20BCE204

IRS Practical 3

Text Preprocessing using NLTK. Visualization:

- Word Cloud
- Histogram of top N frequent terms

In [11]:
```
import numpy as np
import pandas as pd
import sklearn
from sklearn.feature_extraction.text import CountVectorizer
import nltk

from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer

import matplotlib.pyplot as plt
# %matplotlib inline
from wordcloud import WordCloud
```

In [1]:
```
pol1 = open("pol1.txt")

pol1_str=pol1.read()

pol1_str
```

Out[1]:
```
'India’s ruling Bharatiya Janata Party (BJP) scored a landslide election wi
n in Prime Minister Narendra Modi’s home state of Gujarat on Thursday, in a
boost for the leader and his Hindu-nationalist party ahead of the 2024 gene
ral election.The BJP took 156 of 182 total seats in the Gujarat state assem
bly following voting earlier this month, marking the party’s best-ever perf
ormance in the state, a longtime BJP stronghold.The main opposition Indian
National Congress won 17 seats, while the Aam Aadmi Party (AAP) took five.
'
```

In [2]:
```
pol=[]

for i in range(10):
    poly=open(f"pol{i+1}.txt",encoding="utf8")
    pol.append(poly.read())


print(len(pol))
print(pol[1])
```

```
10
When world leaders at the Group of 20 summit in Bali, Indonesia, issued a j
oint statement condemning Russia’s war in Ukraine, a familiar sentence stoo
d out from the 1,186-page document.“Today’s era must not be of war,” it sai
d, echoing what Indian Prime Minister Narendra Modi told Russian leader Vla
dimir Putin during a face-to-face meeting in September.Media and officials
in the country of 1.3 billion were quick to claim the inclusion as a sign t
hat the world’s largest democracy had played a vital role in bridging diffe
rences between an increasingly isolated Russia, and the United States and i
ts allies.“How India united G20 on PM Modi’s idea of peace,” ran a headline
in the Times of India, the country’s largest English-language paper. “The P
rime Minister’s message that this is not the era of war… resonated very dee
ply across all the delegations and helped bridge the gap across different p
arties,” India’s Foreign Secretary Vinay Kwatra told reporters Wednesday.
```

In [4]:

```
import nltk
import numpy as np

nltkpol=[]

for i in range(10):
    nltk_pol=nltk.word_tokenize(pol[i])
    nltkpol.append(nltk_pol)

print(nltkpol[2])
```

```
['The', 'declaration', 'came', 'as', 'Indonesian', 'President', 'Joko', 'Wi
dodo', 'handed', 'over', 'the', 'G20', 'presidency', 'to', 'Modi', ',', 'wh
o', 'will', 'host', 'the', 'next', 'leaders', '’', 'summit', 'in', 'the', '
Indian', 'capital', 'New', 'Delhi', 'in', 'September', '2023', '—', 'about'
, 'six', 'months', 'before', 'he', 'is', 'expected', 'to', 'head', 'to', 't
he', 'polls', 'in', 'a', 'general', 'election', 'and', 'contest', 'the', 'c
ountry', '’', 's', 'top', 'seat', 'for', 'a', 'third', 'time', '.', 'As', '
New', 'Delhi', 'deftly', 'balances', 'its', 'ties', 'to', 'Russia', 'and',
'the', 'West', ',', 'Modi', ',', 'analysts', 'say', ',', 'is', 'emerging',
'as', 'a', 'leader', 'who', 'has', 'been', 'courted', 'by', 'all', 'sides',
',', 'winning', 'him', 'support', 'at', 'home', ',', 'while', 'cementing',
'India', 'as', 'an', 'international', 'power', 'broker.', '“', 'The', 'dome
stic', 'narrative', 'is', 'that', 'the', 'G20', 'summit', 'is', 'being', 'u
sed', 'as', 'a', 'big', 'banner', 'in', 'Modi', '’', 's', 'election', 'camp
aign', 'to', 'show', 'he', '’', 's', 'a', 'great', 'global', 'statesmen', '
,', '”', 'said', 'Sushant', 'Singh', ',', 'a', 'senior', 'fellow', 'at', 'N
ew', 'Delhi-based', 'think', 'tank', 'Center', 'for', 'Policy', 'Research',
'.', '“', 'And', 'the', 'current', 'Indian', 'leadership', 'now', 'sees', '
themselves', 'as', 'a', 'powerful', 'country', 'seated', 'at', 'the', 'high
', 'table', '.', '”']
```

