**Chapter 2**

**Exercise 4:**

1945-Truman.txt

men: 2

women: 2

people: 10

1946-Truman.txt

men: 12

women: 7

people: 49

…

2005-GWBush.txt

men: 8

women: 11

people: 18

2006-GWBush.txt

men: 7

women: 7

people: 22

It looks like over time, the usage of “men” decreased, the usage of “women” increased, and “people” fluctuated wildly but remained the same overall

**Exercise 5:**

water.n.01: binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent

* Substance Meronyms: [ ('hydrogen.n.01'), ('oxygen.n.01')]
* Substance Holonyms: [ ('body\_of\_water.n.01'), ('ice.n.01'), ('ice\_crystal.n.01'), ('perspiration.n.01'), ('snowflake.n.01'), ('tear.n.01')]

computer.n.01:a machine for performing calculations automatically

* Part Meronyms: [ ('busbar.n.01'), ('cathode-ray\_tube.n.01'), ('central\_processing\_unit.n.01'), ('chip.n.07'), ('computer\_accessory.n.01'), ('computer\_circuit.n.01'), ('data\_converter.n.01'), ('disk\_cache.n.01'), ('diskette.n.01'), ('hardware.n.03'), ('keyboard.n.01'), ('memory.n.04'), ('monitor.n.04'), ('peripheral.n.01')]
* Part Holonyms: [ ('platform.n.03')]

**Exercise 7:**

Text 1 Moby Dick:

gledy - piggledy whale statements , however authentic , in these extracts , for

lave ? Tell me that . Well , then , however the old sea - captains may order me

ea - captains may order me about -- however they may thump and punch me about ,

needs be the sign of " The Trap ." However , I picked myself up and hearing a

the conclusion that such an idea , however wild , might not be altogether unwa

most obstreperously . I observed , however , that one of them held somewhat al

ade on the sea . In a few minutes , however , he was missed by his shipmates ,

bag ' s mouth . This accomplished , however , he turned round -- when , good he

te man into a purplish yellow one . However , I had never been in the South Sea

tle in the matter of my bedfellow . However , a good laugh is a mighty good thi

ight of the water it had absorbed . However , hat and coat and overshoes were o

pulpit , it had not escaped me that however convenient for a ship , these joint

lf baptized again . For the nonce , however , he proposed to sail about , and s

own and comrade ' s bill ; using , however , my comrade ' s money . The grinni

in to say it was on the starboard . However , by dint of beating about a little

a supper for us both on one clam ?" However , a warm savory steam from the kitc

owners till all is ready for sea . However , it is always as well to have a lo

fectly as he was known to me then . However , my thoughts were at length carrie

I got down our traps , resolving , however , to sleep ashore till the last . B

em !" " No need of profane words , however great the hurry , Peleg ," said Bil

a pilot . I was comforting myself , however , with the thought that in pious Bi

isely -- who knows ? Certain I am , however , that a king ' s head is solemnly

o scientific description . As yet , however , the sperm whale , scientific or p

IZONTAL TAIL . There you have him . However contracted , that definition is the

several varieties , most of which , however , are little known . Broad – nosed

…

Text 9 The Man who was Thursday:

m within this almost saintly oval , however , his face projected suddenly broad

ttle man with the black beard was , however , still contracted with something l

rd . By the sheer rush of routine , however , the motion would have been put an

retary said -- " We are wandering , however , from the point . The only questio

wax lady in evening dress . Snow , however , began to thicken and fall fast ;

y or trance . He was not inclined , however , to feel in this case any very com

. Somewhere behind Red Lion Court , however , he noticed a place where some ene

ouched his milk . One possibility , however , kept him hopeful and yet helpless

his satellites , that their faces , however fierce or sinister , became gradual

ow . Before I could fully recover , however , two or three of these admirers ra

cks during the conversation , which however unobtrusively performed --" " Syme

simplicity , " go to bed !" Syme , however , sat up in bed for a considerable

, if we are to act on it . I will , however , relate the occurrence in detail ,

rst considerable hurt . Dr . Bull , however , having been carefully coached by

. At the end of the four minutes , however , they saw that the Colonel was rig

of Attila . Swiftly as they rode , however , the whole rank still kept well to

he had three motor - cars . These , however , he seemed to use very sparingly ,

the hostility of the place . Now , however , the windows in the houses began o

, " might even call it unpleasant . However , I suppose those lights out in the

h the pier . Two or three figures , however , stood on the beginning of the sto

ondon . Twenty minutes afterwards , however , he came back and made quite a cla

d not see . Before the three cabs , however , could reach up to the spot , he h

before they came to Baker Street , however , he was seen to throw something fa

e a bulky parcel . On examination , however , its bulk was found to consist of

eation of the sun and moon . Here , however , they reckoned from a Christian Su

Moby Dick uses “however” as a sentence starter much more than The Man who was Thursday. It also seems to include “however” in a phrase rather than a single word more than The Man who was Thursday.

**Exercise 9:**

Comparing The Book of Genesis to “Chat Corpus”

Both of these texts are of similar length, with The Book of Genesis having length 44764 and Chat Corpus having length 45010. However, Chat Corpus has much more diverse vocabulary. The Book of Genesis has 2789 unique words, while The Chat Corpus has 6066. Ignoring common words like “it” and punctuation, some of the most common words used in Genesis include “unto”, “thou”, “shall”, “father”, “land”, and “Jacob”. These are all clearly old-ish English or religious words. On the other hand, the most frequent uncommon words in Chat Corpus include “JOIN”, “PART”, “lol”, “hi”, and “…”. JOIN and PART can be assumed to be chat functions, and “lol” and “hi” are both informal, highlighting the differences in “genre” between these two texts.

One interesting word that both texts used was “messenger”. In Genesis, it was used as such:

“we did unto him . And they sent a messenger unto Joseph , saying , Thy father”

While in Chat corpus it was used like:

“) lol U35 yeah i still have him on messenger JOIN JOIN he said i used to op wit”

These are clearly very different contexts

**Exercise 12:**

Using just the len method, the number of unique entries in the CMU dict is 123455. By filtering words to only those with alphabetic letters, the number of entries is 115533. By using the first method, the ratio of entries that have multiple pronunciations is 7.688%.

**Exercise 17:**

I got the following words as the most common for text 1: Moby Dick using the stopwords corpus

[('whale', 906), ('one', 889), ('like', 624), ('upon', 538), ('man', 508), ('ship', 507), ('Ahab', 501), ('ye', 460), ('old', 436), ('sea', 433), ('would', 421), ('head', 335), ('though', 335), ('boat', 330), ('time', 324), ('long', 318), ('said', 302), ('yet', 300), ('still', 299), ('great', 293), ('two', 285), ('seemed', 283), ('must', 282), ('Whale', 282), ('last', 277), ('way', 269), ('Stubb', 255), ('see', 253), ('Queequeg', 252), ('little', 247), ('round', 242), ('whales', 237), ('say', 237), ('three', 237), ('men', 236), ('thou', 232), ('may', 230), ('us', 228), ('every', 222), ('much', 218), ('could', 215), ('Captain', 215), ('first', 210), ('side', 208), ('hand', 205), ('ever', 203), ('Starbuck', 196), ('never', 195), ('good', 192), ('white', 191)]

This is case sensitive and has repeats, so here’s the most common 50 without repeats:

[('whale', 1226), ('one', 921), ('like', 647), ('upon', 566), ('man', 527), ('ship', 518), ('ahab', 511), ('ye', 472), ('sea', 455), ('old', 450), ('would', 432), ('though', 384), ('head', 345), ('yet', 345), ('boat', 336), ('time', 334), ('long', 333), ('captain', 329), ('still', 312), ('great', 306), ('said', 304), ('two', 298), ('must', 283), ('seemed', 283), ('white', 281), ('last', 277), ('see', 272), ('thou', 271), ('way', 271), ('whales', 268), ('stubb', 257), ('queequeg', 252), ('little', 249), ('round', 247), ('three', 245), ('say', 244), ('men', 244), ('sperm', 244), ('may', 240), ('first', 235), ('every', 232), ('well', 230), ('us', 228), ('much', 223), ('could', 216), ('good', 216), ('hand', 214), ('side', 208), ('ever', 206), ('never', 206)]

**Exercise 18:**

Most common bigrams in Monty Python & the Holy Grail including non-alphabetical and not caps-sensitive:

[(('ARTHUR', ':'), 217), (("'", 's'), 140), ((']', '['), 94), (('!', '['), 82), ((':', 'Oh'), 82), (('Oh', ','), 79), (("'", 't'), 77), (('LAUNCELOT', ':'), 76), (('1', ':'), 75), (('#', '1'), 75), (('.', 'ARTHUR'), 75), (('GALAHAD', ':'), 69), ((':', '['), 67), ((':', 'I'), 65), (('!', 'ARTHUR'), 63), (('FATHER', ':'), 63), (('BEDEVERE', ':'), 61), (('KNIGHT', ':'), 59), ((',', 'I'), 58), ((':', 'No'), 56), ((':', 'What'), 55), ((':', 'Well'), 52), (('I', "'"), 51), (('Well', ','), 50), (('VILLAGER', '#'), 47), (("'", 're'), 41), (('#', '2'), 41), (('2', ':'), 40), (('of', 'the'), 39), (('ROBIN', ':'), 39), (('.', 'I'), 38), ((':', 'Yes'), 38), (('No', ','), 38), (('.', '['), 38), ((',', 'you'), 36), (('Ni', '!'), 36), (('boom', ']'), 35), (('?', 'ARTHUR'), 34), (("'", 'm'), 34), (('[', 'boom'), 34), (('BLACK', 'KNIGHT'), 32), (('GUARD', '#'), 32), (('witch', '!'), 31), ((':', 'You'), 30), (('it', '!'), 30), ((',', 'and'), 30), ((',', 'no'), 30), (('[', 'singing'), 29), (('HEAD', 'KNIGHT'), 29), (('TIM', ':'), 28)]

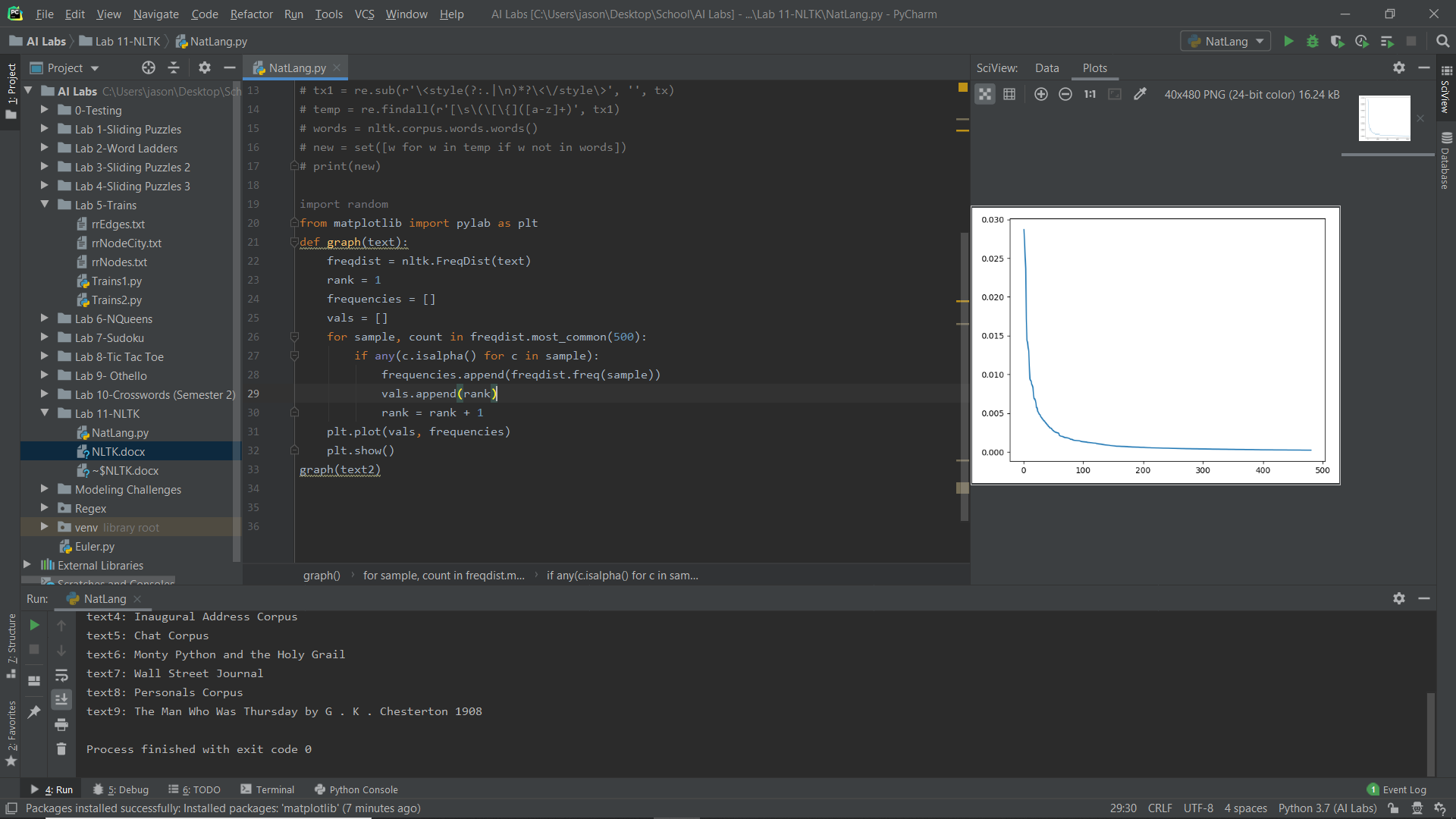
Here are the bigrams of only alphabetical and caps-sensitive:

