DATA SCIENCE REPORT

WAJAHAT MASOOD	SP19-BSE-003
ABDULLAH RAUF	SP19-BSE-007
SAIF UR REHMAN	SP19-BSE-013

Assignment :
Text Data Preprocessing & Analysis

Group 1

Workflow

Data Extraction

Data is been entreated from twitter about the president of Pakistan.

Cleaning of data

Manual cleaning of bogus data

Tokenization of the data

By using NLTK python library we tokenize our data.

Stop Word Removal

By Using GENSIM.PARSING & REMOVE_STOPWORDS we remove the stop words to make analysis better

Stemming

By Using NLTK STEMMER & PORTER STEMMER we Do STEMMING

Workflow Cont...

Lemmatization

By using NLTK STEMMER & WORD NET LEMATIZER we do Lemmatization.

WORD CLOUD

By Using WORD CLOUD LIB, PIL IMAGE, NUMPY, URL LIB, MATPLOT we draw a word cloud

VSM

By using COSIN_SIMILARITY LIB from SK learn. Metrices

N Gram

By Using NKTK we import N gram we take UNI gram, bi gram, Tri Gram up to n gram

TOOLS WE USE

1

PyCharm

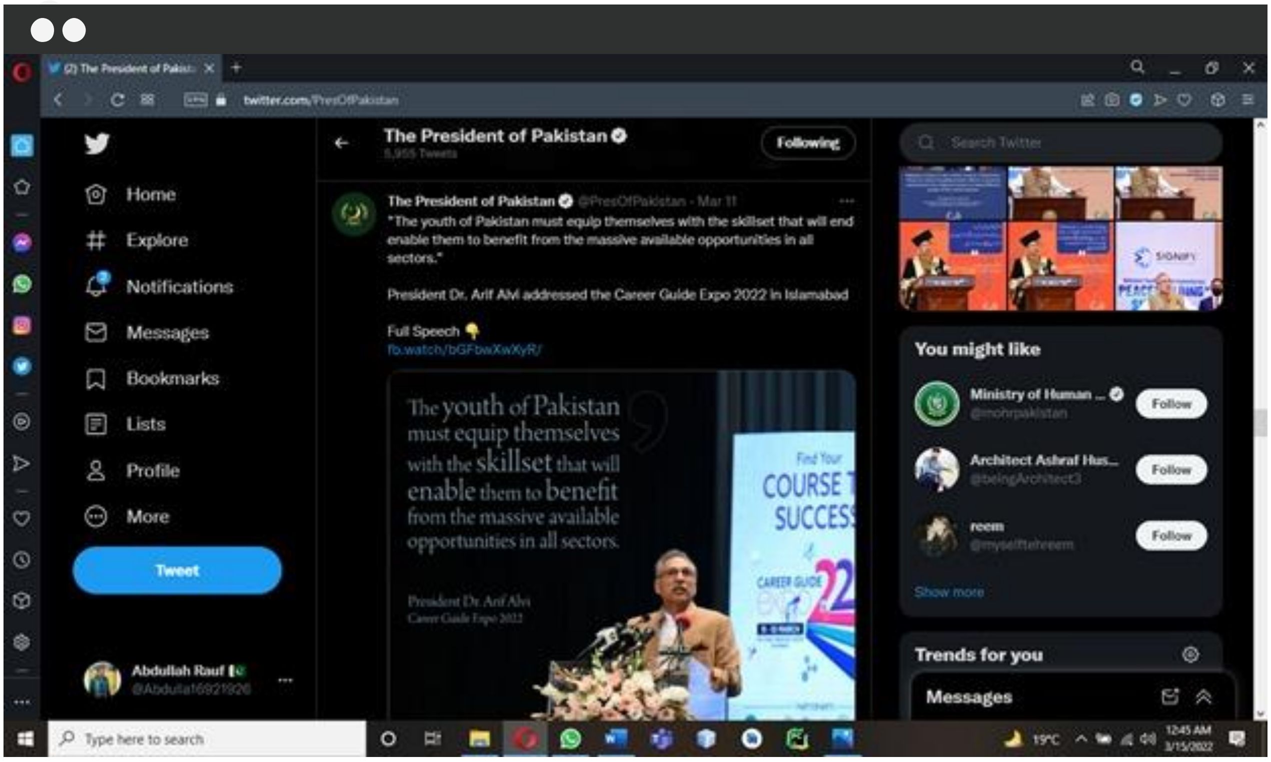
2

Google COLAB

DETAIL OF DATA

We gather Data from twitter (President of Pakistan). We Extract the data and clean it first and than we start our work and its is mentioned in workflow that how we did our wok.

Screen Shot # 00


The President of Pakistan
5,955 Tweets

The President of Pakistan @PresOfPakistan · Mar 11
"The youth of Pakistan must equip themselves with the skillset that will end enable them to benefit from the massive available opportunities in all sectors."

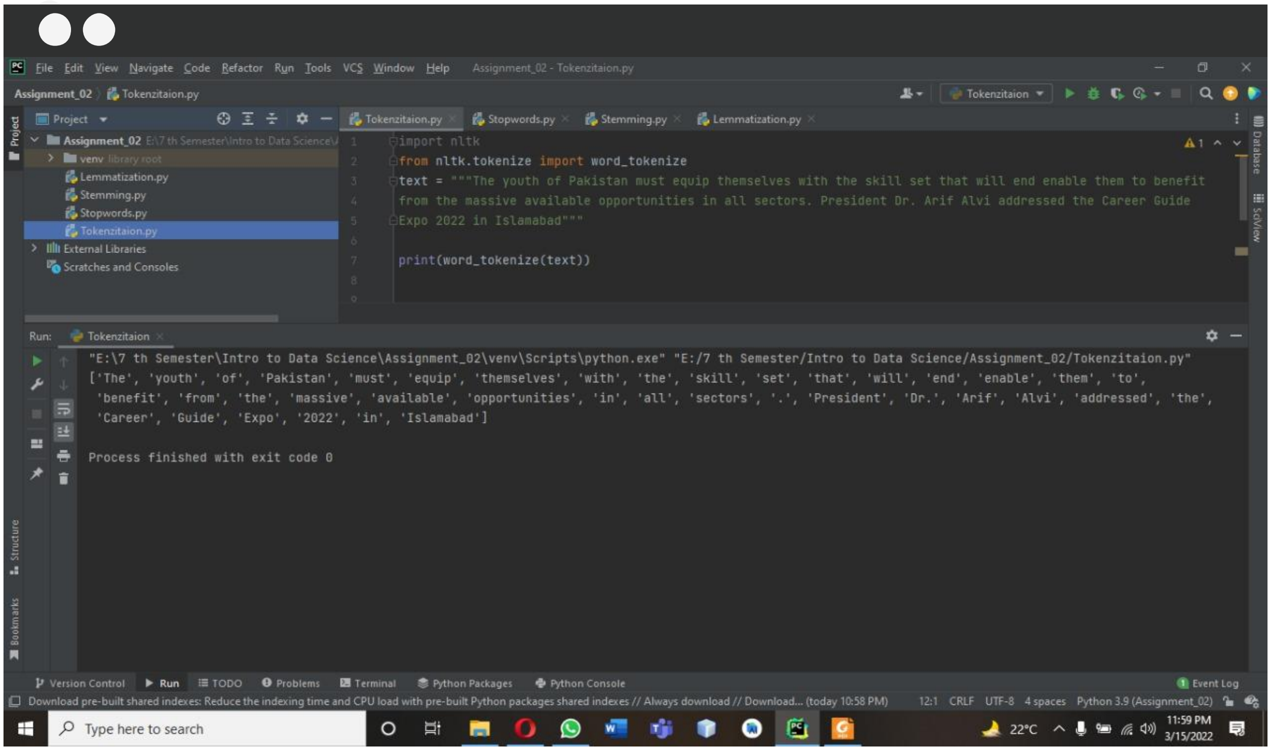
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad

Full Speech 👇
fb.watch/bGFbwXwXyR/

DATA SET

We gather Data from twitter (President of Pakistan). We Extract the data and clean it first and than we start our work and its is mentioned in workflow that how we did our wok.

Screen Shot # 01



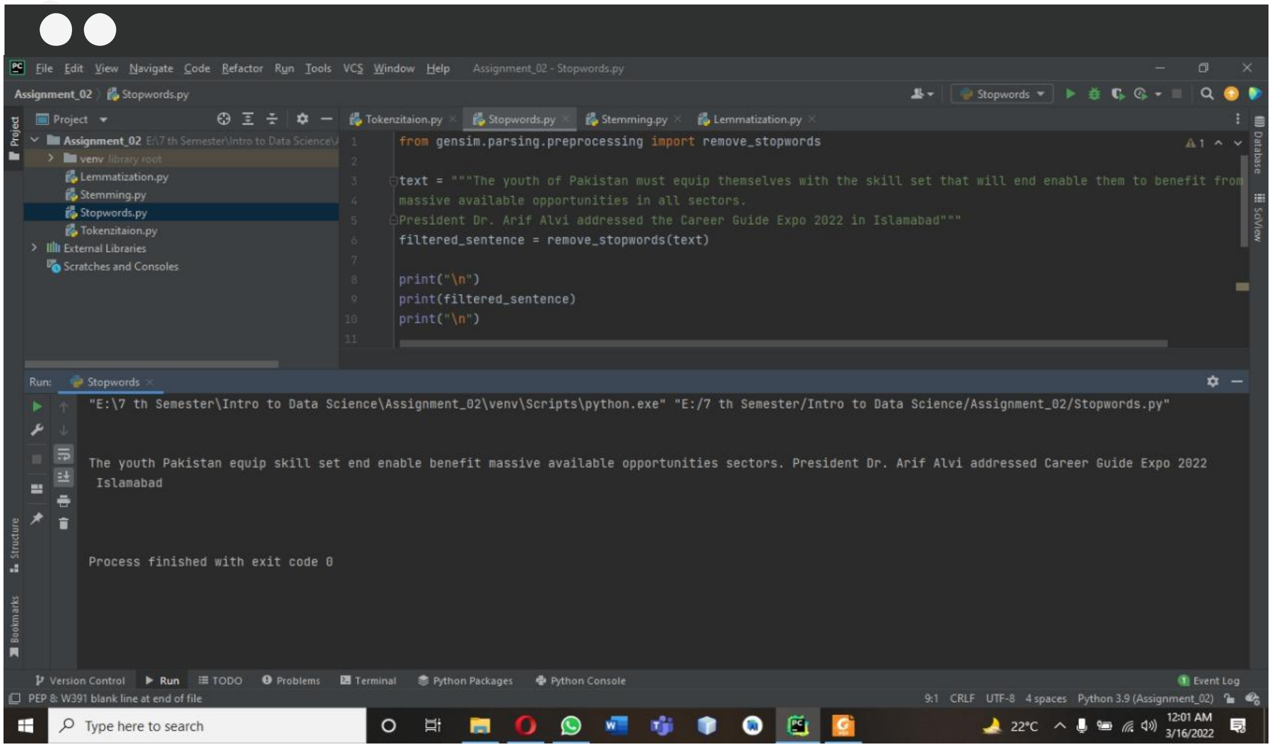
```python
import nltk
from nltk.tokenize import word_tokenize
text = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit
from the massive available opportunities in all sectors. President Dr. Arif Alvi addressed the Career Guide
Expo 2022 in Islamabad"""

print(word_tokenize(text))
```

```
"E:\7 th Semester\Intro to Data Science\Assignment_02\venv\Scripts\python.exe" "E:/7 th Semester/Intro to Data Science/Assignment_02/Tokenzitaion.py"
['The', 'youth', 'of', 'Pakistan', 'must', 'equip', 'themselves', 'with', 'the', 'skill', 'set', 'that', 'will', 'end', 'enable', 'them', 'to',
 'benefit', 'from', 'the', 'massive', 'available', 'opportunities', 'in', 'all', 'sectors', '.', 'President', 'Dr.', 'Arif', 'Alvi', 'addressed', 'the',
 'Career', 'Guide', 'Expo', '2022', 'in', 'Islamabad']

Process finished with exit code 0
```

TOKENIZATION

Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security.

Screen Shot # 02



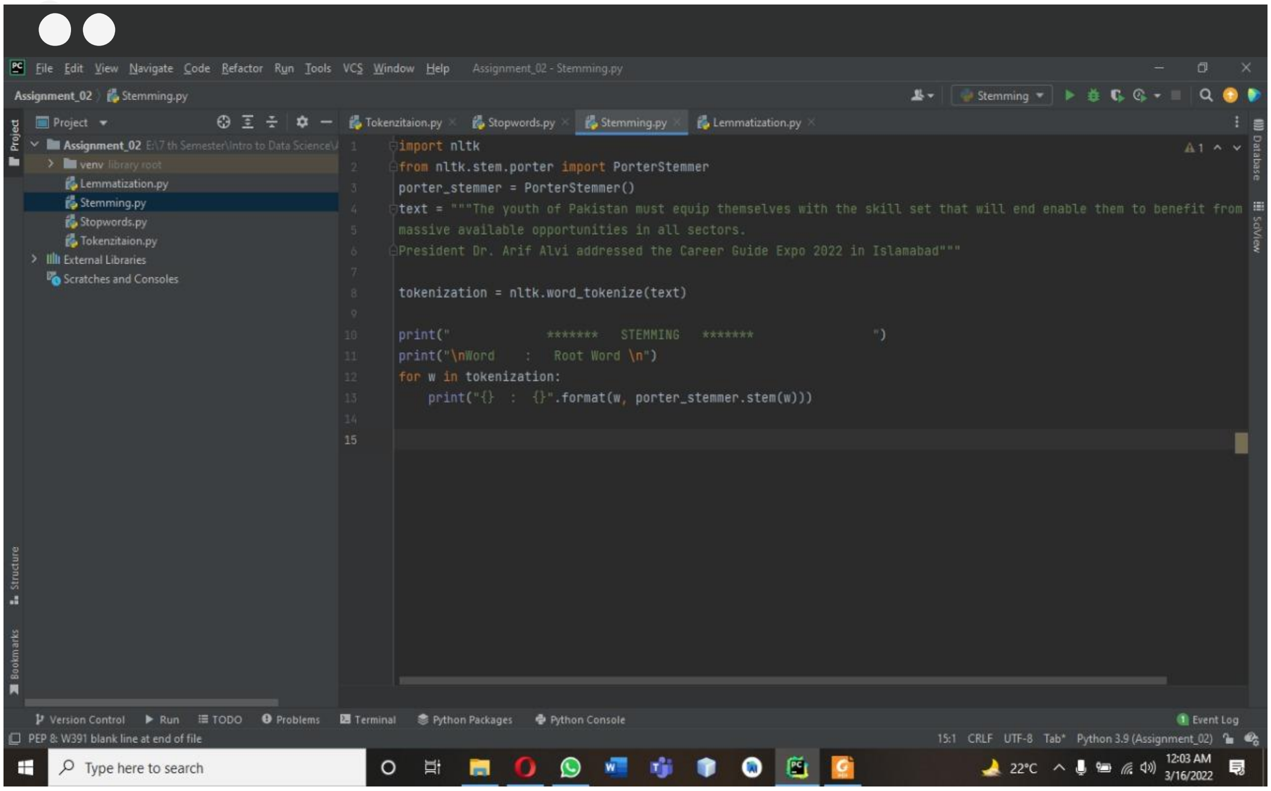
```python
from gensim.parsing.preprocessing import remove_stopwords

text = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit from
massive available opportunities in all sectors.
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad"""
filtered_sentence = remove_stopwords(text)

print("\n")
print(filtered_sentence)
print("\n")
```

```
"E:\7 th Semester\Intro to Data Science\Assignment_02\venv\Scripts\python.exe" "E:/7 th Semester/Intro to Data Science/Assignment_02/Stopwords.py"


The youth Pakistan equip skill set end enable benefit massive available opportunities sectors. President Dr. Arif Alvi addressed Career Guide Expo 2022
 Islamabad



Process finished with exit code 0
```

STOP WORD REMOVAL

Stop word removal is one of the most used preprocessing steps across different NLP applications. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

Screen Shot # 03

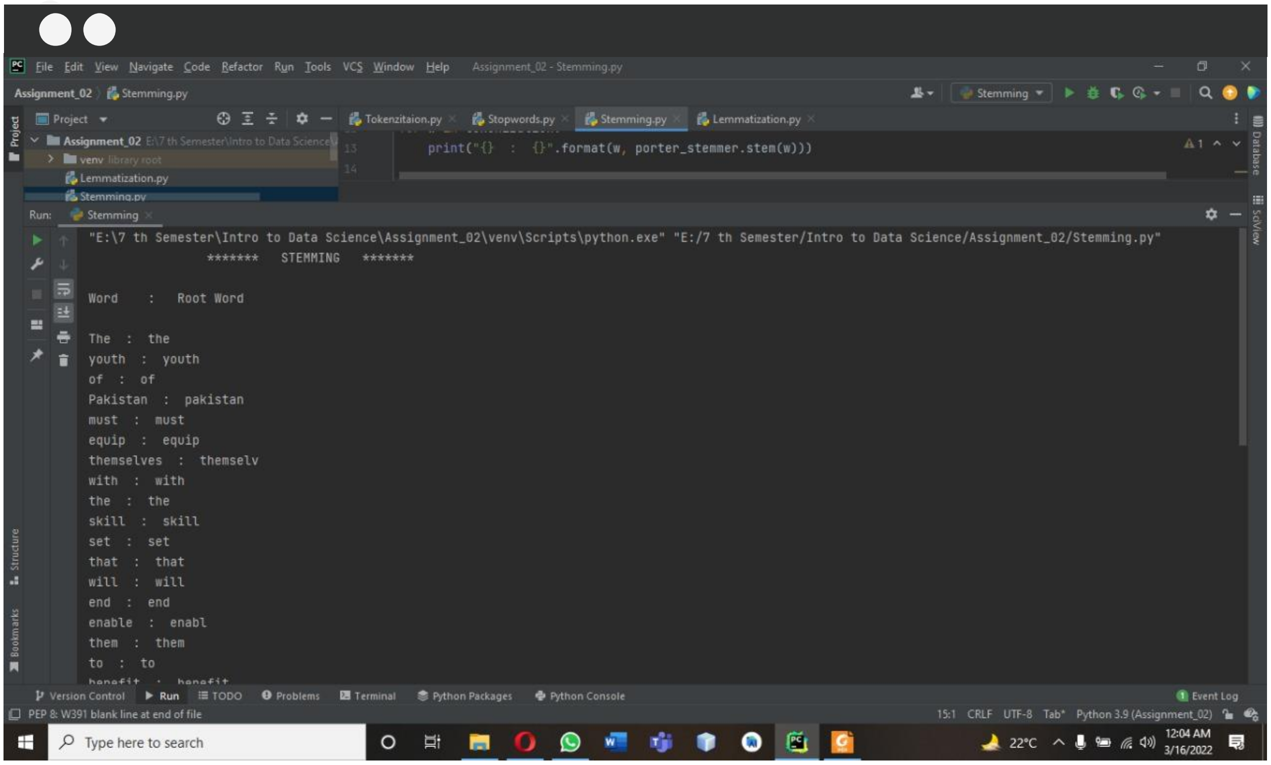


```python
import nltk
from nltk.stem.porter import PorterStemmer
porter_stemmer = PorterStemmer()
text = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit from
massive available opportunities in all sectors.
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad"""

tokenization = nltk.word_tokenize(text)

print("            *******   STEMMING   *******               ")
print("\nWord    :   Root Word \n")
for w in tokenization:
    print("{}  :  {}".format(w, porter_stemmer.stem(w)))

```

STEMMING

In processing unstructured text, stemming is the process of converting multiple forms of the same word into one stem, to simplify the task of analyzing the processed text. For example, in the previous sentence, "processing," "process," and "processed" would all be converted to the single stem "process."

