In [1]:
```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
import warnings
warnings.simplefilter('ignore')
# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save &
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/india-headlines-news-dataset/india-news-headlines.csv
```

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
print (pd.__version__)

from subprocess import check_output
print(check_output(["ls", "../input"]).decode("utf8"))

df = pd.read_csv("/kaggle/input/india-headlines-news-dataset/india-news-headlines.csv", dtype={'publish_date': object})

df['publish_month'] = df.publish_date.str[:6]
df['publish_year'] = df.publish_date.str[:4]
df['publish_month_only'] = df.publish_date.str[4:6]
df['publish_day_only'] = df.publish_date.str[6:8]

df['dt_date'] = pd.to_datetime(df['publish_date'], format='%Y%m%d')
df['dt_month'] = pd.to_datetime(df['publish_month'], format='%Y%m')

print (df.info())
```

```
2.0.3
india-headlines-news-dataset

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3876557 entries, 0 to 3876556
Data columns (total 9 columns):
 #   Column              Dtype
---  ------              -----
 0   publish_date        object
 1   headline_category   object
 2   headline_text       object
 3   publish_month       object
 4   publish_year        object
 5   publish_month_only  object
 6   publish_day_only    object
 7   dt_date             datetime64[ns]
 8   dt_month            datetime64[ns]
dtypes: datetime64[ns](2), object(7)
memory usage: 266.2+ MB
None
```

In [3]:
```
df.head()
```

Out[3]:

	publish_date	headline_category	headline_text	publish_month	publish_year	publish_month_only	publish_day_only	dt_date	dt_month
0	20010102	unknown	Status quo will not be disturbed at Ayodhya; s...	200101	2001	01	02	2001-01-02	2001-01-01
1	20010102	unknown	Fissures in Hurriyat over Pak visit	200101	2001	01	02	2001-01-02	2001-01-01
2	20010102	unknown	America's unwanted heading for India?	200101	2001	01	02	2001-01-02	2001-01-01
3	20010102	unknown	For bigwigs; it is destination Goa	200101	2001	01	02	2001-01-02	2001-01-01
4	20010102	unknown	Extra buses to clear tourist traffic	200101	2001	01	02	2001-01-02	2001-01-01

In [4]:
```
df.describe(include='all')
```

Out[4]:

	publish_date	headline_category	headline_text	publish_month	publish_year	publish_month_only	publish_day_only	dt_date	dt_month
count	3876557	3876557	3876557	3876557	3876557	3876557	3876557	3876557	3876557
unique	8170	1024	3604755	270	23	12	31	NaN	NaN
top	20141010	india	Straight Answers	201605	2016	05	22	NaN	NaN
freq	706	307371	6723	21698	255910	333880	128443	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014-03-01 10:55:27.614066176	2014-02-14 17:07:22.869173760
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2001-01-02 00:00:00	2001-01-01 00:00:00
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2010-08-27 00:00:00	2010-08-01 00:00:00
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014-08-08 00:00:00	2014-08-01 00:00:00
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2018-05-28 00:00:00	2018-05-01 00:00:00
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2023-06-30 00:00:00	2023-06-01 00:00:00

Extracting the headlines for the year 2023

In [5]:

```
df['publish_year'] = pd.to_numeric(df['publish_year'], errors='coerce')
df_2023 = df[df['publish_year'] == 2023]
df_2023.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 90076 entries, 3786481 to 3876556
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   publish_date        90076 non-null  object
 1   headline_category   90076 non-null  object
 2   headline_text       90076 non-null  object
 3   publish_month       90076 non-null  object
 4   publish_year        90076 non-null  int64
 5   publish_month_only  90076 non-null  object
 6   publish_day_only    90076 non-null  object
 7   dt_date             90076 non-null  datetime64[ns]
 8   dt_month            90076 non-null  datetime64[ns]
dtypes: datetime64[ns](2), int64(1), object(6)
memory usage: 6.9+ MB
```

WordClouds- Unigrams, Bigrams, trigrams and ngrams

In [6]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

In [7]:
```
all_headlines = ' '.join(df_2023['headline_text'].dropna())
```

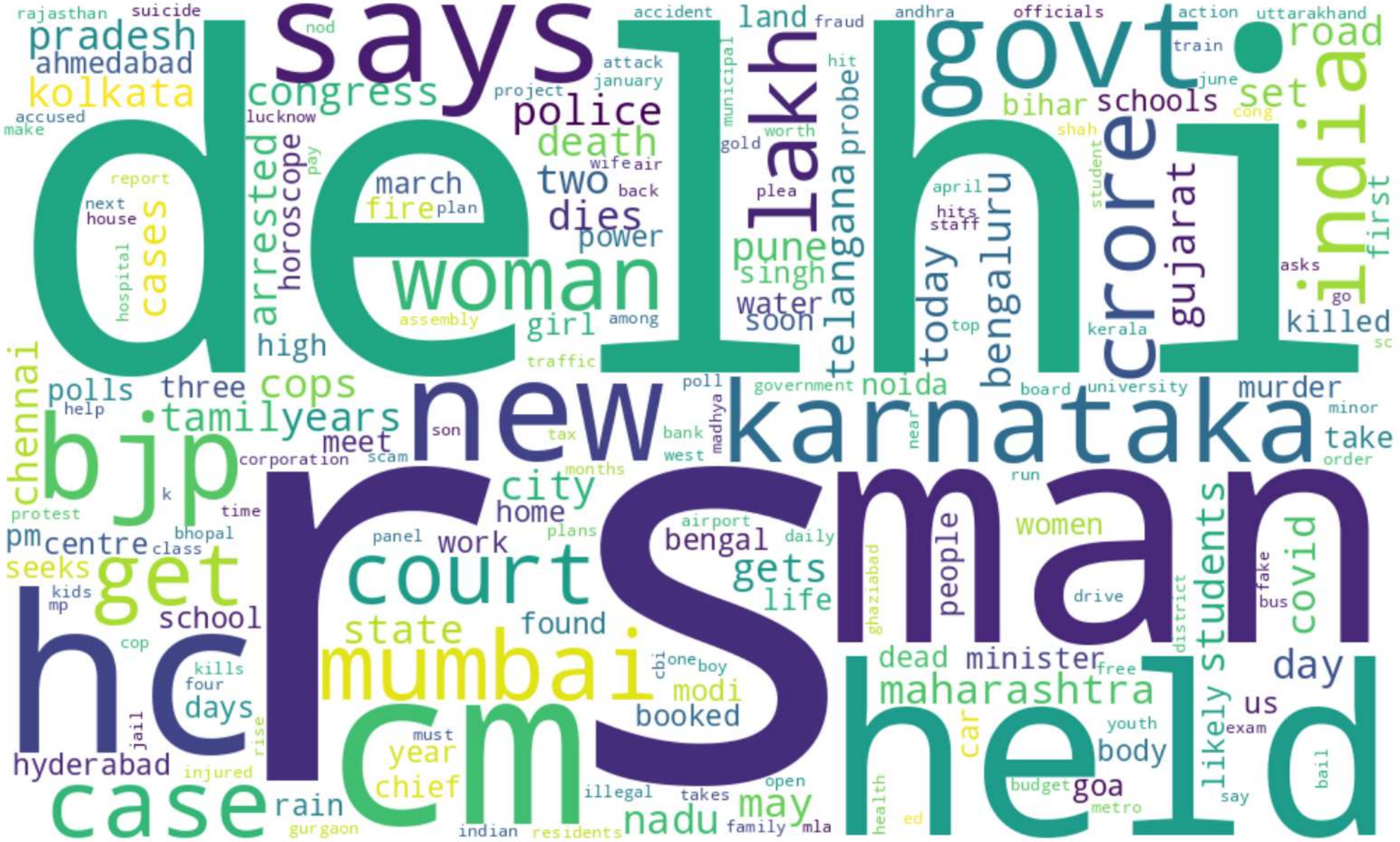
In [8]:
```
words = word_tokenize(all_headlines)
```

In [9]:
```
stop_words = set(stopwords.words('english'))
filtered_words = [word.lower() for word in words if word.isalpha() and word.lower() not in stop_words]
```

In [10]:
```
word_freq = pd.Series(filtered_words).value_counts()
```

In [11]:
```
wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(word_freq)
```

In [12]:
```
plt.figure(figsize=(24, 12))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

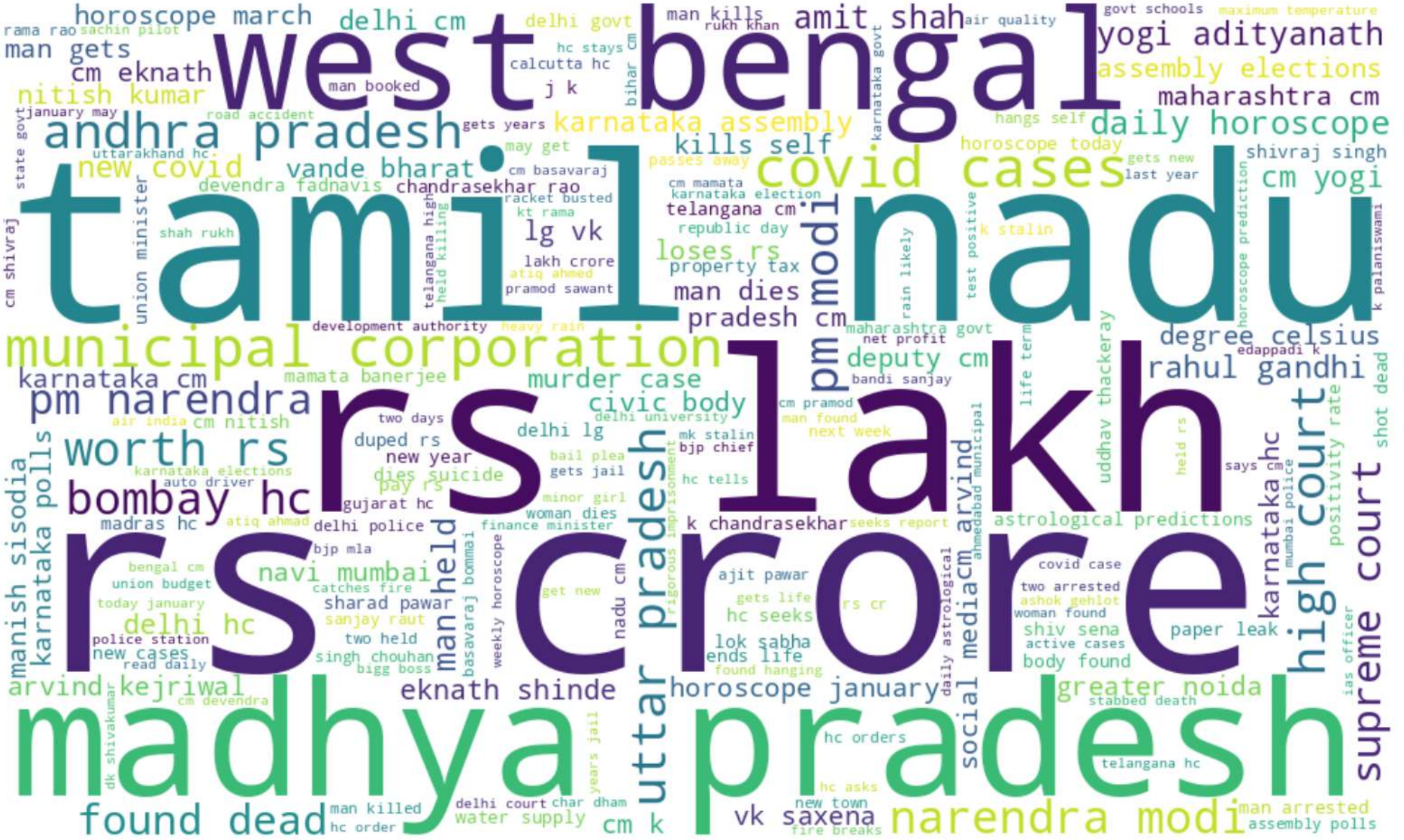


In [13]:
```
from nltk import bigrams
bigrams_list = list(bigrams(filtered_words))
```

In [14]:
```
bigram_strings = [' '.join(bigram) for bigram in bigrams_list]
bigram_freq = pd.Series(bigram_strings).value_counts()
```

In [15]:
```
bigram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(bigram_freq)
```

In [16]:
```
plt.figure(figsize=(24, 12))
plt.imshow(bigram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



In [17]:
```
from nltk import ngrams

trigrams_list = list(ngrams(filtered_words, 3))

trigram_strings = [' '.join(trigram) for trigram in trigrams_list]

trigram_freq = pd.Series(trigram_strings).value_counts()

trigram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(trigram_freq)
```

In [18]:
```
plt.figure(figsize=(24, 12))
plt.imshow(trigram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Trigram Word Cloud')
plt.show()
```

Trigram Word Cloud

In [19]:
```
quadgrams_list = list(ngrams(filtered_words, 4))

quadgram_strings = [' '.join(quadgram) for quadgram in quadgrams_list]

quadgram_freq = pd.Series(quadgram_strings).value_counts()

quadgram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(quadgram_freq)
```

In [20]:
```
plt.figure(figsize=(24, 12))
plt.imshow(quadgram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Quadgram Word Cloud')
plt.show()
```

In [21]:

```
num_words = 8 #change the no of words here
ngrams_list = list(ngrams(filtered_words, num_words))
ngram_strings = [' '.join(ngram) for ngram in ngrams_list]
ngram_freq = pd.Series(ngram_strings).value_counts()


ngram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(ngram_freq)

plt.figure(figsize=(24, 12))
plt.imshow(ngram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title(f'{num_words}-gram Word Cloud')
plt.show()
```

Sentiment Analysis

In [22]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [23]:
```
sia = SentimentIntensityAnalyzer()
df_2023.loc[:, 'compound'] = df_2023['headline_text'].apply(lambda x: sia.polarity_scores(x)['compound'])
```

In [24]:
```
df_2023['sentiment'] = df_2023['compound'].apply(lambda x: 'positive' if x >= 0 else 'negative')
```

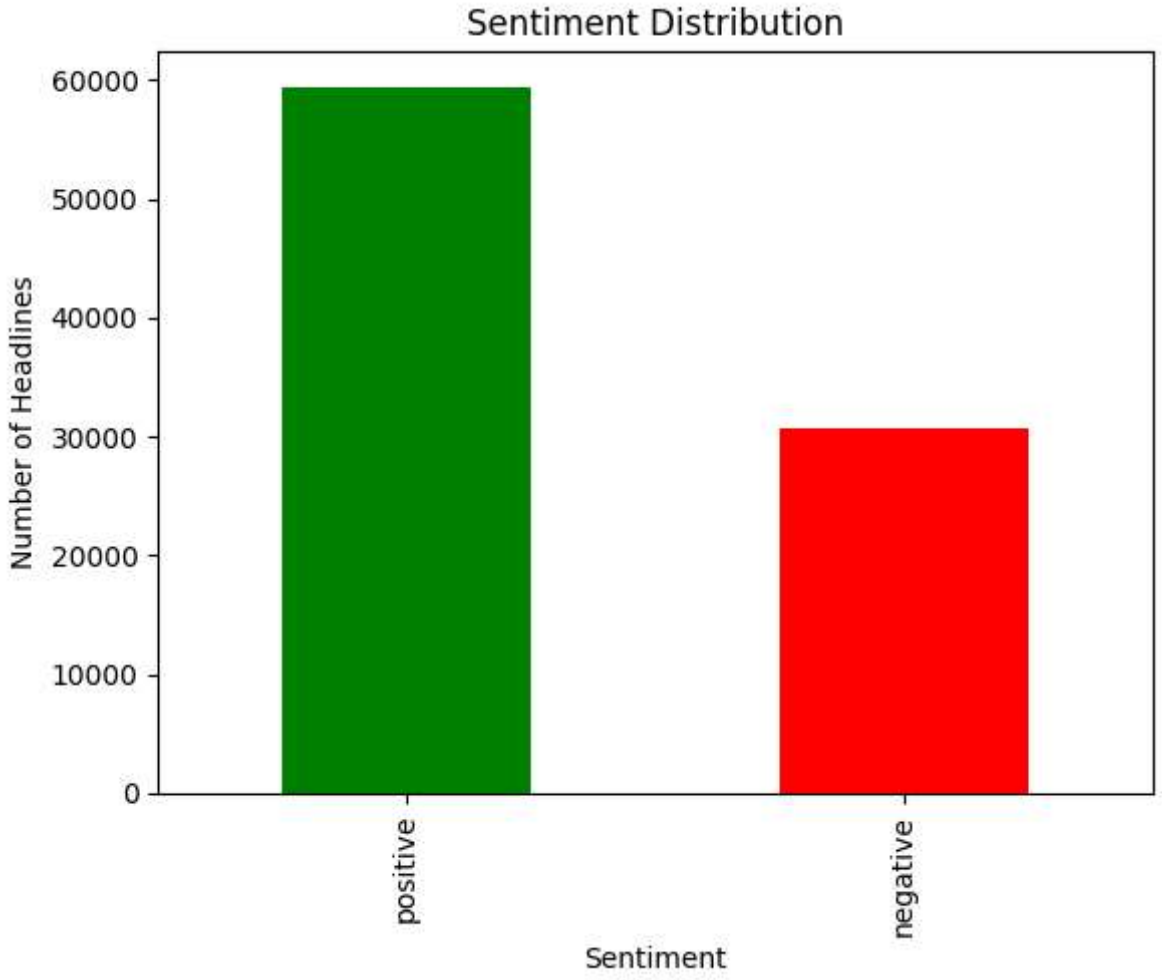
In [25]:
```
df_2023[['headline_text', 'compound', 'sentiment']]
```

Out[25]:

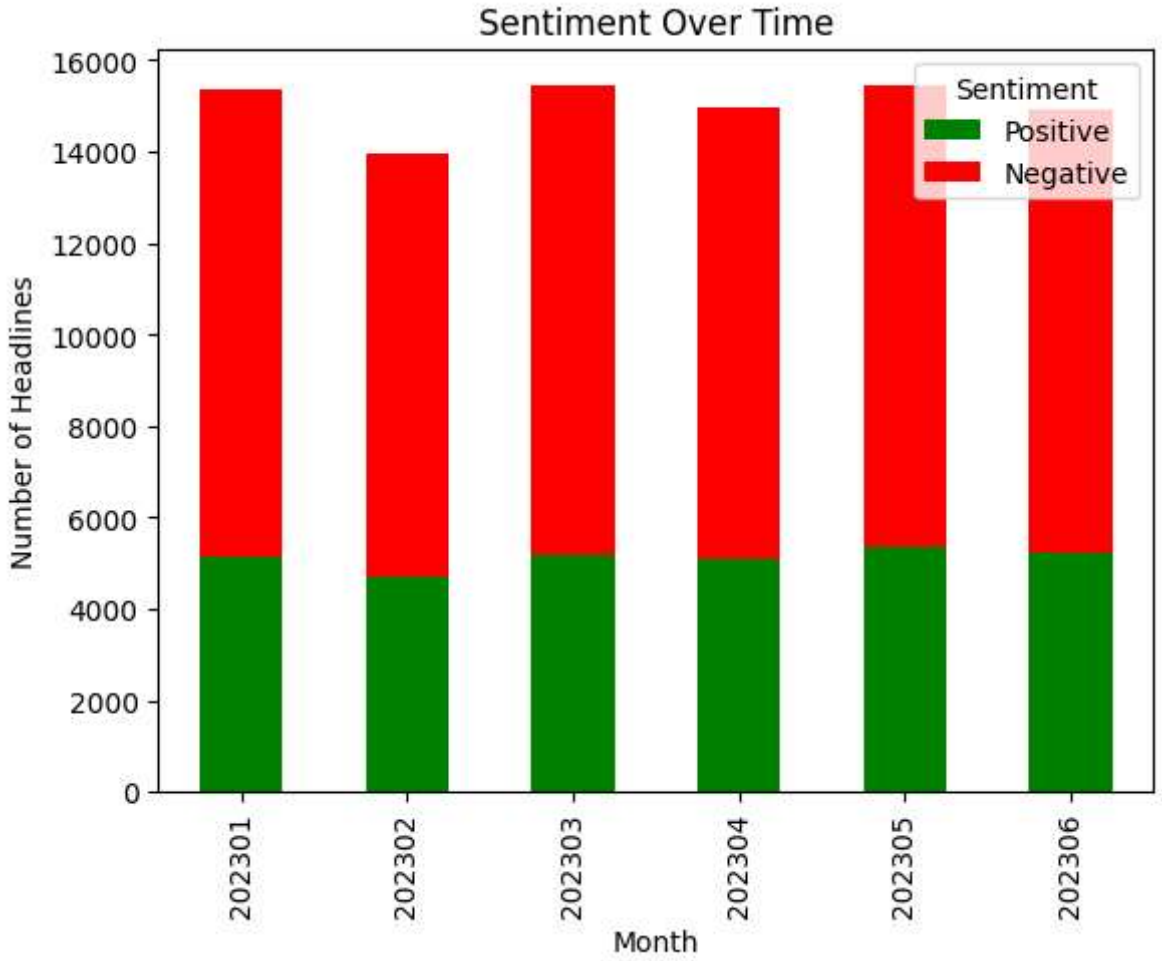
	headline_text	compound	sentiment
3786481	happy new year 2023 memes messages wishes 10 f...	0.8957	positive
3786482	happy new year quotes wishes and messages to s...	0.8625	positive
3786483	Today's Panchang; 1 January 2023: Auspicious T...	0.2023	positive
3786484	Today's Panchang; 1 January 2023: Auspicious T...	0.2023	positive
3786485	Aries Today's Rashifal - 1 January 2023: Your ...	0.5719	positive
...	...	...	...
3876552	10 PIs move HC over thwarted seniority	-0.0258	negative
3876553	Govt notifies award in memory of Parrikar for ...	0.5423	positive
3876554	After youth's death; PWD installs crash barrie...	-0.7650	negative
3876555	Authorities not acting against CRZ violations	-0.5267	negative
3876556	Technicians to hold trial run of mini-EVs in P...	0.0000	positive

90076 rows × 3 columns

In [26]:
```
sentiment_distribution = df_2023['sentiment'].value_counts()
sentiment_distribution.plot(kind='bar', color=['green', 'red'])
plt.title('Sentiment Distribution')
plt.xlabel('Sentiment')
plt.ylabel('Number of Headlines')
plt.show()
```



In [27]:
```
monthly_sentiment = df_2023.groupby('publish_month')['sentiment'].value_counts().unstack().fillna(0)
monthly_sentiment.plot(kind='bar', stacked=True, color=['green', 'red'])
plt.title('Sentiment Over Time')
plt.xlabel('Month')
plt.ylabel('Number of Headlines')
plt.legend(title='Sentiment', loc='upper right', labels=['Positive', 'Negative'])
plt.show()
```

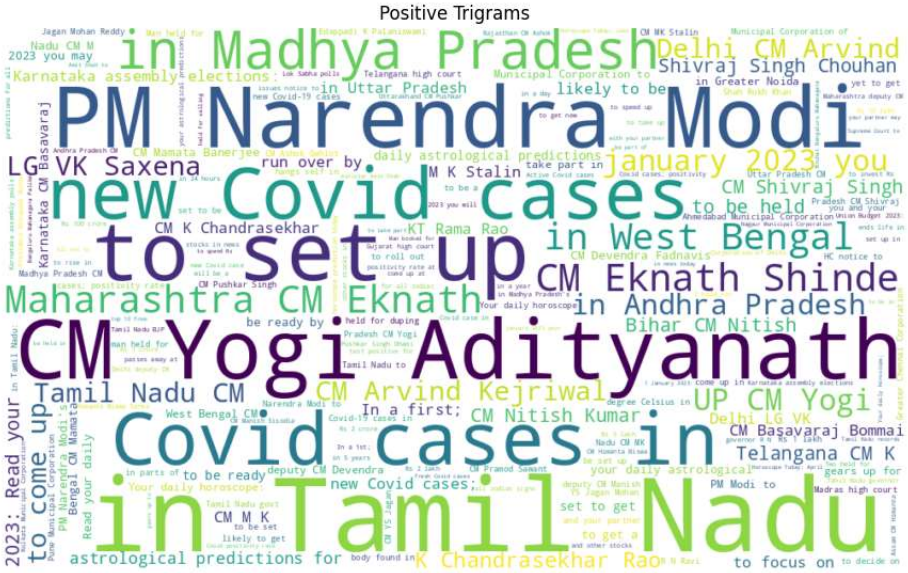


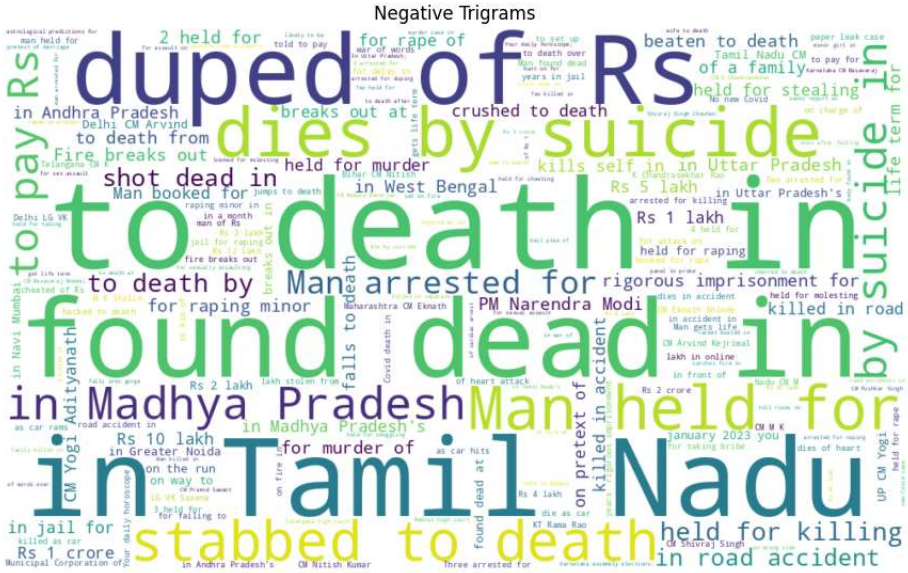
In [28]:
```
top_positive_headlines = df_2023[df_2023['sentiment'] == 'positive'].nlargest(5, 'compound')[['headline_text', 'compound']]
top_negative_headlines = df_2023[df_2023['sentiment'] == 'negative'].nsmallest(5, 'compound')[['headline_text', 'compound']]
print("Top Positive Headlines:")
print(top_positive_headlines)
print("\nTop Negative Headlines:")
print(top_negative_headlines)
```

```
Top Positive Headlines:
                                             headline_text  compound
3807307  Happy Hug Day 2023: Best Messages; Quotes; Wis...    0.9509
3808308  happy valentines day 2023 quotes sayings messa...    0.9468
3807804  Happy Kiss Day 2023: Best Messages; Quotes; Wi...    0.9460
3819280  happy holi 2023 quotes wishes messages status ...    0.9451
3845198  Winning players' respect more important than S...    0.9419

Top Negative Headlines:
                                             headline_text  compound
3863142  Odisha three-train accident: 288 dead in India...   -0.9666
3815400  MLA Umesh Pal's murder: UP cops shoot dead acc...   -0.9584
3794419  Child's death due to manja: 'Not an accident; ...   -0.9565
3823064  Ghaziabad: Dad dead; mom abandoned her; 4-year...   -0.9545
3859318  Realtor brutally murdered at his residence in ...   -0.9545
```

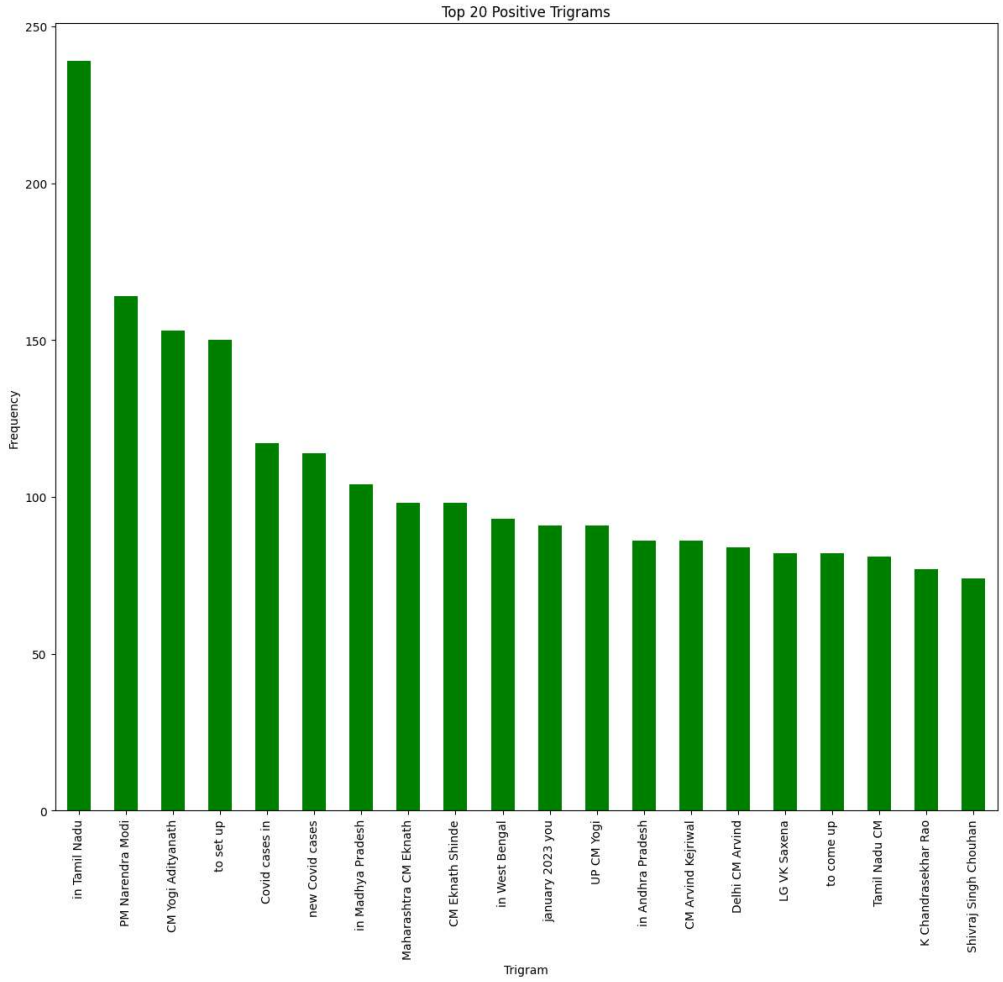
In [29]:
```
positive_words = ' '.join(df_2023[df_2023['sentiment'] == 'positive']['headline_text'])
negative_words = ' '.join(df_2023[df_2023['sentiment'] == 'negative']['headline_text'])
```

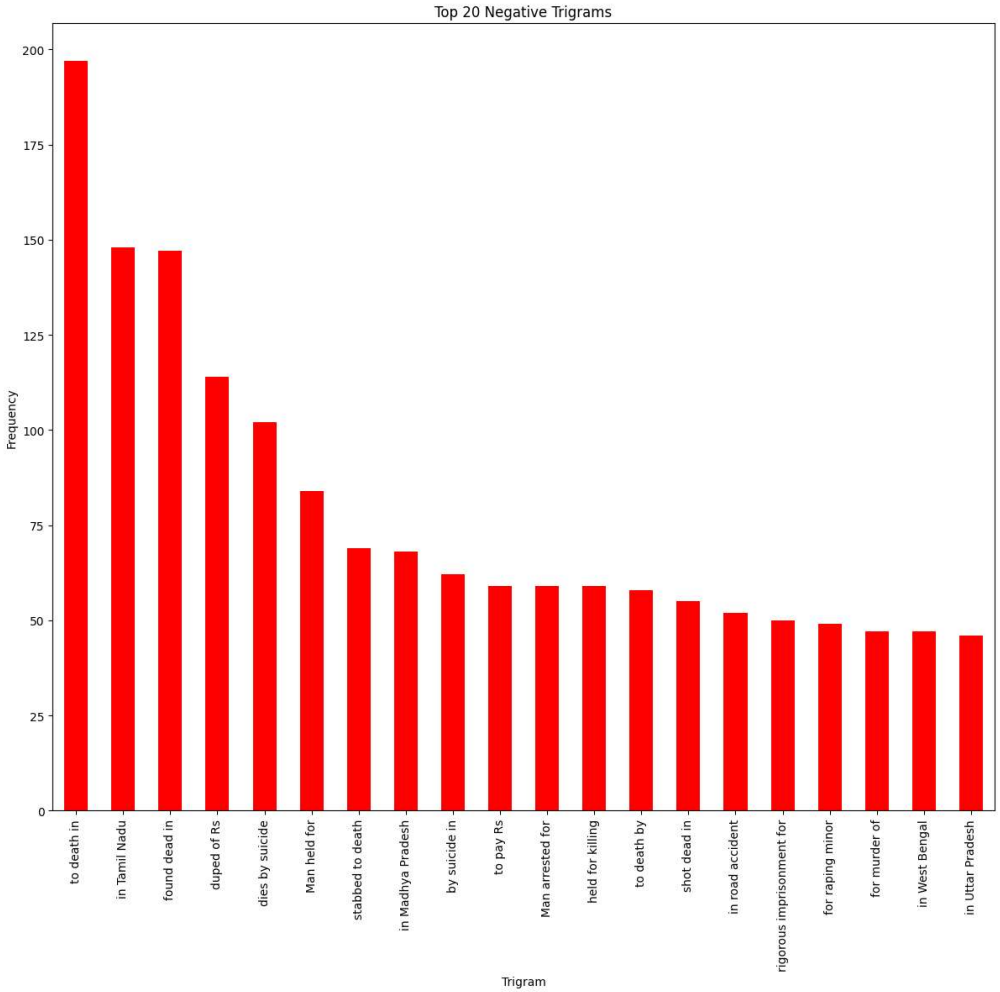
In [30]:
```
positive_trigrams_list = list(ngrams(positive_words.split(), 3))
negative_trigrams_list = list(ngrams(negative_words.split(), 3))

positive_trigram_strings = [' '.join(trigram) for trigram in positive_trigrams_list]
negative_trigram_strings = [' '.join(trigram) for trigram in negative_trigrams_list]

positive_trigram_freq = pd.Series(positive_trigram_strings).value_counts()
negative_trigram_freq = pd.Series(negative_trigram_strings).value_counts()


positive_trigram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(positive_trigram_freq)
negative_trigram_wordcloud = WordCloud(width=1000, height=600, background_color='white').generate_from_frequencies(negative_trigram_freq)

plt.figure(figsize=(24, 12))
plt.subplot(1, 2, 1)
plt.imshow(positive_trigram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Positive Trigrams')

plt.subplot(1, 2, 2)
plt.imshow(negative_trigram_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Negative Trigrams')

plt.show()
```





In [31]:
```
plt.figure(figsize=(24, 12))

plt.subplot(1, 2, 1)
positive_trigram_freq.head(20).plot(kind='bar', color='green')
plt.title('Top 20 Positive Trigrams')
plt.xlabel('Trigram')
plt.ylabel('Frequency')

plt.subplot(1, 2, 2)
negative_trigram_freq.head(20).plot(kind='bar', color='red')
plt.title('Top 20 Negative Trigrams')
plt.xlabel('Trigram')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```





Topic Modelling

In [32]:
```
from sklearn.feature_extraction.text import CountVectorizer
from nltk.corpus import stopwords

corpus = df_2023['headline_text']

stop_words = 'english'
vectorizer = CountVectorizer(stop_words=stop_words, token_pattern=r'\b\w+\b')
X = vectorizer.fit_transform(corpus)
```

In [33]:
```
from sklearn.decomposition import LatentDirichletAllocation
num_topics = 20
lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)
lda_model.fit(X)
```

Out[33]:
```
LatentDirichletAllocation(n_components=20, random_state=42)
```

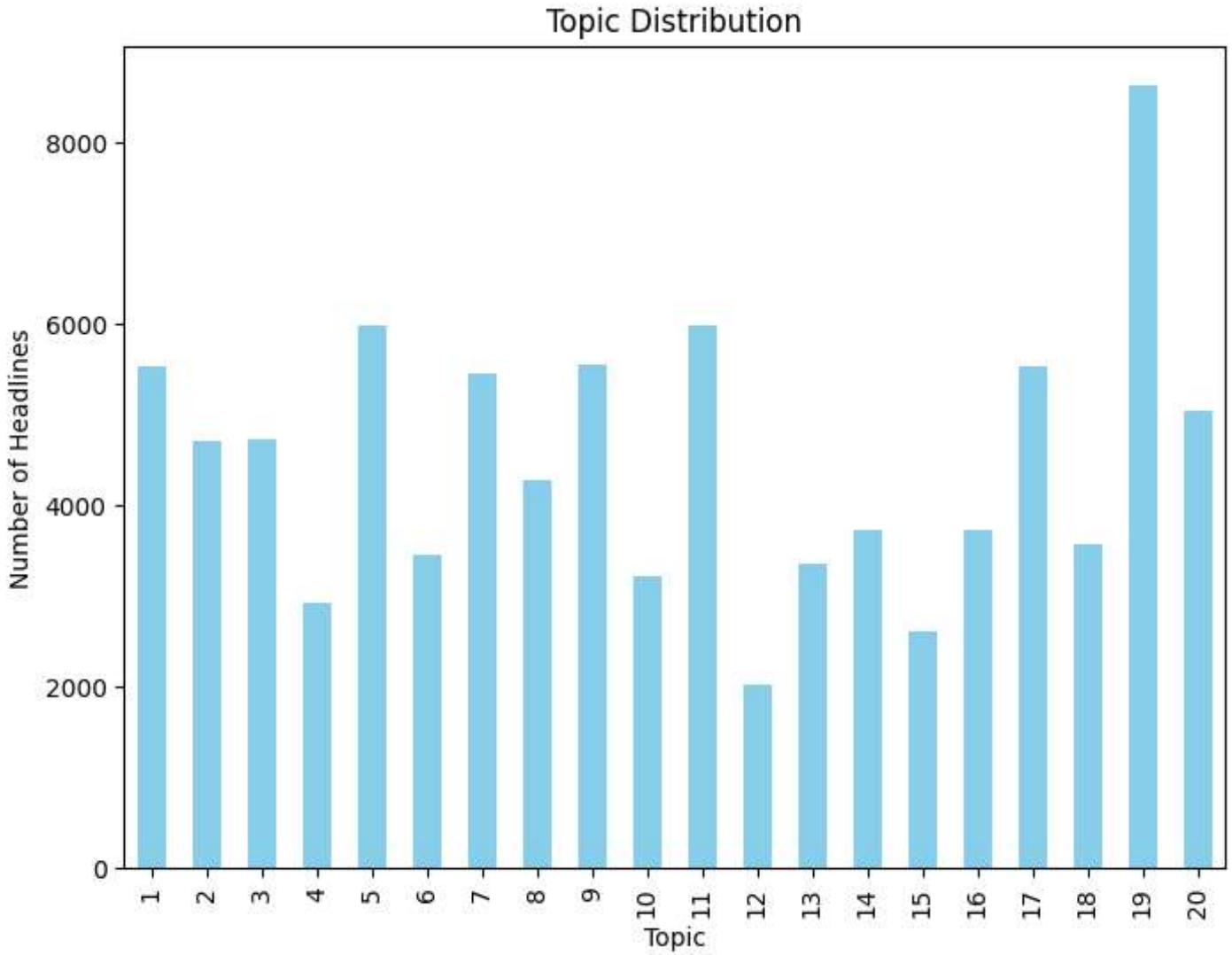
In [34]:
```
feature_names = vectorizer.get_feature_names_out()
num_top_words = 10
topics = {}

for topic_idx, topic in enumerate(lda_model.components_):
    top_words_idx = topic.argsort()[:-num_top_words - 1:-1]
    top_words = [feature_names[i] for i in top_words_idx]
    topics[f"Topic {topic_idx + 1}"] = top_words

df_topics = pd.DataFrame(topics)
df_topics
```

Out[34]:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
0	year	cm	students	pradesh	new	s	covid	2023	hc	state	killed	k	tamil	s	ahead	s
1	old	govt	bengal	madhya	tax	bihar	cases	today	court	govt	dies	paper	nadu	delhi	monsoon	pm
2	man	minister	schools	andhra	road	corporation	1	horoscope	govt	staff	cops	leak	s	woman	final	modi
3	s	says	west	uttar	soon	municipal	new	likely	high	security	2	fadnavis	waste	hospital	mumbai	cm
4	girl	t	school	s	city	body	2	march	sc	iit	car	reddy	m	kolkata	ukraine	metro
5	death	new	health	years	work	delhi	3	january	delhi	save	road	telangana	stalin	food	t	telangana
6	yr	delhi	class	china	green	kumar	years	rain	order	team	man	j	social	water	s	maharashtra
7	woman	karnataka	exam	gets	parking	cm	000	days	shah	meghalaya	injured	devendra	e	residents	chennai	narendra
8	minor	yogi	govt	indian	safety	civic	19	daily	bombay	service	dead	maharashtra	india	day	senior	rao
9	held	adityanath	board	singh	traffic	nitish	5	day	seeks	assam	bus	east	governor	dog	state	shinde

In [35]:
```
df_2023.loc[:, 'topic'] = lda_model.transform(X).argmax(axis=1) + 1
```

In [36]:
```
topic_distribution = df_2023['topic'].value_counts().sort_index()
plt.figure(figsize=(8, 6))
topic_distribution.plot(kind='bar', color='skyblue')
plt.title('Topic Distribution')
plt.xlabel('Topic')
plt.ylabel('Number of Headlines')
plt.show()
```



In [37]:
```
topic_19_headlines = df_2023[df_2023['topic'] == 19]['headline_text']
topic_19_corpus = topic_19_headlines.values
vectorizer_topic_19 = CountVectorizer(stop_words='english', token_pattern=r'\b\w+\b')
X_topic_19 = vectorizer_topic_19.fit_transform(topic_19_corpus)

feature_names_topic_19 = vectorizer_topic_19.get_feature_names_out()
top_words_topic_19 = pd.Series(X_topic_19.sum(axis=0).A1, index=feature_names_topic_19)
top_words_topic_19 = top_words_topic_19.sort_values(ascending=False)
top_words_topic_19.head(20)
```

Out[37]:
```
rs          2835
held        1860
lakh        1351
crore       1287
2           1106
man          905
s            827
arrested     775
1            758
3            696
5            595
mumbai       586
000          562
woman        495
4            494
worth        462
cops         406
police       401
delhi        364
fraud        351
dtype: int64
```

In [38]:
```
for topic_num in range(1, num_topics + 1):
    print(f"Topic {topic_num}:")
    print(df_2023[df_2023['topic'] == topic_num]['headline_text'].head(10))
    print("\n")
```

```
Topic 1:
3786513                    Ushering in a year of wellness !
3786516                    Ushering in a year of wellness !
3786528    I can't imagine anyone but Sohel as Lucky Laks...
3786537    Man-animal conflict claims another life; 2022 ...
3786539    22-year-old held for killing wife; friend over...
3786564                           Man dies in self accident
3786607       Mumbai turns into party zone on New Year's Eve
3786612    Doctor gets 10-year jail for raping colleague ...
3786618    8-year-old boy abused by seniors in south Delh...
3786633    19-year-old girl stabbed to death at home by y...
Name: headline_text, dtype: object


Topic 2:
3786546    Government's duty to hone students' talents by...
3786559    Goa police set up centralised monitoring syste...
3786579    New Year gift for Delhi: Licensing norms for r...
3786646    In Kolkata; some bring party home with DJ & ca...
3786651    Haridwar-based Gurukul Kangri University VC re...
3786677     Govt nod for Sabarimala airport land acquisition
3786680    kerala saji cherian to return to pinarayi vija...
3786704    Chennai Sangaman to be held as govt event: DMK...
3786709        Now; govt eyes better Rule of Law Index score
3786744    Alok Raj new vigilance DG as Bihar govt shuffl...
Name: headline_text, dtype: object


Topic 3:
3786483    Today's Panchang; 1 January 2023: Auspicious T...
3786484    Today's Panchang; 1 January 2023: Auspicious T...
3786545    Overuse of cellphone by kids a key complaint a...
3786589    2023 to witness launch of India's first digita...
3786602    HSC exam from February 21; SSC exam from March...
3786673    Jains in Ahmedabad to hold protest rally on Ja...
3786674    New Year gift: Promotion for UP govt teachers ...
3786701               Maths maven reveals properties of 2023
3786720    Apply online for e-pass to enter Madras HC pre...
3786737           Kerala: Idukki diocese plans joint protest
Name: headline_text, dtype: object


Topic 4:
3786534    reds issue fresh threat to atram call him oppo...
3786549    Mechanised sweeping at Delhi's Connaught Place...
3786563        Papal Flag at half-mast after ex-pope's death
```

```
3786595    Karnataka HC increases aid for girl who lost l...
3786648    Pakistan flag attached to green balloons found...
3786656    1400 tipplers caught in valsad on new years ev...
3786685    Merchant Navy engineer dies as car falls into ...
3786732    Andhra Pradesh CM YS Jagan Mohan Reddy extends...
3786739    Chavakkad: Olive Ridley turns up for nesting; ...
3786748    On last day of 2022; Uttar Pradesh CM Yogi Adi...
Name: headline_text, dtype: object


Topic 5:
3786543    Bharat Jodo Yatra to cross riot-hit areas in n...
3786552    Menace of illegal parking calls for need of to...
3786560          Work of road widening in Porvorim has begun
3786566    Works under way to make entertainment society ...
3786567        Calangute shack owners hope for uptick in biz
3786568     Road safety campaign likely this month in Nashik
3786571    Followers of five Pillars Church threaten to g...
3786576    Isro lines up big-ticket Sun; Chandrayaan-3; G...
3786580    Isro lines up big-ticket Sun; Chandrayaan-3; G...
3786593    Building curbs near naval units cut; will boos...
Name: headline_text, dtype: object


Topic 6:
3786530    my hasselblad camera became an intimate extens...
3786531    Bigg Boss Kannada 9 winner: Roopesh Shetty lif...
3786542    Municipal Corporation of Delhi holds concert f...
3786548    Ignoring state nod; Nagpur Municipal Corporati...
3786553    Bigg Boss 16: MC Stan performs his first live ...
3786557    Exclusive: Bigg Boss Kannada 9 winner Roopesh ...
3786628    Anjani Kumar takes charge; says will ensure po...
3786681    Heinous crimes drop in UP's four police commis...
3786696    Contractor ends life in Tumakuru; MLC says PWD...
3786699    5kg/month food grain scheme for 81 crore peopl...
Name: headline_text, dtype: object


Topic 7:
3786491    Scorpio Astrology Predictions - 1 January 2023...
3786506    Scorpio Astrology Predictions - 1 January 2023...
3786529    Musicians trash airlines over the mistreatment...
3786540    Delhi sees first December without cold wave in...
3786547            Tiger deaths down from 127 to 116 in 2022
3786554      2022 most peaceful for J&K in last 4 years: DGP
3786587    Assam merges 4 new districts with 4 others ahe...
```

```
3786591    After pandemic lull; Puneites ring in New Year...
3786599    Maharashtra property department's December rev...
3786614      Rs 50;000-crore investments in Maharashtra soon
Name: headline_text, dtype: object


Topic 8:
3786485    Aries Today's Rashifal - 1 January 2023: Your ...
3786486    Taurus Today's Horoscope Prediction - 1 Januar...
3786487    Cancer Horoscope Predictions Today - 1 January...
3786488    Leo Detailed Horoscope - 1 January 2023: Your ...
3786489    Virgo Daily Horoscope - 1 January 2023: This i...
3786490    Libra Daily Rashifal - 1 January 2023: Today s...
3786492    Sagittarius Free Horoscope - 1 January 2023: Y...
3786493    Capricorn Daily Horoscope Free - 1 January 202...
3786494    Aquarius Today's Free Daily Horoscope - 1 Janu...
3786495    Pisces Today's Horoscope - 1 January 2023: Thi...
Name: headline_text, dtype: object


Topic 9:
3786535    PM Narendra Modi may skip ISC; ex-Senate membe...
3786538    NGT chief; Delhi LG VK Saxena take boat ride o...
3786551    Pope emeritus Benedict XVI will be remembered ...
3786558    Yet another govt employee arrested by SIT for ...
3786574    Eight national awards to IMA Goa for its activ...
3786590    Home minister Shah didn't heed Rahul security ...
3786610    Retired Bombay HC judge complains against JJ H...
3786616    Issue warrant against cop for his lack of resp...
3786632    Telangana high court junks teacher plea to qua...
3786653    Man accused of rape can't be arrested till cou...
Name: headline_text, dtype: object


Topic 10:
3786533    Caller threatens to blow up Rashtriya Swayamse...
3786565          2 persons die in Vasco after cardiac arrest
3786584    Fix incorrect map of India asap: Minister to W...
3786604    Karnataka govt warns temples against barring D...
3786626    In a spot: Leopard runs loose as animal space ...
3786643    Forensic team at accident site; examines Risha...
3786683    This locksmith's key to quick buck: Moonlighti...
3786706    RTPCR a must for passengers from 6 nations: Ka...
3786791              Shivai prevail over Sanjeevani for crown
3786796    Greens with the help of Cidco save wetland fro...
Name: headline_text, dtype: object
```

```
Topic 11:
3786561             Colvale worker succumbs to burn injuries
3786583         9 killed as SUV collides with bus in Gujarat
3786596          4 killed in firecracker blast near Namakkal
3786639    173 drink-driving & 147 helmetless riding case...
3786640    Pedestrian dead; 4 hurt in year-end accidents ...
3786649    Couple die in fire at Ahmedabad eye care hospital
3786664    Speeding; heavy vehicles turn Wipro Circle int...
3786669    Deep in debt; upa sarpanch Balineni Tirupati d...
3786672    Two die as granite blocks from lorry fall on a...
3786691    Woman dies after bike rams speed-breaker in Ch...
Name: headline_text, dtype: object


Topic 12:
3786601            Two drowned in Uyyakondan canal in Trichy
3786668    Farmer suicides down by 300% since 2014 in Tel...
3786764    Journey towards 'bangaru Telangana' will conti...
3786780    MoD order boon to thousands of stranded Ghatko...
3786809    Journey towards 'bangaru T' will continue: BRS...
3786884                      Manisha wins kickboxing bronze
3786958      Forest fire in Charmadi Ghat has greens worried
3786959    Kudla rings in New Year through music; dance; ...
3787029    Two die at turtling site in Goa's Netravali wi...
3787041    As revelry ends; last-minute flights out of Go...
Name: headline_text, dtype: object


Topic 13:
3786573            Troops interact with Merces; Curca locals
3786581         China trying to destroy Buddhism: Dalai Lama
3786670    Gold plating works will not obstruct darshan a...
3786804                                Forest Flame triumphs
3786807    Tamil Nadu CM M K Stalin launches full-day 'an...
3786859    Cabinet decides to register space park as a so...
3786864    Children's workshop concludes with a bang at T...
3786883    Gujarat-based companies outperform markets wit...
3786912                        CAIT to facilitate exporters
3786933                    Pawan; Indumathi win gold medals
Name: headline_text, dtype: object


Topic 14:
3786536    Highest suicide attempts in 2022; ambulance da...
```

```
3786550    G20 summit: Municipal Corporation of Delhi to ...
3786577     IS claims deadly attack in Egypt Suez Canal city
3786585    Rishabh Pant gets plastic surgery on forehead;...
3786592    Doctor flies down from Germany to Mumbai to he...
3786609    German vet flies to Mumbai to help with dog's ...
3786666    Asked to shut down; admin bulldozes 85 tanning...
3786707    Lucknow Municipal Corporation may take 2 years...
3786755    Jharkhand woman's severed head found in pond i...
3786759    Delhi HC notice to DUSIB on shelter at Kashmer...
Name: headline_text, dtype: object


Topic 15:
3786615    I live on highway; if I don't help; who will: ...
3786647    Cops book 2 flyers for Bangkok-Kolkata Thai fl...
3786688    Undone by war; this Russia-Ukraine love story ...
3786717     Chennai: This councillor is ahead of corporation
3786730    Many students can't utilise state scholarship ...
3786742      Dedicated to making trophies; for over a decade
3786803                     Euphoric best for Chennai feature
3786829    Pubs put a cap on pegs; help summon app cabs t...
3786836            Sahitya Akademihonour for 'Birbal' author
3786855    Disclaimer on UADD website: 'No warranties' on...
Name: headline_text, dtype: object


Topic 16:
3786555    Alia Bhatt: The way 2022 unfolded for me; it f...
3786630    Maharashtra minister Abdul Sattar says talk at...
3786724        30 'Nightingales' commissioned in Indian army
3786726    chhattisgarh cm bhupesh baghel meets pm modi s...
3786760    Kerala's pro-people's policies a model for oth...
3786769    Is Malana a child of Alexander's army? Gene st...
3786842     Traffic diversion for Metro rail stn for a month
3786888    Vacant-post details of librarians sought in Gu...
3786889                CM makes surprise visit to 3 villages
3786918                Mandya puts NY curbs in tourist spots
Name: headline_text, dtype: object


Topic 17:
3786511                  Power has gone to BJP's head: Patole
3786575    'Massive undercurrent'; 2024 win tough for BJP...
3786578    'Massive undercurrent'; 2024 win tough for BJP...
3786582    No tie-up with JD(S); PM Modi will lead Karnat...
3786588    No tie-up with JD(S); PM Modi will lead Karnat...
```

```
3786598    Rahul misleading nation; army is doing its job...
3786619    Tighter security at RSS HQ in Nagpur after cal...
3786625    Not just Karnataka; Maharashtra border with Te...
3786652    Congress MLAs start taking back their resignat...
3786708    No problem if Rahul Gandhi is PM candidate: Bi...
Name: headline_text, dtype: object


Topic 18:
3786481    happy new year 2023 memes messages wishes 10 f...
3786482    happy new year quotes wishes and messages to s...
3786497    Happy New Year 2023: Images; Quotes; Wishes; M...
3786527    Kinder; gentler; easier 2023: That's the theme...
3786532    How to make achievable resolutions; experts sh...
3786556            What's your new year resolution for 2023?
3786569            What's your new year resolution for 2023?
3786623    Bengaluru's Domlur Layout residents end the ye...
3786637      Nuclear scientist Dinesh Shukla is head of AERB
3786654    Mumbai Metropolitan Region Authority-like body...
Name: headline_text, dtype: object


Topic 19:
3786541    Delhi gym owner's ex-staffer who 'stole' Rs 10...
3786562             Sanquelim man held in job scam gets bail
3786586         Kingpin of Bihar hooch tragedy held in Delhi
3786622      Noida gang promised MBBS seats; cheated over 50
3786624    On pillion; he planned to snatch Rs 24 lakh fr...
3786627    In Bengaluru; 3 arrested with Rs 88 lakh in de...
3786636    Cable case: 3 Bengaluru cops under scanner in ...
3786650    Kids' cooperative bank in Gujarat: Big-time sa...
3786667    New Year's gift: 234 techies among 1;157 UP Po...
3786679    Man nabbed for killing friend over petty issue...
Name: headline_text, dtype: object


Topic 20:
3786570    Oppn doubts Naik's intention; says resignation...
3786594       Radhe Maa discharged in domestic violence case
3786600    NIA secures conviction in all 38 cases decided...
3786603    Wolves gone; leopards rule the roost in Pune's...
3786606    Constable files case on objectionable video of...
3786613    Fake transperson gets 7 years jail for rape & ...
3786617    Tunisha Sharma suicide case: Actor Sheezan Kha...
3786657    Centre nod to procure mandua; Uttarakhand's in...
3786662    Businessman kidnapping case: NSA against prime...
```

```
3786663    Cops looking for 2 suspects in Udaipur paper l...
Name: headline_text, dtype: object
```

In [39]:
```
df_2023
```

Out[39]:

	publish_date	headline_category	headline_text	publish_month	publish_year	publish_month_only	publish_day_only	dt_date	dt_month	compound	sentim
3786481	20230101	life-style.events	happy new year 2023 memes messages wishes 10 f...	202301	2023	01	01	2023-01-01	2023-01-01	0.8957	posi
3786482	20230101	life-style.events	happy new year quotes wishes and messages to s...	202301	2023	01	01	2023-01-01	2023-01-01	0.8625	posi
3786483	20230101	astrology.horoscope	Today's Panchang; 1 January 2023: Auspicious T...	202301	2023	01	01	2023-01-01	2023-01-01	0.2023	posi
3786484	20230101	astrology.panchang	Today's Panchang; 1 January 2023: Auspicious T...	202301	2023	01	01	2023-01-01	2023-01-01	0.2023	posi
3786485	20230101	astrology.horoscope	Aries Today's Rashifal - 1 January 2023: Your ...	202301	2023	01	01	2023-01-01	2023-01-01	0.5719	posi
...	...	...	...	...	...	...	...	...	...	...	
3876552	20230630	city.goa	10 PIs move HC over thwarted seniority	202306	2023	06	30	2023-06-30	2023-06-01	-0.0258	nega
3876553	20230630	city.goa	Govt notifies award in memory of Parrikar for ...	202306	2023	06	30	2023-06-30	2023-06-01	0.5423	posi
3876554	20230630	city.goa	After youth's death; PWD installs crash barrie...	202306	2023	06	30	2023-06-30	2023-06-01	-0.7650	nega
3876555	20230630	city.goa	Authorities not acting	202306	2023	06	30	2023-06-30	2023-06-01	-0.5267	nega

	publish_date	headline_category	headline_text	publish_month	publish_year	publish_month_only	publish_day_only	dt_date	dt_month	compound	sentim
			against CRZ violations								
3876556	20230630	city.goa	Technicians to hold trial run of mini-EVs in P...	202306	2023	06	30	2023-06-30	2023-06-01	0.0000	posi

