LING 408/508: Computational Techniques for Linguists

Lecture 26

Today's Topics

- Reminder: Homework 10:
  - have everyone gotten the GET and POST methods working okay?
- Onto the next topic…

Next Programming Language

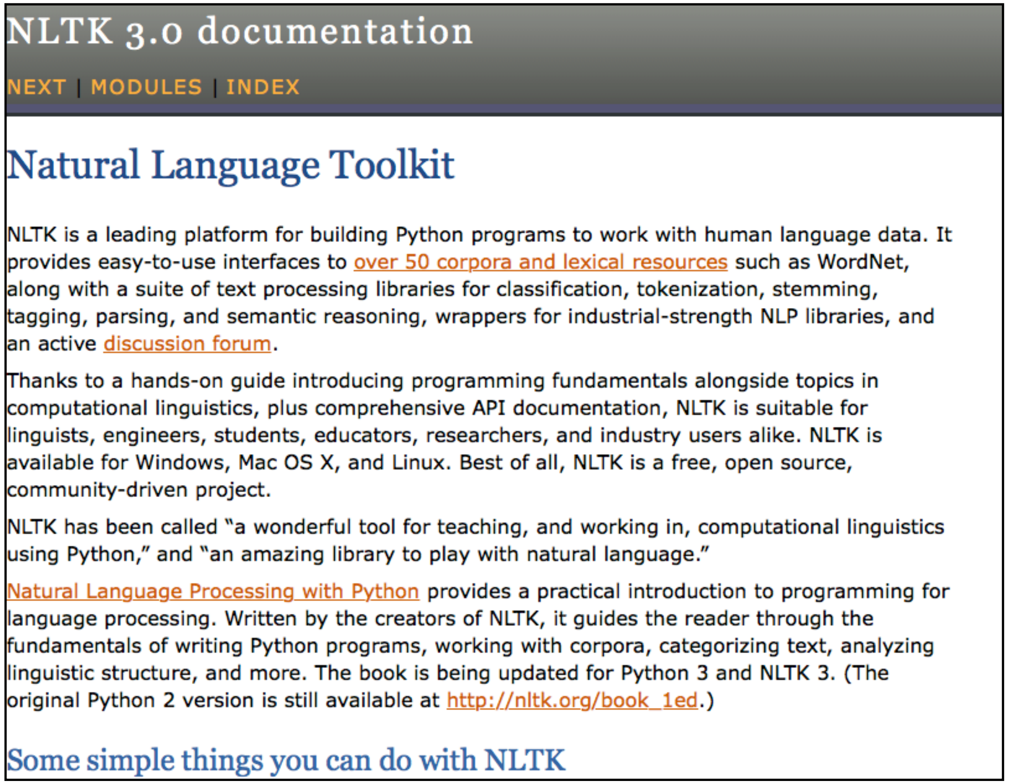
So far, in this course you've been exposed to:

1. Binary language of the CPU: encoding of numbers and characters (unicode)
2. Linux (Ubuntu under VirtualBox)
3. command line: bash shell
4. bash shell scripting: `#!/bin/bash`
5. awk and regular expressions (regex)
6. html/css
7. Javascript + DOM
8. Web client/server model

Now we switch for the rest of the semester to a regular programming course:

- Python

Why Python?



Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

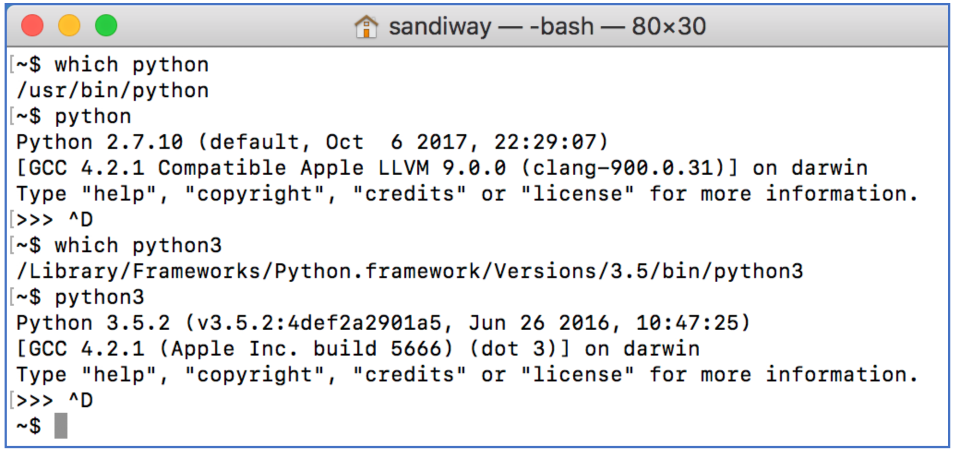
Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

Some simple things you can do with NLTK

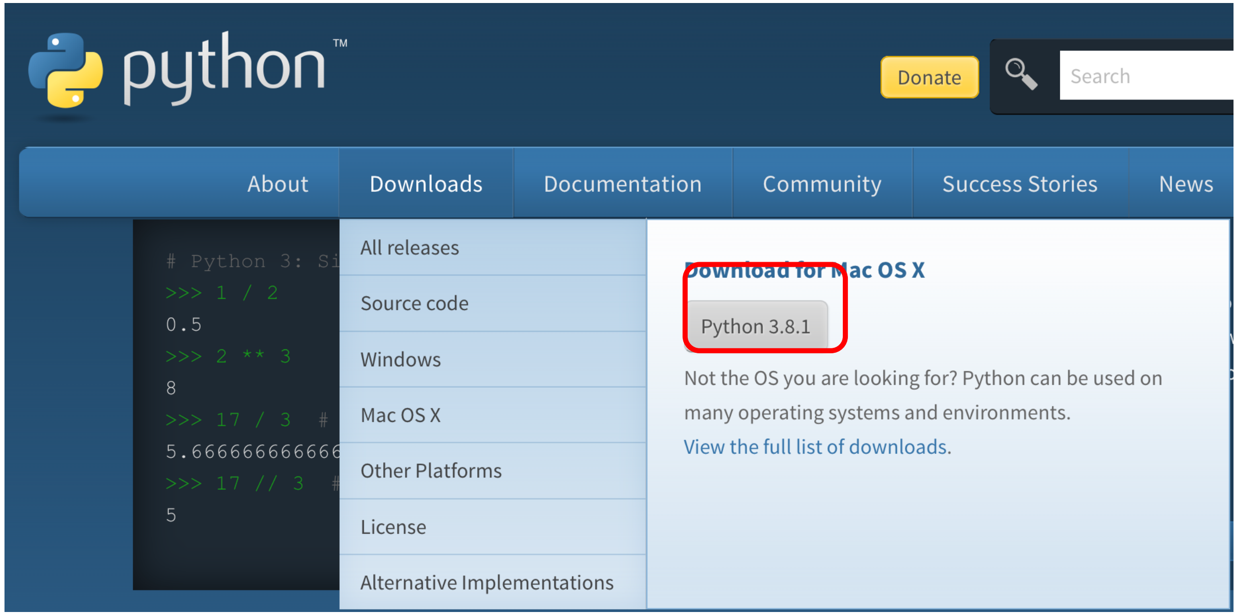
- Installed by default on OSX
  - but you should install and use python3



```
~$ which python
/usr/bin/python
~$ python
Python 2.7.10 (default, Oct  6 2017, 22:29:07)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.31)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> ^D
~$ which python3
/Library/Frameworks/Python.framework/Versions/3.5/bin/python3
~$ python3
Python 3.5.2 (v3.5.2:4def2a2901a5, Jun 26 2016, 10:47:25)
[GCC 4.2.1 (Apple Inc. build 5666) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> ^D
~$ 
```

Programming

- python.org: install python3 on your laptop (if not already installed)



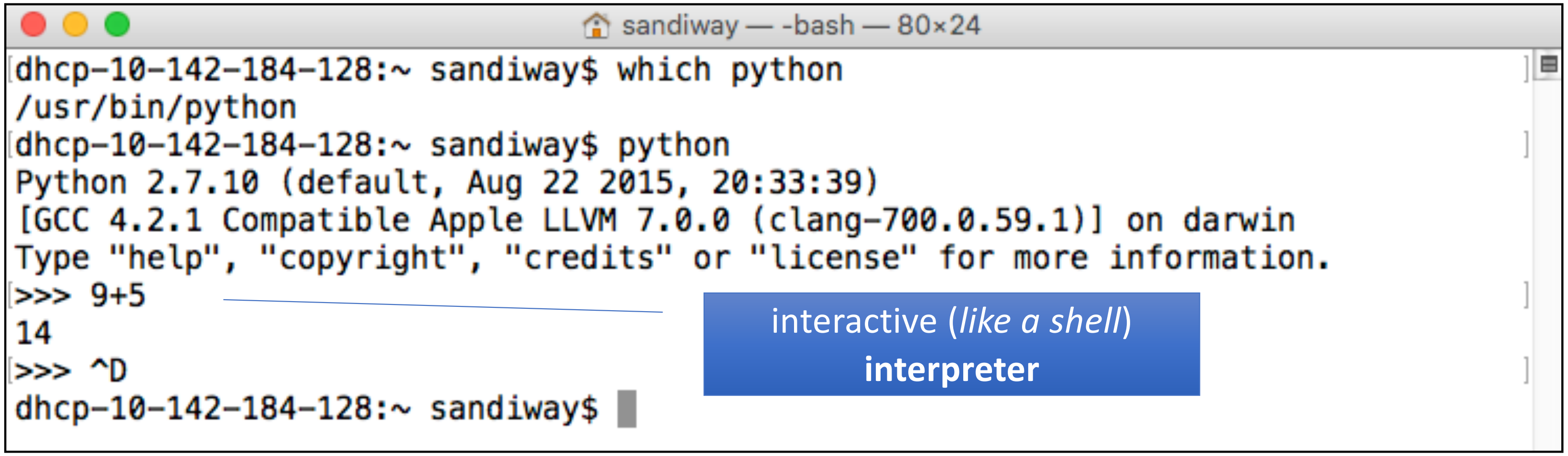
```
# Python 3: Si
>>> 1 / 2
0.5
>>> 2 ** 3
8
>>> 17 / 3  #
5.666666666666
>>> 17 // 3  #
5
```

Not the OS you are looking for? Python can be used on many operating systems and environments.
View the full list of downloads.

do **not** install Python 2.7

Python

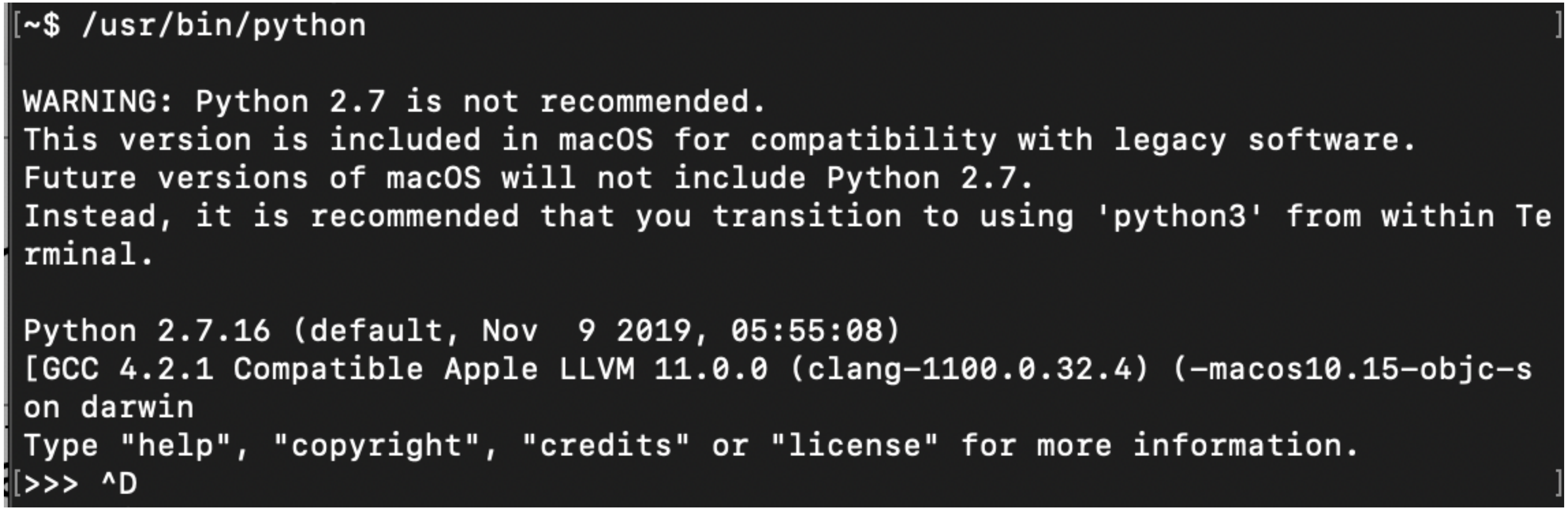
- Python 2.7 Installed by default on MacOS prior to MacOS Catalina
  - Terminal command: **python**



```
dhcp-10-142-184-128:~ sandiway$ which python
/usr/bin/python
dhcp-10-142-184-128:~ sandiway$ python
Python 2.7.10 (default, Aug 22 2015, 20:33:39)
[GCC 4.2.1 Compatible Apple LLVM 7.0.0 (clang-700.0.59.1)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> 9+5
14
>>> ^D
dhcp-10-142-184-128:~ sandiway$
```

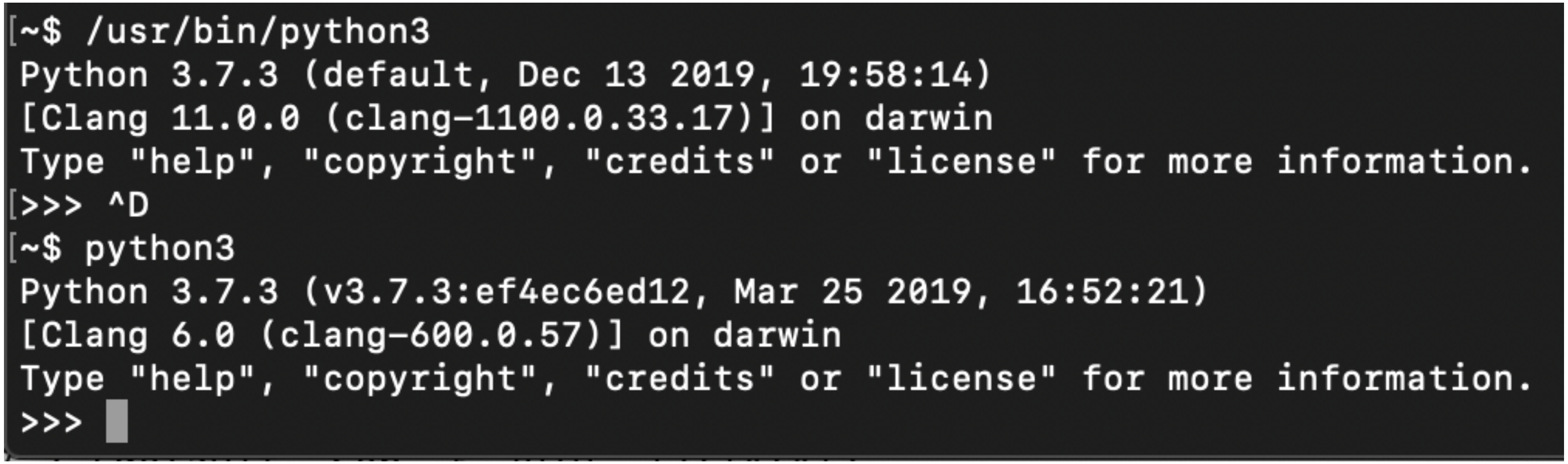
interactive (*like a shell*) **interpreter**

Python: MacOS Catalina



```
[~$ /usr/bin/python                                                          ]

WARNING: Python 2.7 is not recommended.
This version is included in macOS for compatibility with legacy software.
Future versions of macOS will not include Python 2.7.
Instead, it is recommended that you transition to using 'python3' from within Te
rminal.

Python 2.7.16 (default, Nov  9 2019, 05:55:08)
[GCC 4.2.1 Compatible Apple LLVM 11.0.0 (clang-1100.0.32.4) (-macos10.15-objc-s
on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> ^D                                                                      ]
```