In [6]:

```
punpol=[]

for i in range(10):
    pun_pol=[c for c in nltkpol[i] if c.isalpha() ]
    punpol.append(pun_pol)
# fil_text= [w for w in nltkpol[0] if w not in spw_words]


print(punpol[1])
```

```
['When', 'world', 'leaders', 'at', 'the', 'Group', 'of', 'summit', 'in', 'B
ali', 'Indonesia', 'issued', 'a', 'joint', 'statement', 'condemning', 'Russ
ia', 's', 'war', 'in', 'Ukraine', 'a', 'familiar', 'sentence', 'stood', 'ou
t', 'from', 'the', 'Today', 's', 'era', 'must', 'not', 'be', 'of', 'war', '
it', 'said', 'echoing', 'what', 'Indian', 'Prime', 'Minister', 'Narendra',
'Modi', 'told', 'Russian', 'leader', 'Vladimir', 'Putin', 'during', 'a', 'm
eeting', 'in', 'and', 'officials', 'in', 'the', 'country', 'of', 'billion',
'were', 'quick', 'to', 'claim', 'the', 'inclusion', 'as', 'a', 'sign', 'tha
t', 'the', 'world', 's', 'largest', 'democracy', 'had', 'played', 'a', 'vit
al', 'role', 'in', 'bridging', 'differences', 'between', 'an', 'increasingl
y', 'isolated', 'Russia', 'and', 'the', 'United', 'States', 'and', 'its', '
How', 'India', 'united', 'on', 'PM', 'Modi', 's', 'idea', 'of', 'peace', 'r
an', 'a', 'headline', 'in', 'the', 'Times', 'of', 'India', 'the', 'country'
, 's', 'largest', 'paper', 'The', 'Prime', 'Minister', 's', 'message', 'tha
t', 'this', 'is', 'not', 'the', 'era', 'of', 'resonated', 'very', 'deeply',
'across', 'all', 'the', 'delegations', 'and', 'helped', 'bridge', 'the', 'g
ap', 'across', 'different', 'parties', 'India', 's', 'Foreign', 'Secretary'
, 'Vinay', 'Kwatra', 'told', 'reporters', 'Wednesday']
```

In [7]:

```
from nltk.corpus import stopwords as spw
spw_words=spw.words('english')

filtpol=[]

for i in range(10):
    filt_pol=[w for w in punpol[i] if w not in spw_words]
    filtpol.append(filt_pol)
# fil_text= [w for w in nltkpol[0] if w not in spw_words]


print(filtpol[1])
```

```
['When', 'world', 'leaders', 'Group', 'summit', 'Bali', 'Indonesia', 'issue
d', 'joint', 'statement', 'condemning', 'Russia', 'war', 'Ukraine', 'famili
ar', 'sentence', 'stood', 'Today', 'era', 'must', 'war', 'said', 'echoing',
'Indian', 'Prime', 'Minister', 'Narendra', 'Modi', 'told', 'Russian', 'lead
er', 'Vladimir', 'Putin', 'meeting', 'officials', 'country', 'billion', 'qu
ick', 'claim', 'inclusion', 'sign', 'world', 'largest', 'democracy', 'playe
d', 'vital', 'role', 'bridging', 'differences', 'increasingly', 'isolated',
'Russia', 'United', 'States', 'How', 'India', 'united', 'PM', 'Modi', 'idea
', 'peace', 'ran', 'headline', 'Times', 'India', 'country', 'largest', 'pap
er', 'The', 'Prime', 'Minister', 'message', 'era', 'resonated', 'deeply', '
across', 'delegations', 'helped', 'bridge', 'gap', 'across', 'different', '
parties', 'India', 'Foreign', 'Secretary', 'Vinay', 'Kwatra', 'told', 'repo
rters', 'Wednesday']
```

In [8]:

```
strpol=[]

for i in range(10):

    str=""
    for j in range(len(filtpol[i])):
        str+=filtpol[i][j]
        str+=" "
    strpol.append(str)

print(strpol[1])
```

```
When world leaders Group summit Bali Indonesia issued joint statement conde
mning Russia war Ukraine familiar sentence stood Today era must war said ec
hoing Indian Prime Minister Narendra Modi told Russian leader Vladimir Puti
n meeting officials country billion quick claim inclusion sign world larges
t democracy played vital role bridging differences increasingly isolated Ru
ssia United States How India united PM Modi idea peace ran headline Times I
ndia country largest paper The Prime Minister message era resonated deeply
across delegations helped bridge gap across different parties India Foreign
Secretary Vinay Kwatra told reporters Wednesday
```

In [9]:

```
count = CountVectorizer()

matrix = count.fit_transform(strpol)
df = pd.DataFrame(matrix.toarray(), columns=count.get_feature_names())
df
```

Out[9]:

	aadmi	aam	aap	abdel	abvp	according	accused	achieve	across	address	...	win
0	1	1	1	0	0	0	0	0	0	0	...	1
1	0	0	0	0	0	0	0	0	2	0	...	0
2	0	0	0	0	0	0	0	0	0	0	...	0
3	0	0	0	0	0	0	0	0	0	0	...	0
4	0	0	0	0	2	0	1	0	0	0	...	0
5	0	0	0	0	0	0	0	0	0	0	...	0
6	0	0	0	1	0	0	0	0	0	0	...	0
7	0	0	0	0	0	0	0	1	0	0	...	0
8	0	0	0	0	0	1	0	1	0	0	...	0
9	0	0	0	0	0	0	0	0	0	1	...	0

10 rows × 551 columns

In [10]:

```
dfwc=df.T
dfwc
```

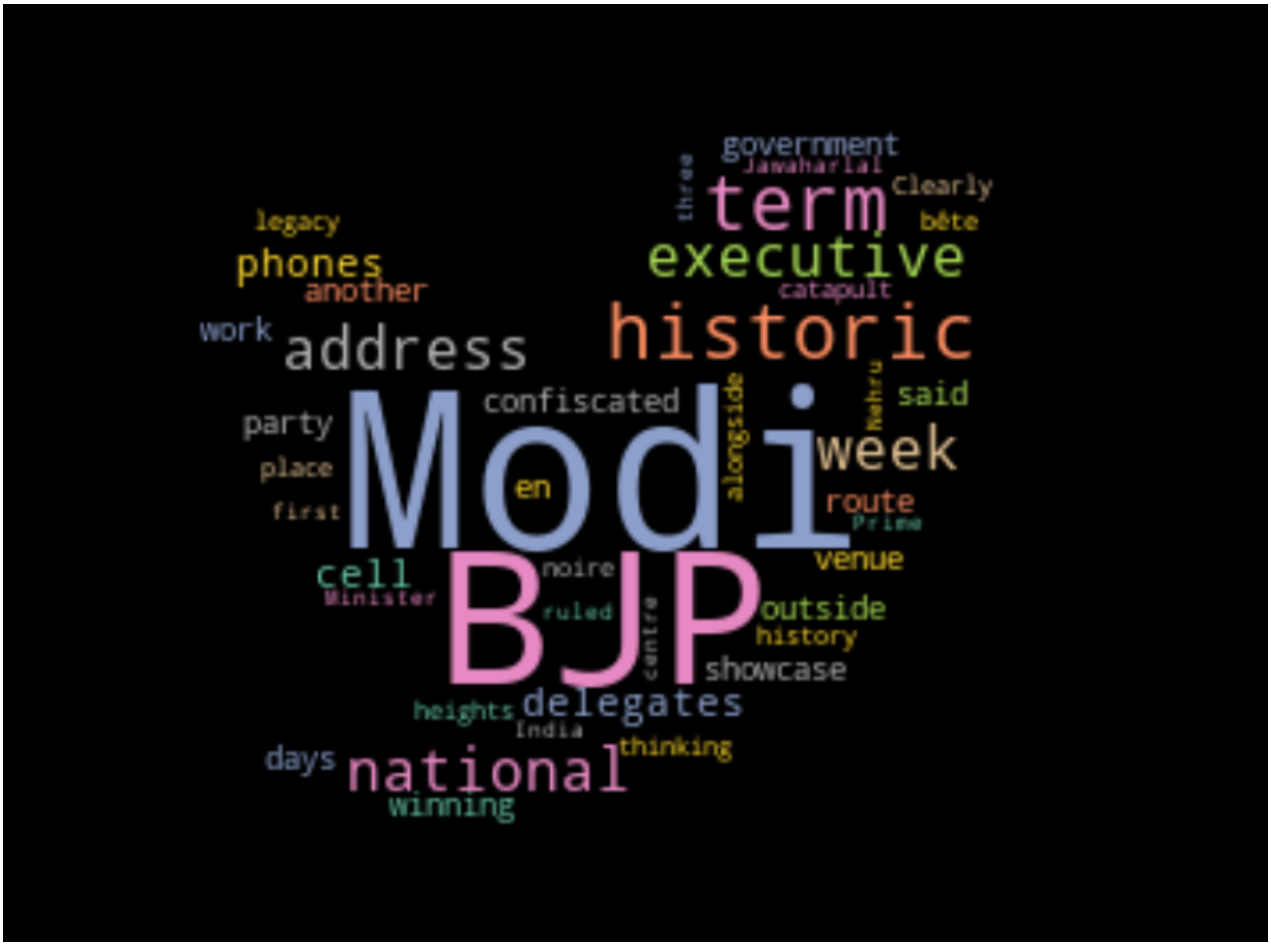
Out[10]:

	0	1	2	3	4	5	6	7	8	9
aadmi	1	0	0	0	0	0	0	0	0	0
aam	1	0	0	0	0	0	0	0	0	0
aap	1	0	0	0	0	0	0	0	0	0
abdel	0	0	0	0	0	0	1	0	0	0
abvp	0	0	0	0	2	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
work	0	0	0	0	0	0	0	0	0	1
world	0	2	0	4	0	0	0	0	0	0
would	0	0	0	0	0	0	0	0	0	1
year	0	0	0	0	0	0	1	1	0	0
years	0	0	0	0	0	1	0	0	0	0

551 rows × 10 columns

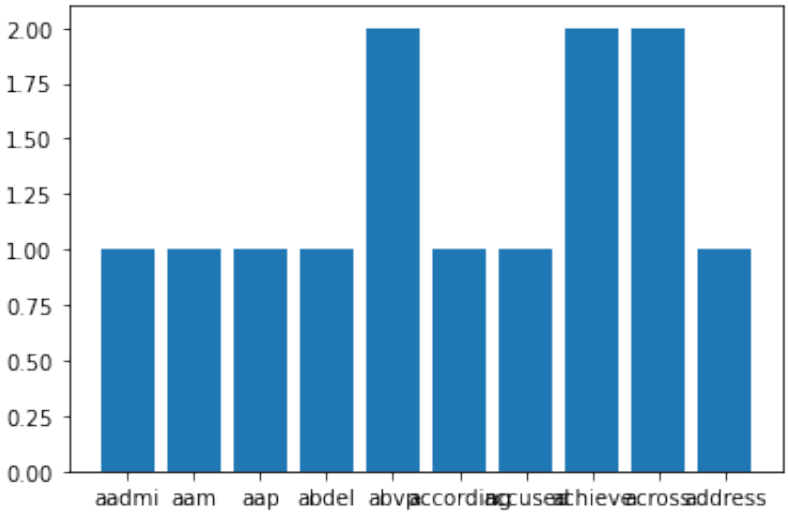
In [12]:

```
# wc= WordCloud().generate_from_frequencies(dfwc[1])
import pandas as pd
import matplotlib.pyplot as plt
from PIL import Image
import numpy as np

mask = np.array(Image.open('twitter2.png'))
for i in range(10):
    wc = WordCloud(width=1600 , height=800,mask=mask,background_color="Bla


plt.figure(figsize=(20,10),facecolor='k')
plt.imshow(wc,interpolation='bilinear')
plt.axis('off')
# plt.tight_layout (pad=0)
plt.show()
```



In [13]:

```
term=df.columns
freqpol=[]
termt=[]
for i in range(10):
    x=0
    termt.append(term[i])
    for j in range(10):
        x+=df[term[i]][j]

#         print(df[i].sum)
#     print(x)
    freqpol.append(x)
#     freqpol.append
# print(term)
plt.bar(termt,freqpol)
plt.show()
```


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
aadmi
aam
aap
abdel
abvp
accordi
acquse
achieve
across
address

In [ ]: