

# NATURAL LANGUAGE PROCESSESING

NAME: RAHUL GUDIVADA

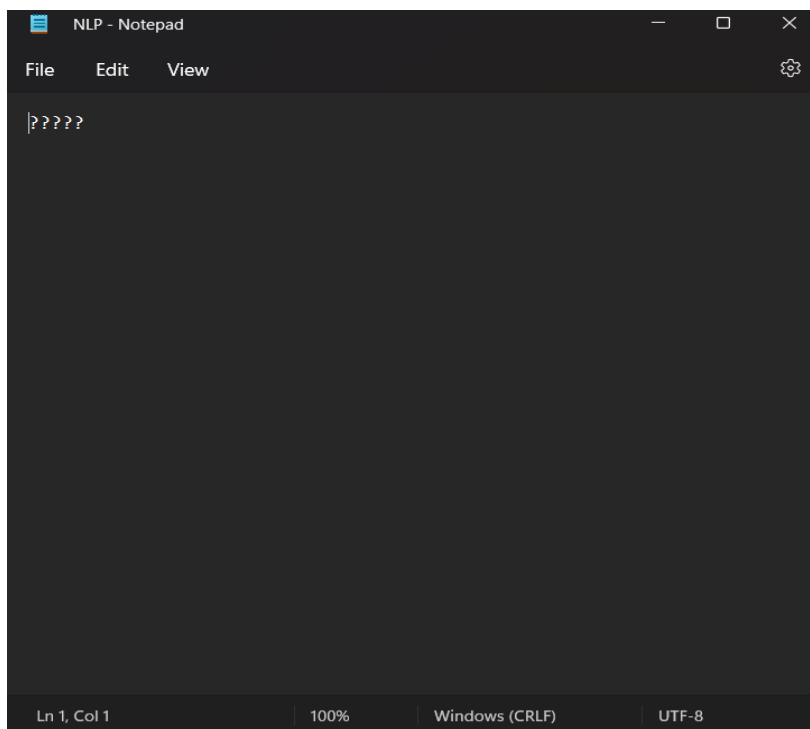
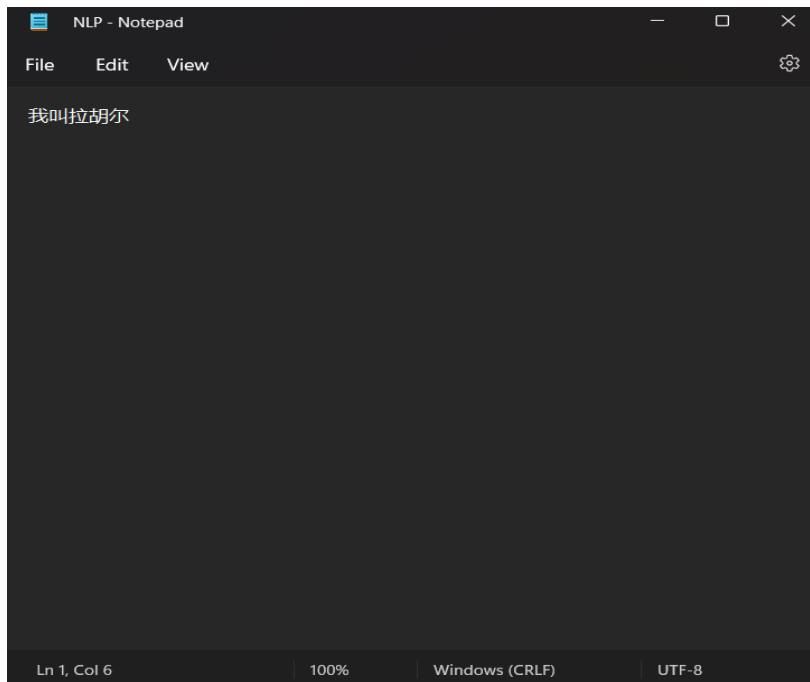
REG NO: 19BCE2469

DATE: 15/07/2022

---

## Task 1:

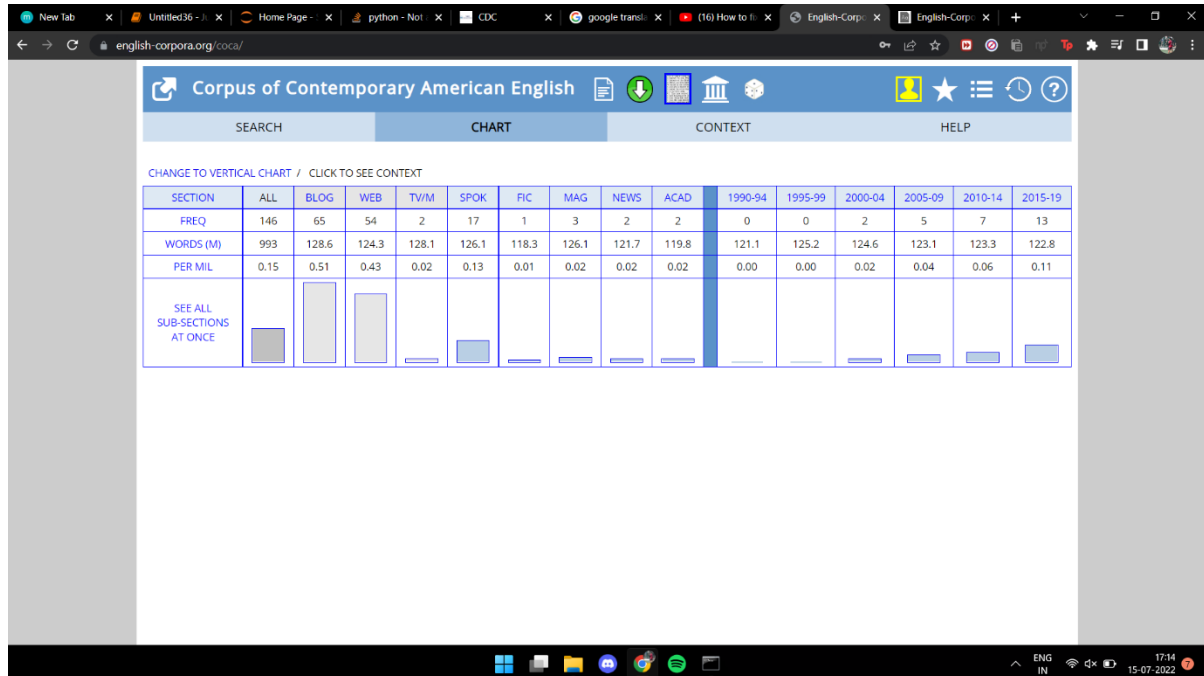
Encoding:



## Task 2:

### 2.1

#### COCA, Story from Data



From the graph, we can clearly see that Barrack Obama wasn't famous during the time period of 1990-94 till 2000-2004. But slowly slowly he started gaining popularity as he started entering politics. From these graphs we can see that from 2005 onwards, people started looking for him more in internet and how famous he became.

We can also see the frequency that is the frequency of him being in blogs is 65, whereas in WEB it is 54.

We can also see how many times his name as a WORD(M) is used in Blogs which is 65, whereas in WEB it is 54.

## 2.2

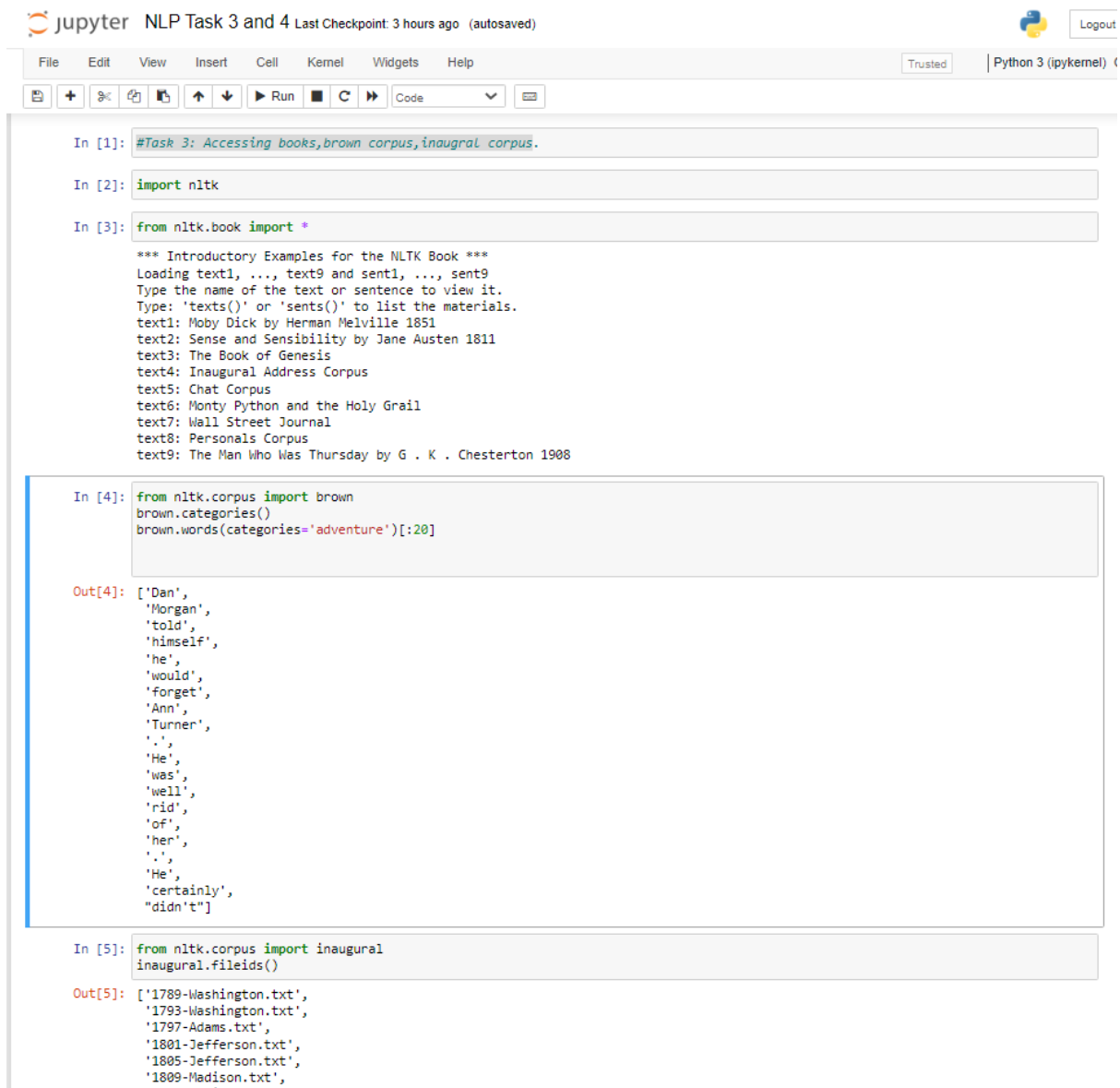
## Concordance

The screenshot displays the Corpus of Contemporary American English (COCA) website interface. The top navigation bar includes tabs for SEARCH, FREQUENCY, CONTEXT, and OVERVIEW. The main content area shows a search result for the term "Barrack Obama". The search parameters are: FIND SAMPLE: 100, PAGE: << < 1 / 2 > >>. The results are displayed in a table with columns for Rank, Year, Source, and Text. The text column shows the context of the search term within various news articles and blog posts from 2012. The interface also includes a sidebar with a "CLICK FOR MORE CONTEXT" button and a "HELP" button. The bottom of the page shows the URL: <https://www.english-corpora.org/coca/v4.asp?l=5029462&ID=1189410097>.

