Chapter 6 - Ex 6: WordClouds

Part 1:

- Cho dữ liệu Youtube04-Eminem.csv là dữ liệu được lấy từ UCI Machine Learning Repository, trong đó chứa các YouTube comment cho các video của các nghệ sỹ nổi tiếng.
- Hãy bỏ các STOPWORD
- Vẽ WordClouds cho dữ liệu này

Part 2:

- Sử dụng vietnamese stop word từ : https://github.com/stopwords/vietnamese-stopwords (https://github.com/stopwords/vietnamese-stopwords) => tạo thành set các stop words
- Đọc file ngon_tu_quang_cao.txt => đưa nội dung vào biến text
- Bổ sung thêm một số từ không quan trọng vào stopwords
- Vẽ wordclouds
- Chọn hình làm wc_mask phù hợp => vẽ wordclouds với wc_mask

Part 1:

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
```

In [2]:
```
# Reads 'Youtube04-Eminem.csv' file
df = pd.read_csv(r"youtube/Youtube04-Eminem.csv", encoding ="latin-1")
df.head()
```

Out[2]:

	COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
0	z12rwfnyyrbsefonb232i5ehdxzkjzjs2	Lisa Wellas	NaN	+447935454150 lovely girl talk to me xxxï»¿	1
1	z130wpnwwnyuetxcn23xf5k5ynmkdpjrj04	jason graham	2015-05-29T02:26:10.652000	I always end up coming back to this song ï»¿	0
2	z13vsfqirtavjvu0t22ezrgzyorwxhpf3	Ajkal Khan	NaN	my sister just received over 6,500 new <a rel=...	1
3	z12wjzc4eprnvja4304cgbbizuved35wxcs	Dakota Taylor	2015-05-29T02:13:07.810000	Coolï»¿	0
4	z13xjfr42z3uxdz2223gx5rrzs3dt5hna	Jihad Naser	NaN	Hello I'am from Palestineï»¿	1

In [3]:
```
comment_words = ' '
stopwords = set(STOPWORDS)
#stopwords
```

In [4]:
```
# iterate through the csv file to get content
for content in df.CONTENT:
    # convert content to string
    content = str(content)

    # split the value to list
    elements = content.split()

    # Converts each element into lowercase
    for i in range(len(elements)):
        elements[i] =  elements[i].lower()
    # make content words from elements
    for words in elements:
        comment_words = comment_words + words + ' '
```

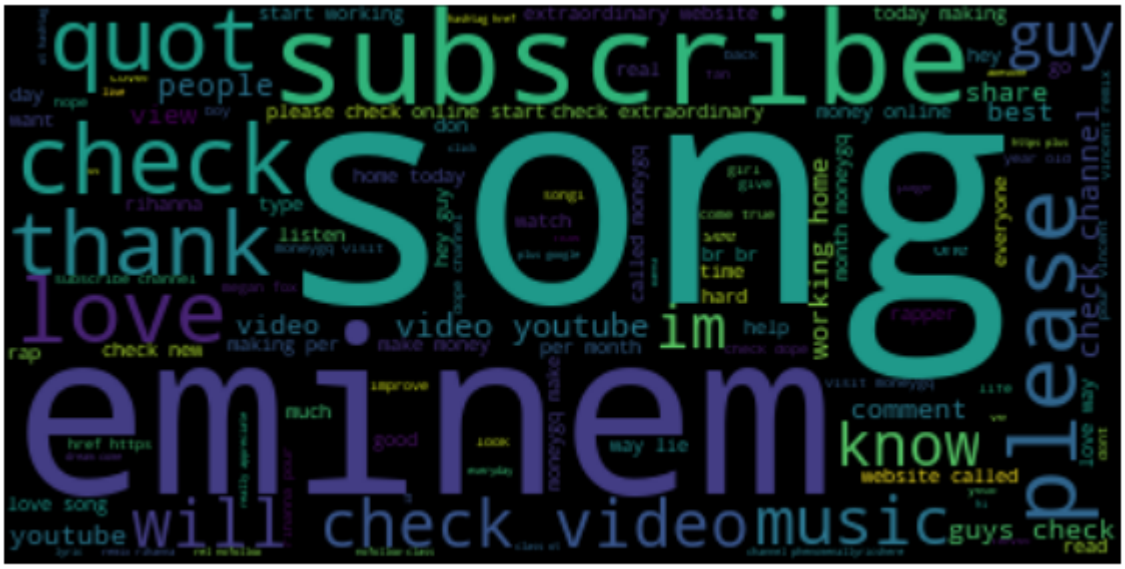
In [5]:
```
# comment_words
```

In [6]:
```
# instantiate a word cloud object
wc = WordCloud(
    background_color='black',
    max_words=1000,
    stopwords=stopwords
)

# generate the word cloud
wc.generate(comment_words)
```

Out[6]:
```
<wordcloud.wordcloud.WordCloud at 0x12d3e8f5fd0>
```

In [7]:
```
# 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 (https://github.com/stopwords/vietnamese-stopwords) => tạo thành set các stop words
- Đọc file ngon_tu_quang_cao.txt => đưa nội dung vào biến text
- Bổ sung thêm một số từ không quan trọng vào stopwords
- Vẽ wordclouds
- Chọn hỉnh làm wc_mask phù hợp => vẽ wordclouds với wc_mask

In [8]:
```
stopwords = set()
f = open("vietnamese-stopwords.txt", "r", encoding='utf-8')
for line in f:
    word = f.readline()
    stopwords.add(word.replace('\n',''))
f.close()
```

In [9]:
```
# stopwords
```

In [10]:
```
text = ""
f = open("ngon_tu_quang_cao.txt", "r", encoding='utf-8')
for line in f:
     # convert content to string
    line = str(f.readline())
    content = line.replace('\n','')

    # split the value to list
    elements = content.split()

    # Converts each element into lowercase
    for i in range(len(elements)):
        elements[i] =  elements[i].lower()
    # make content words from elements
    for words in elements:
        text = text + words + ' '
f.close()
```

In [11]:
```
text
```

Out[11]:
```
'giới thiệu hiện nay, đâu đâu, khi ở nhà hay ra ngoài đường, trong rạp hát, rạp
chiếu phim, bến xe, ga tàu, trường học… đều đập vào mắt mọi người những hình ản
h và thông tin quảng cáo đủ loại và đủ màu sắc của nhiều mặt hàng sản phẩm. bên
cạnh những yếu tố hình ảnh, màu sắc, âm thanh, thì yếu tố ngôn từ cũng cần được
quan tâm chú ý nhiều hơn và việc sử dụng ngôn từ như thế nào cho phù hợp với tí
nh dân tộc việt nam, có tính hấp dẫn, hiệu quả đã trở thành một nghệ thuật thực
sự. thế nhưng, công tác quản lý dịch vụ này vẫn còn rất nhiều điều bất cập, các
vấn đề thiết yếu về ngôn ngữ, văn hóa và pháp lý chưa phù hợp với văn hóa dân t
ộc việt nam. chính vì vậy, hơn lúc nào hết, việc nghiên cứu ngôn ngữ quảng cáo
là rất cần thiết trong các phương tiện quảng cáo hiện nay. quảng cáo chính là m
ột trong những cầu nối quan trọng giữa nhà sản xuất và người tiêu dùng. thông q
ua những chất liệu khác nhau như hình ảnh, màu sắc, âm thanh, ngôn từ… của quản
g cáo, nhà sản xuất giới thiệu sản phẩm của mình đến với khác hàng. ở đây, “ngô
n từ quảng cáo”, hiểu theo nghĩa cụ thể, đó là tất cả các ký tự chữ viết có nội
dung được thể hiện trên một mẩu quảng cáo theo một kịch bản hoặc một chiến lược
quảng cáo nhất định, nhằm đạt đến một hay nhiều mục đích do nhà quảng cáo đề r
a. đặc điểm của ngôn từ quảng cáo là ngắn gọn, dễ nhớ, tạo ấn tượng và phải phù
hợp với kết cấu ngôn ngữ. với tiếng việt, ngoài việc tận dụng những từ đồng âm,
đồng nghĩa, thanh điệu, vần… còn xét về tính ẩn dụ, thậm xưng, nhân hóa… nó khô
ng nhất thiết phải được cấu tạo thành câu hoàn chỉnh và có nhiều câu như đối vớ
i ngôn từ của một văn bản hành chính hay khoa học, mà thường chỉ là những cụm t
ừ ngắn gọn, có nội dung cô đọng, hàm súc và đôi khi được thẩm mỹ hóa bằng các b
iện pháp nghệ thuật và đặc biệt phải phù hợp văn hóa việt nam ví dụ như “tết là
m điều hay, vận may nhân khắp” trong quảng cáo sản phẩm “omo”. như vậy, qua việ
c phân tích ngôn từ trong những mẫu quảng cáo trên, chúng ta thấy được phần nào
những cách thức sử dụng, vai trò và ý nghĩa của ngôn từ trong việc quảng cáo hi
ện nay, cũng như các đặc điểm, xu hướng của việc sử dụng ngôn từ trong quảng cá
o. ngôn ngữ cũng có những quy luật phát triển riêng. bản thân ngôn ngữ trong qu
á trình tồn tại luôn có những biến thể, cách sử dụng ngôn ngữ mới. tuy nhiên, n
hững biến thể đó trong ngôn ngữ cộng đồng cần được đặt trong phạm vi văn hóa dâ
n tộc truyền thống. ngôn từ là một bộ phận, một thành tố rất quan trọng và gần
như không thể thiếu được trong quảng cáo. nó đóng vai trò to lớn có tính quyết
định đến sự hiệu quả và thành công của một chương trình quảng cáo. việc sáng tạ
o ngôn từ quảng cáo vì vậy trở nên vô cùng quan trọng và được xem như một hoạt
động nghệ thuật thực sự. đã đến lúc chúng ta cần có những chuẩn mực trong văn h
óa sử dụng ngôn từ. chỉ có như vậy, mới gìn giữ được sự trong sáng của tiếng vi
ệt, gìn giữ được bản sắc văn hóa của dân tộc. suy cho cùng, những sản phẩm quản
g cáo không chỉ là để giới thiệu sản phẩm đến người tiêu dùng mà thông qua đó c
huyển tải ý nghĩa, thông điệp và đôi khi chuyên chở cả một giá trị văn hóa của
một quốc gia. '
```

In [12]:
```
list_of_words = ['và', 'một', 'của', 'có', 'đó', 'rất', 'nào', 'được',
                 'khi', 'thể', 'sự', 'tính', 'trong']
for word in list_of_words:
    stopwords.add(word) # add the less important word to stopwords
# instantiate a word cloud object
wc = WordCloud(
    background_color='black',
    max_words=1000,
    stopwords=stopwords
)

# generate the word cloud
wc.generate(text)
```

Out[12]:
```
<wordcloud.wordcloud.WordCloud at 0x12d3e9ae518>
```

In [13]:
```
# display the word clouds
plt.figure(figsize=(10, 12))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [14]:
```
import numpy as np
from PIL import Image
```

In [15]:
```
# save mask to wc_mask
wc_mask = np.array(Image.open('ad_s.png'))
```

In [16]:
```
plt.imshow(wc_mask, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [17]:
```
# instantiate a word cloud object
wc1 = WordCloud(background_color='white', max_words=1000,
                mask=wc_mask, stopwords=stopwords)
# generate the word cloud
wc1.generate(text)
# display the word cloud
plt.figure(figsize=(10, 12))
plt.imshow(wc1, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [ ]:
```
```