Screen Shot # 04



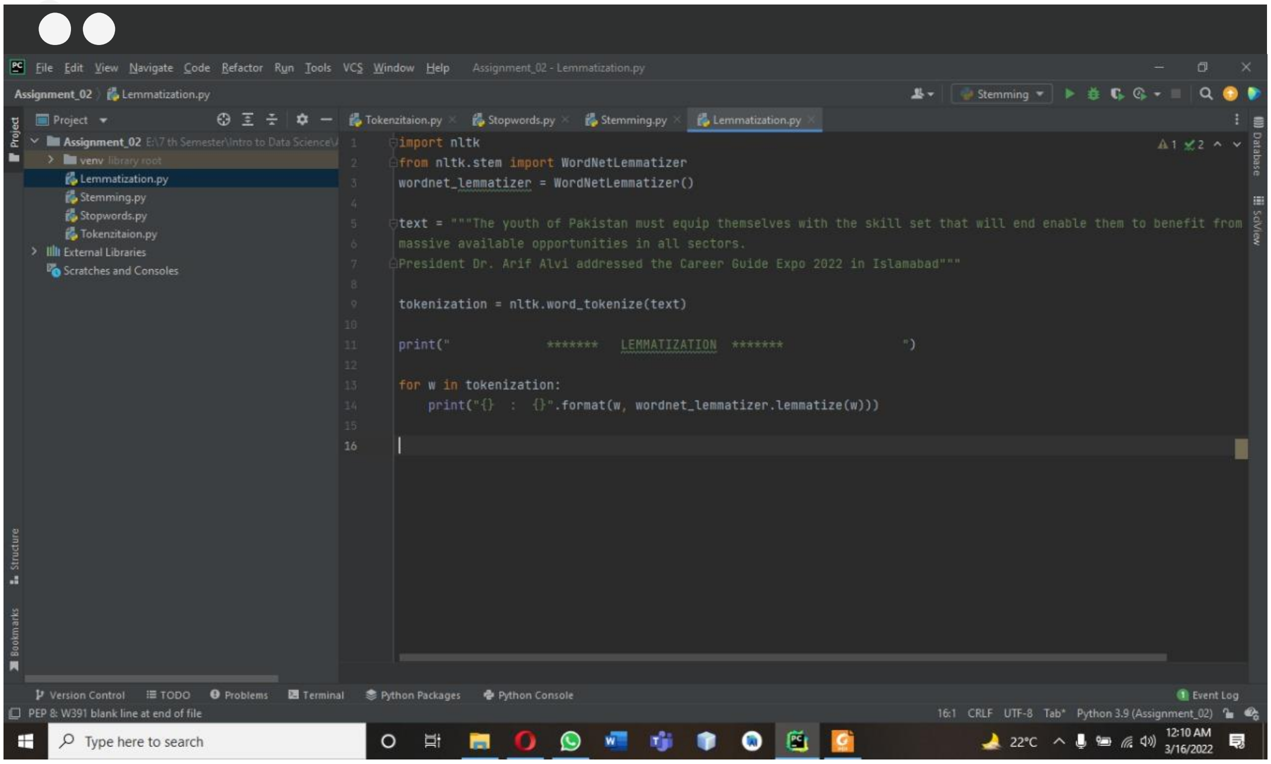
```
    print("{}  :  {}".format(w, porter_stemmer.stem(w)))
```

```
"E:\7 th Semester\Intro to Data Science\Assignment_02\venv\Scripts\python.exe" "E:/7 th Semester/Intro to Data Science/Assignment_02/Stemming.py"
                *******   STEMMING   *******

Word    :   Root Word

The  :  the
youth  :  youth
of  :  of
Pakistan  :  pakistan
must  :  must
equip  :  equip
themselves  :  themselv
with  :  with
the  :  the
skill  :  skill
set  :  set
that  :  that
will  :  will
end  :  end
enable  :  enabl
them  :  them
to  :  to
benefit  :  benefit
```

STEMMING RESULT

In processing unstructured text, stemming is the process of converting multiple forms of the same word into one stem, to simplify the task of analyzing the processed text. For example, in the previous sentence, "processing," "process," and "processed" would all be converted to the single stem "process."

Screen Shot # 05

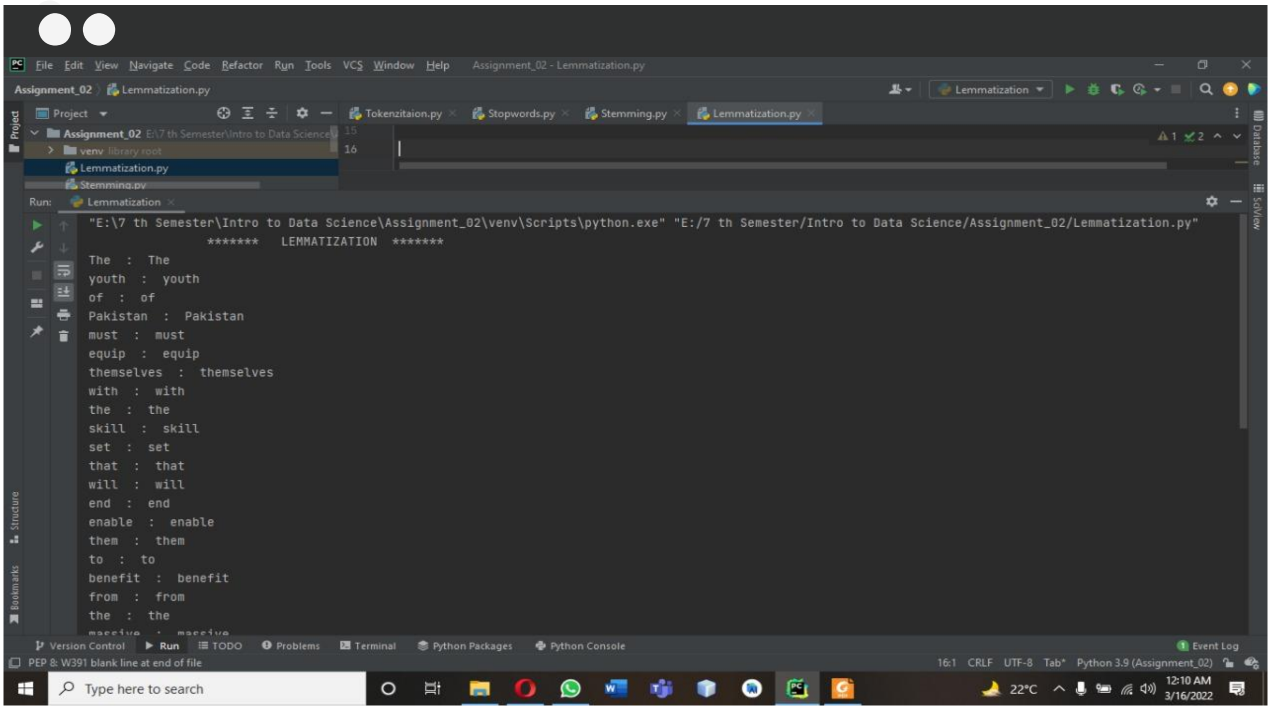


```python
import nltk
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()

text = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit from
massive available opportunities in all sectors.
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad"""

tokenization = nltk.word_tokenize(text)

print("              *******   LEMMATIZATION  *******                ")

for w in tokenization:
    print("{}  :  {}".format(w, wordnet_lemmatizer.lemmatize(w)))

```

LEMMETIZER

Lemmatization is the method to normalize the text documents. The main goal of the text normalization is to keep the vocabulary small and remove the noise(unwanted stuff) which helps to improve the accuracy of many language modeling tasks.

Screen Shot # 06

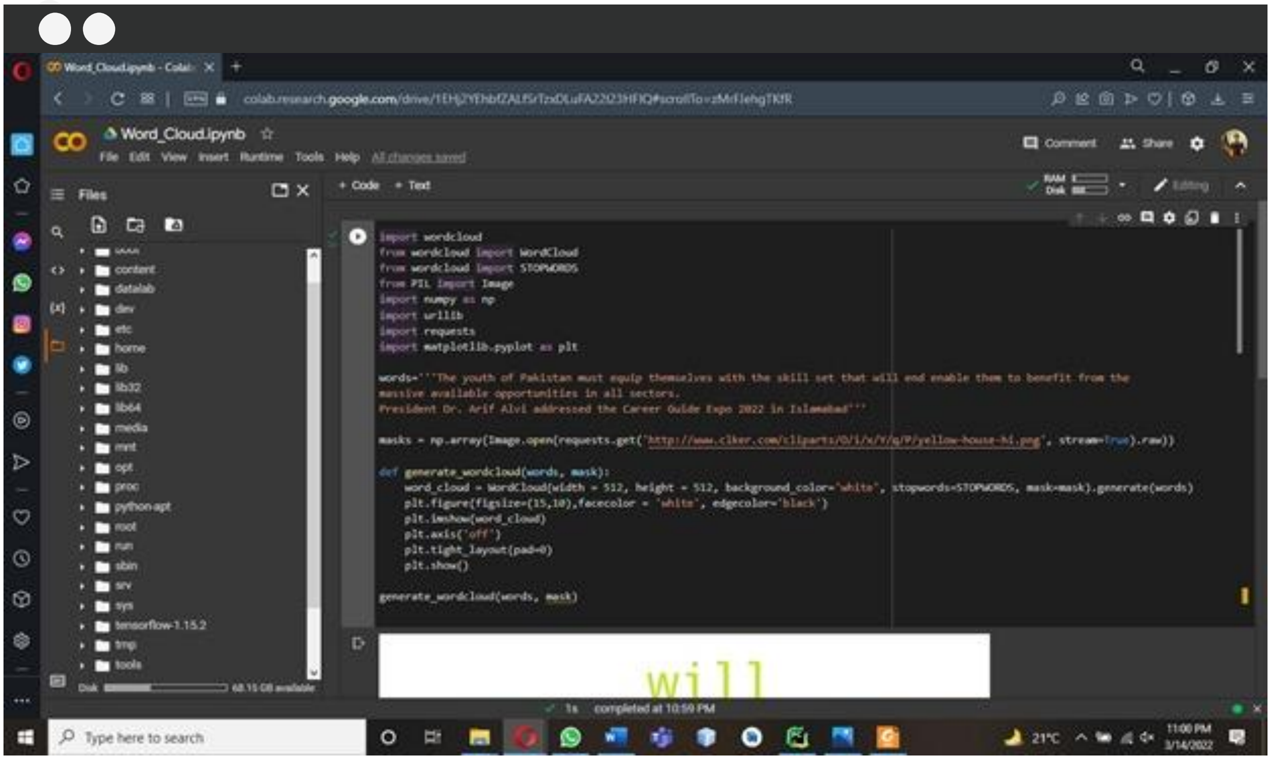


```
"E:\7 th Semester\Intro to Data Science\Assignment_02\venv\Scripts\python.exe" "E:/7 th Semester/Intro to Data Science/Assignment_02/Lemmatization.py"
                *******   LEMMATIZATION  *******
The  :  The
youth  :  youth
of  :  of
Pakistan  :  Pakistan
must  :  must
equip  :  equip
themselves  :  themselves
with  :  with
the  :  the
skill  :  skill
set  :  set
that  :  that
will  :  will
end  :  end
enable  :  enable
them  :  them
to  :  to
benefit  :  benefit
from  :  from
the  :  the
massive  :  massive
```

