Chapter 6 - Exercise 5: WordClouds

Part 1:

```
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS


# Tạo dataframe df từ file dữ liệu Youtube04-Eminem.csv, in vài dòng dữ liệu đầu của df
df = pd.read_csv("data\Youtube04-Eminem.csv", encoding ="latin-1")
df.head()
df['CONTENT'].head(20)
```

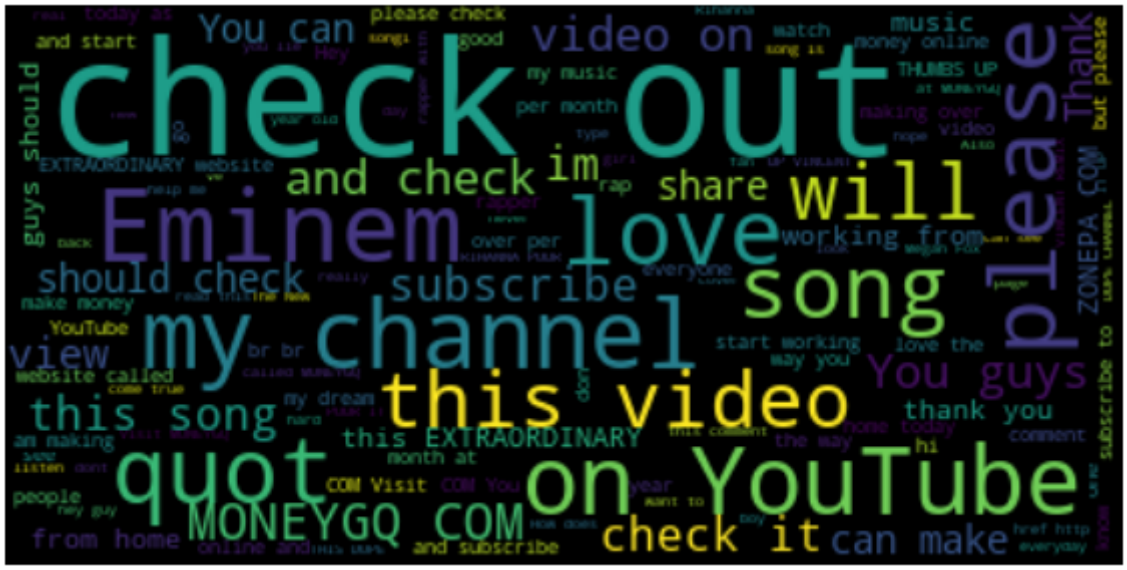
```
0           +447935454150 lovely girl talk to me xxxï»¿
1     I always end up coming back to this song<br />ï»¿
2     my sister just received over 6,500 new <a rel=...
3                                                Cooliï»¿
4                     Hello I&#39;am from Palastineiï»¿
5     Wow this video almost has a billion views! Did...
6     Go check out my rapping video called Four Whee...
7                                  Almost 1 billionï»¿
8                     Aslamu Lykum... From Pakistaniï»¿
9     Eminem is idol for very people in EspaÃ±a and ...
10                      Help me get 50 subs please ï»¿
11                                  i love song :)ï»¿
12    Alright ladies, if you like this song, then ch...
13    The perfect example of abuse from husbands and...
14    The boyfriend was Charlie from the TV show LOS...
15    <a href="https://www.facebook.com/groups/10087...
16             Take a look at this video on YouTube:ï»¿
17            Check out our Channel for nice Beats!!ï»¿
18      Rihanna and Eminem together are unstoppable.ï»¿
19               Check out this playlist on YouTube:ï»¿
Name: CONTENT, dtype: object
```

```
# Hãy bỏ các STOPWORD
stopwords = set(STOPWORDS)
stopwords
```

```
{'a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'all',
 'also',
 'am',
 'an',
 'and',
 'any',
 'are',
```

```
      "aren't",
      'as',
      'at',
      'be',
      'because',
      'been',
      'before',
      'being',
      'below',
      'between',
      'both',
      'but',
      'by',
      'can',
      "can't",
      'cannot',
      'com',
      'could',
      "couldn't",
      'did',
      "didn't",
      'do',
      'does',
      "doesn't",
      'doing',
      "don't",
      'down',
      'during',
      'each',
      'else',
      'ever',
      'few',
      'for',
      'from',
      'further',
      'get',
      'had',
      "hadn't",
      'has',
      "hasn't",
      'have',
      "haven't",
      'having',
      'he',
      "he'd",
      "he'll",

comment_words =' '.join(df['CONTENT'])


# Vẽ WordClouds cho dữ liệu này
wc = WordCloud(
    background_color='black',
    stopwords=stopwords
)

# generate the word cloud
wc.generate(comment_words)

# display the word clouds
```

```
plt.figure(figsize=(10, 12))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Part 2:

```
# Sử dụng vietnamese stop word từ : https://github.com/stopwords/vietnamese-stopwords => t
stopwords = set()
f = open(r"data\vietnamese-stopwords.txt", "r", encoding='utf-8')
for line in f:
    word = f.readline()
    stopwords.add(word.replace('\n',''))
f.close()

stopwords
```

```
{'a ha',
 'ai ai',
 'ai đó',
 'amen',
 'anh ấy',
 'ba ba',
 'ba cùng',
 'ba ngày',
 'ba tăng',
 'bao lâu',
 'bao nả',
 'biết',
 'biết bao nhiêu',
 'biết chừng nào',
 'biết mấy',
 'biết trước',
 'biết đâu',
 'biết đâu đấy',
 'buổi',
 'buổi mới',
 'buổi sớm',
```

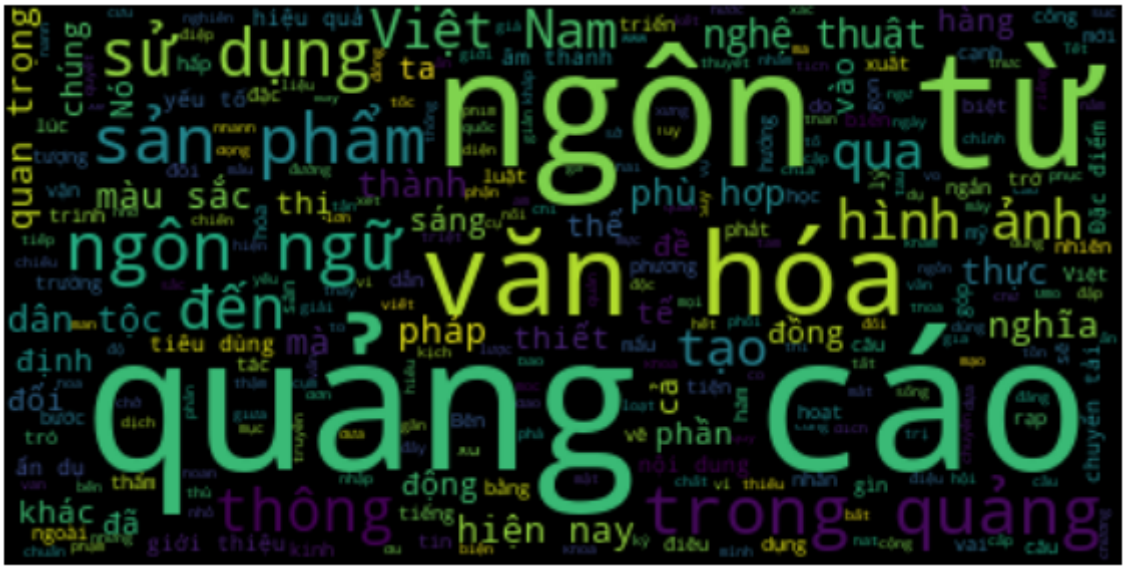
```
      'bà ấy',
      'bài bác',
      'bài cái',
      'bán',
      'bán dạ',
      'bây bẩy',
      'bây giờ',
      'bèn',
      'bên',
      'bên có',
      'bông',
      'bước khỏi',
      'bước đi',
      'bản',
      'bản riêng',
      'bản ý',
      'bất cứ',
      'bất kì',
      'bất kỳ',
      'bất ngờ',
      'bất quá',
      'bất thình lình',
      'bất đồ',
      'bấy chầy',
      'bấy giờ',
      'bấy lâu nay',
      'bấy nhiêu',
      'bập bõm',
      'bắt đầu từ',
      'bằng cứ',
      'bằng người',
      'bằng như',
      'bằng nấy',
      'bằng được',
      'bển',
      'bị',
      'bị vì',
      'bỏ bà',
```

```
# Đọc file ngon_tu_quang_cao.txt => đưa nội dung vào biến text

text = ''
with open(r'data\ngon_tu_quang_cao.txt', 'r', encoding='utf-8') as f:
    text = f.read()
text
```

```
'\ufeffNGÔN TỪ TRONG QUẢNG CÁO\nGiới thiệu\nViệt Nam đang từng bước hội nhập vào nền
```

```
# Bổ sung thêm một số từ không quan trọng vào stopwords
list_of_words = ['và', 'một', 'của', 'có', 'đó', 'rất', 'nào', 'được',
                 'khi', 'thể', 'sự', 'tính', 'trong','cũng','cùng','cho','hay','chỉ']
for word in list_of_words:
    stopwords.add(word)


# Vẽ wordclouds
wc = WordCloud(
```

```
    background_color='black',
    max_words=1000,
    stopwords=stopwords
)

wc.generate(text)

plt.figure(figsize=(10, 12))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```




```
import numpy as np
from PIL import Image


# Chọn hình làm wc_mask phù hợp
wc_mask = np.array(Image.open('ad_s.png'))

plt.imshow(wc_mask, interpolation='bilinear')
plt.axis('off')
plt.show()
```

```
# Vẽ wordclouds với wc_mask
```

```
wc1 = WordCloud(background_color='white', max_words=1000,
                mask=wc_mask, stopwords=stopwords)

wc1.generate(text)

plt.figure(figsize=(10, 12))
plt.imshow(wc1, interpolation='bilinear')
plt.axis('off')
plt.show()
```