[(('OF', 'THE'), 39), (('A', 'WITCH'), 35), (('HA', 'HA'), 34), (('BLACK', 'KNIGHT'), 33), (('HEAD', 'KNIGHT'), 29), (('CLOP', 'CLOP'), 26), (('IN', 'THE'), 26), (('KNIGHTS', 'OF'), 25), (('BURN', 'HER'), 22), (('THE', 'HOLY'), 22), (('I', 'AM'), 21), (('WHAT', 'IS'), 21), (('FRENCH', 'GUARD'), 21), (('MUMBLE', 'MUMBLE'), 20), (('GOING', 'TO'), 19), (('SIR', 'ROBIN'), 19), (('HOLY', 'GRAIL'), 19), (('SQUEAK', 'SQUEAK'), 19), (('OF', 'NI'), 19), (('RUN', 'AWAY'), 18), (('KING', 'ARTHUR'), 17), (('COME', 'ON'), 17), (('ARE', 'YOU'), 16), (('IS', 'YOUR'), 16), (('IT', 'IS'), 15), (('I', 'DON'), 15), (('I', 'HAVE'), 15), (('THE', 'GRAIL'), 15), (('SAW', 'SAW'), 15), (('IS', 'A'), 14), (('WE', 'ARE'), 14), (('SHUT', 'UP'), 14), (('TO', 'TELL'), 14), (('S', 'GOING'), 14), (('I', 'CAN'), 13), (('ALL', 'RIGHT'), 13), (('DO', 'YOU'), 12), (('S', 'NOT'), 12), (('BRING', 'OUT'), 12), (('YOUR', 'DEAD'), 12), (('TO', 'BE'), 12), (('SIR', 'LAUNCELOT'), 12), (('CARTOON', 'CHARACTER'), 12), (('THANK', 'YOU'), 12), (('M', 'NOT'), 11), (('OUT', 'YOUR'), 11), (('AND', 'THE'), 11), (('YOU', 'ARE'), 11), (('MY', 'LIEGE'), 11), (('IF', 'YOU'), 11)]