LEMMETIZER OUTPUT

Lemmatization is the method to normalize the text documents. The main goal of the text normalization is to keep the vocabulary small and remove the noise(unwanted stuff) which helps to improve the accuracy of many language modeling tasks.

Screen Shot # 07

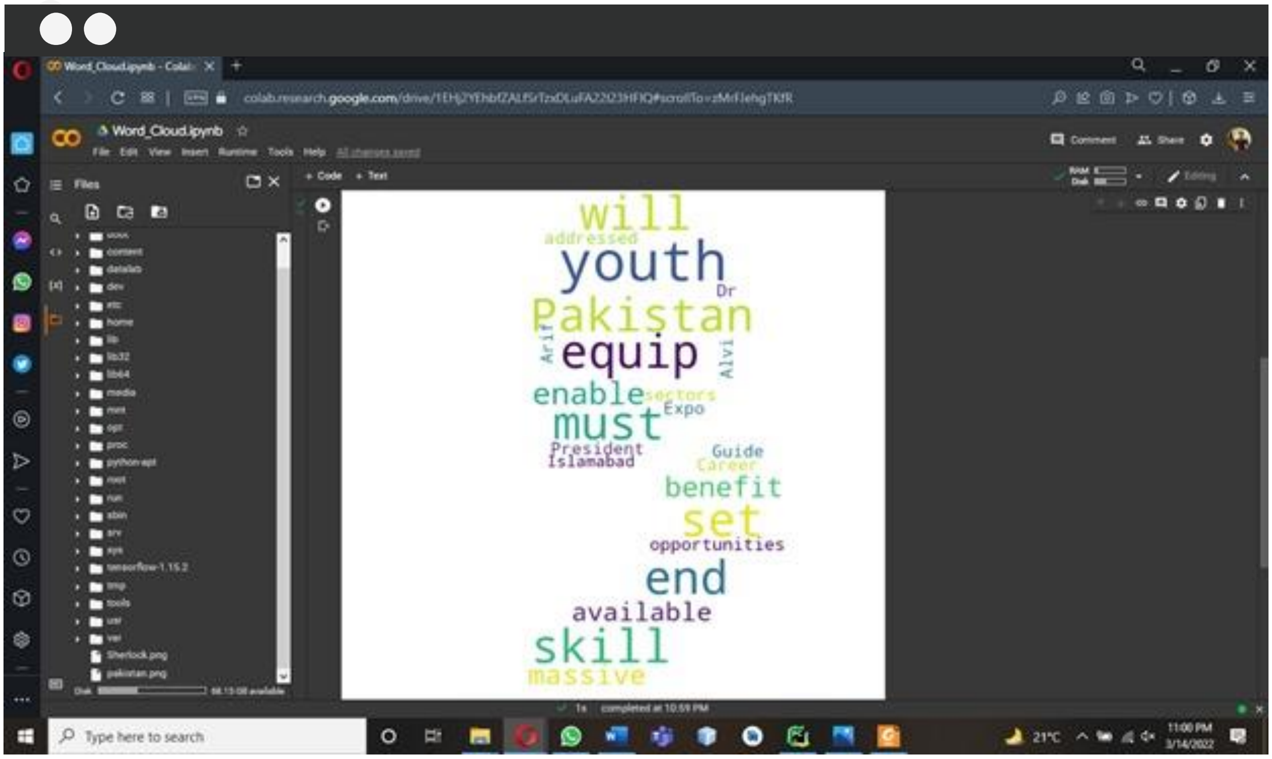


```
import wordcloud
from wordcloud import WordCloud
from wordcloud import STOPWORDS
from PIL import Image
import numpy as np
import urllib
import requests
import matplotlib.pyplot as plt

words='''The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit from the
massive available opportunities in all sectors.
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad'''

masks = np.array(Image.open(requests.get('http://www.clker.com/cliparts/O/i/x/Y/q/P/yellow-house-hi.png', stream=True).raw))

def generate_wordcloud(words, mask):
    word_cloud = WordCloud(width = 512, height = 512, background_color='white', stopwords=STOPWORDS, mask=mask).generate(words)
    plt.figure(figsize=(15,10),facecolor = 'white', edgecolor='black')
    plt.imshow(word_cloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()

generate_wordcloud(words, mask)
```

WORD CLOUD OF DATA SET

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text.

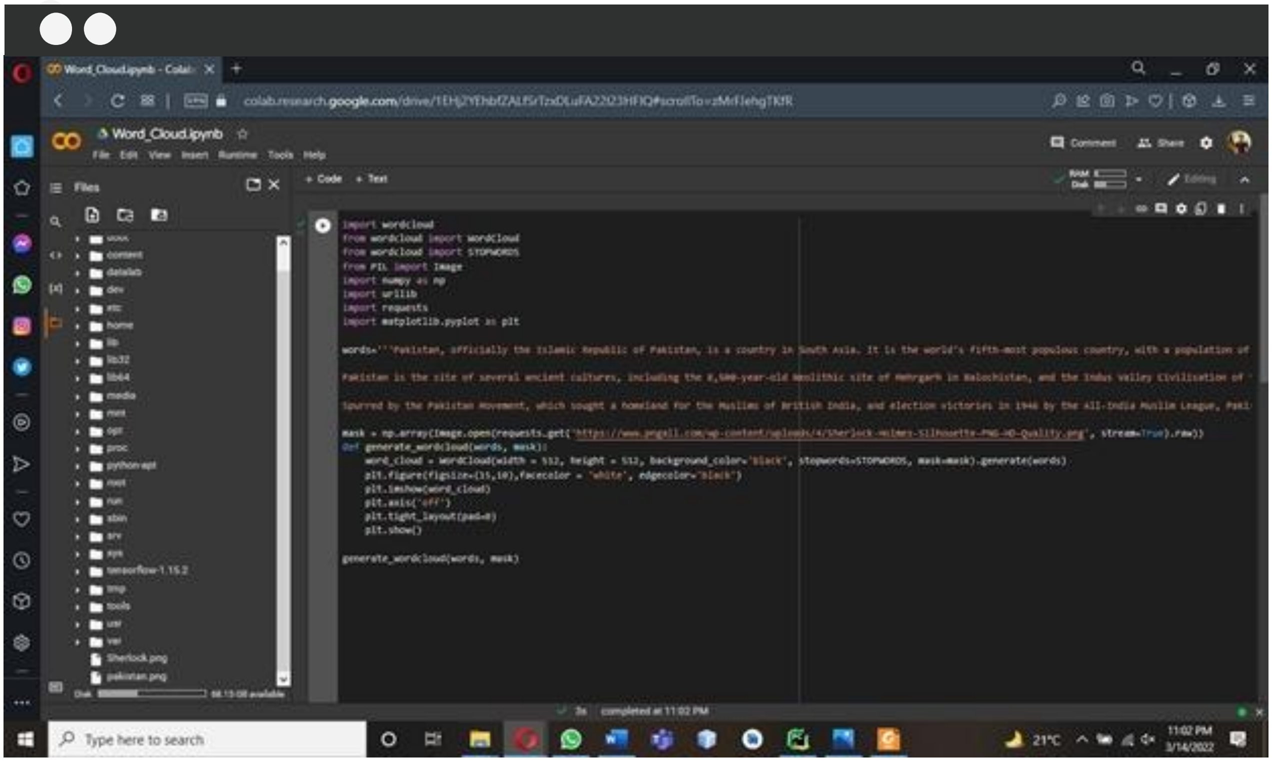
Screen Shot # 08



WORD CLOUD OF DATA SET OUTPUT

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text.