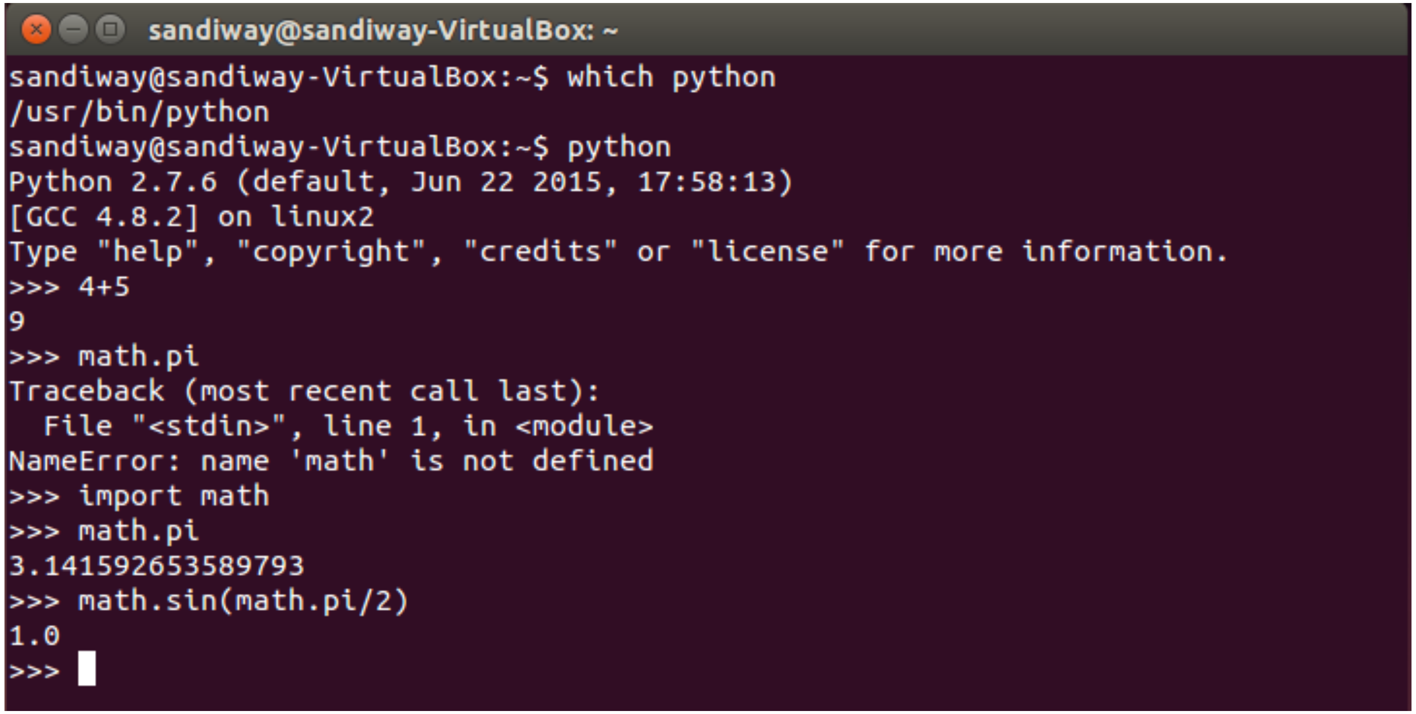
Python: MacOS Catalina



```
[~$ /usr/bin/python3
Python 3.7.3 (default, Dec 13 2019, 19:58:14)
[Clang 11.0.0 (clang-1100.0.33.17)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> ^D
[~$ python3
Python 3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 16:52:21)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

Python

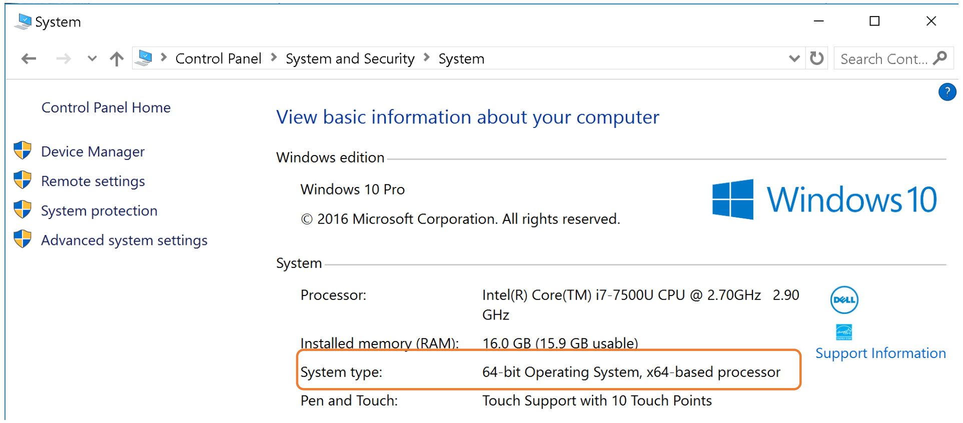
- On Ubuntu:



```
sandiway@sandiway-VirtualBox:~$ which python
/usr/bin/python
sandiway@sandiway-VirtualBox:~$ python
Python 2.7.6 (default, Jun 22 2015, 17:58:13)
[GCC 4.8.2] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> 4+5
9
>>> math.pi
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'math' is not defined
>>> import math
>>> math.pi
3.141592653589793
>>> math.sin(math.pi/2)
1.0
>>> 
```

Windows 10: 32 or 64 bit system?

- Look for System in Control Panel:



View basic information about your computer

Windows edition

Windows 10 Pro

© 2016 Microsoft Corporation. All rights reserved.

System

Processor: Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz

Installed memory (RAM): 16.0 GB (15.9 GB usable)

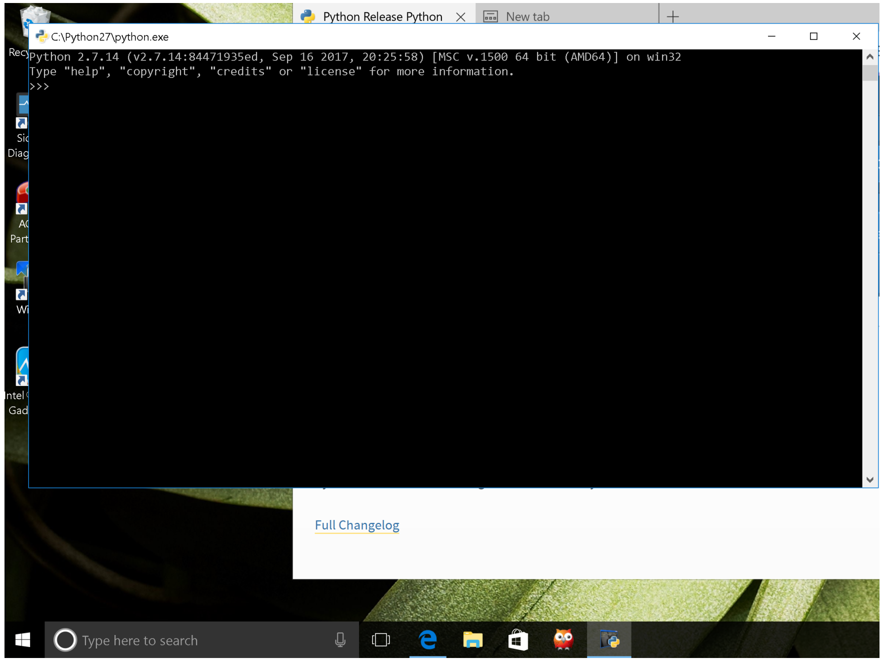
System type: 64-bit Operating System, x64-based processor

Pen and Touch: Touch Support with 10 Touch Points

- Choose correct file:
  - Windows x86-64 executable installer (for 64 bit systems)
  - Windows x86 executable installer (for 32 bit systems)

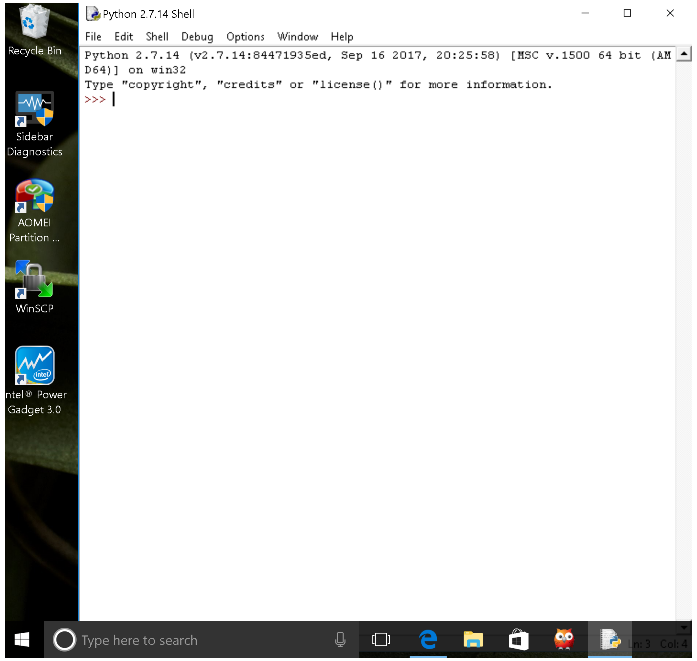
Windows 10: Running Python

Command line (PowerShell ok too)



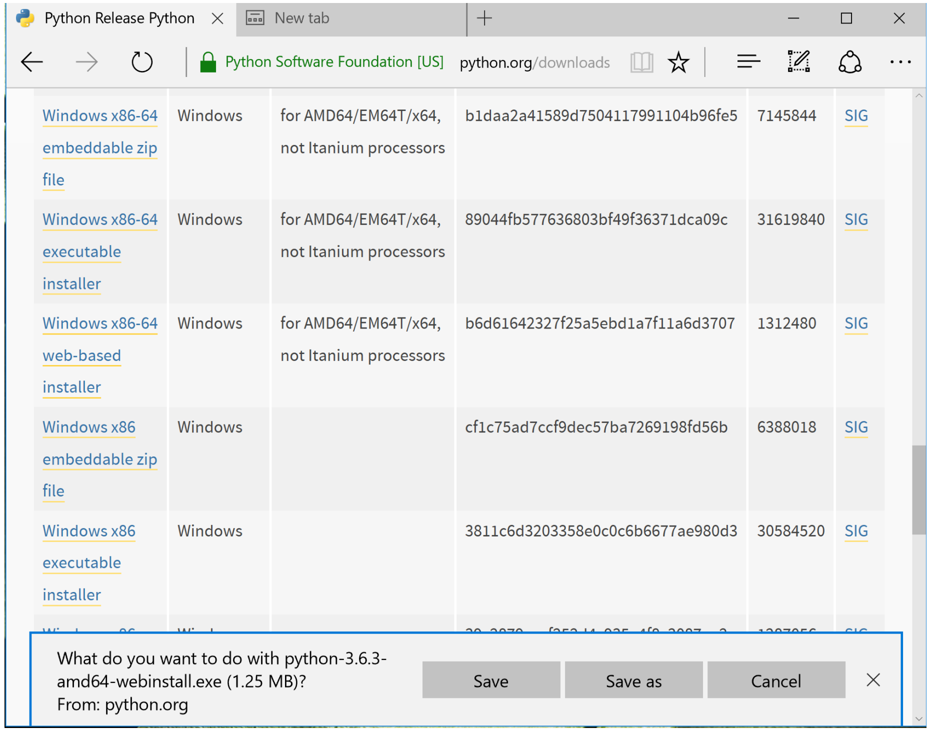
```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:25:58) [MSC v.1500 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

IDLE



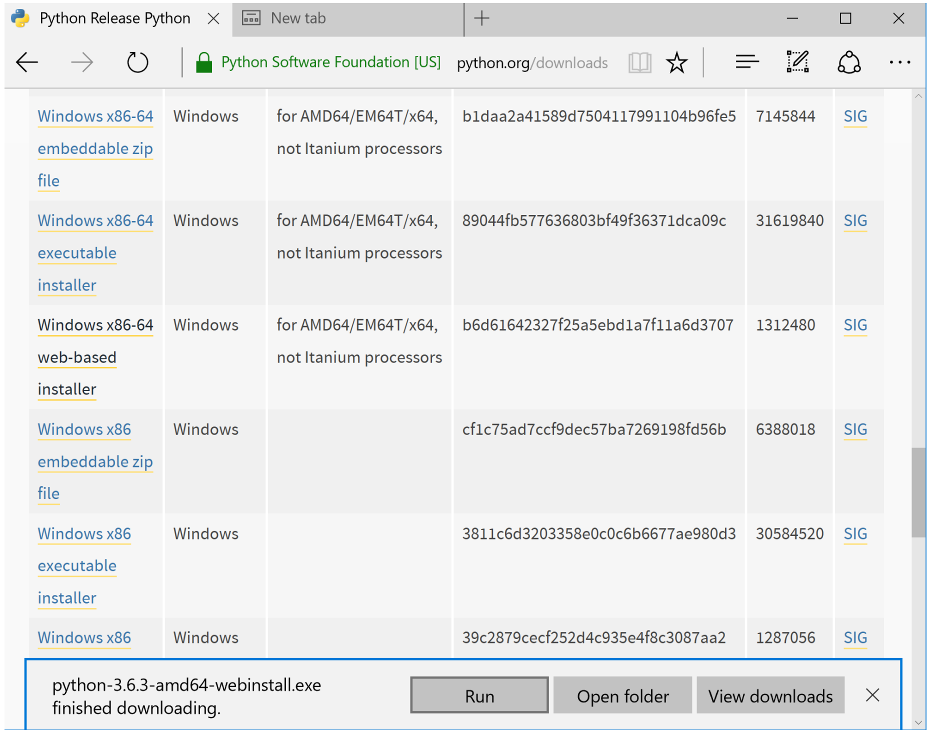
```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:25:58) [MSC v.1500 64 bit (AM
D64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
```

Windows 10: Python 3 download



Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64, not Itanium processors	b1daa2a41589d7504117991104b96fe5	7145844	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64, not Itanium processors	89044fb577636803bf49f36371dca09c	31619840	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64, not Itanium processors	b6d61642327f25a5ebd1a7f11a6d3707	1312480	SIG
Windows x86 embeddable zip file	Windows		cf1c75ad7ccf9dec57ba7269198fd56b	6388018	SIG
Windows x86 executable installer	Windows		3811c6d3203358e0c0c6b6677ae980d3	30584520	SIG

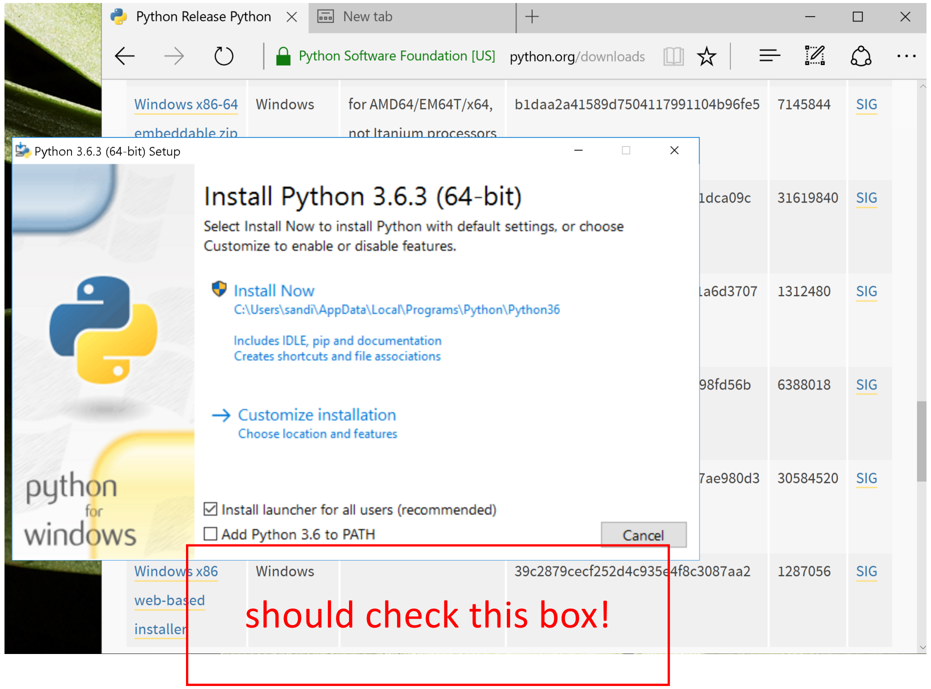
What do you want to do with python-3.6.3-amd64-webinstall.exe (1.25 MB)?
From: python.org



Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64, not Itanium processors	b1daa2a41589d7504117991104b96fe5	7145844	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64, not Itanium processors	89044fb577636803bf49f36371dca09c	31619840	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64, not Itanium processors	b6d61642327f25a5ebd1a7f11a6d3707	1312480	SIG
Windows x86 embeddable zip file	Windows		cf1c75ad7ccf9dec57ba7269198fd56b	6388018	SIG
Windows x86 executable installer	Windows		3811c6d3203358e0c0c6b6677ae980d3	30584520	SIG
Windows x86	Windows		39c2879cecf252d4c935e4f8c3087aa2	1287056	SIG

python-3.6.3-amd64-webinstall.exe finished downloading.

Windows 10: Running the Python 3 installer



Windows x86-64 embeddable zip | Windows | for AMD64/EM64T/x64, not Itanium processors | b1daa2a41589d7504117991104b96fe5 | 7145844 | SIG

1dca09c | 31619840 | SIG

1a6d3707 | 1312480 | SIG

98fd56b | 6388018 | SIG

7ae980d3 | 30584520 | SIG

Windows x86 web-based installer | Windows | 39c2879cecf252d4c935e4f8c3087aa2 | 1287056 | SIG

Python 3.6.3 (64-bit) Setup

Install Python 3.6.3 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

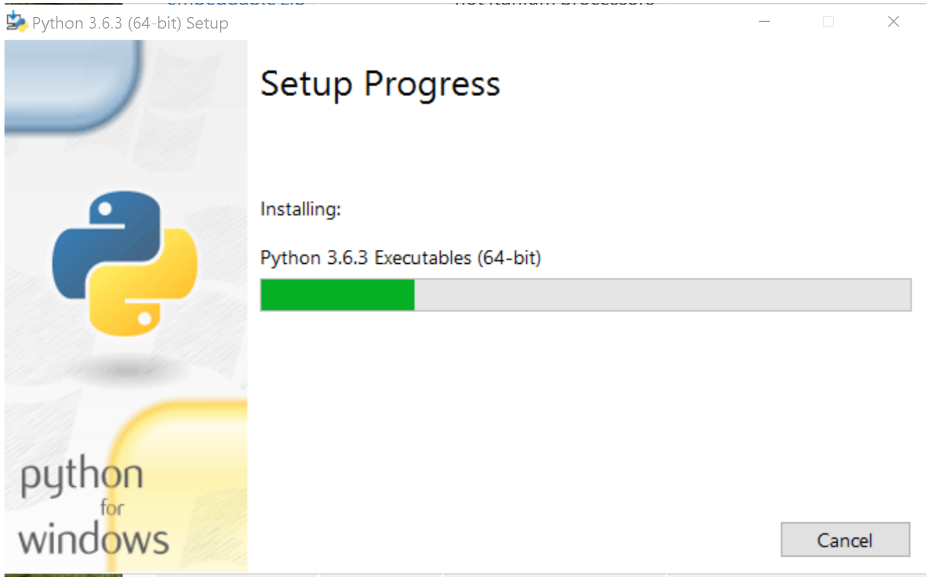
Install Now
C:\Users\sandi\AppData\Local\Programs\Python\Python36

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)
☐ Add Python 3.6 to PATH

should check this box!



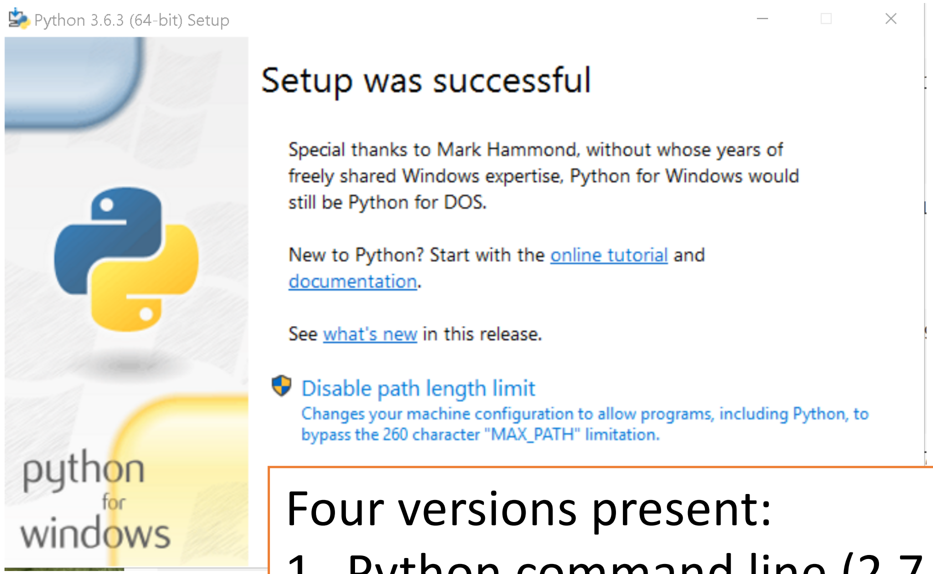
Python 3.6.3 (64-bit) Setup

Setup Progress

Installing:

Python 3.6.3 Executables (64-bit)

Windows 10: Finish install, start Python



Python 3.6.3 (64-bit) Setup

Setup was successful

Special thanks to Mark Hammond, without whose years of freely shared Windows expertise, Python for Windows would still be Python for DOS.

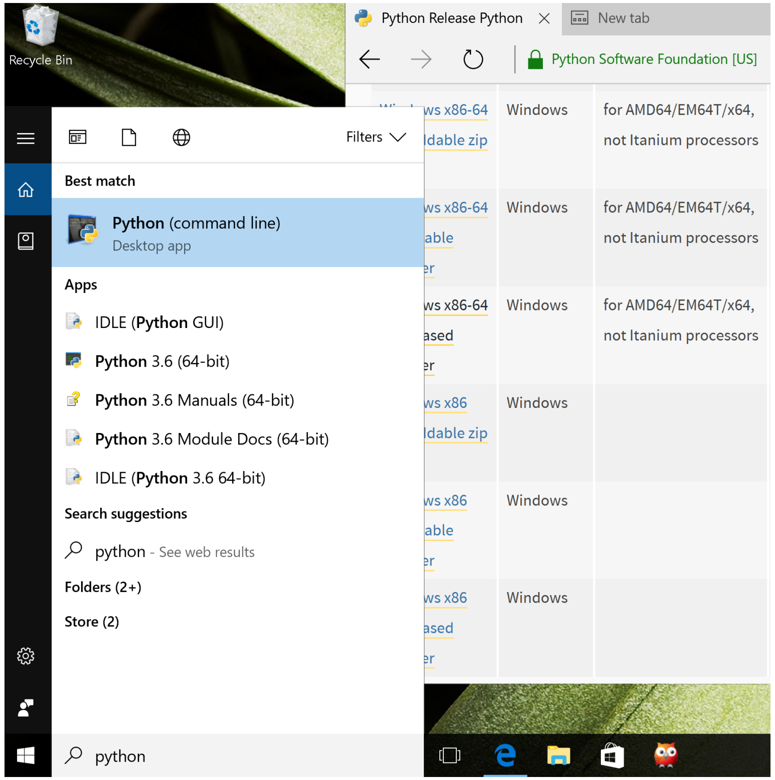
New to Python? Start with the online tutorial and documentation.

See what's new in this release.

Disable path length limit
Changes your machine configuration to allow programs, including Python, to bypass the 260 character "MAX_PATH" limitation.

Four versions present:

1. Python command line (2.7.x)
2. IDLE Python (2.7.x)
3. Python 3.6
4. IDLE Python 3.6



Best match

Python (command line)
Desktop app

Apps

IDLE (Python GUI)

Python 3.6 (64-bit)

Python 3.6 Manuals (64-bit)

Python 3.6 Module Docs (64-bit)

IDLE (Python 3.6 64-bit)

Search suggestions

python - See web results

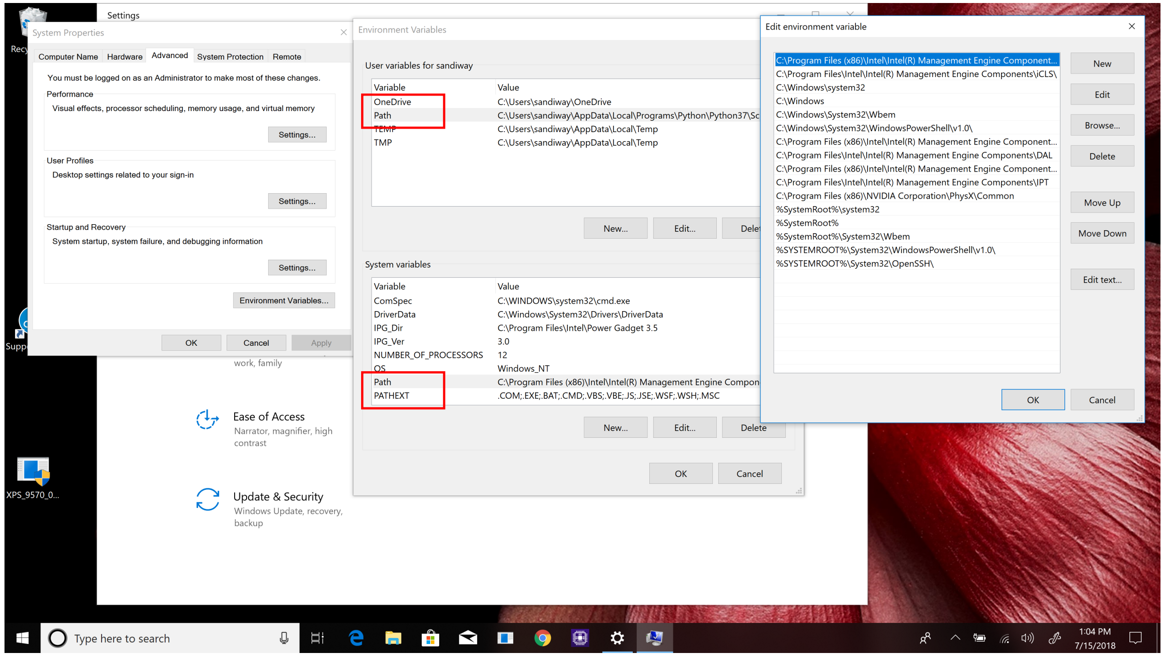
Folders (2+)

Store (2)

python

Windows 10: Environment Variables

if you need to manually add the directory for the Python executable to your PATH



You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles

Desktop settings related to your sign-in

Startup and Recovery

System startup, system failure, and debugging information

User variables for sandiway

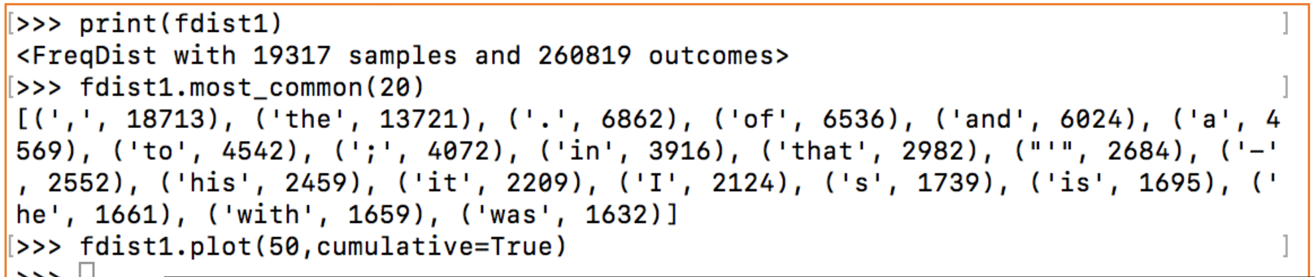
Variable	Value
OneDrive	C:\Users\sandiway\OneDrive
Path	C:\Users\sandiway\AppData\Local\Programs\Python\Python37\Sc
TEMP	C:\Users\sandiway\AppData\Local\Temp
TMP	C:\Users\sandiway\AppData\Local\Temp

System variables

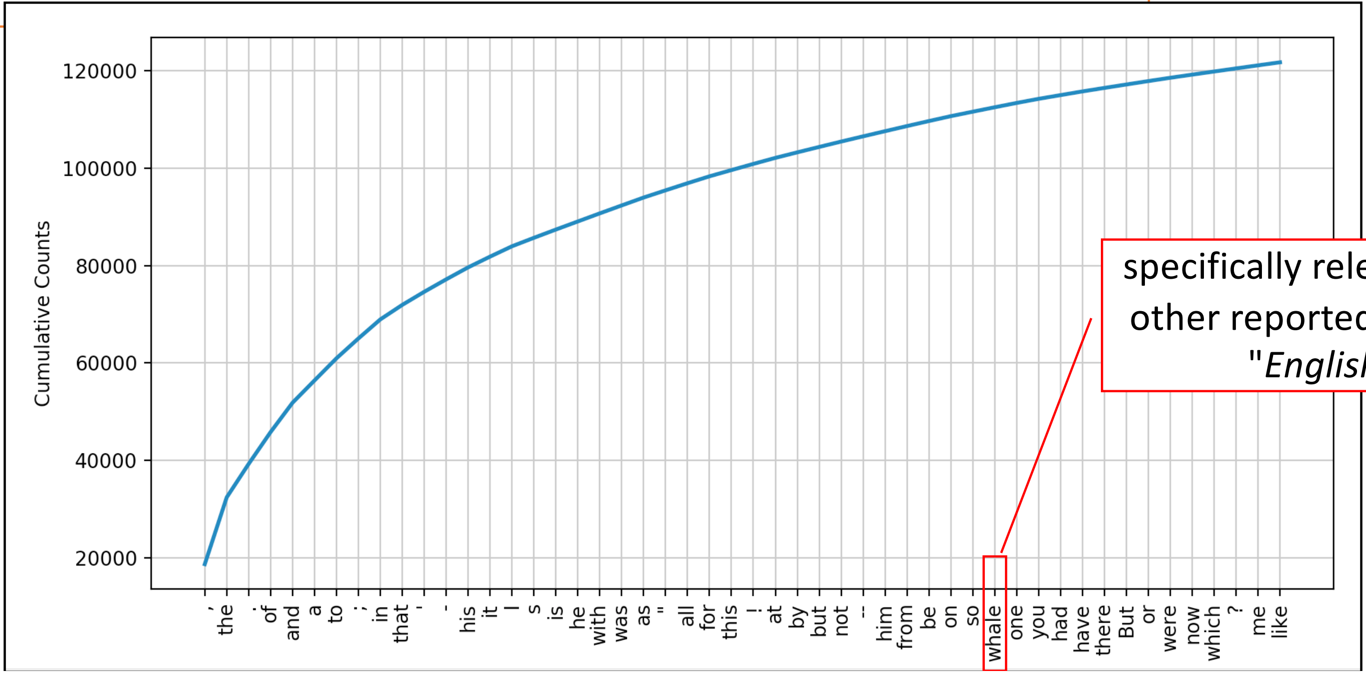
Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
IPG_Dir	C:\Program Files\Intel\Power Gadget 3.5
IPG_Ver	3.0
NUMBER_OF_PROCESSORS	12
OS	Windows_NT
Path	C:\Program Files (x86)\Intel\Intel(R) Management Engine Compon
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC

```
C:\Program Files (x86)\Intel\Intel(R) Management Engine Component...
C:\Program Files\Intel\Intel(R) Management Engine Components\iCLS\
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Program Files (x86)\Intel\Intel(R) Management Engine Component...
C:\Program Files\Intel\Intel(R) Management Engine Components\DAL
C:\Program Files (x86)\Intel\Intel(R) Management Engine Component...
C:\Program Files\Intel\Intel(R) Management Engine Components\IPT
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
```

Distribution of words in *Moby Dick*



```
>>> print(fdist1)
<FreqDist with 19317 samples and 260819 outcomes>
>>> fdist1.most_common(20)
[(',', 18713), ('the', 13721), ('.', 6862), ('of', 6536), ('and', 6024), ('a', 4
569), ('to', 4542), (';', 4072), ('in', 3916), ('that', 2982), ("'", 2684), ('-'
, 2552), ('his', 2459), ('it', 2209), ('I', 2124), ('s', 1739), ('is', 1695), ('
he', 1661), ('with', 1659), ('was', 1632)]
>>> fdist1.plot(50,cumulative=True)
>>> 
```

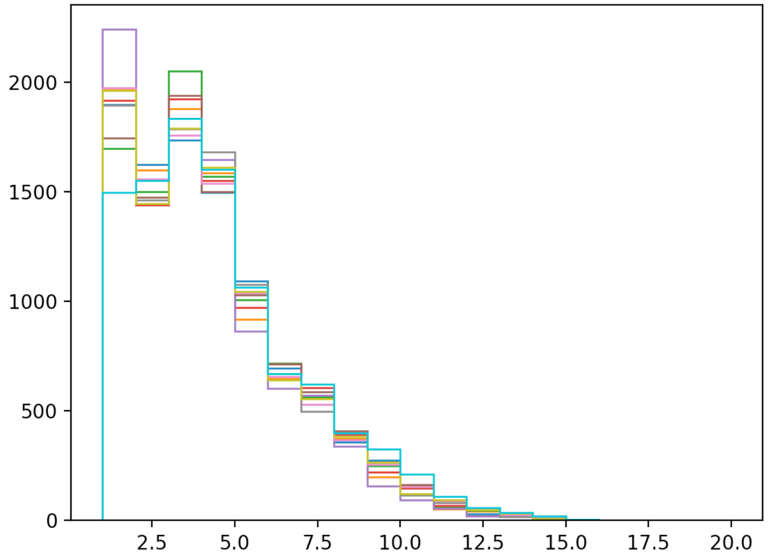


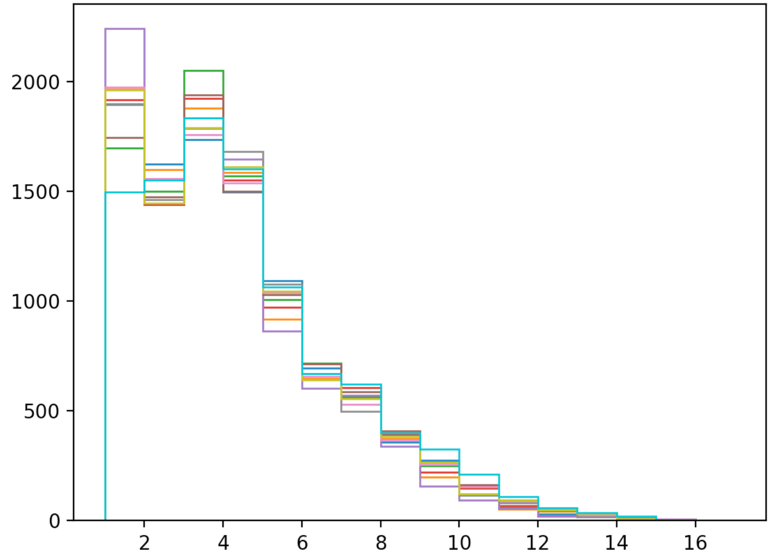
specifically relevant to Moby Dick; other reported words are generic "*English plumbing*"

Stylometry: compare word length distribution

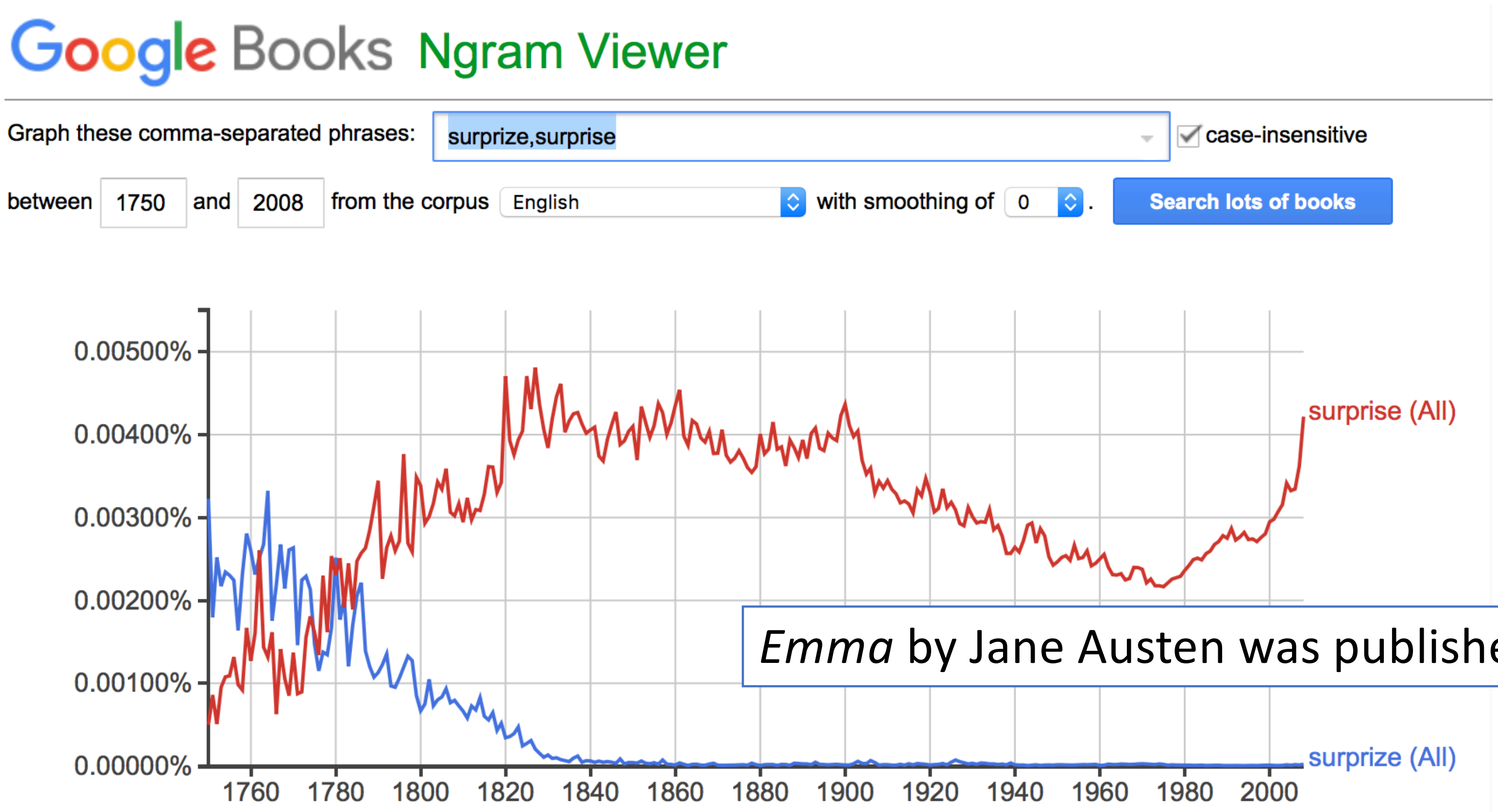
```
len1s = [len1[i*10000:i*10000+10000] for i in range(10)]
for l in len1s:
    plt.hist(l, bins=np.arange(min(l),max(l)+1), histtype='step')
plt.show()
```

Forensic linguistics





Google: relative frequency of two spellings



Google Books Ngram Viewer

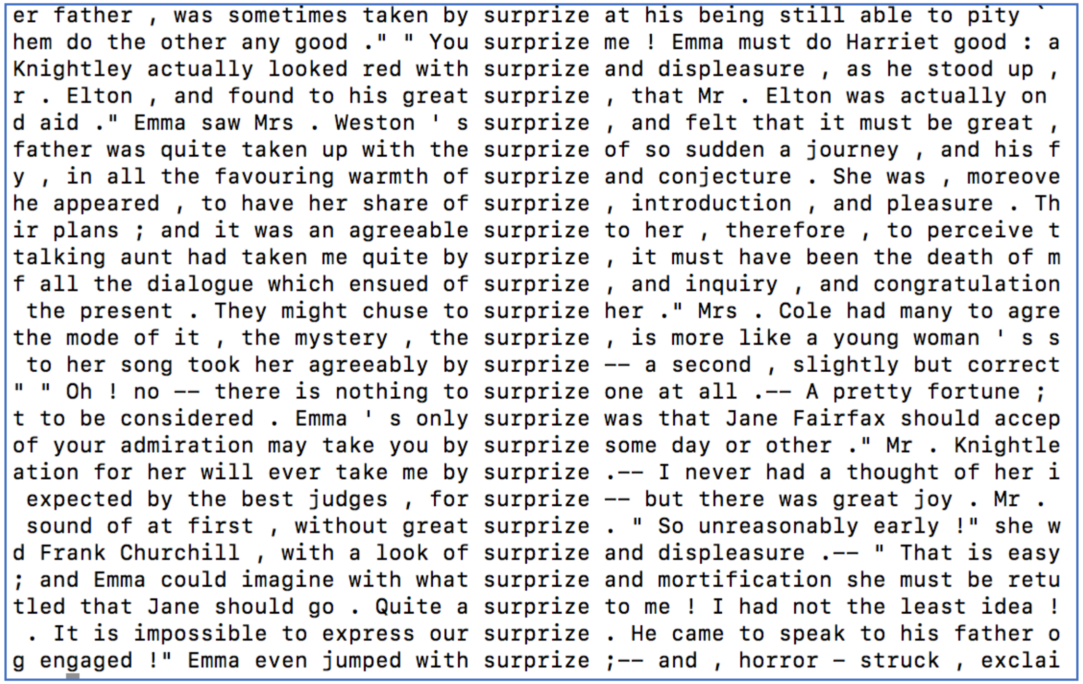
Graph these comma-separated phrases: surprize,surprise ☑ case-insensitive

between 1750 and 2008 from the corpus English with smoothing of 0.

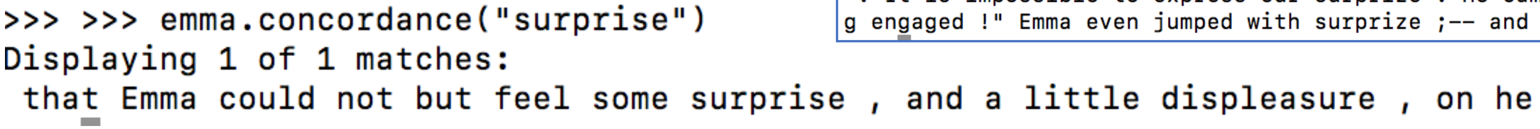
Emma by Jane Austen was published in 1815

Concordance

```
>>> import nltk
>>> emma = nltk.Text(nltk.corpus.gutenberg.words('austen-emma.txt'))
>>> emma.concordance("surprize")
Displaying 25 of 37 matches:
```

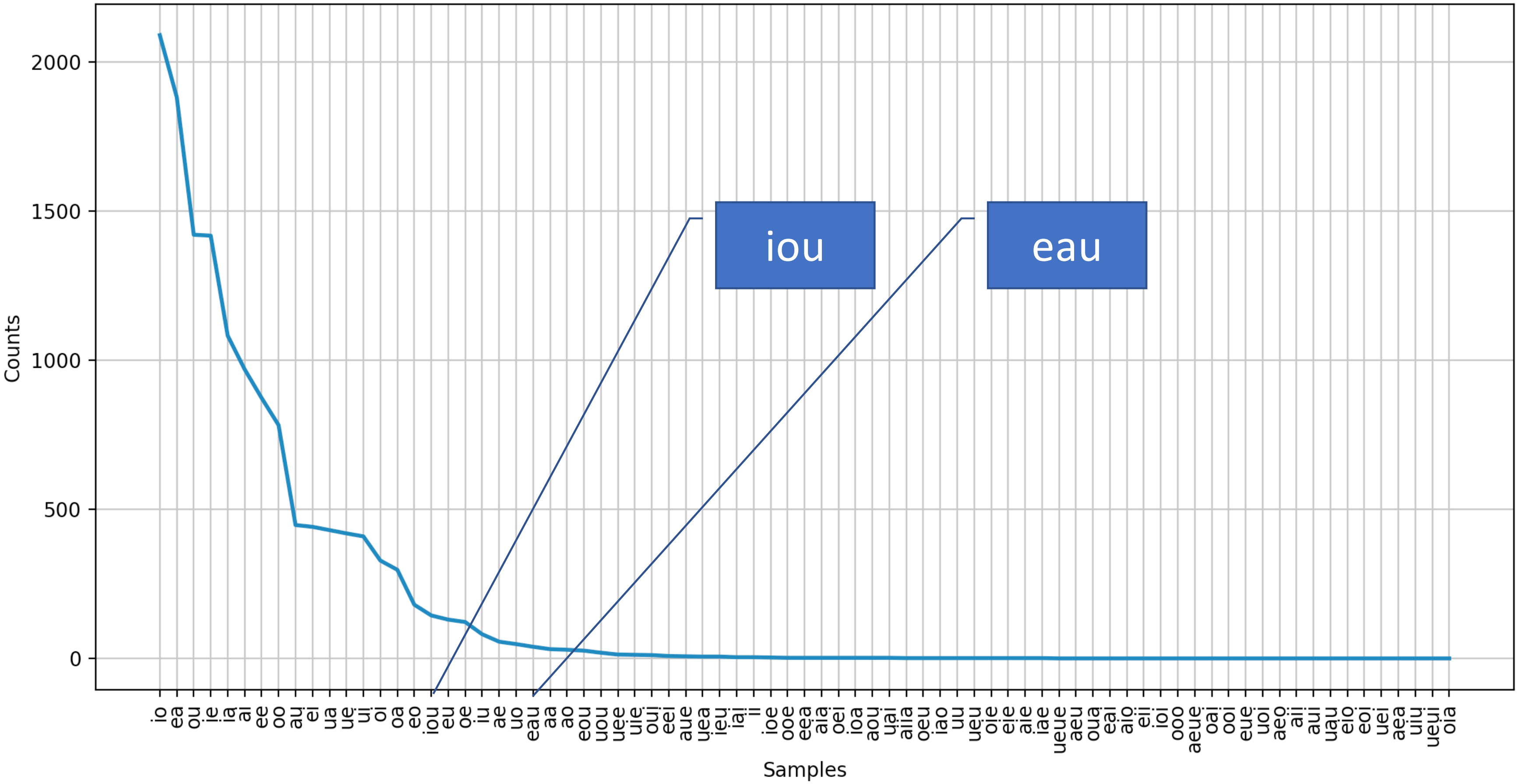


```
er father , was sometimes taken by surprize at his being still able to pity `
hem do the other any good ." " You surprize me ! Emma must do Harriet good : a
Knightley actually looked red with surprize and displeasure , as he stood up ,
r . Elton , and found to his great surprize , that Mr . Elton was actually on
d aid ." Emma saw Mrs . Weston ' s surprize , and felt that it must be great ,
father was quite taken up with the surprize of so sudden a journey , and his f
y , in all the favouring warmth of surprize and conjecture . She was , moreove
he appeared , to have her share of surprize , introduction , and pleasure . Th
ir plans ; and it was an agreeable surprize to her , therefore , to perceive t
talking aunt had taken me quite by surprize , it must have been the death of m
f all the dialogue which ensued of surprize , and inquiry , and congratulation
 the present . They might chuse to surprize her ." Mrs . Cole had many to agre
the mode of it , the mystery , the surprize , is more like a young woman ' s s
 to her song took her agreeably by surprize -- a second , slightly but correct
" " Oh ! no -- there is nothing to surprize one at all .-- A pretty fortune ;
t to be considered . Emma ' s only surprize was that Jane Fairfax should accep
of your admiration may take you by surprize some day or other ." Mr . Knightle
ation for her will ever take me by surprize .-- I never had a thought of her i
 expected by the best judges , for surprize -- but there was great joy . Mr .
 sound of at first , without great surprize . " So unreasonably early !" she w
d Frank Churchill , with a look of surprize and displeasure .-- " That is easy
; and Emma could imagine with what surprize and mortification she must be retu
tled that Jane should go . Quite a surprize to me ! I had not the least idea !
 . It is impossible to express our surprize . He came to speak to his father o
g engaged !" Emma even jumped with surprize ;-- and , horror - struck , exclai
```



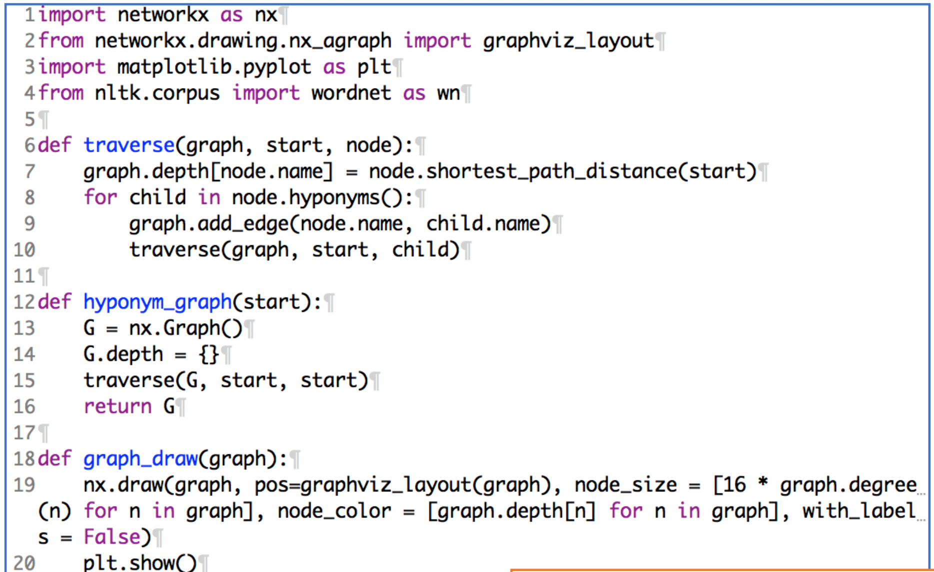
```
>>> >>> emma.concordance("surprise")
Displaying 1 of 1 matches:
 that Emma could not but feel some surprise , and a little displeasure , on he
```

Counting frequency of occurrences of sequences of vowels in English



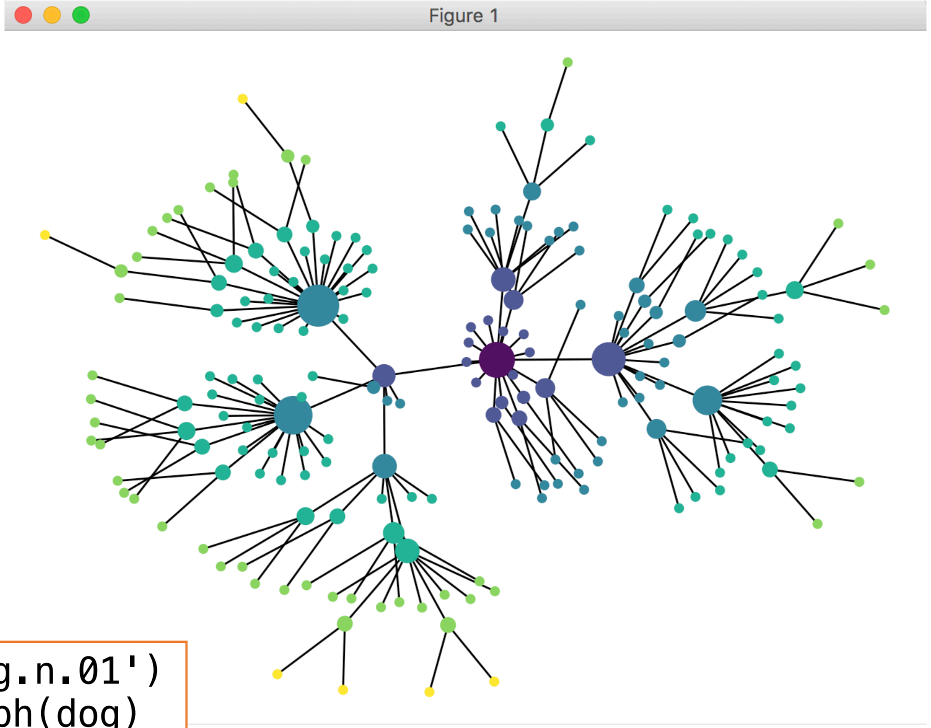
WordNet relations: types of dogs

`graph.py` on course website



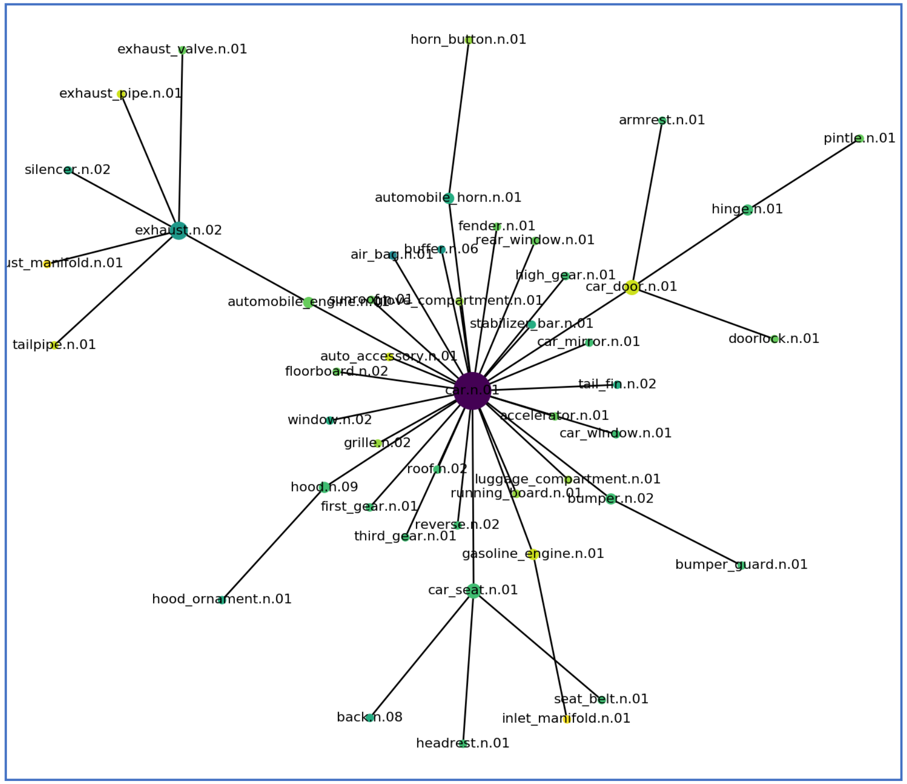
```python
import networkx as nx
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.pyplot as plt
from nltk.corpus import wordnet as wn

def traverse(graph, start, node):
    graph.depth[node.name] = node.shortest_path_distance(start)
    for child in node.hyponyms():
        graph.add_edge(node.name, child.name)
        traverse(graph, start, child)

def hyponym_graph(start):
    G = nx.Graph()
    G.depth = {}
    traverse(G, start, start)
    return G

def graph_draw(graph):
    nx.draw(graph, pos=graphviz_layout(graph), node_size = [16 * graph.degree(n) for n in graph], node_color = [graph.depth[n] for n in graph], with_labels = False)
    plt.show()
```

```python
dog = wn.synset('dog.n.01')
graph = hyponym_graph(dog)
graph_draw(graph)
```



WordNet relations: parts of a car

```
from nltk.corpus import wordnet as wn
c = wn.synset('car.n.01')
g = graph(c, 'part_meronyms')
graph_draw(g)
```



Term Programming Project

It's time to think about the end of the semester …

- Propose a programming project
  - half of your grade
- Send your proposal to me
- *If I give up the go-ahead, you can start working on it …*