicine

Clinical Research Information Officer, UCSD Health

mhogarth@health.ucsd.edu

co-IOR of MED277: "Introduction to Biomedical Natural Language Processing"



Goals of Lab 2

- Learn to use NLTK ‘tokenizers’ to identify sentences
- Use SpaCy for biomedical NLP ‘entity recognition’
- Learn to use the UMLS ‘search’ function to retrieve possible concept unique identifiers (codes), thus performing auto-coding

Expected Time

- 2-4 hours of your time

Tools you will need

PyCharm:

<https://www.jetbrains.com/pycharm/download/>

Sublime editor:

<https://www.sublimetext.com/download>



UC San Diego

Review -- Tokenizing

Text
“The cat sat on the mat.”

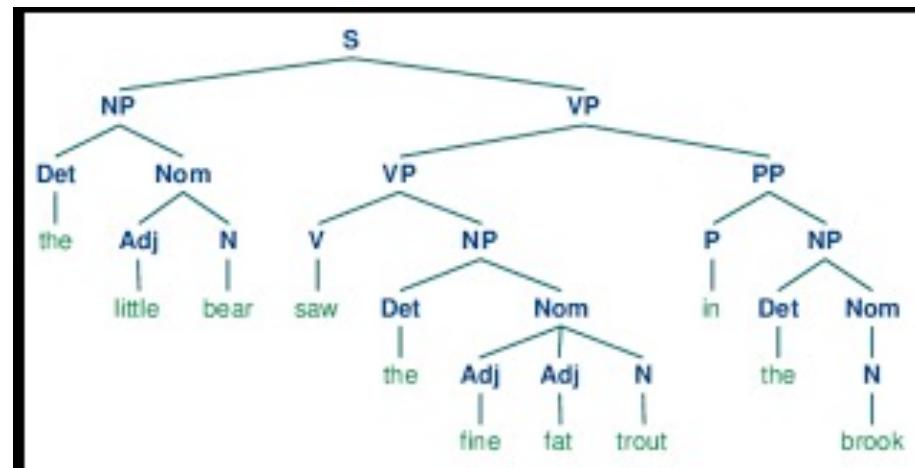


Tokens
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”



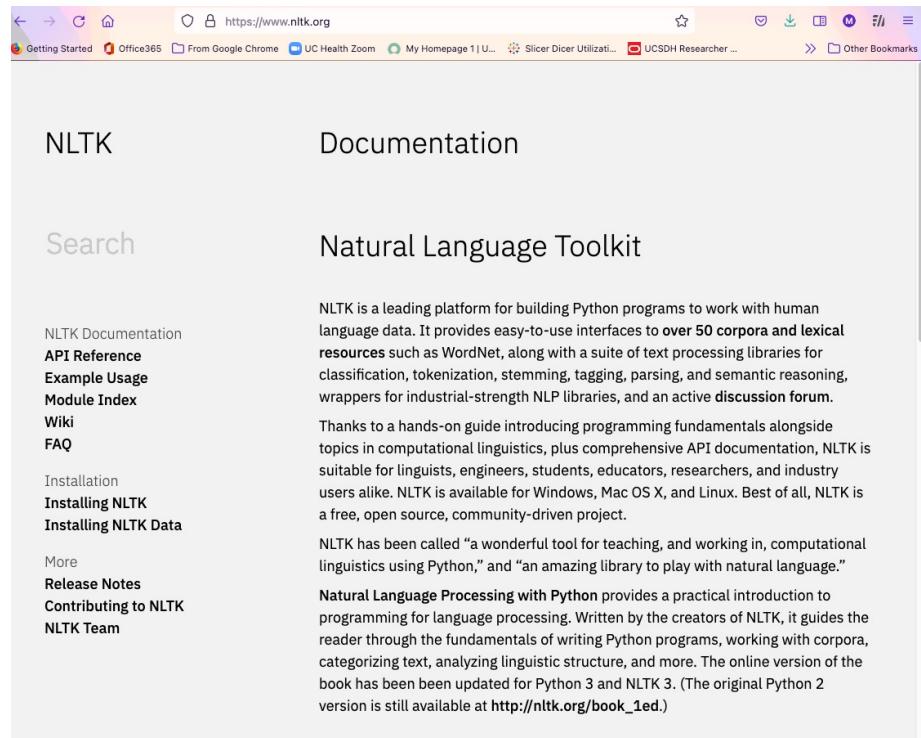
Part of Speech Tagging (POS Tagging)

- Identifying words and their ‘parts of speech’
- Tagging the words with a label that signifies their part of speech
- Uses an algorithm for ‘part of speech’ detection
- Uses a ‘tagset’ as the labeling



What is NLTK

- Natural Language Tool Kit (NLTK)
- Python NLP tools
 - Tokenizer
 - Stemmer, Lemmatizer
 - Sentence “tokenizer”
 - Part of Speech Tagger

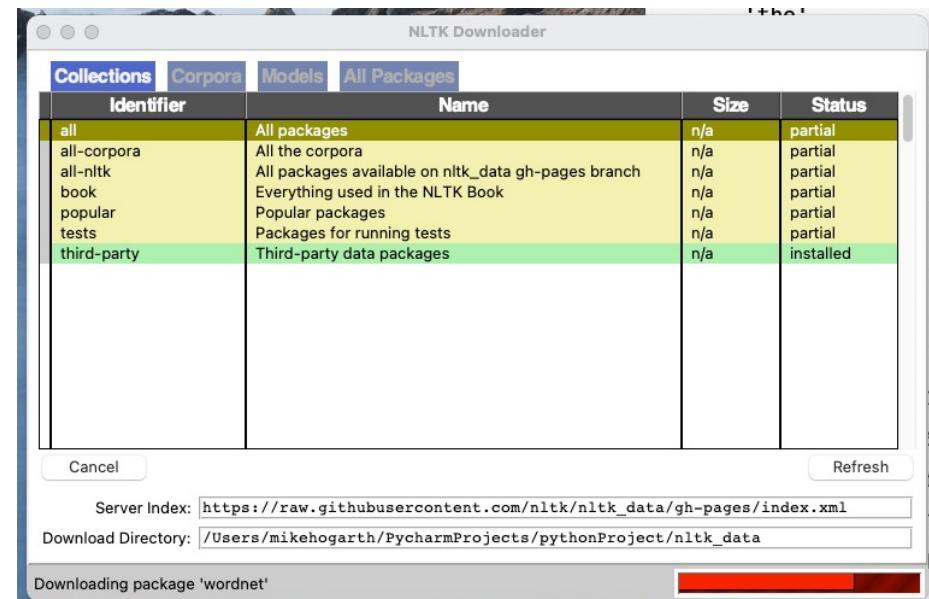


The screenshot shows the homepage of the NLTK website (<https://www.nltk.org>). The page has a clean, modern design with a light gray background. At the top, there's a navigation bar with links like "Getting Started", "Office365", "From Google Chrome", "UC Health Zoom", "My Homepage 1 | U...", "Slicer Dicer Utilizati...", "UCSDH Researcher ...", and "Other Bookmarks". Below the navigation, there are two main sections: "NLTK" on the left and "Documentation" on the right. Under "NLTK", there's a "Search" input field. Under "Documentation", there's a large heading "Natural Language Toolkit" followed by a detailed description of what NLTK is and its features. On the far left, there's a sidebar with links to "NLTK Documentation", "API Reference", "Example Usage", "Module Index", "Wiki", "FAQ", "Installation", "Installing NLTK", "Installing NLTK Data", "More", "Release Notes", "Contributing to NLTK", and "NLTK Team".

Using the NLTK

- Installing the package into your IDE
- Importing the package
 - Import nltk into the environment (in PyCharm)
 - Run nltk.download()
 - Have it save the package data in a place of your choosing

```
import nltk  
nltk.download()
```



Important – need to set environment variable

A new window should open, showing the NLTK Downloader. Click on the File menu and select Change Download Directory. For central installation, set this to `C:\nltk_data` (Windows), `/usr/local/share/nltk_data` (Mac), or `/usr/share/nltk_data` (Unix). Next, select the packages or collections you want to download.

If you did not install the data to one of the above central locations, you will need to set the `NLTK_DATA` environment variable to specify the location of the data. (On a Windows machine, right click on "My Computer" then select Properties > Advanced > Environment Variables > User Variables > New...)

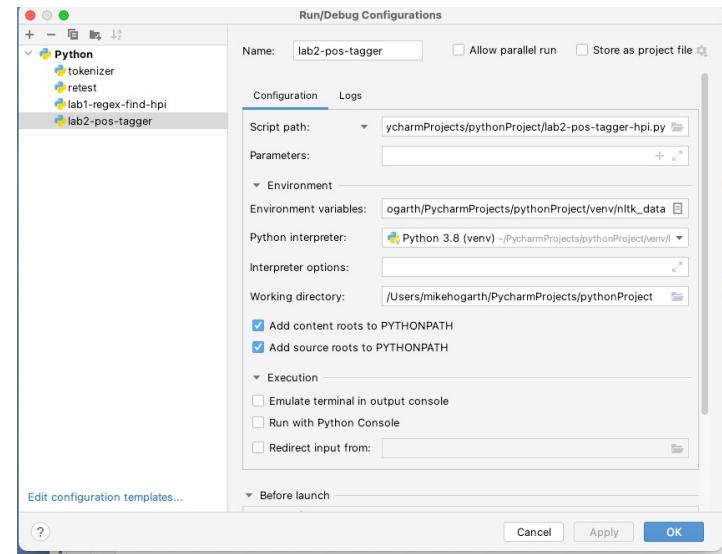
Test that the data has been installed as follows. (This assumes you downloaded the Brown Corpus):

The screenshot shows the PyCharm interface. On the left is the Project tool window displaying a directory structure for a project named 'pythonProject'. Inside 'pythonProject' are several files: lab1-regex-find-hpi.py, lab2-pos-tagger-hpi.py, lab2-hpi.csv, lab1-hpi.csv, lab2-hpi.csv, main.py, mimic_notes.csv, and retest.py. The right side shows the code editor with the file 'lab1-regex-find-hpi.py' open. The code imports nltk and defines a function to find sentences in a CSV file. The code editor has a yellow highlight on the line `print (nltk.word_tokenize(sentences[2]))`. A floating window titled 'Edit Configurations...' lists four configurations: tokenizer, retest, lab1-regex-find-hpi, and lab2-pos-tagger.

```
pythonProject - lab2-pos-tagger-hpi.py
pythonProject / lab1-regex-find-hpi.py | lab2-pos-tagger-hpi.py | lab2-hpi.csv
Edit Configurations...
tokenizer
retest
lab1-regex-find-hpi
lab2-pos-tagger

1 14 1 ^ v

1 import nltk
2 from nltk.corpus import brown
3 def find_sentences(input_text):
4     sentences = []
5     sentences = re.split("\.", input_text)
6     return sentences
7
8 #####
9 fname = "lab2-hpi.csv"
10 with open(fname) as f:
11     input_text = f.read()
12
13 sentences = find_sentences(input_text)
14 print (nltk.word_tokenize(sentences[2]))
15
16 counter=0
17
```



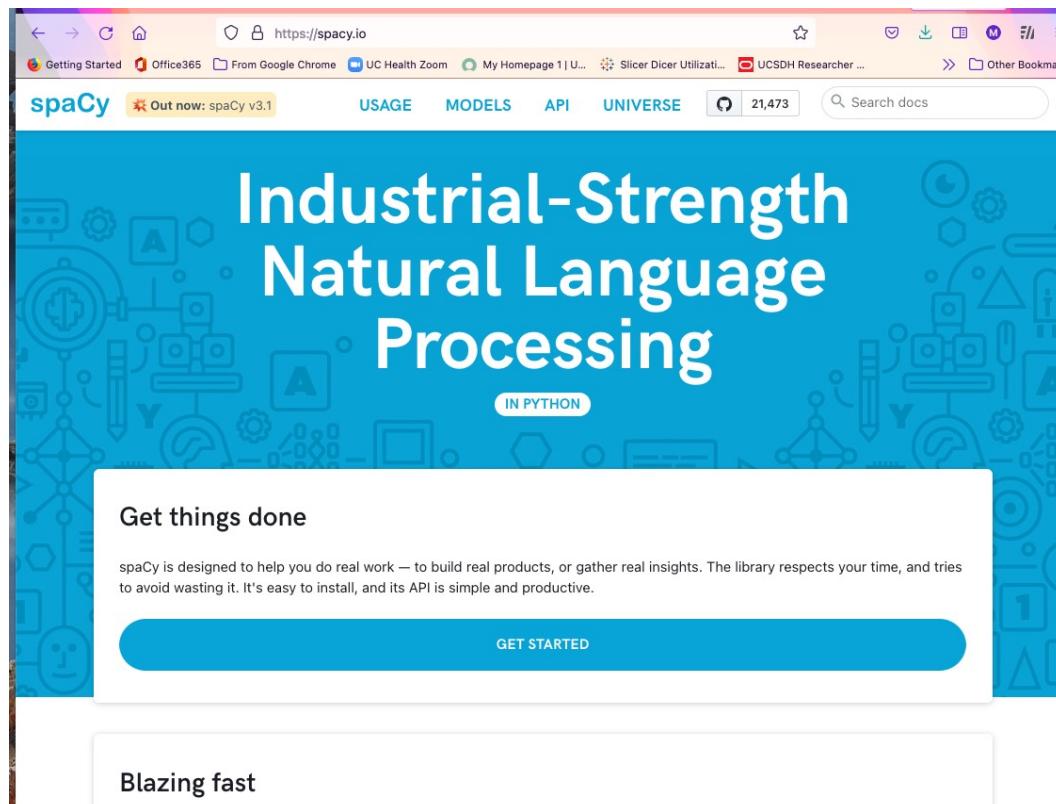
Using the NLTK POS Tagger

```
('During', 'IN')
('his', 'PRP$')
('hospital', 'JJ')
('course', 'NN')
(',', ',')
('the', 'DT')
('patient', 'NN')
('has', 'VBZ')
('become', 'VBN')
('progressively', 'RB')
('hypoxic', 'JJ')
(',', ',')
('initially', 'RB')
('satting', 'VBG')
('mid', 'PRP')
('90s', 'CD')
```

```
nouns = []
for y in sentences:
    y = nltk.word_tokenize(y)
    y = nltk.pos_tag(y)
    noun_tag_list=[ "NN", "NNS", "NNP", "NNPS"]
    for x in y:
        if x[1] in noun_tag_list:
            print(x)
            nouns.append(x[0])
```



What is SpaCy



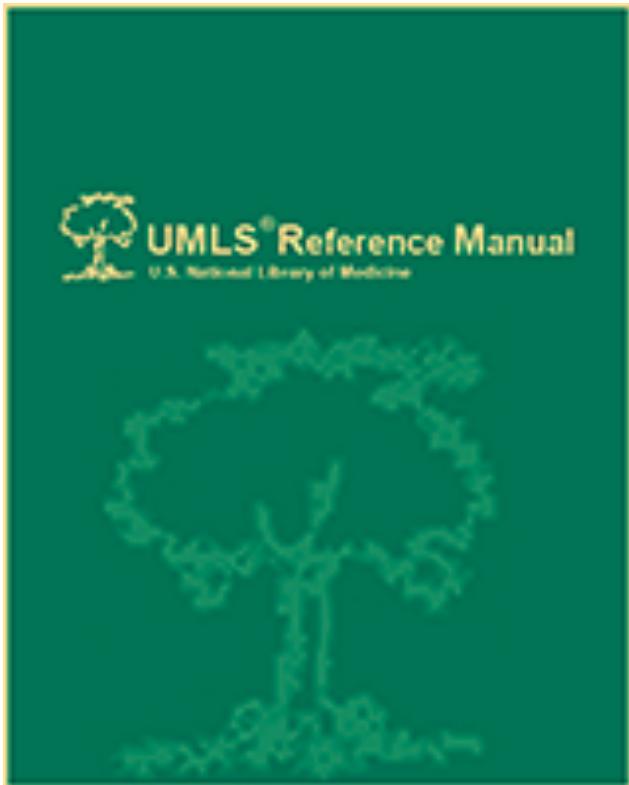
Using spaCy

- Various NLP functions including entity recognition
- Can identify a list of “entities”
 - Does not auto-code them

```
#use scispacy core biomedical corpus/models on input text to find entities
nlp=spacy.load("en_core_sci_sm")
doc=nlp(input_text)
length = len(doc)
x=0
print("\n-----SciSpacy entity extraction-----")
for v in range(len(doc.ents)):
    string=str(doc.ents[v])
    umls_data = get_umls_info(string)
    print("\n-----\n"+string)
    for result in umls_data["results"]:
        if umls_data["results"][0]["ui"] == "NONE":
            break
        print(result["ui"]+"|"+result["name"]+"|"+result["rootSource"])
```



What is the UMLS



- A system managed by the National Library of Medicine
- Published since 1986
- Has 3 main components
 - Metathesaurus
 - Semantic Network
 - SPECIALIST Lexicon
- Integrates 220+ “source vocabularies” (2021)

UMLS Basic Organization

Organized by Concept (unique concept)

- a concept unique identifier (CUI) is provided for the concept
- concepts are given “properties”
 - CUI
 - definition
 - terms
 - synonymous terms are clustered together into a concept
- Relations (semantic network)
 - concepts are related to other concepts
 - properties in relations
 - types of relationship

Example: Addison's Disease

Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

Adrenal gland diseases
Adrenal disorder
Disorder of adrenal gland
Diseases of the adrenal glands

C0001621

MeSH
AOD
Read
SNOMED

D000307
0000005418
C15z.
DB-70000

Adrenal Gland Diseases

Using the UMLS UTS RESTful API

The screenshot shows the UMLS API Technical Documentation homepage. At the top, there is a NIH logo and a search bar. Below the header, there are navigation links for "UMLS API Technical Documentation", "Release Notes", "User Authentication", "Searching the UMLS", "Retrieving UMLS Data", "Retrieving Source-Asserted Data", and "REST API Cookbook". A breadcrumb trail indicates the current page is "UMLS Home » UMLS REST API". The main content area is titled "UMLS REST API Home Page" and discusses the Authentication Service Endpoint. It explains that the authentication service provides methods for retrieving a ticket granting ticket as well as single-use service tickets. Service tickets are needed each time you search or retrieve content from the UMLS REST API. For help with making authentication calls using Postman, it points to a tutorial: "UMLS REST API: Authentication and Calling". To the right, there is a sidebar with a feedback form and links to Postman sample collections and API Terms of Service.

UMLS REST API Home Page

Authentication Service Endpoint

The authentication service provides methods for retrieving a ticket granting ticket as well as single-use service tickets.

Service tickets are needed each time you search or retrieve content from the UMLS REST API.

For help with making authentication calls using Postman, see our tutorial: [UMLS REST API: Authentication and Calling](#).

Base URI	Method Type	Path	Description
https://utslogin.nlm.nih.gov			
	POST	/cas/v1/api-key	Retrieves a Ticket Granting Ticket (TGT)
	POST	/cas/v1/tickets/{TGT}	Retrieves a single-use Service Ticket

We welcome your feedback on our [customer service form](#). Please use "UMLS REST API feedback" in your subject line.

Check out the [Postman sample collections](#), or [code samples in Python, Java, and Perl](#) on Github to help you get started using the UMLS REST API.

[API Terms of Service](#)

Retrieve CUI with text search - exact



The screenshot shows a web browser window with the URL <https://documentation.uts.nlm.nih.gov/rest/search/#uris>. The page title is "Searching the UMLS". The main content area has a heading "Searching the UMLS" and links to "URIs", "Query Parameters", "Sample Output", and "Paging through results". Below this, it says "URIs with /search support the following use cases:" followed by a bulleted list:

- Return a list of CUIs and their names when searching a human readable term.
- Return a list of source-asserted identifiers (codes) and their names when searching a human readable term.
- Map source-asserted identifiers to UMLS CUIs.

Note that 'current' in the URI can always be used to search against the latest UMLS publication. You may use any valid UMLS release back to 2008AA in your URI if you would like to search against a particular version of the UMLS.

A sidebar on the right contains a message: "We welcome your feedback on our [customer service form](#). Please use 'UMLS REST API feedback' in your subject line." It also links to "Postman sample collections" and "code samples in Python, Java, and Perl on Github". A link to "API Terms of Service" is at the bottom of the sidebar.

URIs
The base URI is <https://uts-ws.nlm.nih.gov/rest>

Sample URI	Description	Returned JSON Object classType
/search/current?string=fraction of carpal bone	Retrieves CUIs for a search term	searchResults
/search/current?string=fraction of carpal bone&searchType=exact	Uses 'exact' searching	searchResults
/search/current?string=fraction of carpal bone&sabs=SNOMEDCT_US&returnIdType=code	Returns SNOMEDCT concepts associated with a search term	searchResults
/search/current?string=9468002&inputType=sourceUi&searchType=exact&sabs=SNOMEDCT_US	Returns UMLS CUIs associated with a SNOMEDCT_US concept	searchResults

Questions?



Eastern view from Iron Mountain peak, San Diego (2021)