Screen Shot # 09

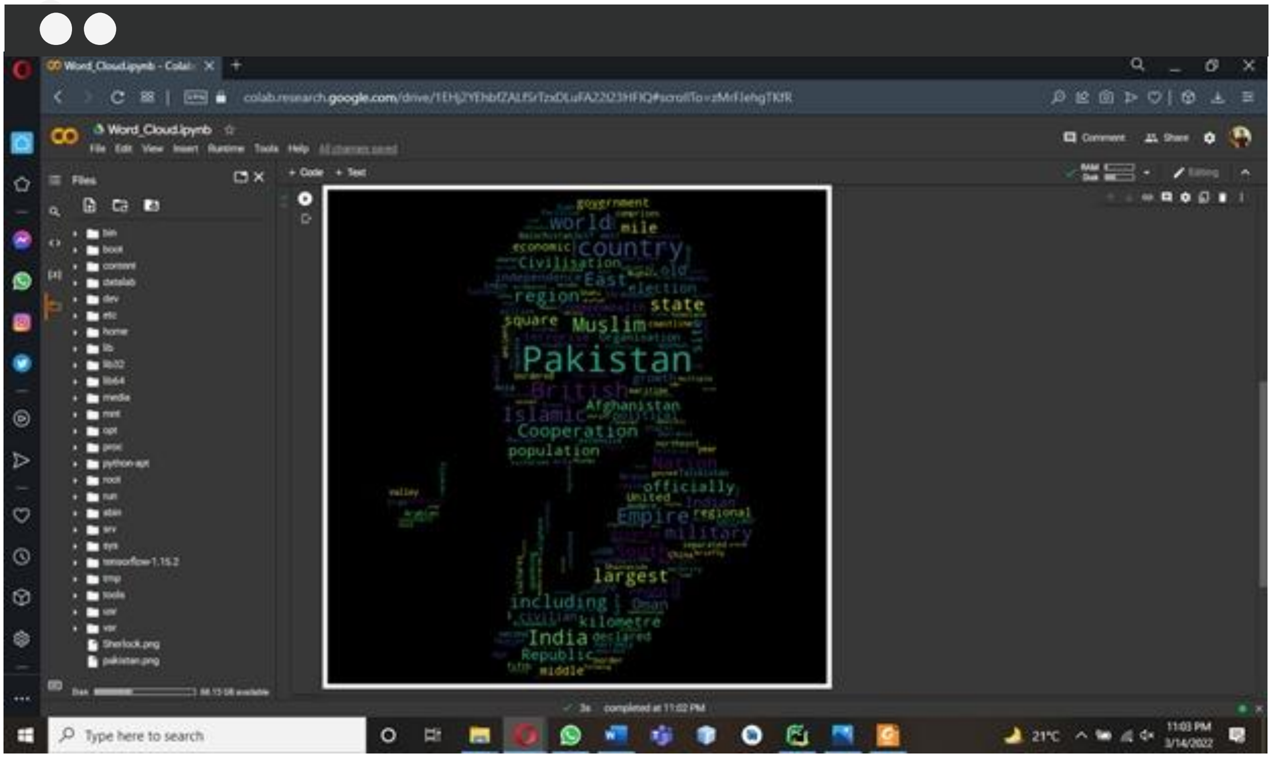


```
import wordcloud
from wordcloud import WordCloud
from wordcloud import STOPWORDS
from PIL import Image
import numpy as np
import urllib
import requests
import matplotlib.pyplot as plt

words='''Pakistan, officially the Islamic Republic of Pakistan, is a country in South Asia. It is the world's fifth-most populous country, with a population of

Pakistan is the site of several ancient cultures, including the 8,500-year-old Neolithic site of Mehrgarh in Balochistan, and the Indus Valley Civilisation of

Spurred by the Pakistan Movement, which sought a homeland for the Muslims of British India, and election victories in 1946 by the All-India Muslim League, Paki

mask = np.array(Image.open(requests.get('https://www.pngall.com/wp-content/uploads/4/Sherlock-Holmes-Silhouette-PNG-HD-Quality.png', stream=True).raw))
def generate_wordcloud(words, mask):
    word_cloud = WordCloud(width = 512, height = 512, background_color='black', stopwords=STOPWORDS, mask=mask).generate(words)
    plt.figure(figsize=(15,10),facecolor = 'white', edgecolor='black')
    plt.imshow(word_cloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()

generate_wordcloud(words, mask)
```

WORD CLOUD OF DATA SET

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. DATA FROM WIKIPEDIA

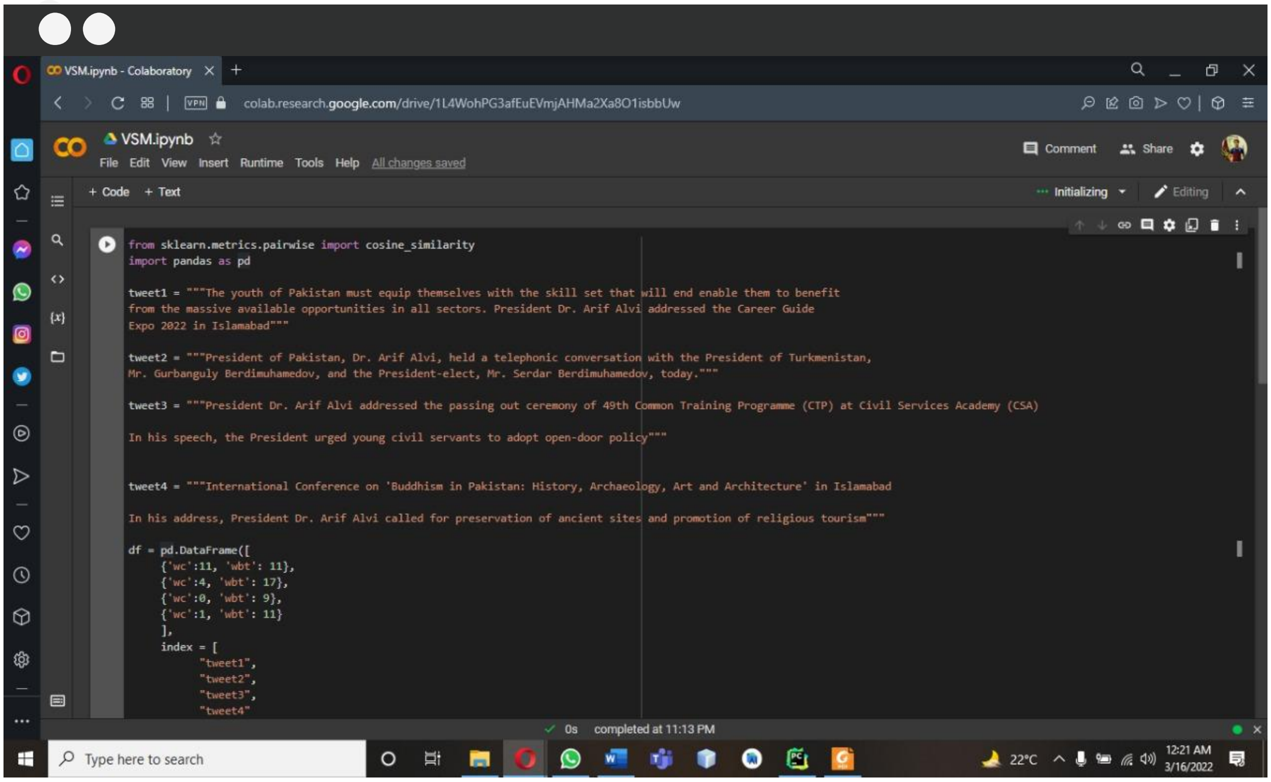
Screen Shot # 10



WORD CLOUD OF DATA SET OUTPUT

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. TWITTER DATA SET

Screen Shot # 11

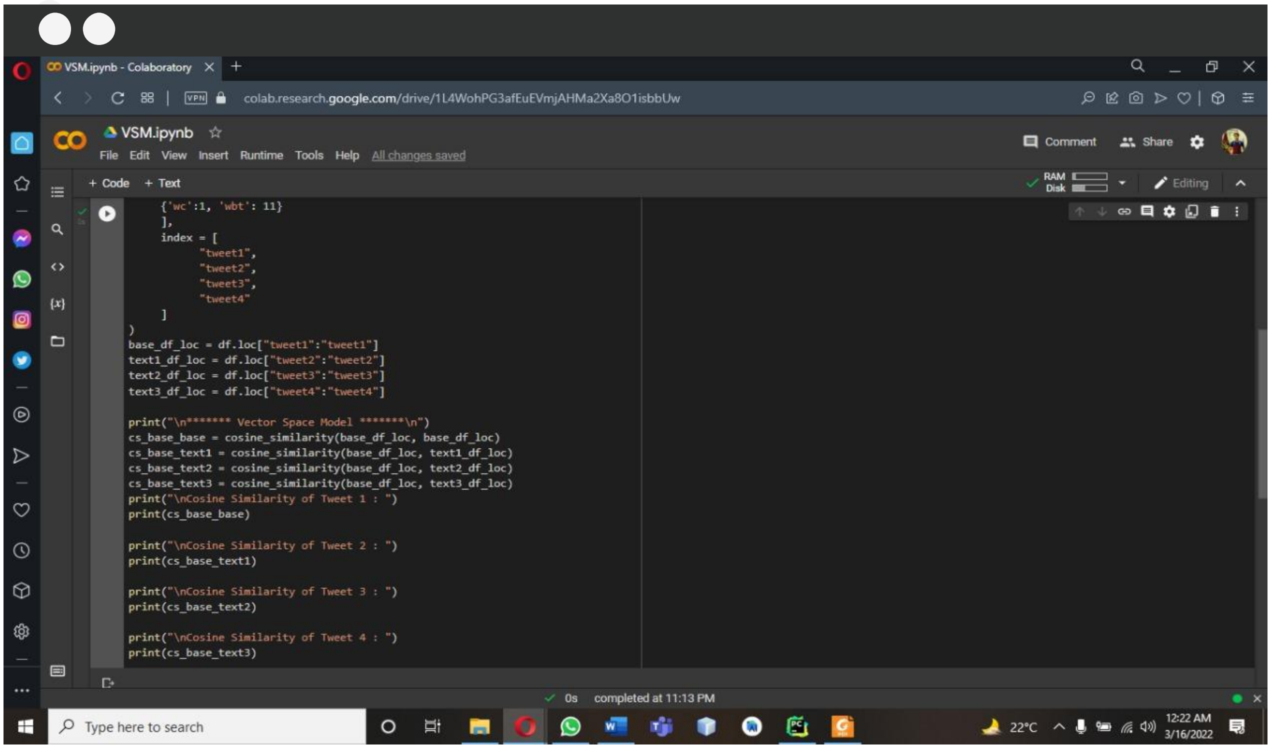


```python
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd

tweet1 = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit
from the massive available opportunities in all sectors. President Dr. Arif Alvi addressed the Career Guide
Expo 2022 in Islamabad"""

tweet2 = """President of Pakistan, Dr. Arif Alvi, held a telephonic conversation with the President of Turkmenistan,
Mr. Gurbanguly Berdimuhamedov, and the President-elect, Mr. Serdar Berdimuhamedov, today."""

tweet3 = """President Dr. Arif Alvi addressed the passing out ceremony of 49th Common Training Programme (CTP) at Civil Services Academy (CSA)

In his speech, the President urged young civil servants to adopt open-door policy"""


tweet4 = """International Conference on 'Buddhism in Pakistan: History, Archaeology, Art and Architecture' in Islamabad

In his address, President Dr. Arif Alvi called for preservation of ancient sites and promotion of religious tourism"""

df = pd.DataFrame([
     {'wc':11, 'wbt': 11},
     {'wc':4, 'wbt': 17},
     {'wc':0, 'wbt': 9},
     {'wc':1, 'wbt': 11}
     ],
     index = [
           "tweet1",
           "tweet2",
           "tweet3",
           "tweet4"
```

VECTOR SPACE MODELING

Vector space models are to consider the relationship between data that are represented by vectors. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Screen Shot # 12

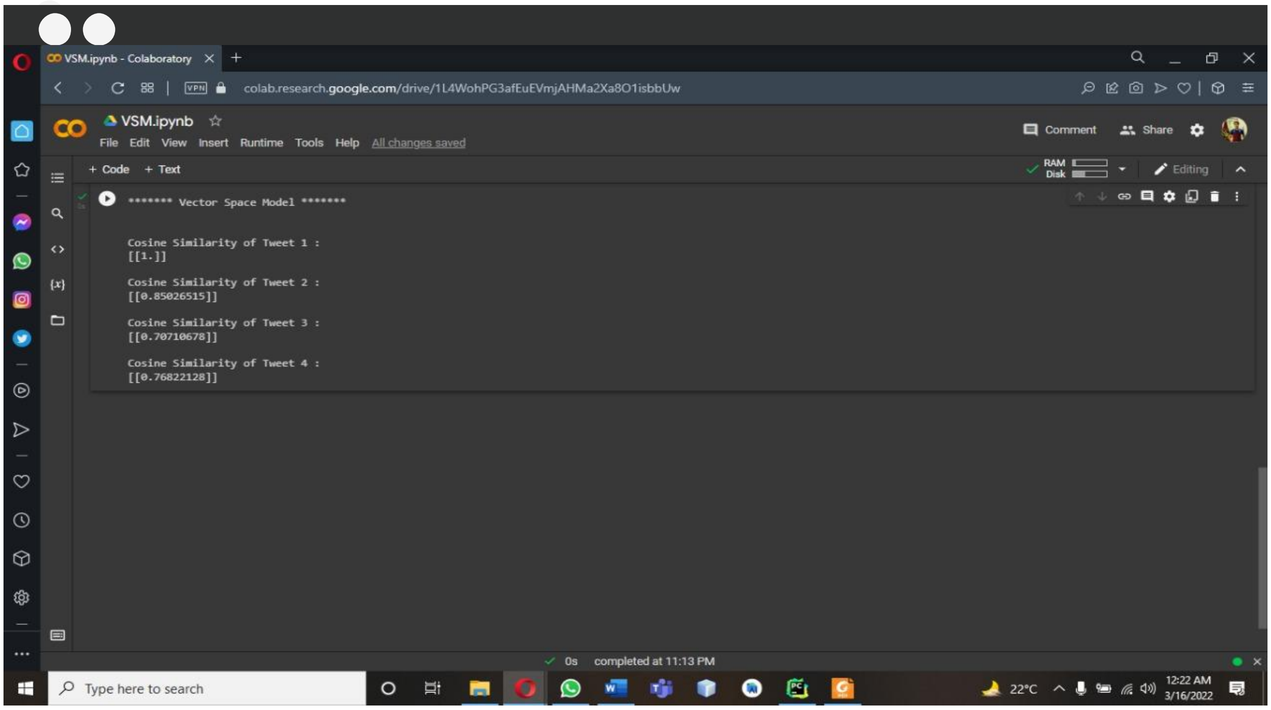


```
        {'wc':1, 'wbt': 11}
        ],
        index = [
              "tweet1",
              "tweet2",
              "tweet3",
              "tweet4"
        ]
)
base_df_loc = df.loc["tweet1":"tweet1"]
text1_df_loc = df.loc["tweet2":"tweet2"]
text2_df_loc = df.loc["tweet3":"tweet3"]
text3_df_loc = df.loc["tweet4":"tweet4"]

print("\n******* Vector Space Model *******\n")
cs_base_base = cosine_similarity(base_df_loc, base_df_loc)
cs_base_text1 = cosine_similarity(base_df_loc, text1_df_loc)
cs_base_text2 = cosine_similarity(base_df_loc, text2_df_loc)
cs_base_text3 = cosine_similarity(base_df_loc, text3_df_loc)
print("\nCosine Similarity of Tweet 1 : ")
print(cs_base_base)

print("\nCosine Similarity of Tweet 2 : ")
print(cs_base_text1)

print("\nCosine Similarity of Tweet 3 : ")
print(cs_base_text2)

print("\nCosine Similarity of Tweet 4 : ")
print(cs_base_text3)
```

VECTOR SPACE MODELING

Vector space models are to consider the relationship between data that are represented by vectors. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Screen Shot # 13

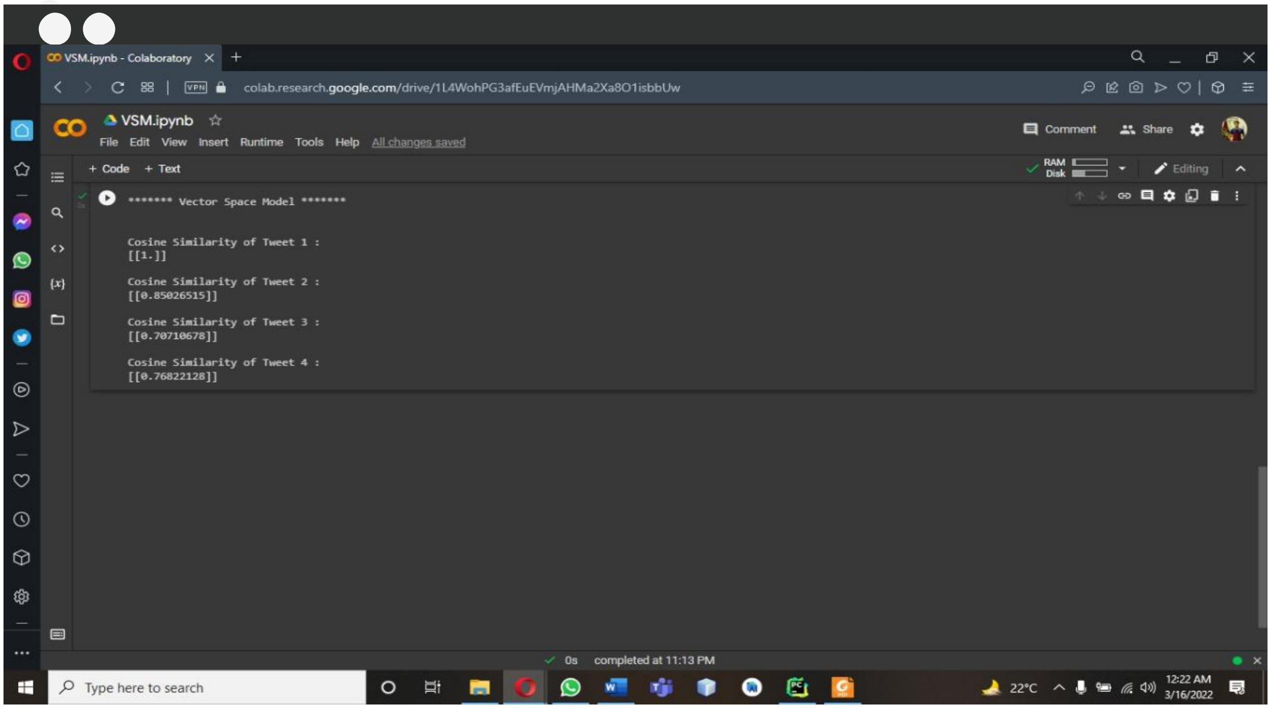


```
******* Vector Space Model *******


Cosine Similarity of Tweet 1 :
[[1.]]

Cosine Similarity of Tweet 2 :
[[0.85026515]]

Cosine Similarity of Tweet 3 :
[[0.70710678]]

Cosine Similarity of Tweet 4 :
[[0.76822128]]
```

VECTOR SPACE MODELING OUTPUT

Vector space models are to consider the relationship between data that are represented by vectors. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Screen Shot # 14

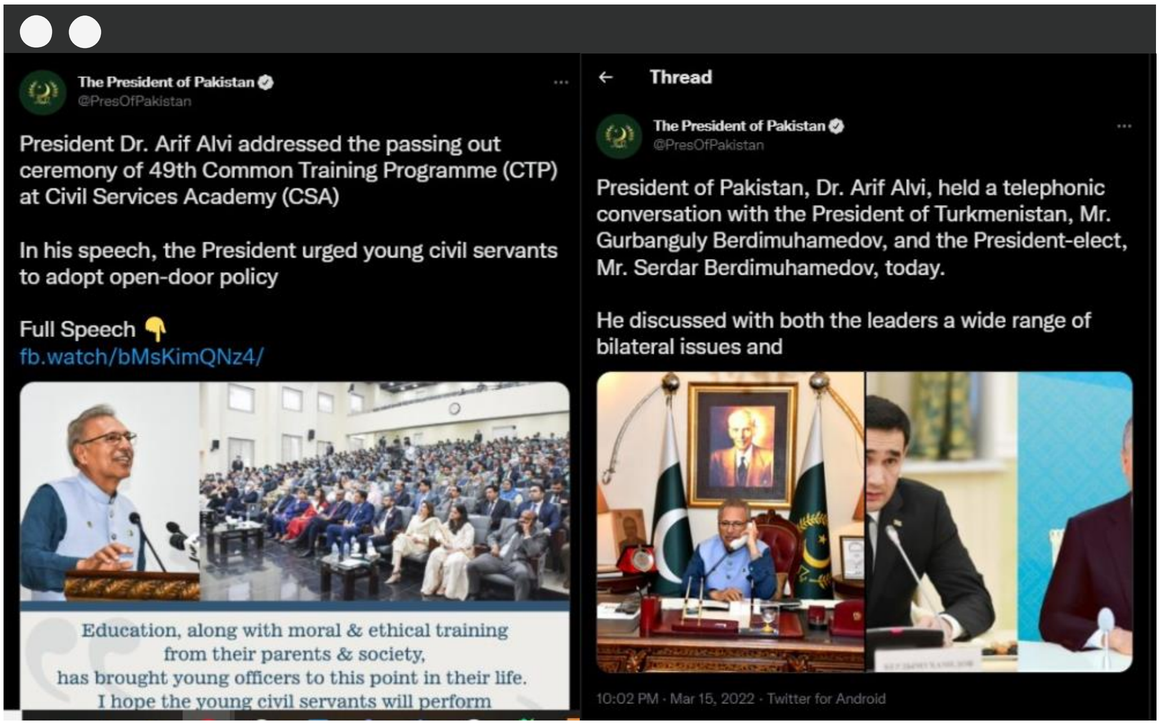


```
******* Vector Space Model *******


Cosine Similarity of Tweet 1 :
[[1.]]

Cosine Similarity of Tweet 2 :
[[0.85026515]]

Cosine Similarity of Tweet 3 :
[[0.70710678]]

Cosine Similarity of Tweet 4 :
[[0.76822128]]
```

