Lab-2.1: Assignment

`Gathering text data with an API`

IMPORTANT: The lab shown here (on the website) is just an HTML version, included for reference. To download the assignment, please navigate to the Labs tab in the Share-point dropdown menu in the website's navigation bar. The relevant assignment can be determined from the folder's name, click on the three dots & select download to get the assignment.

NOTE: It is recommended that you complete this `.ipynb` file in VS-code.

`Submission:`

- Export the completed assignment to HTML or PDF (preferably with Quarto) and upload it to Canvas.
- The final uploaded version should NOT have any code-errors present
- All outputs must be visible in the uploaded version, including code-cell outputs, images, graphs, etc

Assignment-1:

- Read over the `News-API`, `Wikipedia API`, and `Google Scholar` API sections in the lab-demonstration section, if you have not done so already.
  - https://jfh.georgetown.domains/dsan5000/
- Get an API key for the `News-API`: see following link
- **Submission**: Insert your API key below

```
API_KEY='6f3d3278d4fe4915b3cb9a76f4f47ccc'
```

Assignment-2:

- Use the provided `News-API` code as a starting point
- Select `THREE` random topics (e.g. Georgetown, Cats, Clouds) but choose whatever you like
- Query the API to pull text data and store the results in three different dictionaries
- Extract the `title` and `description` text and store for later processing (up to you how you do this)
- Clean the text as needed

```
import requests
import json
import re
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
```

```
baseURL = "https://newsapi.org/v2/everything?"
total_requests=2
verbose=True

# THIS CODE WILL NOT WORK UNLESS YOU INSERT YOUR API KEY IN THE NEXT LINE
API_KEY='6f3d3278d4fe4915b3cb9a76f4f47ccc'
TOPIC=['Premier_League','South_Korea','North_Korea']
```

```
for i in TOPIC:
    j = 0
    URLpost = {'apiKey': API_KEY,
                'q': '+'+ i,
                'sortBy': 'relevancy',
                'totalRequests': 1}

    #print(baseURL)
    # print(URLpost)

    #GET DATA FROM API
    response = requests.get(baseURL, URLpost) #request data from the server
    # print(response.url);
    exec(f'article_{i} = response.json()') #extract txt data from request into json
```

```
def string_cleaner(input_string):
    try:
        out=re.sub(r"""
                    [,.;@#?!&$-]+  # Accept one or more copies of punctuation
                    \ *           # plus zero or more copies of a space,
                    """,
                    " ",          # and replace it with a single space
                    input_string, flags=re.VERBOSE)

        #REPLACE SELECT CHARACTERS WITH NOTHING
        out = re.sub('[’.]+', '', input_string)

        #ELIMINATE DUPLICATE WHITESPACES USING WILDCARDS
        out = re.sub(r'\s+', ' ', out)

        #CONVERT TO LOWER CASE
        out=out.lower()
    except:
        print("ERROR")
        out=''
    return out
```

```
def td_list(dictionary):

    article_list = []

    for i in range(len(dictionary['articles'])):
        words = string_cleaner(dictionary['articles'][i]['title']+dictionary['articles'][i

        article_list.append(words)

    return article_list
```

```
def td_dict(dictionary):

    article_dictionary = {}

    for i in range(len(dictionary['articles'])):
        title = string_cleaner(dictionary['articles'][i]['title'])
        description = string_cleaner(dictionary['articles'][i]['description'])
        article_dictionary.update({title:description})

    return article_dictionary
```

```
a = td_list(article_North_Korea)
b = td_list(article_South_Korea)
c = td_list(article_Premier_League)
```

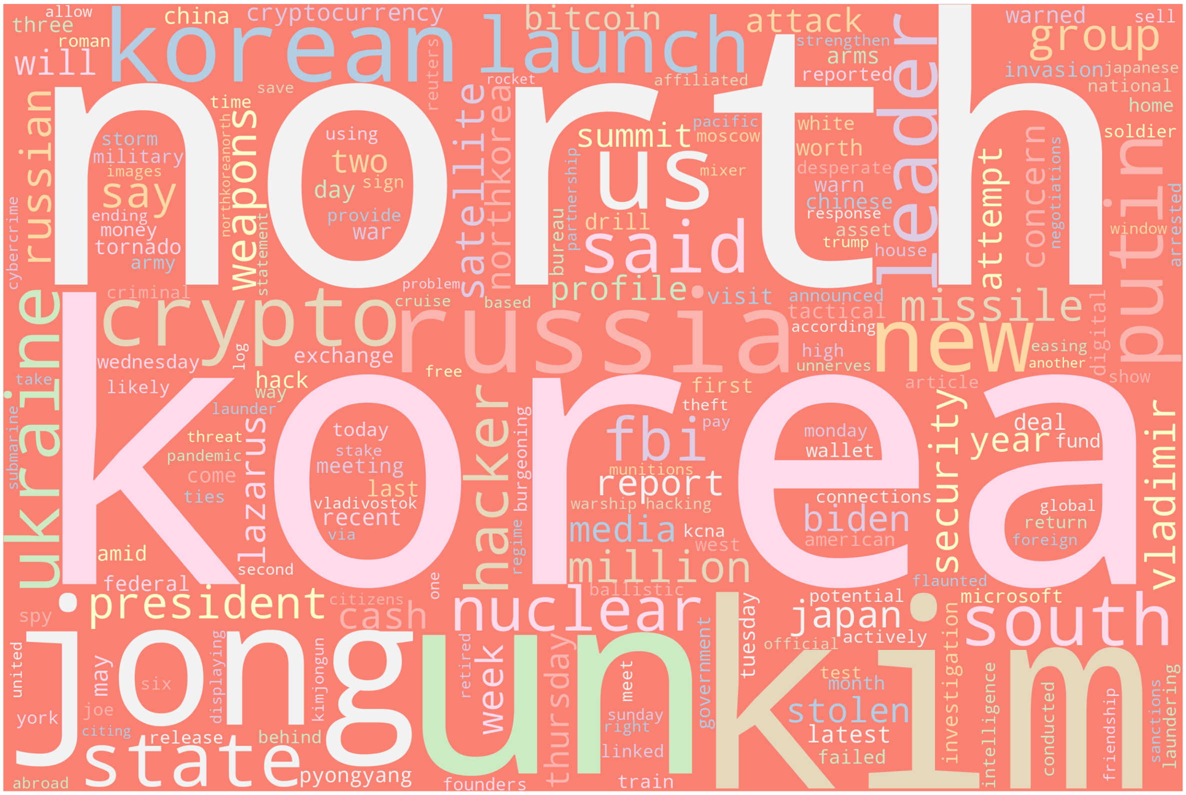
Assignment-3:

- Use the provided `Wikipedia-API` code as a starting point
- For EACH `THREE` of the random topics, create a word cloud for your cleaned `title` and `description` text

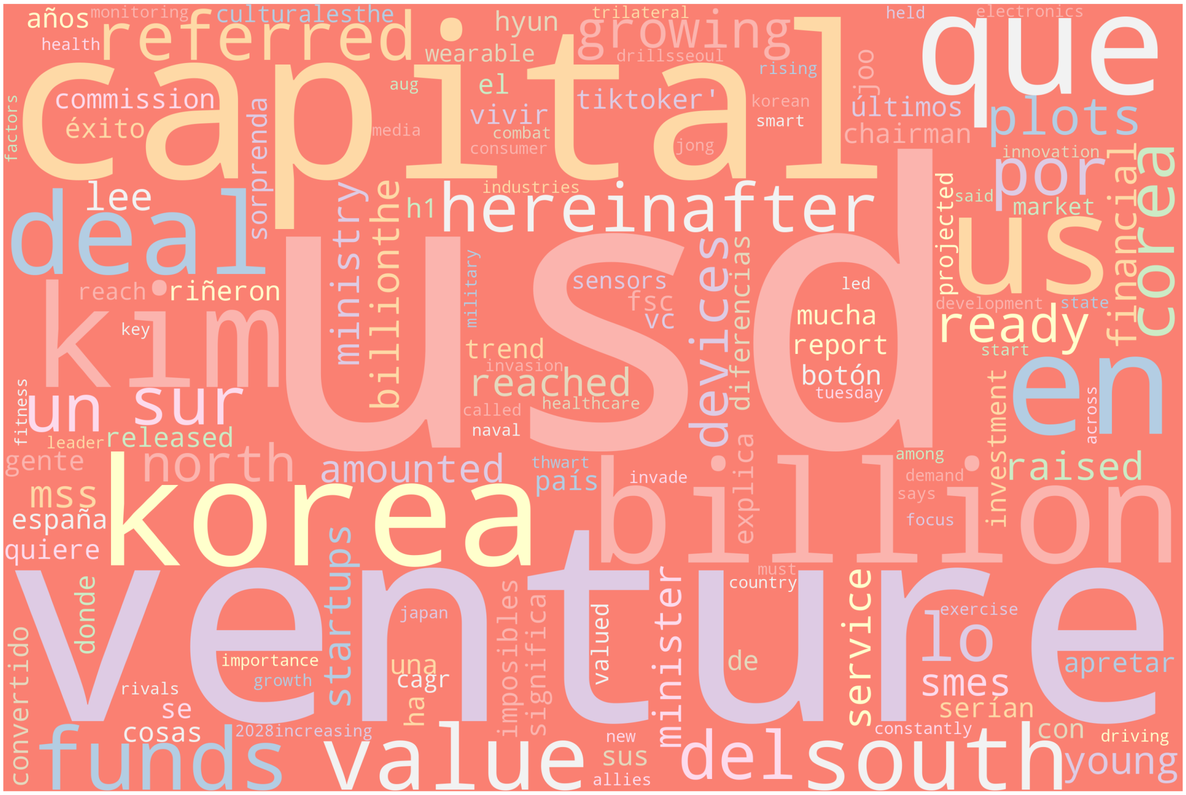
```
def generate_word_cloud(my_text):
    from wordcloud import WordCloud, STOPWORDS
    import matplotlib.pyplot as plt
    # exit()
    # Import package
    # Define a function to plot word cloud
    def plot_cloud(wordcloud):
        # Set figure size
        plt.figure(figsize=(40, 30))
        # Display image
        plt.imshow(wordcloud)
        # No axis details
        plt.axis("off");

    # Generate word cloud
    wordcloud = WordCloud(
        width = 3000,
        height = 2000,
        random_state=1,
        background_color='salmon',
        colormap='Pastel1',
        collocations=False,
        stopwords = STOPWORDS).generate(my_text)
    plot_cloud(wordcloud)
    plt.show()
```

```
North_Korea = str()
for key in a:
    North_Korea += key

South_Korea = str()
for key in b:
    South_Korea += key

Premier_League = str()
for key in c:
    Premier_League += key
```

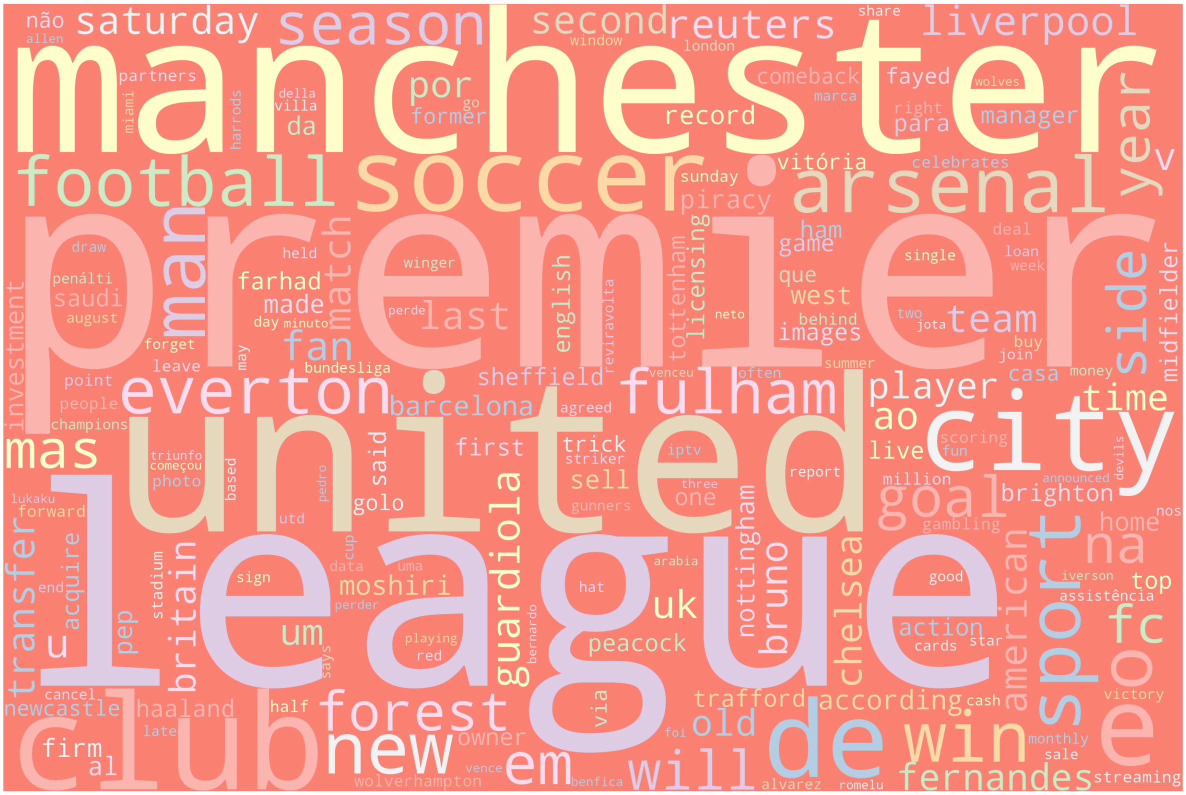
```
generate_word_cloud(North_Korea)
```



```
generate_word_cloud(South_Korea)
```



```
generate_word_cloud(Premier_League)
```


não
saturday
season
second
reuters
share
liverpool
allen
window
london
manchester
partners
della
villa
miami
harrods
go
por
former
da
comeback
fayed
wolves
marca
right
para
manager
record
year
football
soccer
vitória
celebrates
sunday
arsenal
v
piracy
premier
draw
held
ham
game
deal
single
loan
week
man
match
winger
penálti
farhad
saudi
august
made
day
minuto
perde
last
english
reviravolta
tottenham
licensing
que
west
behind
images
neto
two
jota
team
side
midfielder
investment
fan
forget
leave
may
bundesliga
buy
join
summer
point
everton
sheffield
venceu
often
casa
money
people
player
time
champions
barcelona
agreed
fulham
ao
united
city
mas
first
trick
striker
iptv
scoring
fun
live
triunfo
começou
photo
said
based
pedro
report
devils
sell
three
one
million
announced
lukaku
forward
golo
gunners
goal
brighton
nos
utd
guardiola
nottingham
gambling
home
league
data
uma
arabia
end
acquire
sign
moshiri
hat
good
iverson
top
stadium
transfer
cup
perder
na
american
assistência
pep
britain
uk
bruno
chelsea
sport
fc
um
playing
bernardo
peacock
action
cards
star
u
red
says
o
cancel
newcastle
haaland
half
forest
trafford
according
cash
victory
late
via
e
club
firm
al
new
owner
old
foi
de
win
monthly
sale
vence
em
will
benfica
alvarez
romelu
wolverhampton
fernandes
streaming