Rank	Year	Source	Text
1	2012	WEB	huffingtonpost.com . But considering that progressives currently hold the reigns of power (I mean, <b>Barrack Obama</b> is a progressive, right?), wouldn't right-wingers actually b
2	2012	WEB	theblaze.com it regardless of proof being put right in front of them. My God, <b>Barrack Obama</b> (or whatever the hell his name is) is proof! ALL main
3	2012	WEB	abcnews.go.com 18, 2008, 11:48 pm 11:48 pm # What I have learned today: <b>Barrack Obama</b> has a Xerox machine and knows how to use it. Michelle Obama is
4	2012	WEB	...walski.wordpress.com are one thing, really another. All five of these states were won by <b>Barrack Obama</b> four years ago. In order to succeed, the Republican ticket must turn
5	2012	WEB	spectator.org big contributors to Republicans) and 2) Crash the Economy and insure his candidate <b>Barrack Obama</b> is elected President. Wayne 5.13.11 8:57AM # But i
6	2012	WEB	...ashingtonmonthly.com the recession. # Bob Flanagan on March 09, 2012 10:11 AM: # <b>Barrack Obama</b> is in over his head and out of his depth. He is a
7	2012	WEB	glennbeck.com 's sake you better not push fat boy Chris Christie (who went campaigning with <b>Barrack Obama</b> in New Jersey the day after Sandy hit) on me because I w
8	2012	WEB	glennbeck.com dreamers fled. Today a different kind of dream is taking shape in America. <b>Barrack Obama's</b> dream for America is one that rides hard on the back of the
9	2012	WEB	theblaze.com 2004, by the Kenyan Standard Times... Kenyan-born US Senate hopeful, <b>Barrack Obama</b> , appeared set to take over the Illinois Senate seat after his ma
10	2012	WEB	eclectablog.com , catching Bin Laden wasn't his fault. Honestly, do you think that <b>Barrack Obama</b> was the strategist that organized the intel and coordinated the op? No;
11	2012	WEB	cbsnews.com . Sure they are lying through their teeth with the approval of that honest man <b>Barrack Obama</b> not making a peep # This Obama anti Romney ad is so
12	2012	WEB	abcnews.go.com . The only network that challenged Obama was FOX. I do not know this <b>Barrack Obama</b> person. Call me what you like, but his name is too Islamic
13	2012	WEB	zerohedge.com party. The only wasted vote I see is one for either Mitt Romney or <b>Barrack Obama</b> . # So, seeing as both the "Mainstream" candidates work for
14	2012	WEB	...ng.blogs.nytimes.com won by Obama, despite the close race. # In my opinion, President <b>Barrack Obama</b> has done nothing of tremendous value for our country. I respect him
15	2012	WEB	huffingtonpost.com ... # Got ta say a lot of people agree with you. They think <b>Barrack Obama</b> is a Muslim and a non citizen because of his name too. So
16	2012	WEB	clashdaily.com ) began leading prayers at every service specifically for "our President and leader, <b>Barrack Obama</b> ". I will not attend services there again until he is rem
17	2012	WEB	enduswars.org this President must too be called to account and impeachment proceedings should be initiated against <b>Barrack Obama</b> . " 5/15/11 U.S. Policy is Rooted i
18	2012	WEB	politico.com shouldnt matter the color of the Voter - but to Eric Holder and by extension <b>Barrack Obama</b> color is what matters! Obama spent 20 years listening to r
19	2012	WEB	guardianlv.com 73856 # See if the United States of America send <b>Barrack Obama</b> back to the White House for a second term, or will they change
20	2012	WEB	abcnews.go.com 2012, 2:01 pm 2:01 pm # Yet another LAW ignored by the administration of <b>Barrack Obama</b> . This is a violation of Federal Law, and Union Regulations, ye

## Task 3:

### Accessing books, brown corpus, inaugural corpus



The image shows a Jupyter Notebook interface with the title "NLP Task 3 and 4 Last Checkpoint: 3 hours ago (autosaved)". The interface includes a top menu bar with options like File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu is a toolbar with icons for saving, running, and other notebook functions. The notebook contains several code cells:

```
In [1]: #Task 3: Accessing books,brown corpus,inaugral corpus.
```

```
In [2]: import nltk
```

```
In [3]: 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  
text5: Chat Corpus  
text6: Monty Python and the Holy Grail  
text7: Wall Street Journal  
text8: Personals Corpus  
text9: The Man Who Was Thursday by G . K . Chesterton 1908

```
In [4]: from nltk.corpus import brown  
brown.categories()  
brown.words(categories='adventure')[:20]
```

```
Out[4]: ['Dan',  
        'Morgan',  
        'told',  
        'himself',  
        'he',  
        'would',  
        'forget',  
        'Ann',  
        'Turner',  
        '.',  
        'He',  
        'was',  
        'well',  
        'rid',  
        'of',  
        'her',  
        '.',  
        'He',  
        'certainly',  
        "didn't"]
```

```
In [5]: from nltk.corpus import inaugural  
inaugural.fileids()
```

```
Out[5]: ['1789-Washington.txt',  
        '1793-Washington.txt',  
        '1797-Adams.txt',  
        '1801-Jefferson.txt',  
        '1805-Jefferson.txt',  
        '1809-Madison.txt',  
        '1817-Madison.txt']
```

```
'1853-Pierce.txt',
'1857-Buchanan.txt',
'1861-Lincoln.txt',
'1865-Lincoln.txt',
'1869-Grant.txt',
'1873-Grant.txt',
'1877-Hayes.txt',
'1881-Garfield.txt',
'1885-Cleveland.txt',
'1889-Harrison.txt',
'1893-Cleveland.txt',
'1897-McKinley.txt',
'1901-McKinley.txt',
'1905-Roosevelt.txt',
'1909-Taft.txt',
'1913-Wilson.txt',
'1917-Wilson.txt',
'1921-Harding.txt',
'1925-Coolidge.txt',
'1929-Hoover.txt',
'1933-Roosevelt.txt',
'1937-Roosevelt.txt',
'1941-Roosevelt.txt',
'1945-Roosevelt.txt',
'1949-Truman.txt',
'1953-Eisenhower.txt',
'1957-Eisenhower.txt',
'1961-Kennedy.txt',
'1965-Johnson.txt',
'1969-Nixon.txt',
'1973-Nixon.txt',
'1977-Carter.txt',
'1981-Reagan.txt',
'1985-Reagan.txt',
'1989-Bush.txt',
'1993-Clinton.txt',
'1997-Clinton.txt',
'2001-Bush.txt',
'2005-Bush.txt',
'2009-Obama.txt',
'2013-Obama.txt',
'2017-Trump.txt',
'2021-Biden.txt']
```

```
In [6]: inaugural.words(fileids='2009-Obama.txt')
#find for Lincoln, trump
#compare first 5 words of Lincoln and obama and then Lincoln and trump
```

```
Out[6]: ['My', 'fellow', 'citizens', ':', 'I', 'stand', 'here', ...]
```

```
In [7]: inaugural.words(fileids='1861-Lincoln.txt')
```

```
Out[7]: ['Fellow', '-', 'Citizens', 'of', 'the', 'United', ...]
```

```
In [8]: inaugural.words(fileids='2017-Trump.txt')
```

```
Out[8]: ['Chief', 'Justice', 'Roberts', ',', 'President', ...]
```

## Task 4:

Experiencing frequency distribution and conditional frequency distribution.

```
In [9]: #Task 4: Experiencing frequency distribution and conditional frequency distribution.

In [10]: text1='Thor: Love and Thunder is a 2022 American superhero film based on Marvel Comics featuring the character Thor, produced by
fd=nltk.FreqDist(text1.split())
fd

Out[10]: FreqDist({'the': 8, 'and': 7, 'is': 4, 'film': 3, 'Marvel': 3, 'by': 3, 'to': 3, 'Thor': 2, 'Studios': 2, 'Waititi': 2, ...})

In [ ]:

In [11]: from nltk.probability import ConditionalFreqDist
cfd=ConditionalFreqDist((len(word),word)for word in text1.split())
cfd[5]

Out[11]: FreqDist({'Thor': 2, 'based': 1, 'Thor,': 1, 'Taika': 1, 'stars': 1, 'Chris': 1, 'Bale,': 1, 'Tessa': 1, 'film,': 1, 'inner': 1})

In [ ]:
```

## Task 5:

Suggest a Corpus Application

### Corpus Approaches to Social-Media

The language of online communities, linguistic variety in brief media texts, and the use of images in computer-mediated communication are all examined from the perspective of corpora. The collection's in-depth descriptions of the methodological aspects of working with social media corpora are one of its strongest features. The collection includes research using novel and creative research approaches for the analysis of multimodal material and atypical corpus texts, as well as research using conventional corpus linguistic methods to social media data.

Even if social media has become increasingly popular as a news source, there are still concerns about how easily it may disseminate rumors and false information. It has been challenging to systematically examine this phenomenon, however, because it is necessary to gather extensive, objective data and make in-situ assessments of its accuracy. A corpus called CREDBANK was created to fill this gap by methodically fusing computer and human computing. CREDBANK is specifically a corpus of tweets, topics, events, and related human credibility assessments.