VECTOR SPACE MODELING OUTPUT

Vector space models are to consider the relationship between data that are represented by vectors. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Screen Shot # 15


The President of Pakistan
@PresOfPakistan

President Dr. Arif Alvi addressed the passing out ceremony of 49th Common Training Programme (CTP) at Civil Services Academy (CSA)

In his speech, the President urged young civil servants to adopt open-door policy

Full Speech 👇
fb.watch/bMsKimQNz4/

Education, along with moral & ethical training from their parents & society, has brought young officers to this point in their life. I hope the young civil servants will perform

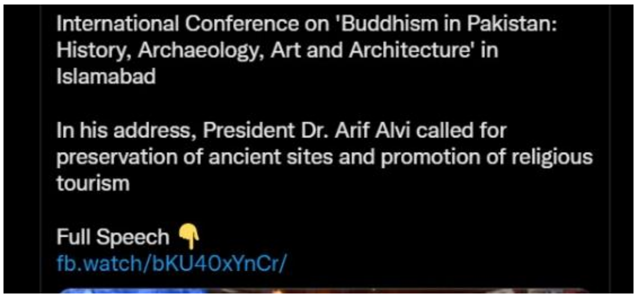
Thread

The President of Pakistan
@PresOfPakistan

President of Pakistan, Dr. Arif Alvi, held a telephonic conversation with the President of Turkmenistan, Mr. Gurbanguly Berdimuhamedov, and the President-elect, Mr. Serdar Berdimuhamedov, today.

He discussed with both the leaders a wide range of bilateral issues and

10:02 PM · Mar 15, 2022 · Twitter for Android


International Conference on 'Buddhism in Pakistan: History, Archaeology, Art and Architecture' in Islamabad

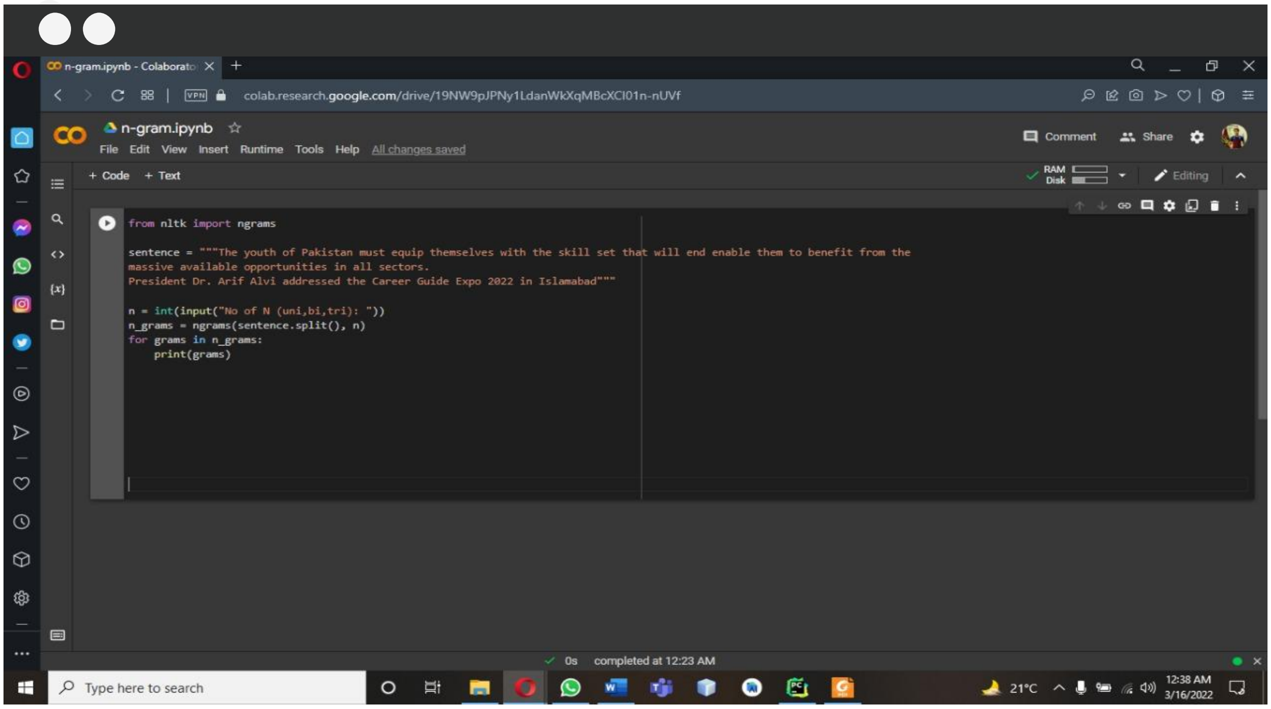
In his address, President Dr. Arif Alvi called for preservation of ancient sites and promotion of religious tourism

Full Speech 👇
fb.watch/bKU40xYnCr/

VSM DATA SET

Vector space models are to consider the relationship between data that are represented by vectors. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Screen Shot # 16

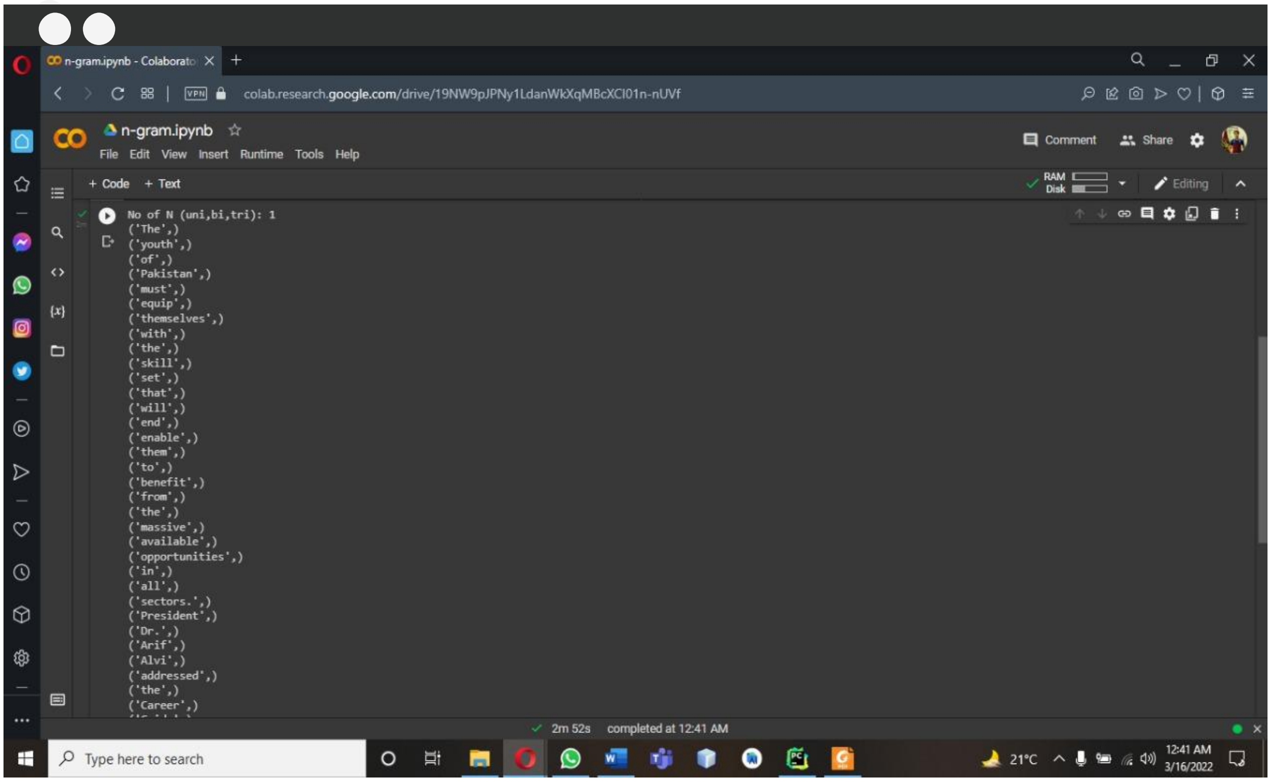


```
from nltk import ngrams

sentence = """The youth of Pakistan must equip themselves with the skill set that will end enable them to benefit from the
massive available opportunities in all sectors.
President Dr. Arif Alvi addressed the Career Guide Expo 2022 in Islamabad"""

n = int(input("No of N (uni,bi,tri): "))
n_grams = ngrams(sentence.split(), n)
for grams in n_grams:
    print(grams)
```

N-GRAM

N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighbouring sequences of items in a document. They come into play when we deal with text data in NLP(Natural Language Processing) tasks.

Screen Shot # 17



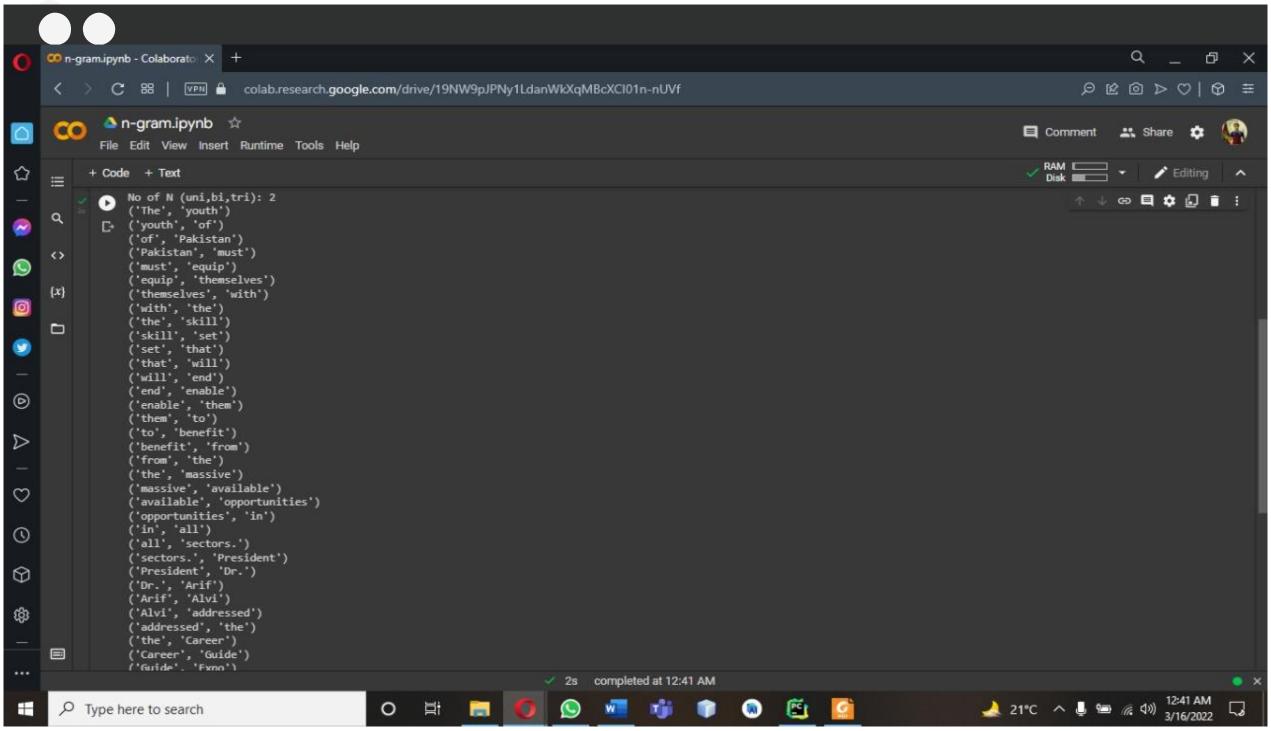
```
No of N (uni,bi,tri): 1
('The',)
('youth',)
('of',)
('Pakistan',)
('must',)
('equip',)
('themselves',)
('with',)
('the',)
('skill',)
('set',)
('that',)
('will',)
('end',)
('enable',)
('them',)
('to',)
('benefit',)
('from',)
('the',)
('massive',)
('available',)
('opportunities',)
('in',)
('all',)
('sectors.',)
('President',)
('Dr.',)
('Arif',)
('Alvi',)
('addressed',)
('the',)
('Career',)
```

n-gram.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
2m 52s completed at 12:41 AM

UNI-GRAM

unigram means taking only one word at a time, bigram means taking two words at a time and trigram means taking three words at a time.

Screen Shot # 18

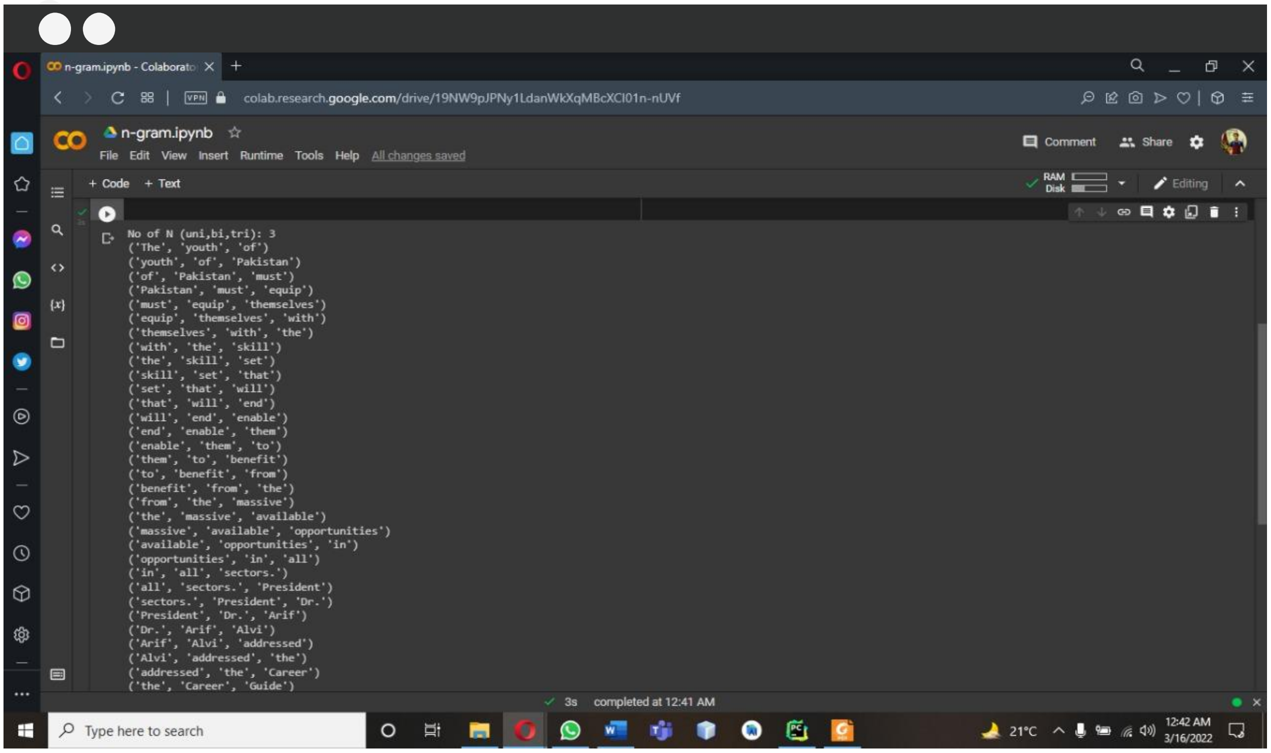

n-gram.ipynb - Colaborato
colab.research.google.com/drive/19NW9pJPNy1LdanWkXqMBcXCI01n-nUVf
n-gram.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
RAM
Disk
Editing
2s completed at 12:41 AM
Type here to search
21°C
12:41 AM
3/16/2022

```
No of N (uni,bi,tri): 2
('The', 'youth')
('youth', 'of')
('of', 'Pakistan')
('Pakistan', 'must')
('must', 'equip')
('equip', 'themselves')
('themselves', 'with')
('with', 'the')
('the', 'skill')
('skill', 'set')
('set', 'that')
('that', 'will')
('will', 'end')
('end', 'enable')
('enable', 'them')
('them', 'to')
('to', 'benefit')
('benefit', 'from')
('from', 'the')
('the', 'massive')
('massive', 'available')
('available', 'opportunities')
('opportunities', 'in')
('in', 'all')
('all', 'sectors.')
('sectors.', 'President')
('President', 'Dr.')
('Dr.', 'Arif')
('Arif', 'Alvi')
('Alvi', 'addressed')
('addressed', 'the')
('the', 'Career')
('Career', 'Guide')
('Guide', 'Expo')
```

BI-GRAM

N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document. They come into play when we deal with text data in NLP(Natural Language Processing) tasks.

Screen Shot # 20


n-gram.ipynb - Colaborato
colab.research.google.com/drive/19NW9pJPNy1LdanWkXqMBcXCl01n-nUVf
n-gram.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
3s completed at 12:41 AM
Type here to search
21°C
12:42 AM 3/16/2022

```
No of N (uni,bi,tri): 3
('The', 'youth', 'of')
('youth', 'of', 'Pakistan')
('of', 'Pakistan', 'must')
('Pakistan', 'must', 'equip')
('must', 'equip', 'themselves')
('equip', 'themselves', 'with')
('themselves', 'with', 'the')
('with', 'the', 'skill')
('the', 'skill', 'set')
('skill', 'set', 'that')
('set', 'that', 'will')
('that', 'will', 'end')
('will', 'end', 'enable')
('end', 'enable', 'them')
('enable', 'them', 'to')
('them', 'to', 'benefit')
('to', 'benefit', 'from')
('benefit', 'from', 'the')
('from', 'the', 'massive')
('the', 'massive', 'available')
('massive', 'available', 'opportunities')
('available', 'opportunities', 'in')
('opportunities', 'in', 'all')
('in', 'all', 'sectors.')
('all', 'sectors.', 'President')
('sectors.', 'President', 'Dr.')
('President', 'Dr.', 'Arif')
('Dr.', 'Arif', 'Alvi')
('Arif', 'Alvi', 'addressed')
('Alvi', 'addressed', 'the')
('addressed', 'the', 'Career')
('the', 'Career', 'Guide')
```

TRI-GRAM

N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document. They come into play when we deal with text data in NLP(Natural Language Processing) tasks.

THE END

REFERENCES

https://www.pngall.com/wp-content/uploads/4/Sherlock-Holmes-Silhouette-PNG-HD-Quality.png

https://www.askpython.com/python/examples/n-grams-python-nltk

https://colab.research.google.com/?utm_source=scs-index

https://twitter.com/PresOfPakistan

https://www.nltk.org