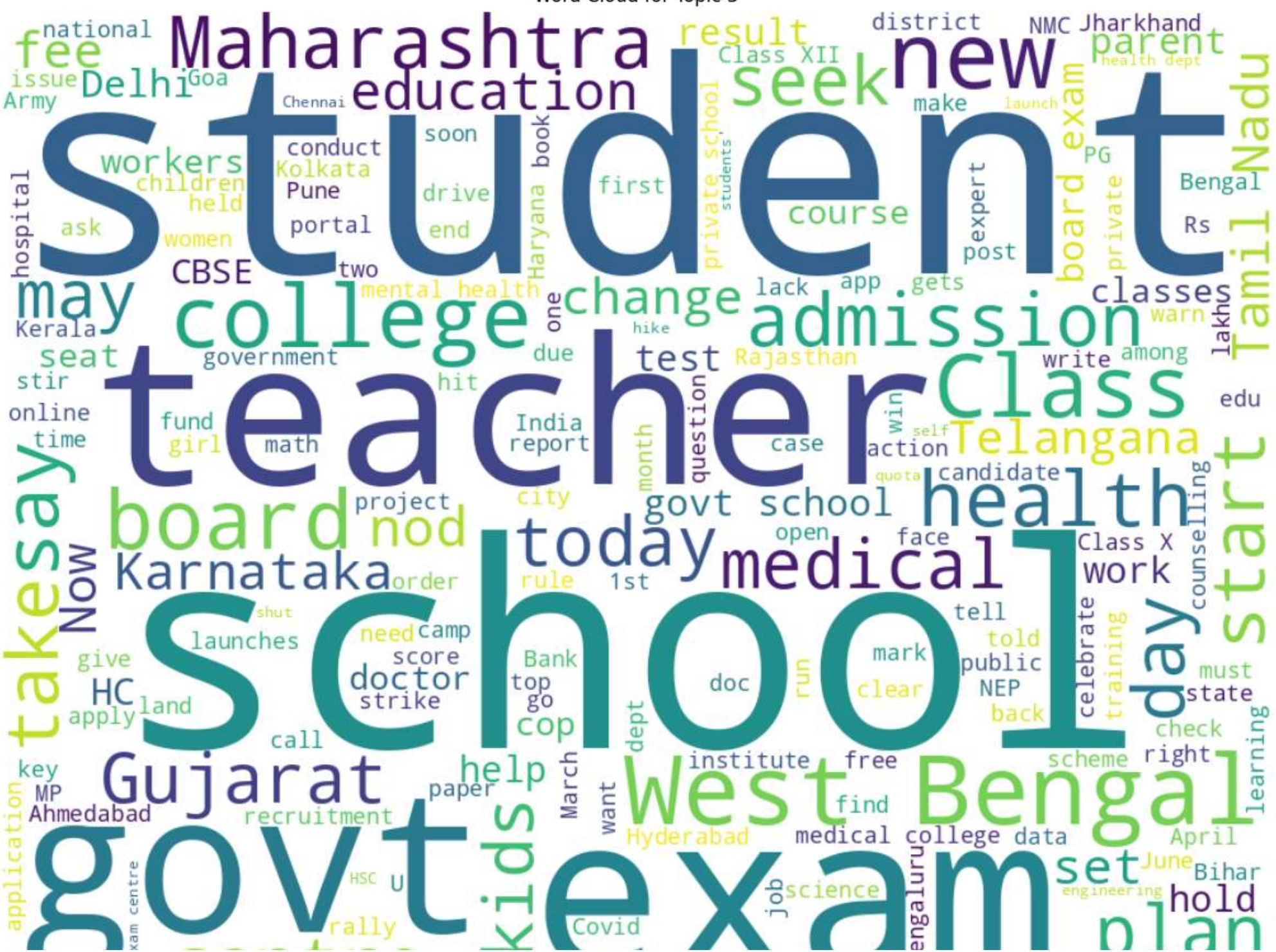
In [40]:
```
num_topics = df_2023['topic'].nunique()
for topic_num in range(1, num_topics + 1):
    topic_headlines = df_2023[df_2023['topic'] == topic_num]['headline_text']
    text_for_wordcloud = ' '.join(topic_headlines)

    wordcloud = WordCloud(width=1000, height=800, background_color='white').generate(text_for_wordcloud)

    plt.figure(figsize=(24, 12))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(f'Word Cloud for Topic {topic_num}')
    plt.show()
```

Word Cloud for Topic 1


Thane
suspected
mother
school
pretext
Patna
charge
drown
parents
Youth
University
doctor
life
district
Woman raped
woman
electrocuted

Word Cloud for Topic 2


rule
boycott
likely
CEO
medical
cost
road
ensure
use
Uttarakhand
launches


Word Cloud for Topic 3
student
teacher
school
govt
exam
Maharashtra
education
new
seek
result
fee
national
Delhi
Goa
Army
issue
Chennai
district
NMC Jharkhand
Class XII
parent
health dept
launch
make
board exam
private school
Haryana book
soon
conduct
workers
Kolkata
children
held
Pune
drive
first
course
PG
Bengal
Rs
Tamil Nadu
hospital
ask
women
portal
end
expert
post
private
students
CBSE
two
may
college
mental health
change
lack
app
gets
classes
admission
Kerala
one
hike
warn
lakhs
seat
government
due
test
Rajasthan
write
among
Class
stir
hit
question
edu
online
fund
India
win
Telangana
time
girl
math
report
case
action
self
quota
candidate
health
Now
say
take
board
nod
project
city
month
govt school
today
medical
start
day
Karnataka
order
rule
1st
open
face
Class X
work
counselling
shut
tell
launches
need
camp
told
give
score
Bank
mark
public
must
HC
doctor
top
doc
run
clear
NEP
state
apply
land
strike
go
back
celebrate
training
call
cop
dept
check
key
Gujarat
help
institute
free
scheme
right
learning
MP
paper
March
West
Bengal
Ahmedabad
recruitment
want
find
application
kids
Hyderabad
medical college data
April
June
set
Bihar
HSC
U
job
science
engineering
hold
Covid
rally
plan


court
centre
Mumbai
University
focus
assessment
year
begin
way
protest
will
certificate
yet
survey
panel
notice

Word Cloud for Topic 4


tak
home
director
unit
Pakistan
space
party
ready
Japan
Russia
family
North
second
threat
time
Goa

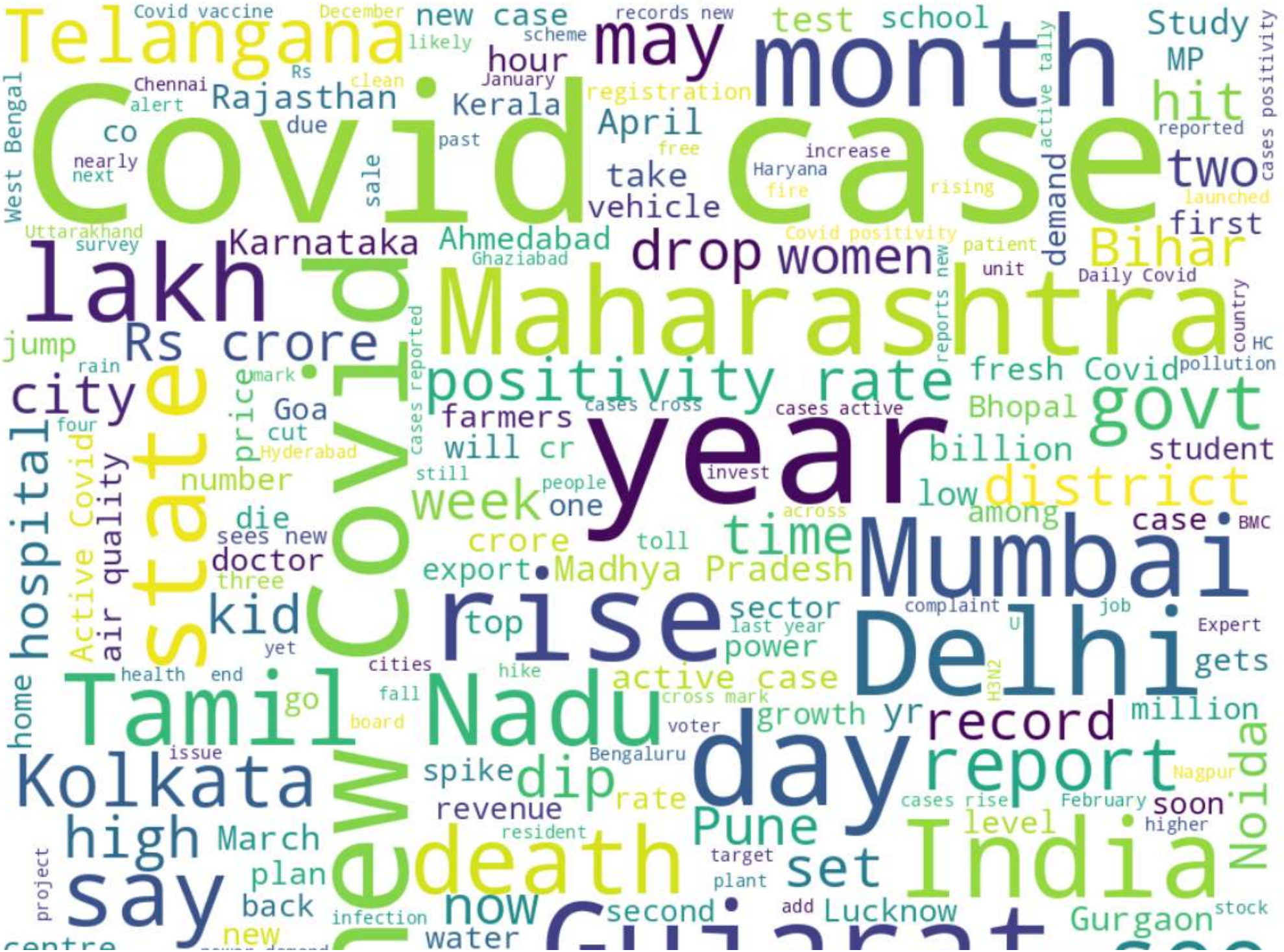
Word Cloud for Topic 5


start
women
people
Uttarakhand
speed
Rs crore
New Town
green
safety
licence
station
unit
must

Word Cloud for Topic 6

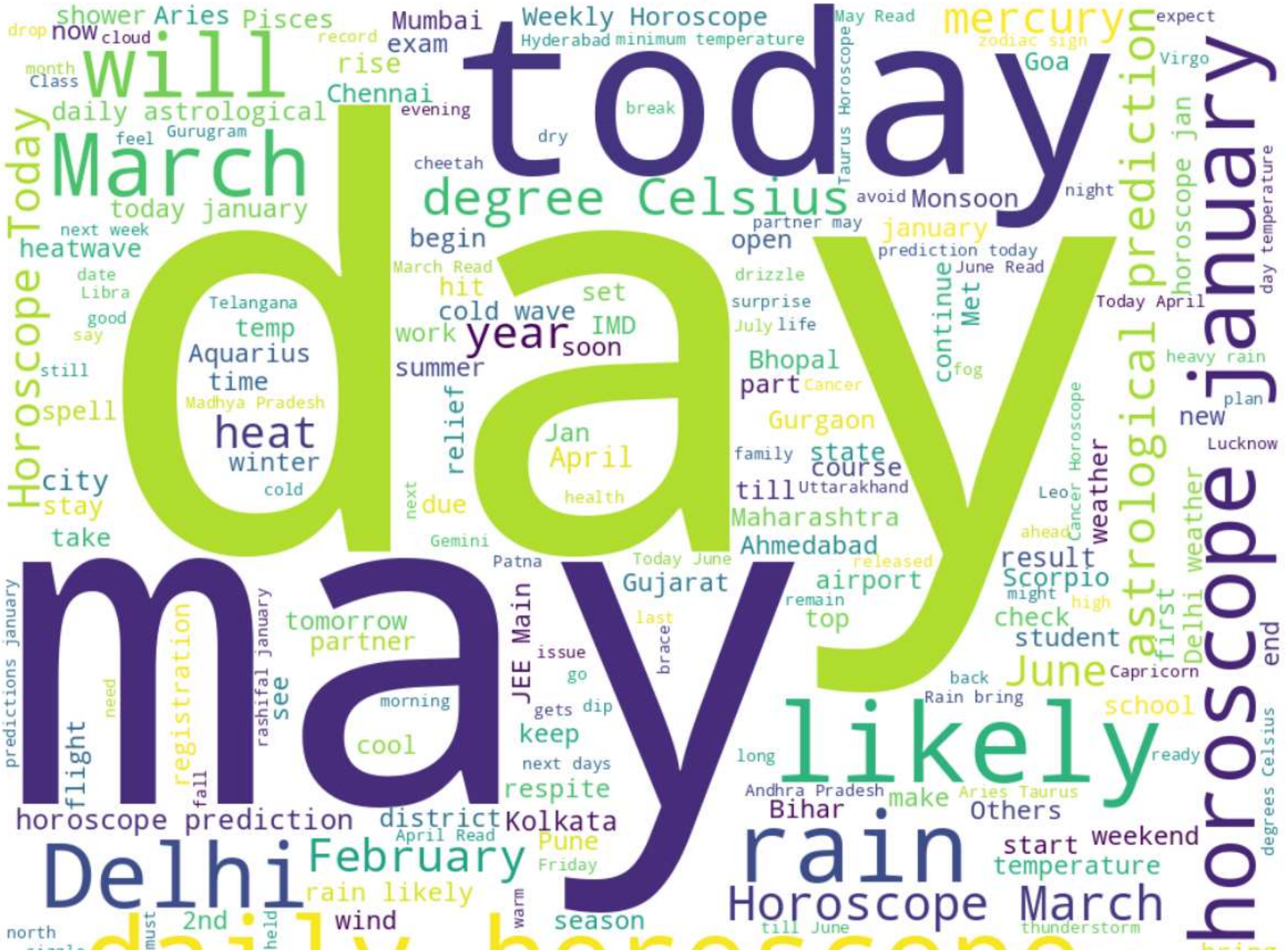

police
birthday
land
BJP
road
Job
start
give
last
Woman
Greater Noida
booked
scheme
time
CM Tejashwi
college
jail

Word Cloud for Topic 7




centre
power demand
highest
food
surge
list
Gujarat
service
claim
See
amid
pre Covid
Covid death
raise
air
found
test positive
drive

Word Cloud for Topic 8




sizzle
daily horoscope
Bengaluru
bring
India
rainfall
week
give
RajasthanUniversity
normal
light rain
maximum temperature

Word Cloud for Topic 9


HC order
transfer
Gujarat
government
slams
info
Telangana HC
AAP
bill
Rukh Khan
VK Saxena

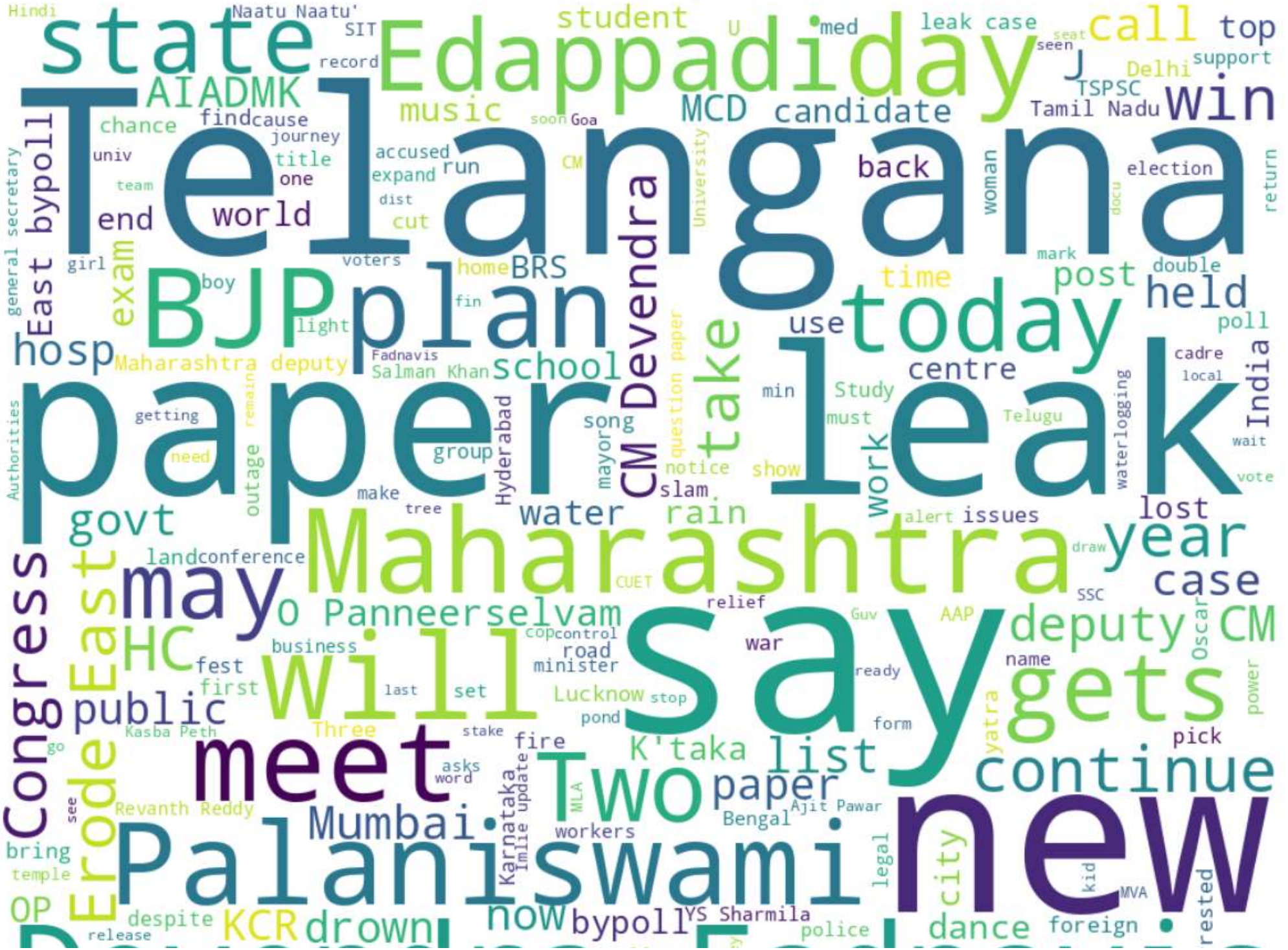
Word Cloud for Topic 10

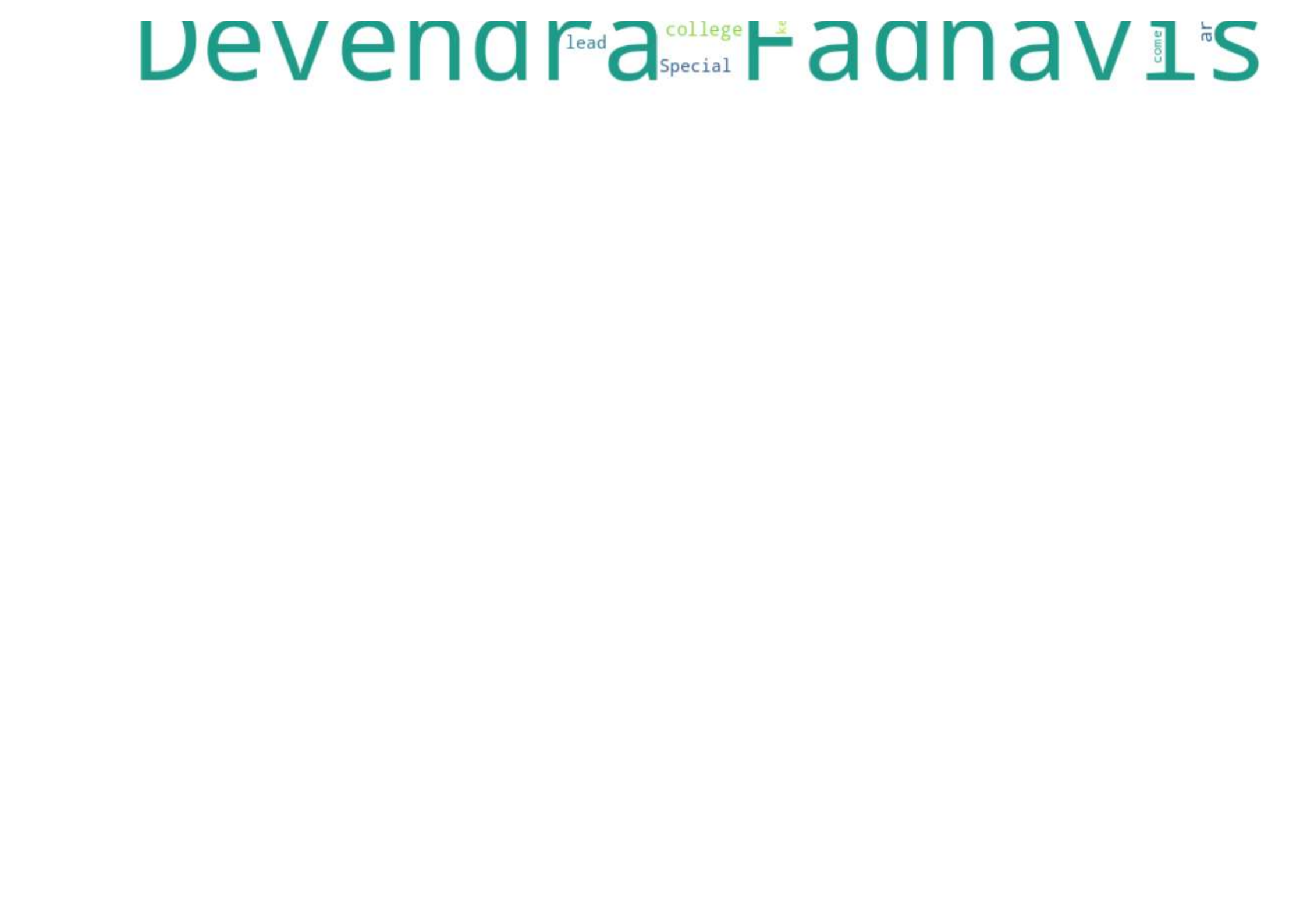

plan
site
Bank
Benga
final
Cen
fund
board
minister
sport
women
policy
G20

Word Cloud for Topic 11

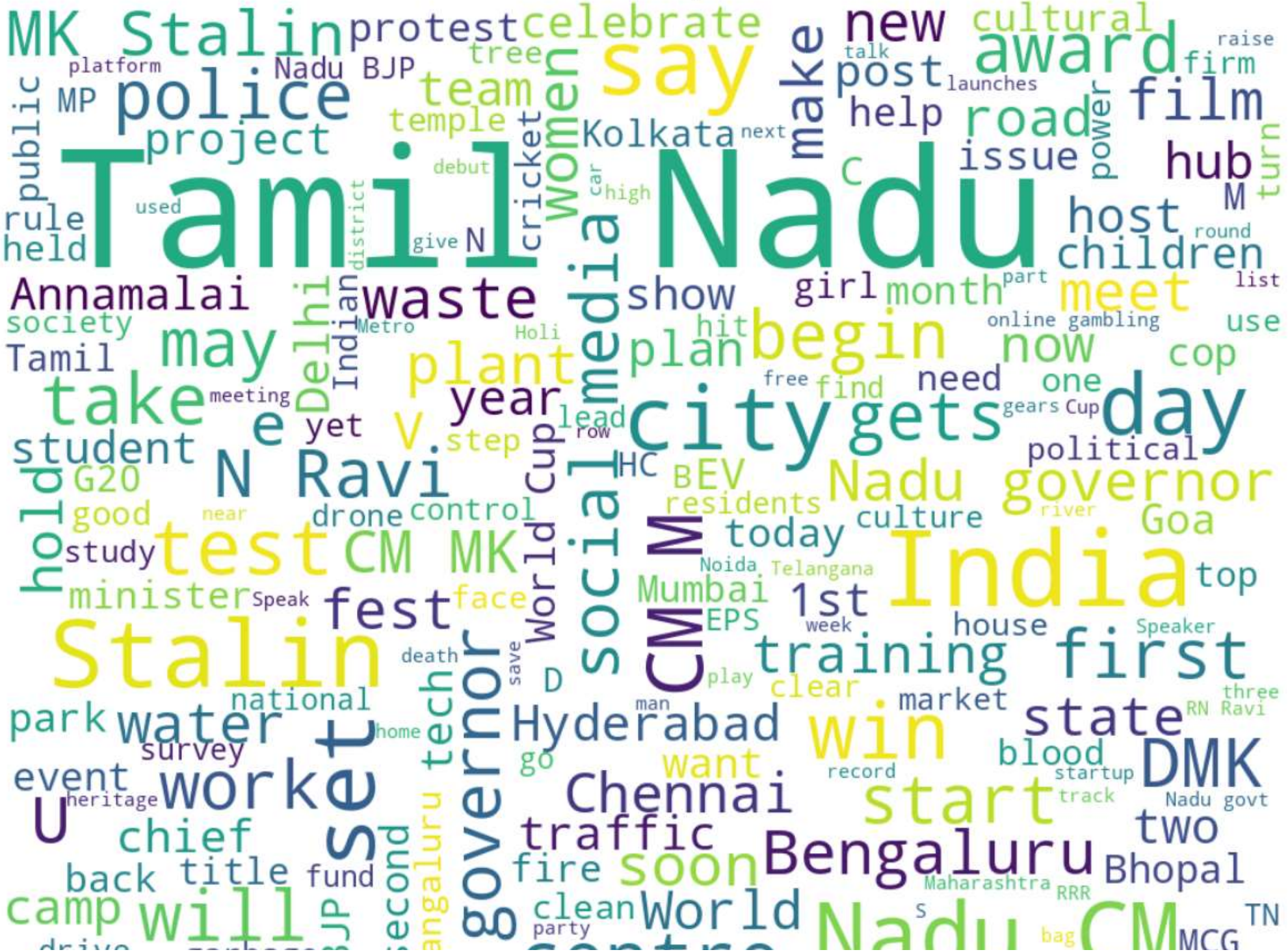

drunk
scooter
body
MP
open
wrong
dist
head
help
collision
hosp
Pune
hospital

Word Cloud for Topic 12



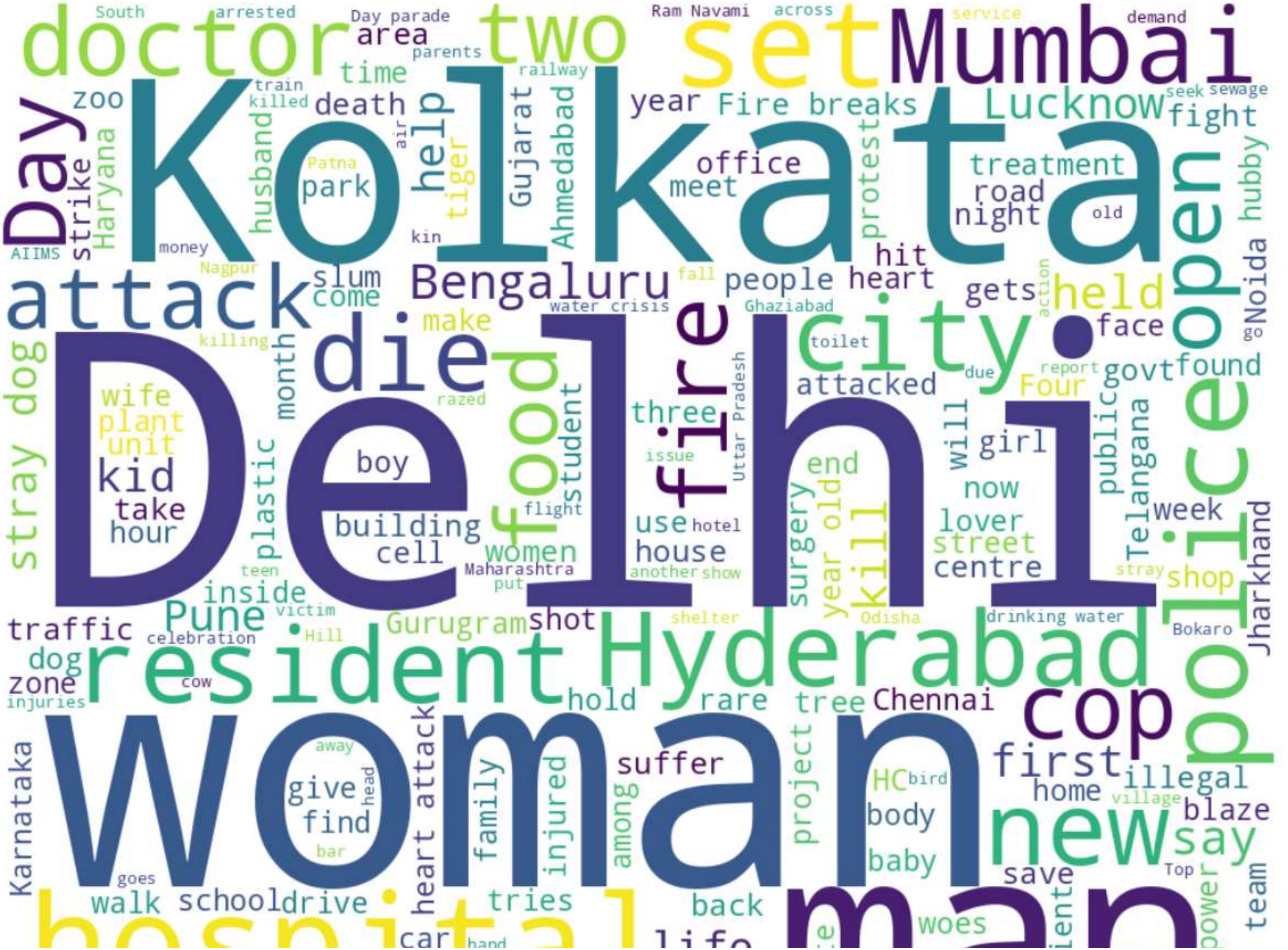

Devendra Fadnavis
lead
college
Special
come
ar

Word Cloud for Topic 13

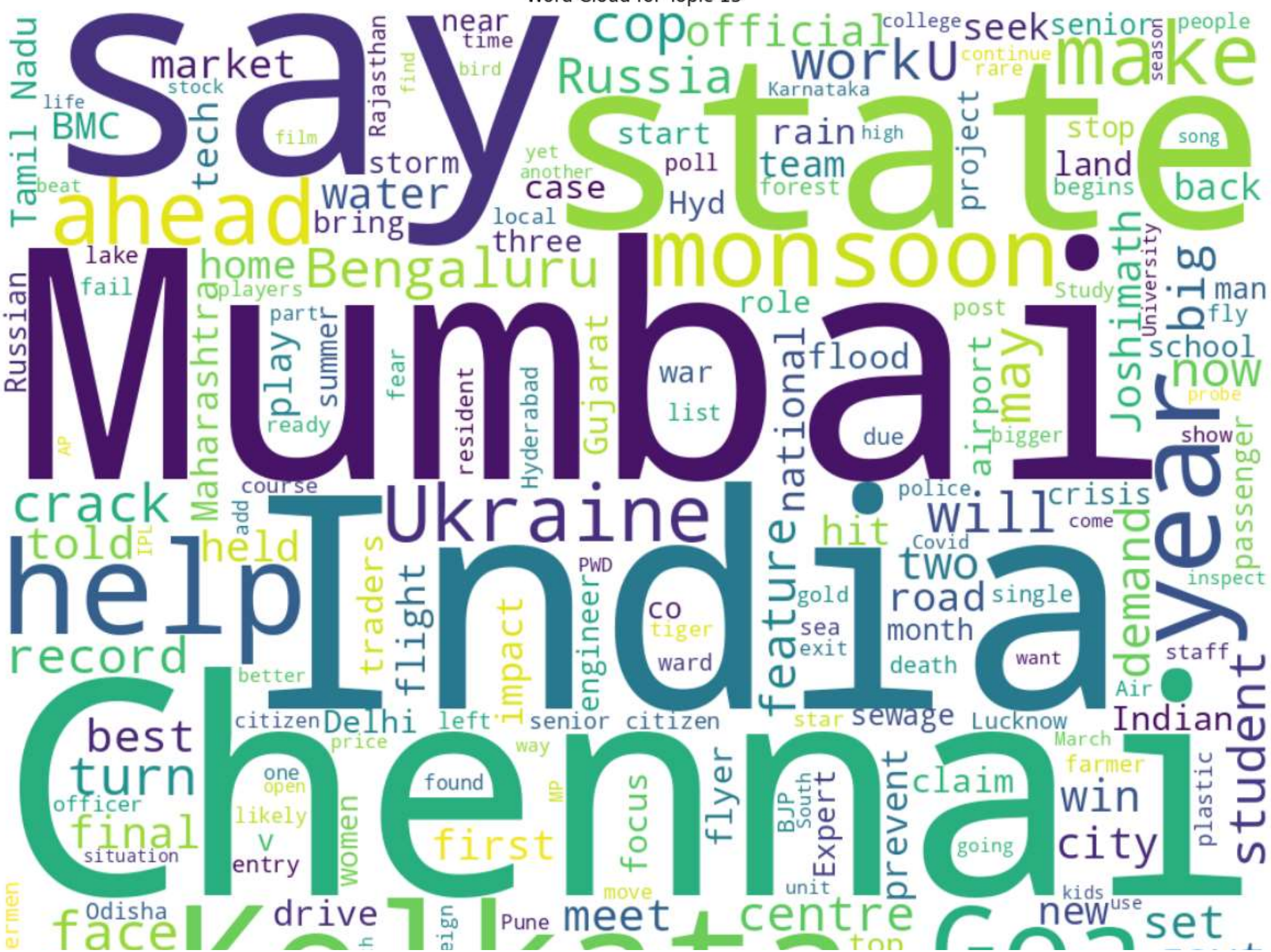



drive
garbage
open
ops
govt
green
centre
Karnataka
star
final
AIADMK

Word Cloud for Topic 14




leopard
booked
Republic Day
water
rise
U
one
stat
pati
may
case
near
Indian

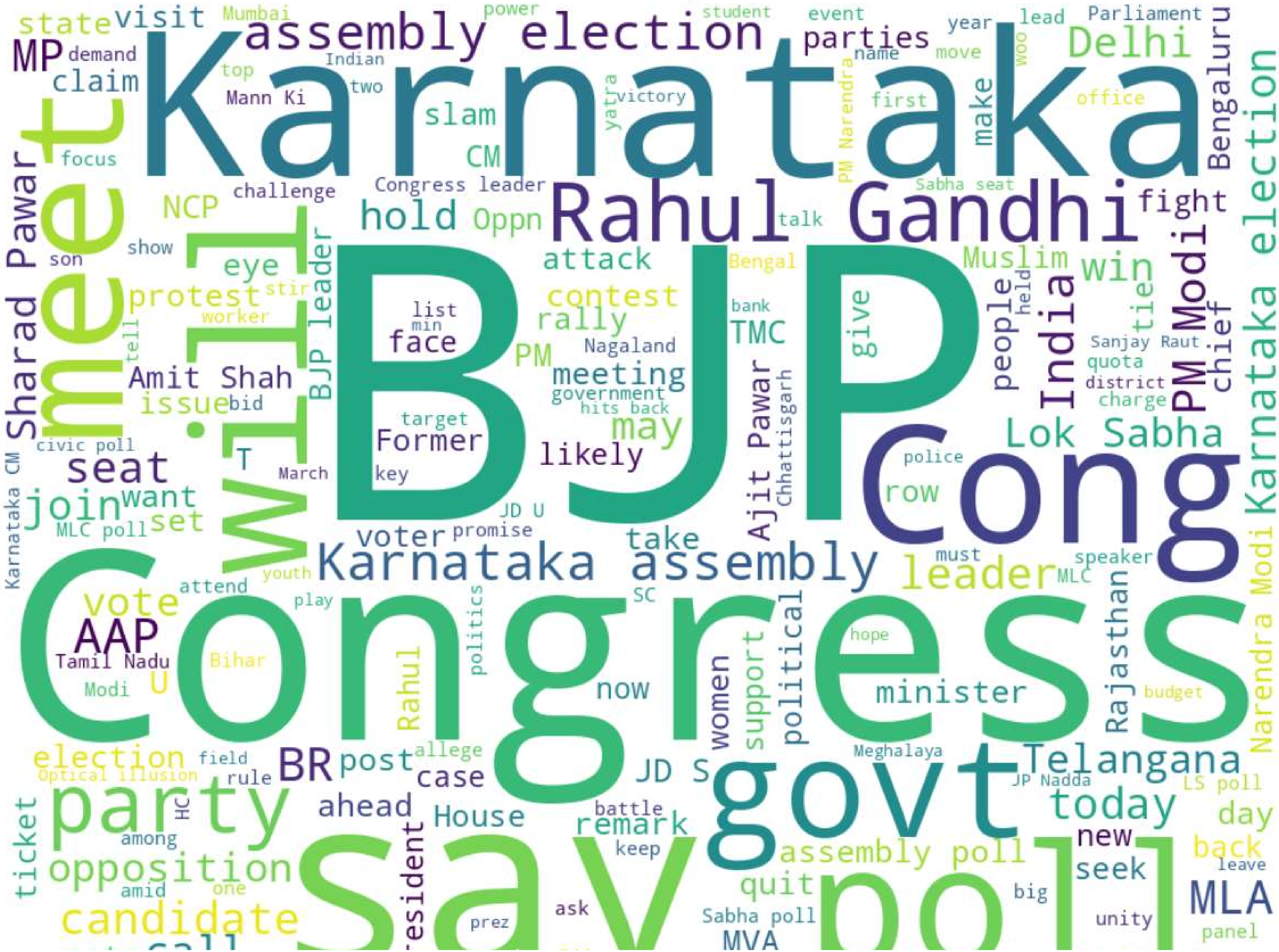

Word Cloud for Topic 15
Say
Mumbai
India
Chennai
Kolkata
state
make
monsoon
year
help
Ukraine
ahead
Bengaluru
Russia
workU
cop
official
seek
senior
market
BMC
tech
water
storm
start
rain
team
case
Hyd
land
back
stop
project
Tamil Nadu
Rajasthan
near
time
bird
stock
life
film
beat
bring
local
three
lake
fail
home
Maharashtra
Russian
players
play
summer
part
ready
fear
resident
Hyderabad
Gujarat
war
list
role
national
flood
airport
may
post