**Exercise 23:**

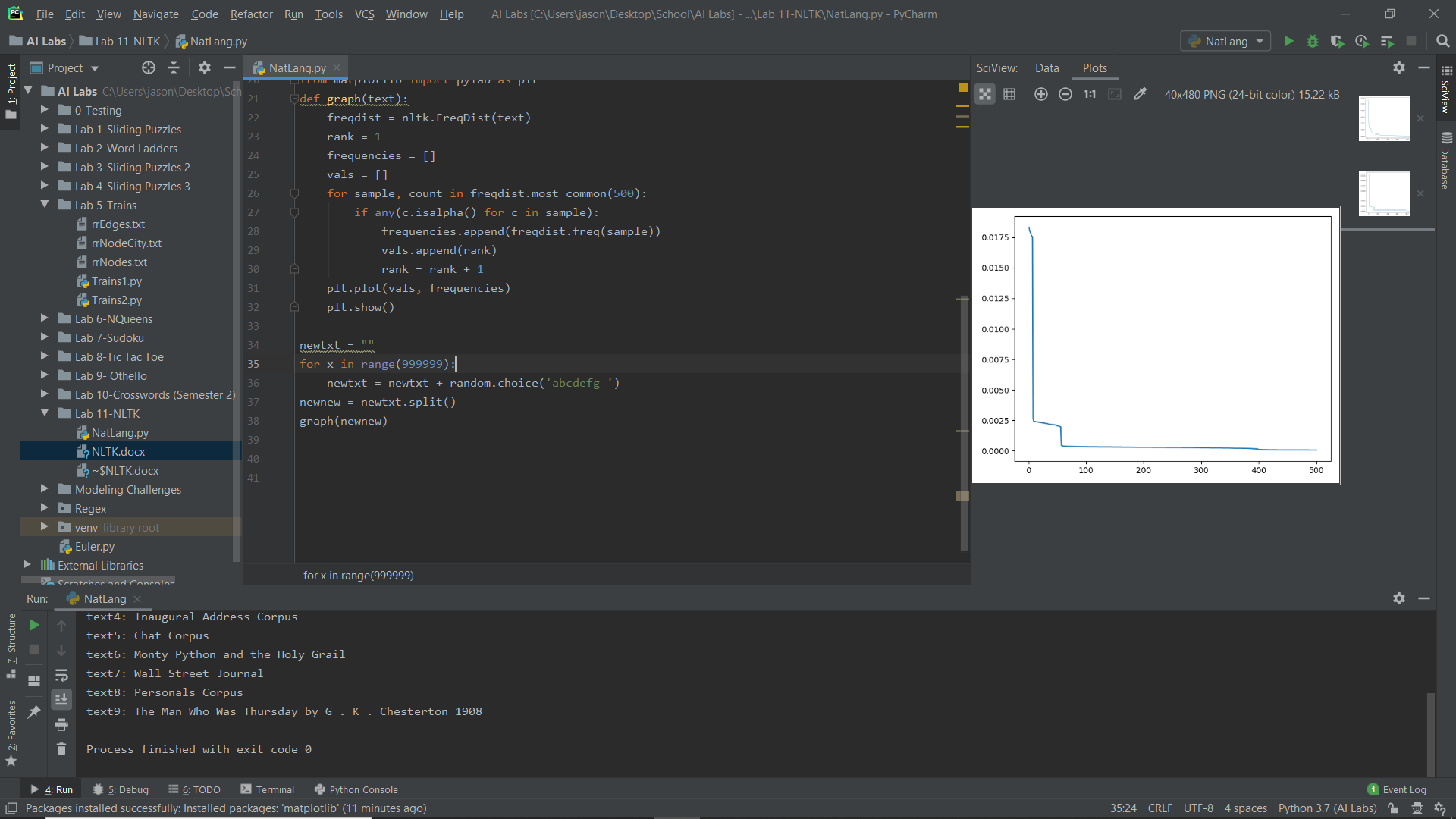
Part A:

from matplotlib import pylab as plt  
def graph(text):  
 freqdist = nltk.FreqDist(text)  
 rank = 1  
 frequencies = []  
 vals = []  
 for sample, count in freqdist.most\_common(500):  
 if any(c.isalpha() for c in sample):  
 frequencies.append(freqdist.freq(sample))  
 vals.append(rank)  
 rank = rank + 1  
 plt.plot(vals, frequencies)  
 plt.show()  
graph(text2)



Part B:

newtxt = ""  
for x in range(999999):  
 newtxt = newtxt + random.choice('abcdefg ')  
newnew = newtxt.split()  
graph(newnew)



It looks like Zipf’s law generally holds up, even in random “words”. However, the random nature of random words causes there to be sudden jumps, presumably when word length changes.

**Exercise 27:**

I kept count of how many of each word type there were along with total synset count for each word type. I got average 1.5455227238638067 for nouns, 2.1866273523545225 for verbs, 1.4336283185840708 for adjectives, and 1.2532916759651864 for adverbs.

**Chapter 3**

**Exercise 20:**

I wrote a short program to go to worldometers.info/coronavirus and find the current number of coronavirus cases.

from urllib import request  
url = "https://www.worldometers.info/coronavirus/coronavirus-cases/"  
response = request.urlopen(url)  
raw = response.read().decode('utf8')  
type(raw)  
pos = raw.find("There are currently")  
text = raw[pos:pos+20] + raw[pos+52:pos+75] + " of coronavirus in the world"  
print(text)

As of 6:50 PM on 3/9/2020, it prints: “There are currently 114,285 confirmed cases of coronavirus in the world”.

**Exercise 22:**

I was having trouble accessing the BeautifulSoup thing, so I did my best without it.

url = "https://www.bbc.co.uk"  
response = request.urlopen(url)  
raw = response.read().decode('utf8')  
type(raw)  
tx = re.sub(r'\<script(?:.|\n)\*?\<\/script\>', '', raw)  
tx1 = re.sub(r'\<style(?:.|\n)\*?\<\/style\>', '', tx)  
temp = re.findall(r'[\s\(\[\{]([a-z]+)', tx1)  
words = nltk.corpus.words.words()  
new = set([w for w in temp if w not in words])  
print(new)

This prints: {'optimised', 'rejected', 'angers', 'https', 'gte', 'datetime', 'ideas', 'src', 'heroes', 'inline', 'legends', 'discovers', 'picks', 'overreacting', 'bbc', 'recipes', 'reveals', 'streamed', 'monks', 'gt', 'services', 'joins', 'downs', 'ups', 'says', 'favourite', 'flatbread', 'meals', 'asynchronously', 'lang', 'gems', 'plays', 'hacks', 'headlines', 'rugby', 'brings', 'missed', 'has', 'mins', 'prices', 'spellcheck', 'stacked', 'doctors', 'laverne', 'results', 'href', 'uk', 'placeholder', 'autocomplete', 'rules', 'paleo', 'sci', 'coronavirus', 'opinions', 'http', 'moments', 'autocapitalize', 'placed', 'stories', 'countries', 'autocorrect', 'tv', 'meets', 'scores', 'hp', 'lte', 'teams', 'charset', 'brewers', 'including', 'gives', 'parents', 'endif', 'modules', 'recommended', 'sites', 'draws', 'prepares', 'browsers', 'symptoms', 'equals', 'lt', 'lockdown', 'highlights', 'tips', 'iplayer', 'attempts', 'things', 'dir', 'complaints', 'html', 'shared'}

As we can see, it still has some JS elements that weren’t filtered out, like “html”. So it isn’t perfect

**Chapter 6**

**Exercise 4:**

contains(doubts) = True pos : neg = 9.8 : 1.0

contains(sans) = True neg : pos = 8.8 : 1.0

contains(hugo) = True pos : neg = 7.8 : 1.0

contains(effortlessly) = True pos : neg = 7.5 : 1.0

contains(dismissed) = True pos : neg = 7.1 : 1.0

contains(mediocrity) = True neg : pos = 6.9 : 1.0

contains(overwhelmed) = True pos : neg = 6.4 : 1.0

contains(fabric) = True pos : neg = 6.4 : 1.0

contains(wires) = True neg : pos = 6.2 : 1.0

contains(wits) = True pos : neg = 5.8 : 1.0

contains(topping) = True pos : neg = 5.8 : 1.0

contains(ugh) = True neg : pos = 5.7 : 1.0

contains(uplifting) = True pos : neg = 5.7 : 1.0

contains(bruckheimer) = True neg : pos = 5.6 : 1.0

contains(bounce) = True neg : pos = 5.6 : 1.0

contains(lang) = True pos : neg = 5.1 : 1.0

contains(understands) = True pos : neg = 4.5 : 1.0

contains(matheson) = True pos : neg = 4.4 : 1.0

contains(controversy) = True pos : neg = 4.3 : 1.0

contains(quicker) = True neg : pos = 4.3 : 1.0

contains(maxwell) = True neg : pos = 4.3 : 1.0

contains(locks) = True neg : pos = 4.3 : 1.0

contains(tsui) = True neg : pos = 4.3 : 1.0

contains(admired) = True pos : neg = 4.2 : 1.0

contains(cronenberg) = True pos : neg = 3.9 : 1.0

contains(derivative) = True neg : pos = 3.8 : 1.0

contains(attorney) = True pos : neg = 3.8 : 1.0

contains(existential) = True pos : neg = 3.7 : 1.0

contains(bandits) = True pos : neg = 3.7 : 1.0

contains(restoring) = True pos : neg = 3.7 : 1.0

Most of these make sense: effortlessly as very positive and mediocrity as very negative seem correct. However, I was somewhat confused by how “doubts” could be positive. Perhaps it’s often used in the context of something like “despite doubts by critics, the movie was good”.