Lab 2: NLTK, SpaCy, and the UMLS

Michael Hogarth, MD, FACP, FACMI
Professor, Biomedical Informatics, Dept of Medicine
Clinical Research Information Officer, UCSD Health
mihogarth@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

UC San Diego

Expected Time

- 2-4 hours of your time

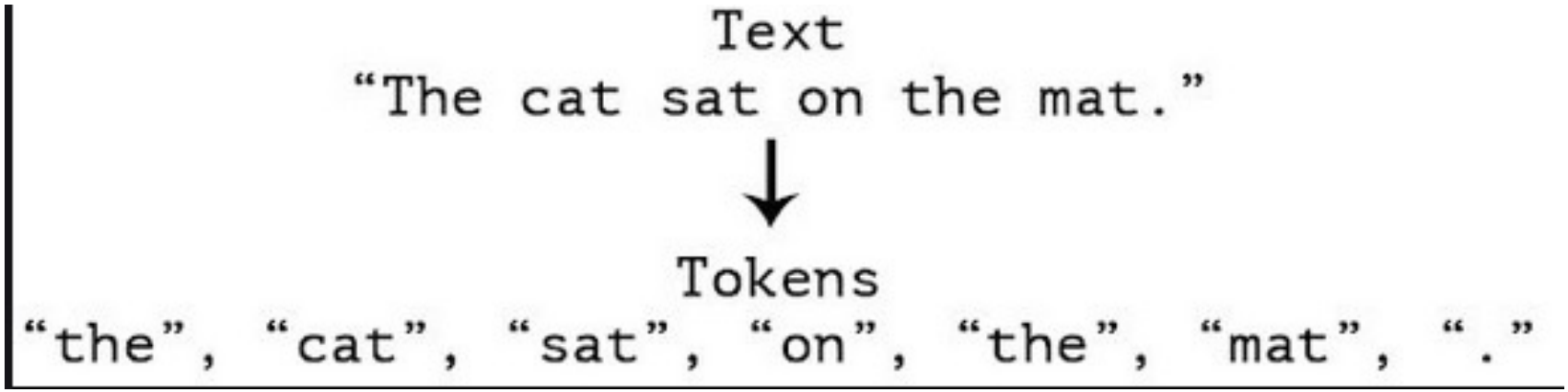
UC San Diego

Tools you will need

PyCharm:
https://www.jetbrains.com/pycharm/download/

Sublime editor:
https://www.sublimetext.com/download

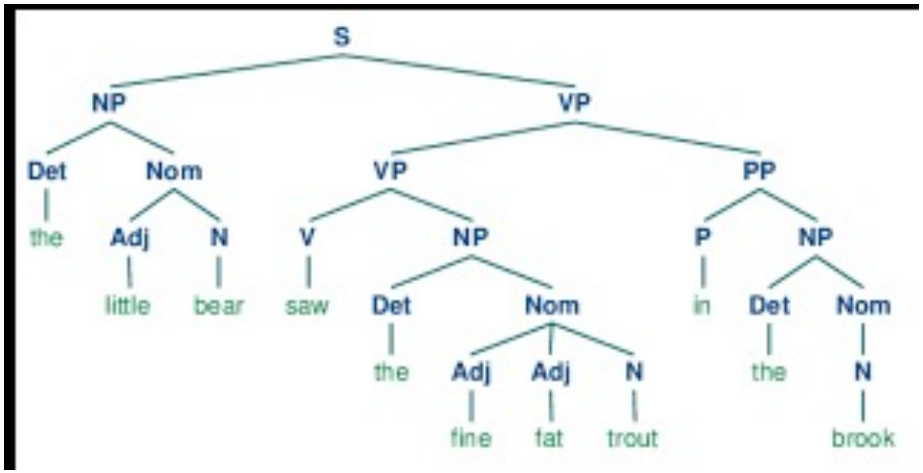
Review -- Tokenizing



UC San Diego

Part of Speech Tagging (POS Tagging)

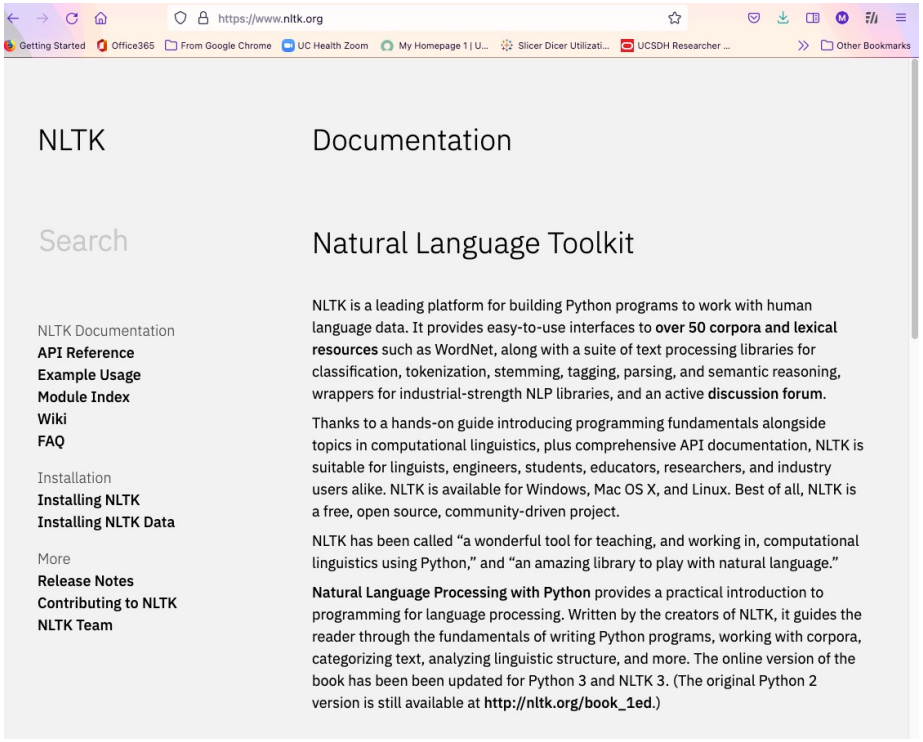
- Identifying words and their ‘parts of speech’
- Tagging the words with a label that signifies their part of speech
- Uses an algorithm for ‘part of speech’ detection
- Uses a ‘tagset’ as the labeling



UC San Diego

What is NLTK

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

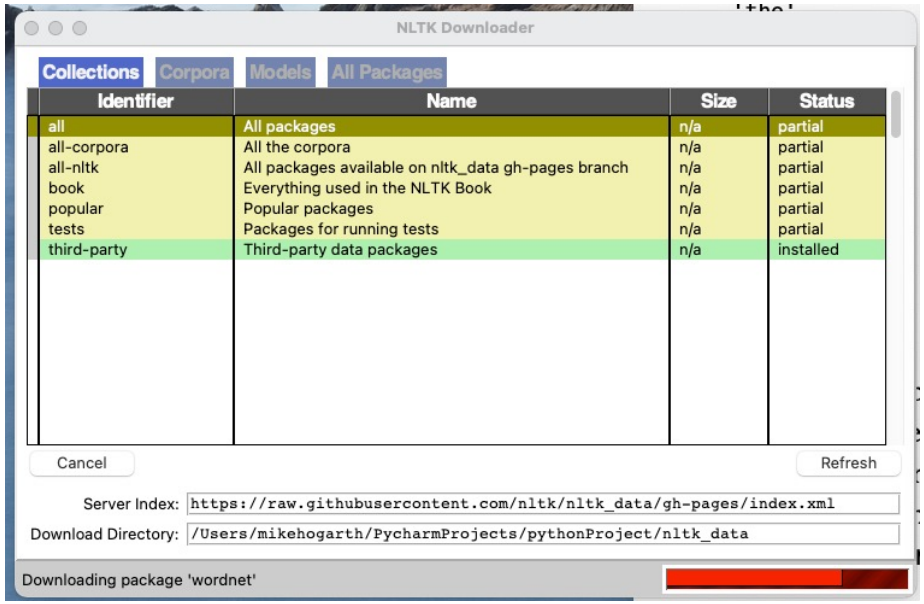


UC San Diego

Using the NLTK

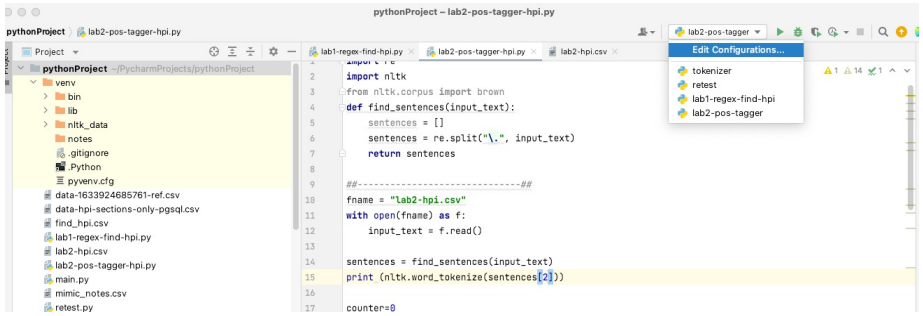
- Installing the package into your IDE
- Importing the package
  - Import nltk into the environment (in PyCharm)
  - Run nltk.download()
  - Have it save the package data in a place of your choosing

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



UC San Diego

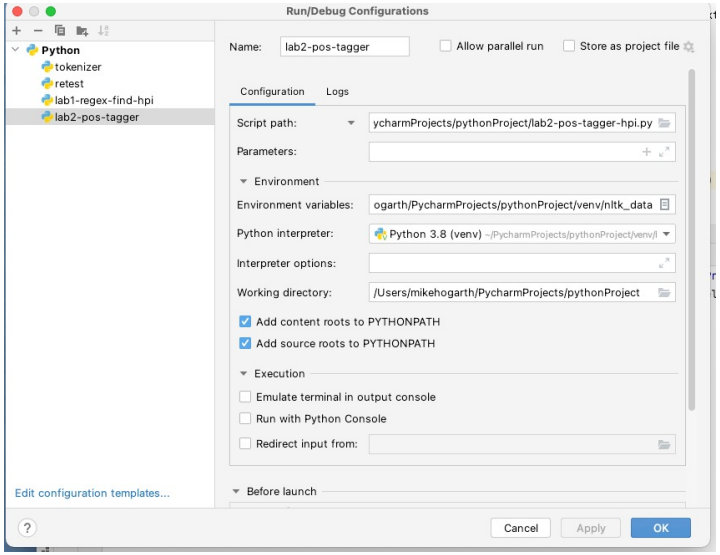
Important – need to set environment variable



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):



UC San Diego

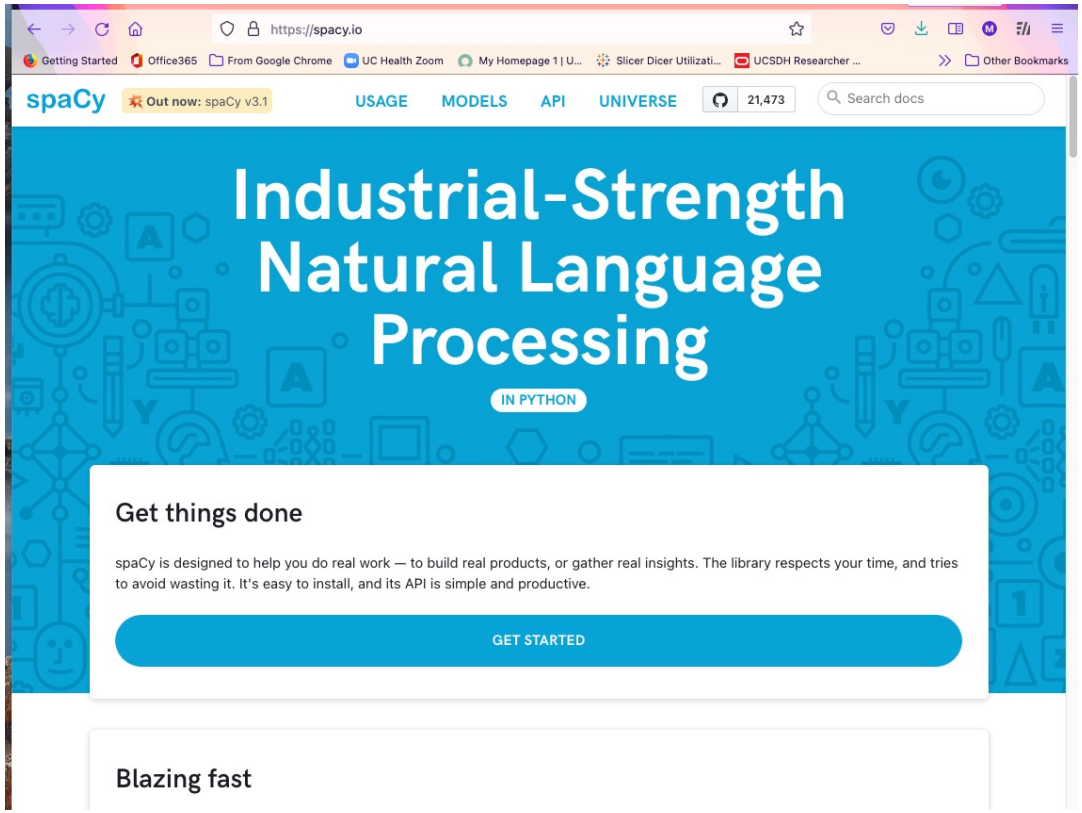
Using the NLTK POS Tagger

```
('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])
```

UC San Diego

What is SpaCy



UC San Diego

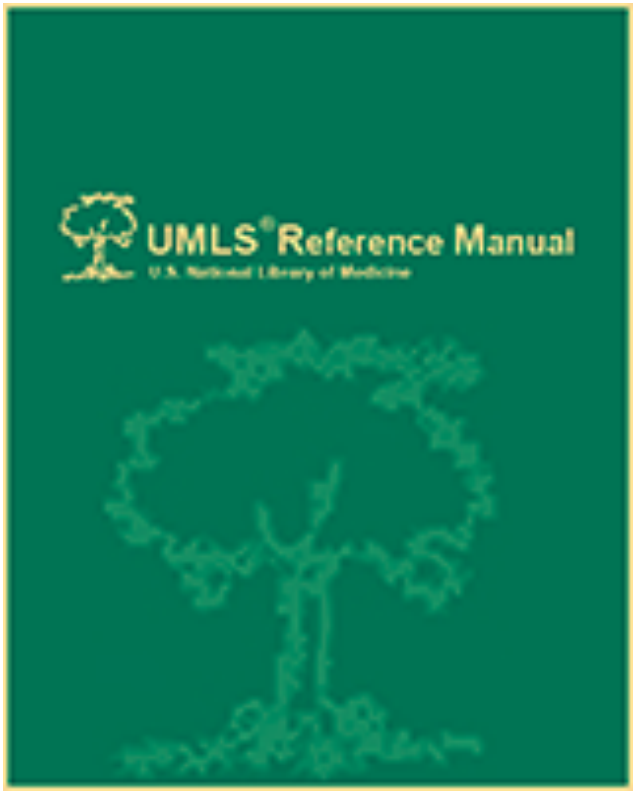
Using spaCy

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

```
#use scispacy core biomed 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"])
```

UC San Diego

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)

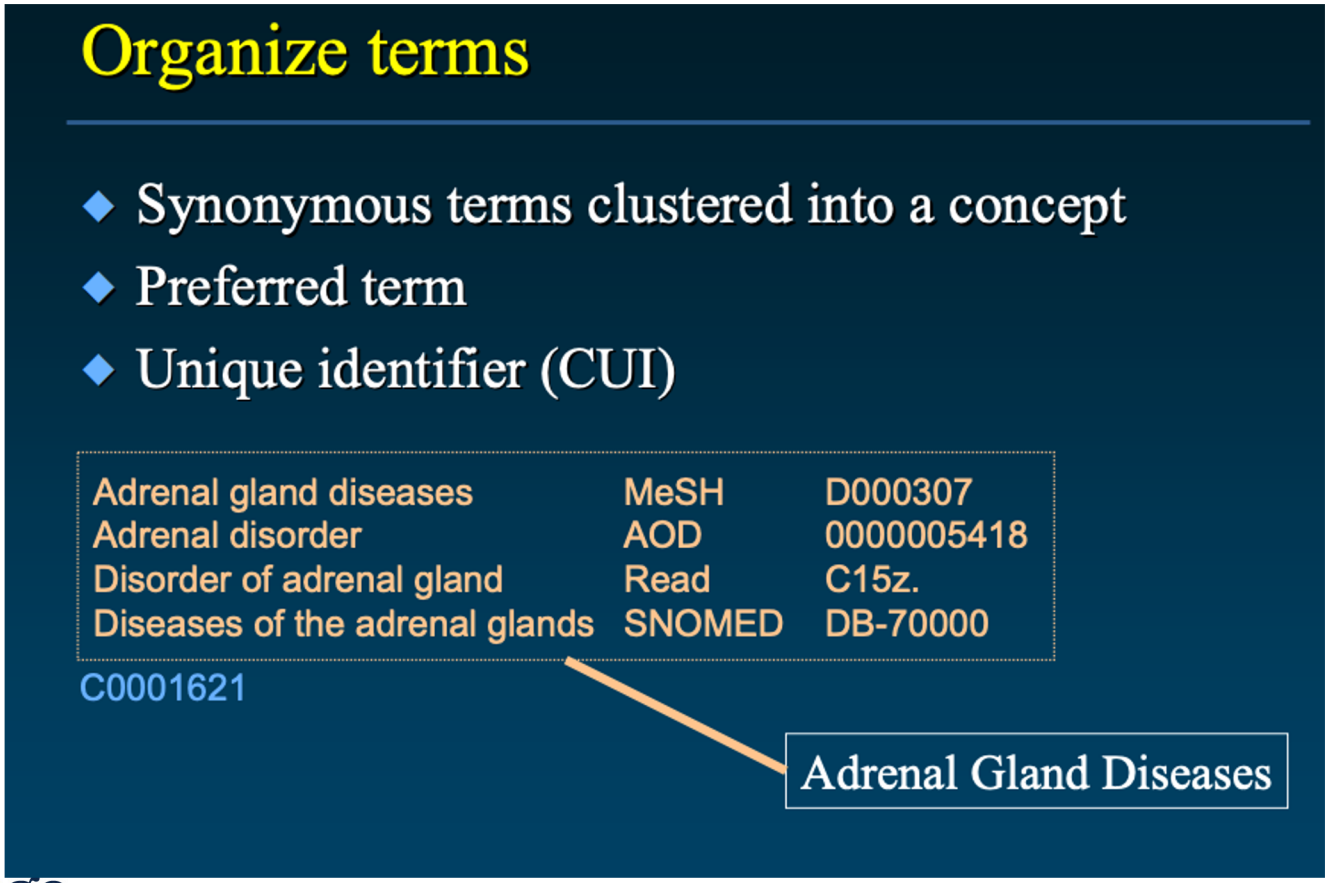
UC San Diego

UMLS Basic Organization

Organized by Concept (unique concept)

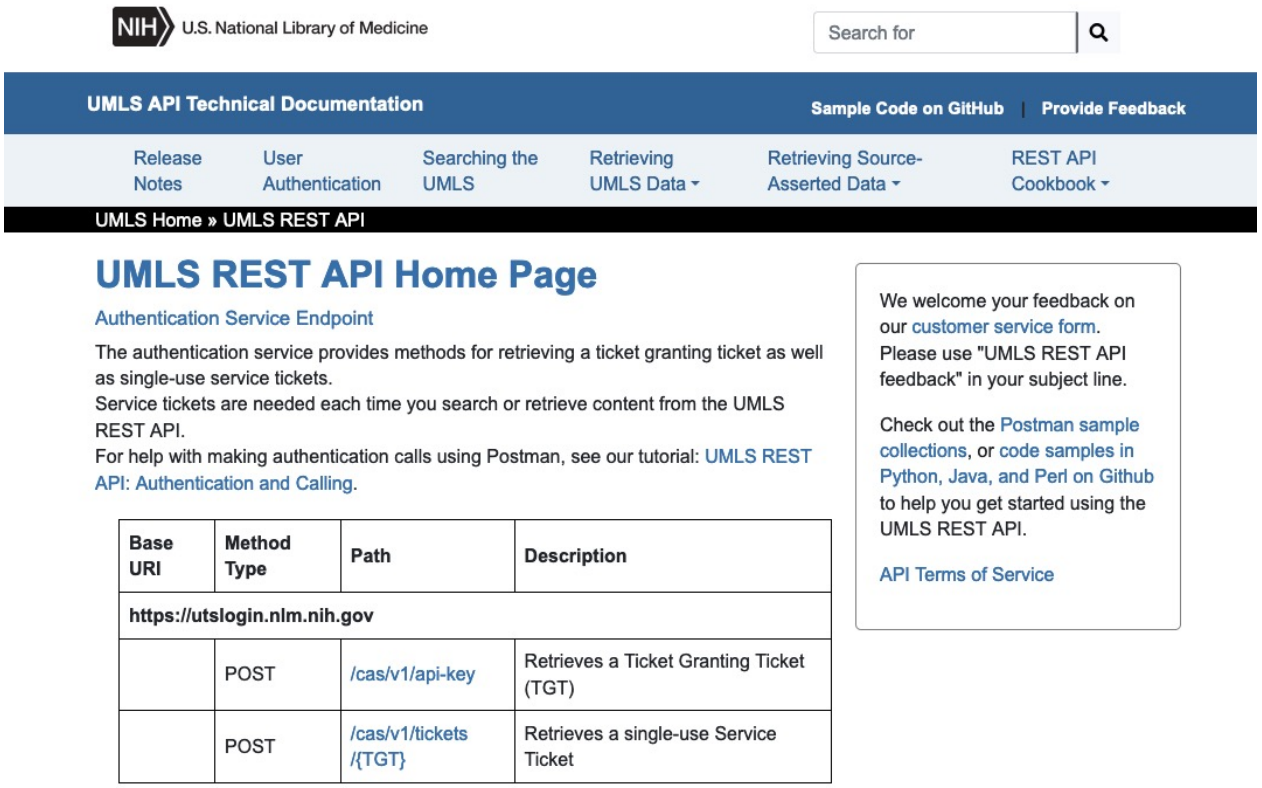
- a concept unique identifier (CUI) is provided for the concept
- concepts are given “properties”
  - CUI
  - definition
  - terms
    - synonymous terms are clustered together into a concept
- Relations (semantic network)
  - concepts are related to other concepts
  - properties in relations
    - types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

Example: Addison's Disease



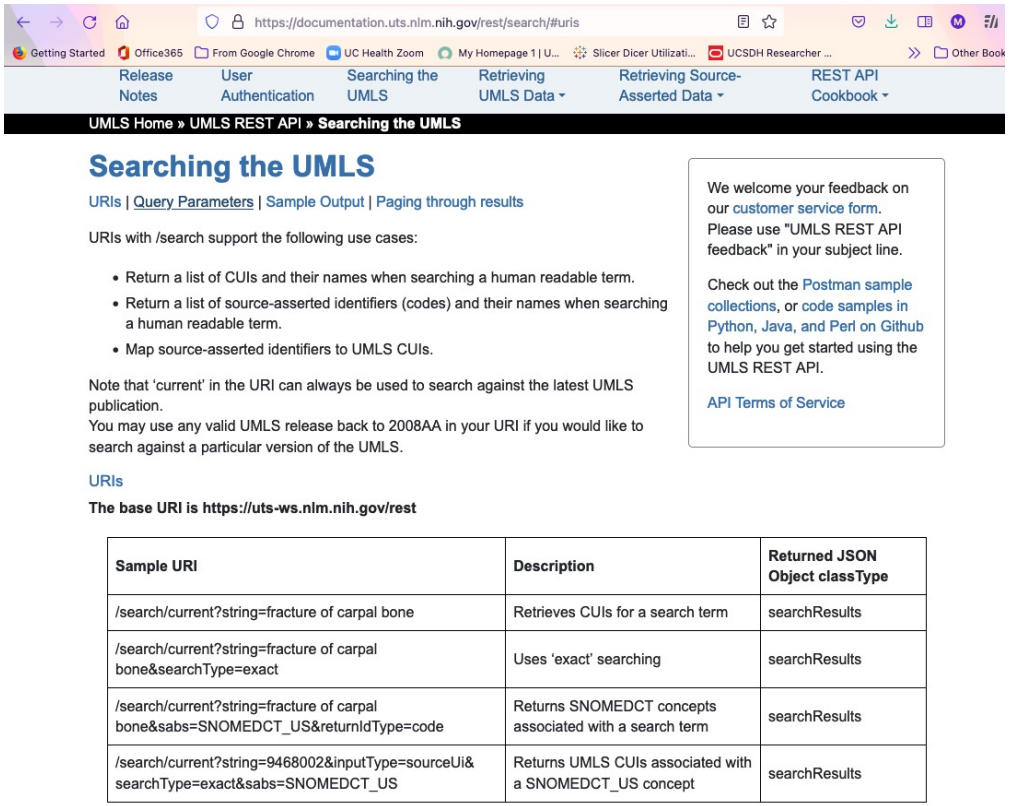
UC San Diego

Using the UMLS UTS RESTful API



UC San Diego

Retrieve CUI with text search - exact



https://documentation.uts.nlm.nih.gov/rest/search/#uris

Release Notes | User Authentication | Searching the UMLS | Retrieving UMLS Data | Retrieving Source-Asserted Data | REST API Cookbook

UMLS Home » UMLS REST API » **Searching the UMLS**

Searching the UMLS

URIs | Query Parameters | Sample Output | Paging through results

URIs with /search support the following use cases:

- 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.

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

URIs

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

Sample URI	Description	Returned JSON Object classType
/search/current?string=fracture of carpal bone	Retrieves CUIs for a search term	searchResults
/search/current?string=fracture of carpal bone&searchType=exact	Uses 'exact' searching	searchResults
/search/current?string=fracture 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

UC San Diego

Questions?

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

UC San Diego