Study
University
Joshimath
big
man
fly
school
now
probe
show
bigger
due
police
crisis
will
hit
two
road
month
death
demand
feature
impact
flight
traders
engineer
PWD
gold
co
tiger
ward
sea
exit
single
Covid
come
passenger
inspect
staff
want
Air
Indian
student
Lucknow
March
farmer
plastic
sewage
star
left
senior citizen
citizen
Delhi
price
way
crack
told
held
IPL
add
course
record
better
best
turn
final
officer
situation
one
open
likely
V
entry
found
first
focus
women
MP
flyer
prevent
claim
win
city
BJP
South
Expert
going
unit
move
kids
use
new
set
face
Odisha
drive
meet
centre
Pune
top
sign
rare
continue
college
season
people
Karnataka
high
song
yet
another
forest
begins
poll
AP


fish
exam
day
drain
house
solar
react
fore
kotkata
pass
keep
wind
village
goa
govt
event
take
look

Word Cloud for Topic 16

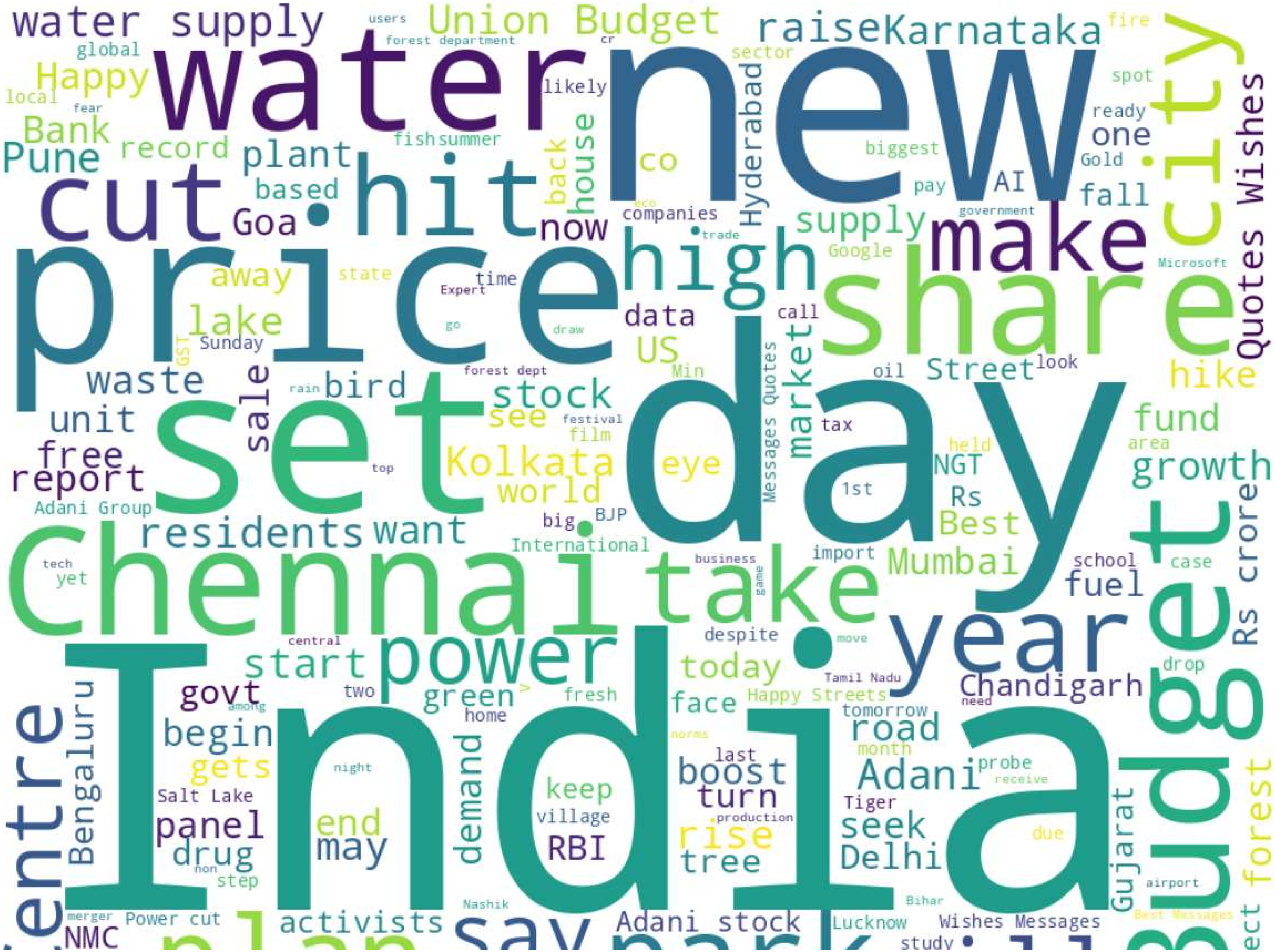

state
Sena UBT
give
launch
March
day
protest
Chandrasekhar
Rao
city
Mumbai
Secunderabad Cantonment

Word Cloud for Topic 17



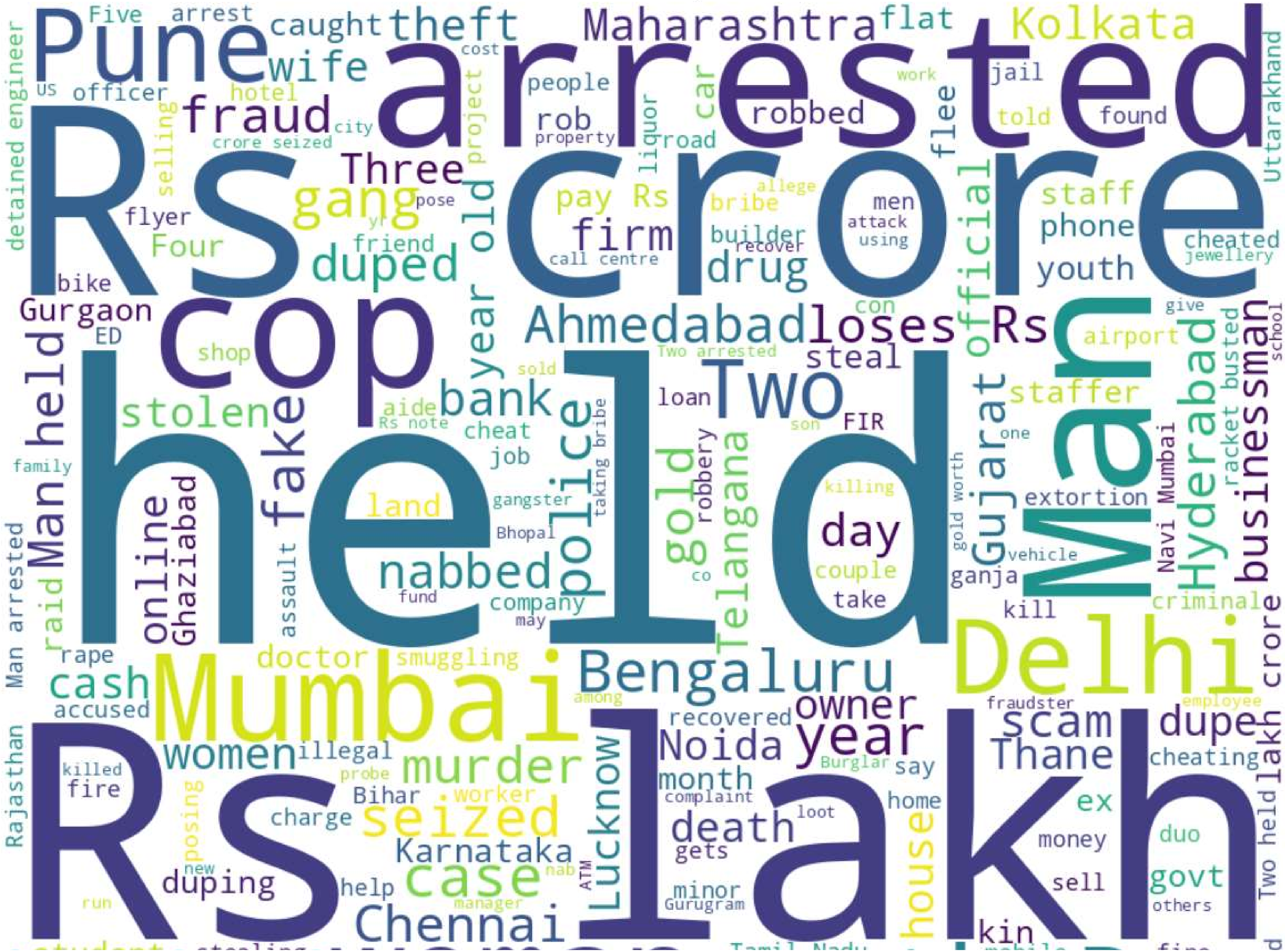

neta
plan
BJP MLA
Maharashtra
member
gets
Lingayat
ex
Siddaramaiah
campaign
stage
second
job
threat
Karnataka poll
CM Siddaramaiah
Bajrang Dal

Word Cloud for Topic 18



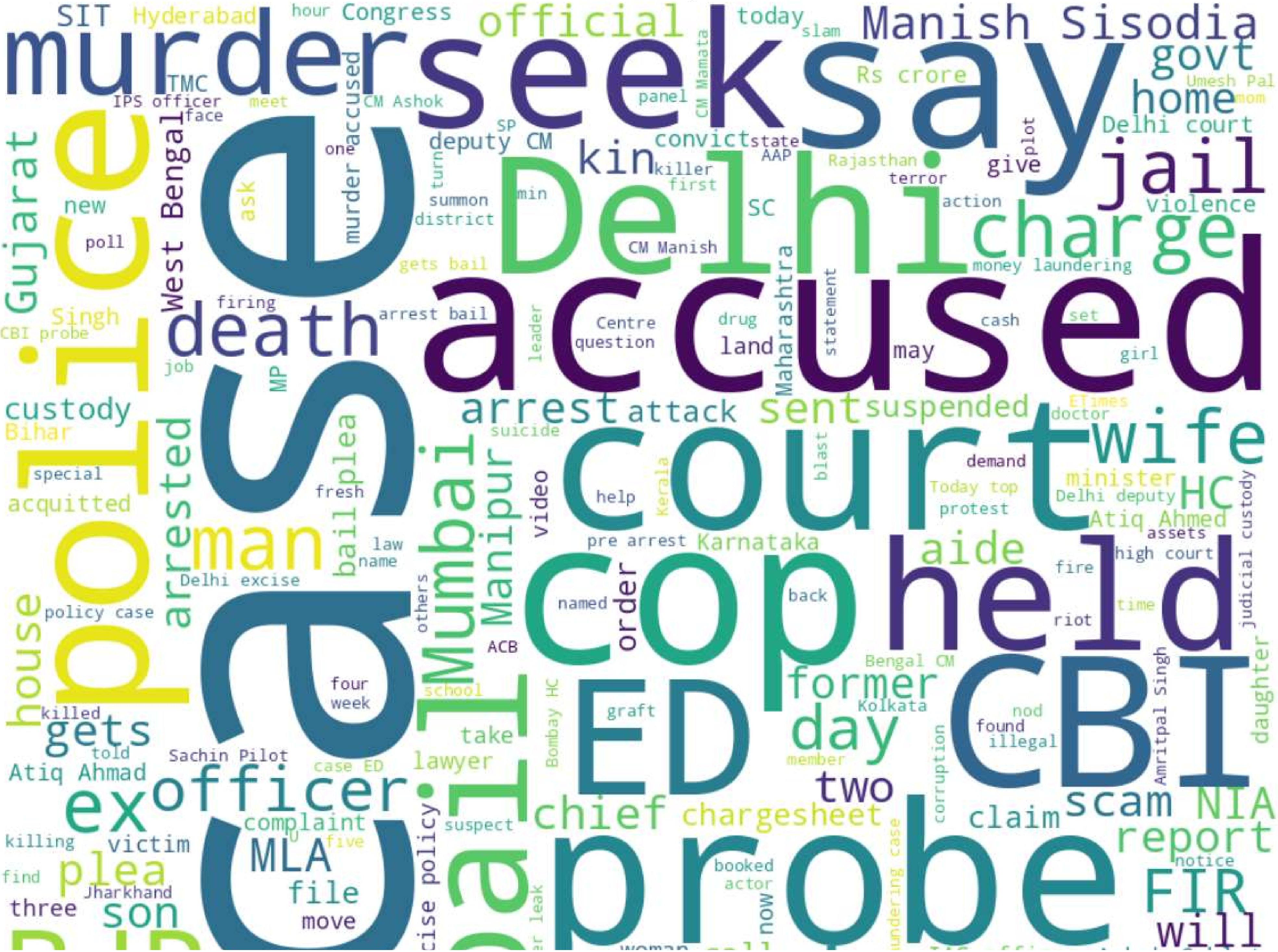

reserve
Image
Noida
close
Maharashtra
first
passes away
Municipal Corporation
Congress
lower

Word Cloud for Topic 19




booked
woman
worth
Rs
student
stealing
Tamil Nadu
mobile
fine
pay
trader
techie

Word Cloud for Topic 20



Named Entity Recognition (NER)

In [41]:
```
import nltk
from nltk import word_tokenize, pos_tag, ne_chunk
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

```
[nltk_data] Downloading package maxent_ne_chunker to
[nltk_data]     /usr/share/nltk_data...
[nltk_data]   Package maxent_ne_chunker is already up-to-date!
[nltk_data] Downloading package words to /usr/share/nltk_data...
[nltk_data]   Package words is already up-to-date!
```

Out[41]:
```
True
```

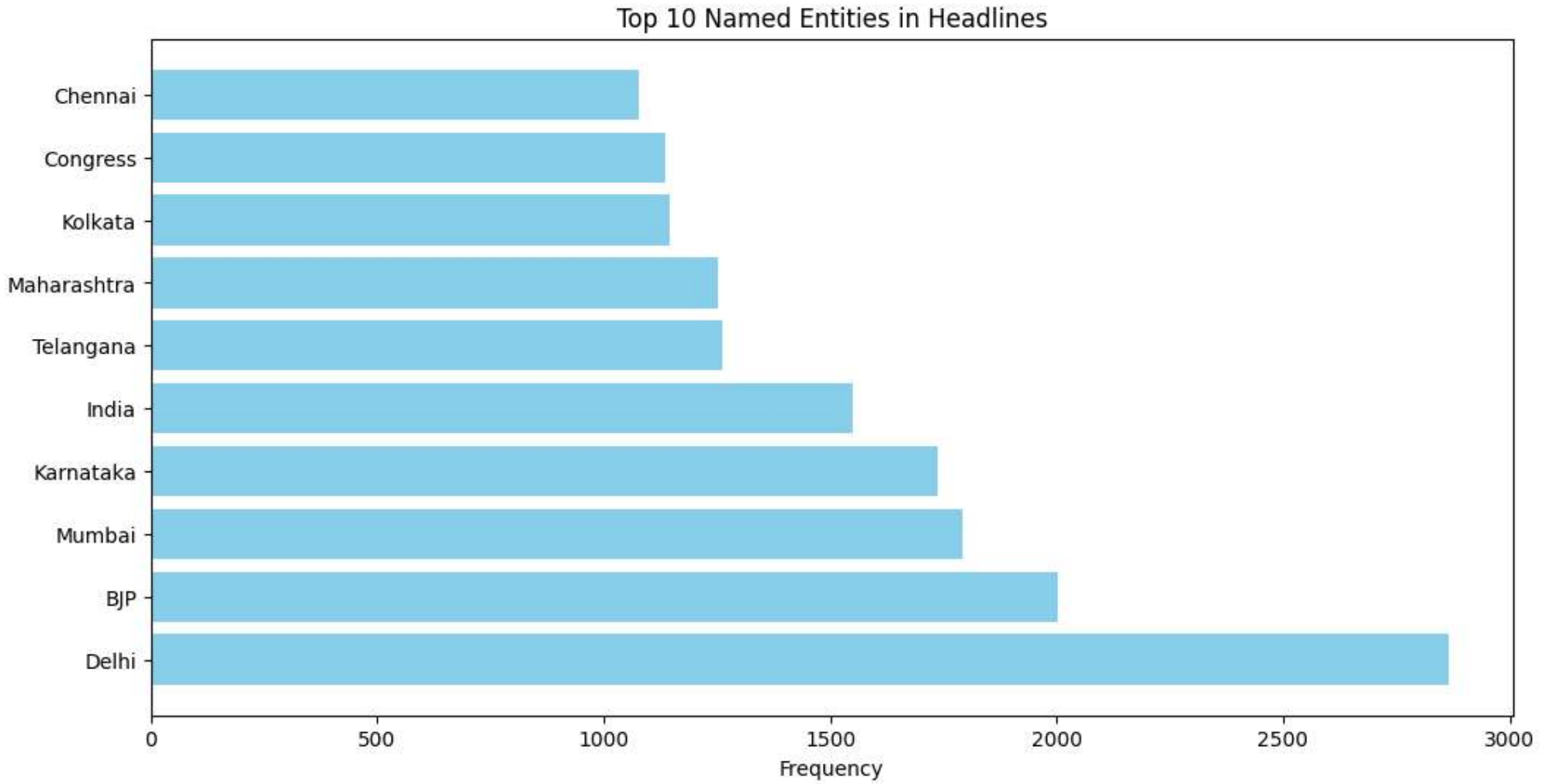
In [42]:
```
for i in range(5):
    sample_headline = df_2023['headline_text'].iloc[i]

    words = word_tokenize(sample_headline)
    tags = pos_tag(words)
    tree = ne_chunk(tags)

    entities = []
    for subtree in tree:
        if isinstance(subtree, nltk.Tree):
            entity = " ".join([token[0] for token in subtree.leaves()])
            entities.append(entity)

    print(f"Named Entities in Headline {i + 1}:")
    print(entities)
    print("\n" + "-"*50 + "\n")
```

```
Named Entities in Headline 1:
[]

------------------------------------------------

Named Entities in Headline 2:
[]

------------------------------------------------

Named Entities in Headline 3:
['Panchang', 'Important']

------------------------------------------------

Named Entities in Headline 4:
['Panchang', 'Important']

------------------------------------------------

Named Entities in Headline 5:
['Rashifal']

------------------------------------------------
```

In [43]:
```python
from collections import Counter
```

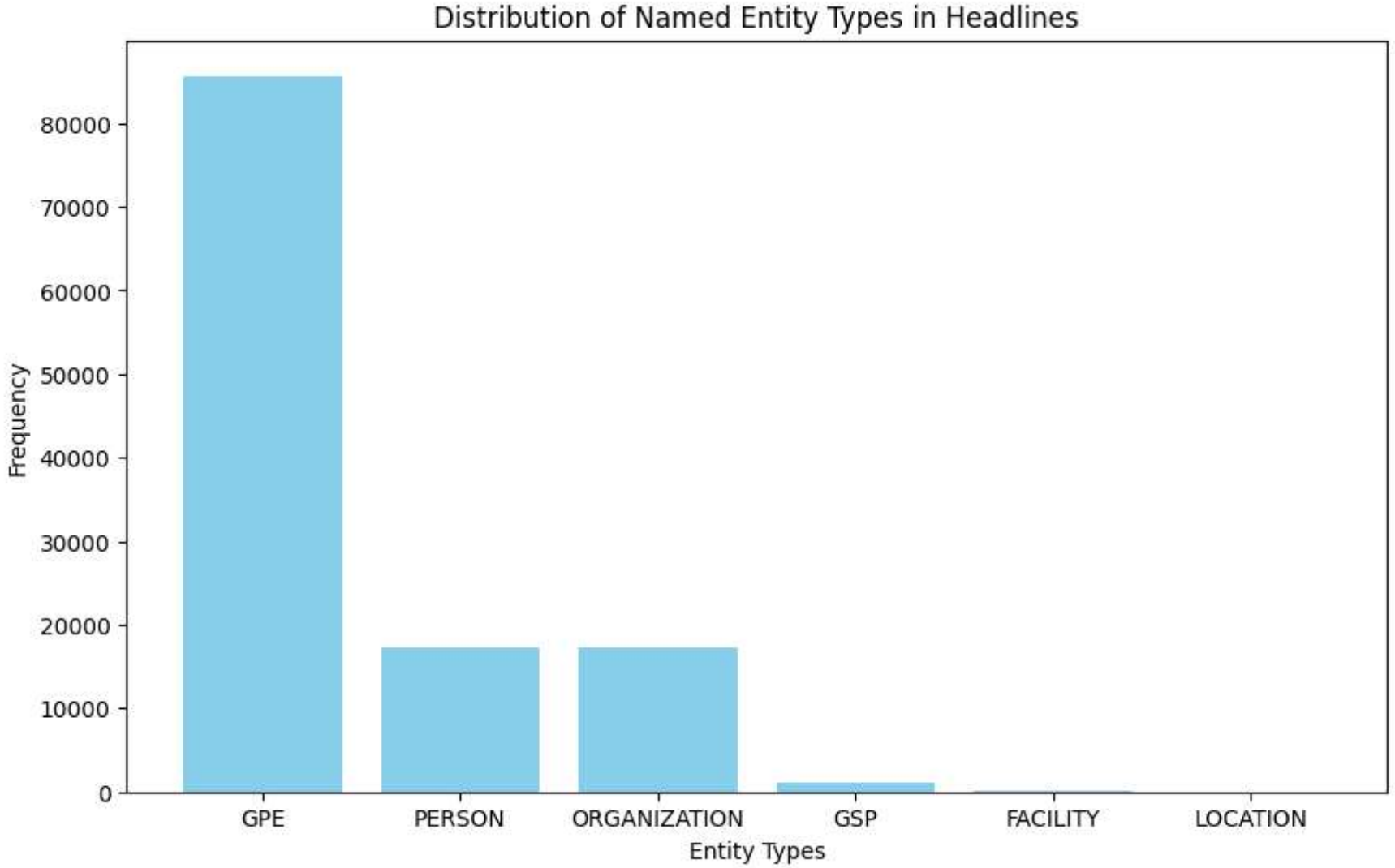
In [44]:
```python
all_entities = []
for headline in df_2023['headline_text']:
    words = word_tokenize(headline)
    tags = pos_tag(words)
    tree = ne_chunk(tags)

    entities = [entity for subtree in tree if isinstance(subtree, nltk.Tree)
                for entity in [" ".join([token[0] for token in subtree.leaves()])]]
    all_entities.extend(entities)

entity_counts = Counter(all_entities)

top_entities = entity_counts.most_common(10)

entities, counts = zip(*top_entities)
plt.figure(figsize=(12, 6))
plt.barh(entities, counts, color='skyblue')
plt.xlabel('Frequency')
```

```
plt.title('Top 10 Named Entities in Headlines')
plt.show()
```



In [46]:
```
entity_types = [entity[1] if len(entity) > 1 else "Unknown" for entity in entity_counts.keys()]
entity_type_counts = Counter(entity_types)
entity_type_counts = Counter()
for entity in all_entities:
    words = word_tokenize(entity)
    tags = pos_tag(words)
    tree = ne_chunk(tags)
    entity_type = [subtree.label() for subtree in tree if isinstance(subtree, nltk.Tree)]
    if entity_type:
        entity_type_counts[entity_type[0]] += 1
```

```
plt.figure(figsize=(10, 6))
plt.bar(entity_type_counts.keys(), entity_type_counts.values(), color='skyblue')
plt.xlabel('Entity Types')
plt.ylabel('Frequency')
plt.title('Distribution of Named Entity Types in Headlines')
plt.show()
```



In [ ]: