In [1]:

```
import re
```

In [2]:

```
pattern=r'\d+'
text="007 is the 008 jersy number of dhoni"
match=re.match(pattern,text)
print(match.group())
```

```
007
```

In [14]:

```
pattern=r'\w+'
text="the 008 jersy number of dhoni"
match=re.match(pattern,text)
print(match)
```

```
<re.Match object; span=(0, 3), match='the'>
```

In [17]:

```
pattern=r'\d+'
text="the 008 jersy number of dhoni"
match=re.search(pattern,text)
print(match.group())
```

```
008
```

In [18]:

```
pattern=r'\d+'
text="the 008 jersy 007 number of dhoni"
match=re.findall(pattern,text)
print(match)
```

```
['008', '007']
```

In [19]:

```
pattern=r'\w+'
text="the 008 jersy 007 number of dhoni"
match=re.findall(pattern,text)
print(match)
```

```
['the', '008', 'jersy', '007', 'number', 'of', 'dhoni']
```

In [20]:

```
import re
pattern=r'\d+'
text="ti like you so much 007"
new=re.sub(pattern,"pooja",text)
new
```

Out[20]:

```
't i like you so much pooja'
```

In [1]:

```
import re
```

In [20]:

```
text = """
Hello world! Contact us at info@example.com or support123@company.org. Foll
Visit <a href="http://example.com">our website</a> for more details. This i
"""
```

In [21]:

```
emails=re.findall(r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]=\.[a-zA-Z]{2,}',text)
print(emails)
```

```
[]
```

In [22]:

```
hashtags=re.findall(r'#\w+',text)
print(hashtags)
```

```
['#AI', '#MachineLearning']
```

In [23]:

```
import re

text = "your_text_containing_emails_here" # Make sure to define the 'text'
emails = re.findall(r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}', text
print(emails)
```

```
[]
```

In [24]:

```
text_no_numbers=re.sub(r'\d+','',text)
text_no_numbers
```

Out[24]:

```
'your_text_containing_emails_here'
```

In [25]:

```
clean_text=re.sub(r"<.*?>",'',text_no_numbers)
print(clean_text)
```

```
your_text_containing_emails_here
```

In [26]:
```
def clean_text(text):
    # Step 1: Extract all email addresses
    emails = re.findall(r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}',



    # Step 2: Extract all hashtags
    hashtags = re.findall(r'#\w+', text)
     # Step 3: Remove all numbers
    text_no_numbers = re.sub(r'\d+', '', text)

    # Step 4: Normalize whitespace (remove extra spaces)
    text_normalized = re.sub(r'\s+', ' ', text_no_numbers).strip()

    # Step 5: Remove any HTML tags
    text_no_html = re
# Step 5: Remove any HTML tags
    text_no_html = re.sub(r'<.*?>', '', text_normalized)

    return {
        "emails": emails,
        "hashtags": hashtags,
        "clean_text": text_no_html
    }

# Test case
text = """
Hello world! Contact us at info@example.com or support123@company.org. Foll
Visit <a href="http://example.com">our website</a> for more details. This i
"""

result = clean_text(text)
result = clean_text(text)
print("Emails Found:", result['emails'])          # Output: ['info@example.
print("Hashtags Found:", result['hashtags'])      # Output: ['#AI', '#Machi
print("Cleaned Text:", result['clean_text'])      # Output: "Hello world! C
```

```
Emails Found: ['info@example.com', 'support123@company.org']
Hashtags Found: ['#AI', '#MachineLearning']
Cleaned Text: Hello world! Contact us at info@example.com or support@compa
ny.org. Follow us on social media: #AI #MachineLearning. Visit our website
for more details. This is a test with number .
```

In [ ]:
```
!pip install module wordCloud
```

In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt

from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

In [28]:
```
dataset = "http://data.insideairbnb.com/canada/on/toronto/2023-03-09/data/l
df = pd.read_csv(dataset)
```

In [2]:
```
!pip install dataset
```

```
Collecting dataset
  Obtaining dependency information for dataset from https://files.pythonho
sted.org/packages/9f/4d/f74a514b5c4efb5c1546160597715cd6096273d7173b36a318
7d2afb663a/dataset-1.6.2-py2.py3-none-any.whl.metadata (https://files.pyth
onhosted.org/packages/9f/4d/f74a514b5c4efb5c1546160597715cd6096273d7173b36
a3187d2afb663a/dataset-1.6.2-py2.py3-none-any.whl.metadata)
  Downloading dataset-1.6.2-py2.py3-none-any.whl.metadata (1.9 kB)
Requirement already satisfied: sqlalchemy<2.0.0,>=1.3.2 in c:\users\pooja
reddy\anaconda3\lib\site-packages (from dataset) (1.4.39)
Collecting alembic>=0.6.2 (from dataset)
  Obtaining dependency information for alembic>=0.6.2 from https://files.p
ythonhosted.org/packages/54/7e/ac0991d1745f7d755fc1cd381b3990a45b404b4d008
fc75e2a983516fbfe/alembic-1.14.1-py3-none-any.whl.metadata (https://files.
pythonhosted.org/packages/54/7e/ac0991d1745f7d755fc1cd381b3990a45b404b4d00
8fc75e2a983516fbfe/alembic-1.14.1-py3-none-any.whl.metadata)
  Downloading alembic-1.14.1-py3-none-any.whl.metadata (7.4 kB)
Collecting banal>=1.0.1 (from dataset)
  Obtaining dependency information for banal>=1.0.1 from https://files.pyt
honhosted.org/packages/ae/c4/7f6e6a539cc6b2da4da3b6a58d5e6f9342c870522ee46
d41f8cbd2156953/banal-1.0.6-py2.py3-none-any.whl.metadata (https://files.p
ythonhosted.org/packages/ae/c4/7f6e6a539cc6b2da4da3b6a58d5e6f9342c870522ee
46d41f8cbd2156953/banal-1.0.6-py2.py3-none-any.whl.metadata)
  Downloading banal-1.0.6-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting Mako (from alembic>=0.6.2->dataset)
  Obtaining dependency information for Mako from https://files.pythonhoste
d.org/packages/1e/bf/7a6a36ce2e4cafdfb202752be68850e22607fccd692847c45c1ae
3c17ba6/Mako-1.3.8-py3-none-any.whl.metadata (https://files.pythonhosted.o
rg/packages/1e/bf/7a6a36ce2e4cafdfb202752be68850e22607fccd692847c45c1ae3c1
7ba6/Mako-1.3.8-py3-none-any.whl.metadata)
  Downloading Mako-1.3.8-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: typing-extensions>=4 in c:\users\pooja redd
y\anaconda3\lib\site-packages (from alembic>=0.6.2->dataset) (4.7.1)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\pooja reddy\an
aconda3\lib\site-packages (from sqlalchemy<2.0.0,>=1.3.2->dataset) (2.0.1)
Requirement already satisfied: MarkupSafe>=0.9.2 in c:\users\pooja reddy\a
naconda3\lib\site-packages (from Mako->alembic>=0.6.2->dataset) (2.1.1)
Downloading dataset-1.6.2-py2.py3-none-any.whl (18 kB)
Downloading alembic-1.14.1-py3-none-any.whl (233 kB)
   ---------------------------------------- 0.0/233.6 kB ? eta -:--:--
   ---------------------------------------- 233.6/233.6 kB 7.2 MB/s eta 0:
00:00
Downloading banal-1.0.6-py2.py3-none-any.whl (6.1 kB)
Downloading Mako-1.3.8-py3-none-any.whl (78 kB)
   ---------------------------------------- 0.0/78.6 kB ? eta -:--:--
   ---------------------------------------- 78.6/78.6 kB ? eta 0:00:00
Installing collected packages: banal, Mako, alembic, dataset
Successfully installed Mako-1.3.8 alembic-1.14.1 banal-1.0.6 dataset-1.6.2
```

In [7]:

```
import pandas as pd
dataset = pd.read_csv('tweets.csv', encoding = 'ISO-8859-1')
dataset.head(3)
```

Out[7]:

	Unnamed: 0	X	text	favorited	favoriteCount	replyToSN	created	truncated
0	1	1	RT @rssurjewala: Critical question: Was PayTM ...	False	0	NaN	2016-11-23 18:40:30	False
1	2	2	RT @Hemant_80: Did you vote on #Demonetization...	False	0	NaN	2016-11-23 18:40:29	False
2	3	3	RT @roshankar: Former FinSec, RBI Dy Governor,...	False	0	NaN	2016-11-23 18:40:03	False

In [8]:

```
dataset.shape
```

Out[8]:

```
(14940, 16)
```

In [9]:

```
pd.read_csv?
```

In [11]:

```
def gen_freq(text):
    word_list = []
    for tw_words in text.split():
        word_list.extend(tw_words)
    word_freq = pd.Series(word_list).value_counts()
    word_freq[:10]
    return word_freq
```

In [12]:

```
word_freq = gen_freq(dataset.text.str)
word_freq
```

Out[12]:

```
RT                           11053
to                            7650
is                            5152
in                            4491
the                           4331
                             ...
#News                            1
notes|                           1
https://t.co/ECl4oIzdHA (https://t.co/ECl4oIzdHA)        1
https://t.co/9MjFtLtCtR (https://t.co/9MjFtLtCtR)        1
https://t.co/hwgqjbqgvG (https://t.co/hwgqjbqgvG)        1
Length: 19601, dtype: int64
```

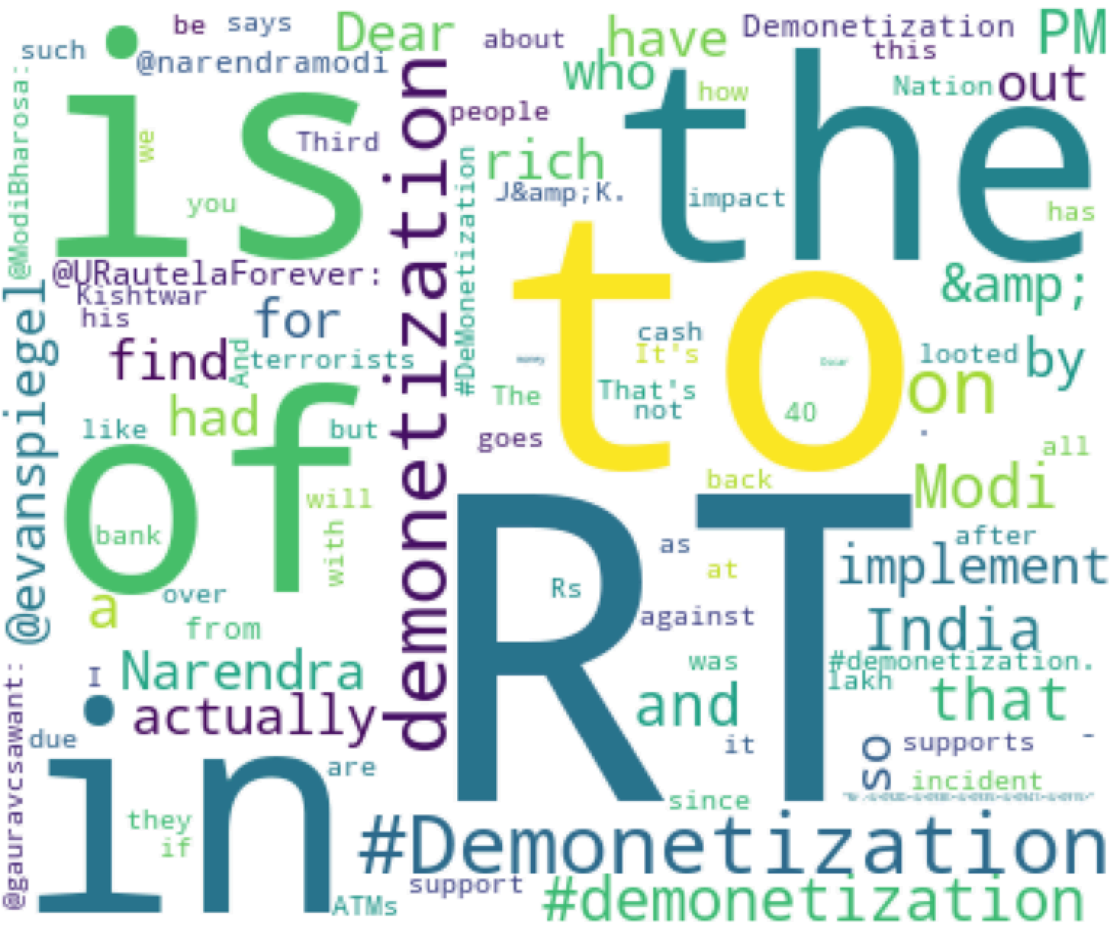
In [14]:

```
!pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\users\pooja reddy\anaconda3
\lib\site-packages (1.9.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\pooja reddy\anacon
da3\lib\site-packages (from wordcloud) (1.24.3)
Requirement already satisfied: pillow in c:\users\pooja reddy\anaconda3\li
b\site-packages (from wordcloud) (9.4.0)
Requirement already satisfied: matplotlib in c:\users\pooja reddy\anaconda
3\lib\site-packages (from wordcloud) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\pooja reddy\an
aconda3\lib\site-packages (from matplotlib->wordcloud) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\pooja reddy\anacon
da3\lib\site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\pooja reddy\a
naconda3\lib\site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\pooja reddy\a
naconda3\lib\site-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\pooja reddy\ana
conda3\lib\site-packages (from matplotlib->wordcloud) (23.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\pooja reddy\an
aconda3\lib\site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\pooja redd
y\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\pooja reddy\anaconda3
\lib\site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.1
6.0)
```

In [18]:

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud
wc = WordCloud(width=400, height=330, max_words=200,background_color='white
plt.figure(figsize=(12, 8))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



In [19]:

```
import re
def clean_text(text):
    text = re.sub(r'RT', '', text)
    text = re.sub(r'&amp;', '', text)
    text = re.sub(r'[?!.;:,#@-]', '', text)
    text = text.lower()
    text = re.sub(r'\d+', '', text)
    text = re.sub(r'%', '', text)
    return text
```

In [20]:
```
from wordcloud import STOPWORDS
print(STOPWORDS)
```

```
{"he'll", "aren't", "they'd", 'to', 'there', 'about', 'through', 'too', 'w
hom', 'also', 'like', 'had', "won't", "they'll", 'his', 'further', 'only',
"why's", "she's", 'are', 'the', "i've", 'from', "i'll", 'would', 'should',
'yourself', 'with', 'but', 'itself', "she'd", "hasn't", 'all', 'hence', "h
e's", "let's", 'yourselves', 'at', 'could', 'until', 'do', 'our', 'them',
"they're", "where's", 'my', "shouldn't", 'a', 'those', 'themselves', "woul
dn't", 'we', "couldn't", "hadn't", 'very', 'its', "when's", 'she', 'out',
"we're", 'other', 'com', 'each', 'and', 'both', 'hers', "it's", 'me', 'ove
r', 'some', 'yours', 'just', "who's", 'else', 'be', 'below', 'an', 'shal
l', 'because', 'or', 'where', 'under', 'ought', "we'd", 'which', 'while',
'down', 'however', 'her', 'him', 'k', 'ourselves', "can't", "don't", 'he',
'how', 'ours', "wasn't", 'when', 'once', "she'll", 'r', "haven't", 'their
s', 'before', 'after', 'any', 'why', 'as', "what's", 'what', 'doing', 'int
o', "i'm", 'of', "here's", 'it', 'above', 'than', 'have', 'is', 'therefor
e', 'again', "weren't", 'am', 'has', "i'd", "isn't", 'such', 'here', 'no
t', 'against', 'having', 'you', "they've", "we've", 'few', 'nor', 'ever',
"mustn't", 'in', "how's", 'were', 'between', 'was', "shan't", "he'd", "the
re's", 'that', 'these', 'own', "you'd", 'more', 'they', 'on', 'their', 'of
f', 'most', 'your', "didn't", 'up', "you're", 'for', "we'll", 'i', 'been',
'this', "you'll", 'himself', 'myself', 'can', 'did', 'then', 'get', 'www',
"that's", 'same', 'since', 'herself', 'by', 'otherwise', 'during', 'http',
'no', 'so', 'being', 'who', 'does', 'cannot', "you've", 'if', "doesn't"}
```

In [ ]:
```
text = dataset.text.apply(lambda x: clean_text(x))
word_freq = gen_freq(text.str)
word_freq = word
```

In [1]:
```
import pandas as pd

text = ['Sarah lives in a hut in the village.',
        'She has an apple tree in her backyard.',
        'The apples are red in colour.']

df = pd.DataFrame(text, columns=['Sentence'])

df
```

Out[1]:

	Sentence
0	Sarah lives in a hut in the village.
1	She has an apple tree in her backyard.
2	The apples are red in colour.

In [ ]: