Chapter 7 - Ex1: NLP - TripAdvisor Review

In [ ]:
```
# from google.colab import drive
# drive.mount("/content/gdrive", force_remount=True)
# %cd '/content/gdrive/My Drive/MDS5_2022/Practice_2022/Chapter7/'
```

```
Mounted at /content/gdrive
/content/gdrive/My Drive/MDS5_2022/Practice_2022/Chapter7
```

In [ ]:
```
import pandas as pd
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
import matplotlib.pyplot as plt
```

* Dữ liệu đọc ra từ file 'review_full_text_tripadvisor.xlsx' đã được tiền xử lý.
* Bạn hãy drop tất cả các cột kết quả sau đó tự làm phần tiền xử lý liệt kê dưới đây:

- Ráp cột title + full_content => thành cột mới **title_content**
- Từ cột rating tạo cột **rating_New** theo số phía sau: vd bubble_50 -> 5
- Từ cột rating_new => tạo cột **label** theo tiêu chí >=4: like, <4: not_like/ hoặc theo tiêu chí: <=2: not_like, 3: neutral, >=4: like
- Từ cột title_content -> tạo cột **text** theo các bước đã được hướng dẫn trong phần **Tiền xử lý dữ liệu tiếng Việt** để có dữ liệu xử lý (có thể bổ sung, hiệu chỉnh cho phù hợp với bộ dữ liệu này)
- Dùng wordcloud để trực quan hóa dữ liệu 'text' theo từng loại (like/not_like...)

Chú ý: Các function cần thiết cho việc tiền xử lý dữ liệu Tiếng Việt nên để vào một file Viet_lib.py để gọi sử dụng khi cần

In [ ]:
```
df = pd.read_excel('review_full_text_tripadvisor.xlsx')
df.head(2)
```

Out[3]:

	hotel_name	customer_name	title	full_content	rating	rating_New	label	title_cont
0	Hotel des Arts Saigon Mgallery	Anh Tuấn L	Quá Tuyệt Vời Khi Ở Des Arts Sài Gòn	#HôtelDesArtsSaiGon là một sự trải nghiệm tuyệ...	bubble_50	5	like	Quá Tu Vời Kh Des Arts G #HôtelD
1	Hotel des Arts Saigon Mgallery	TRƯƠNG BẰNG	Đáng đồng tiền!	Dịch vụ cao cấp, phong cách chuyên nghiệp & tậ...	bubble_50	5	like	Đáng đ tiền!. Dịch cao c phong c

In [ ]:
```
df.shape
```

Out[4]:
```
(78319, 9)
```

In [ ]:
```
# Datasub
df_sub = df[['text', 'label']]
```

In [ ]:
```
df_sub.head(2)
```

Out[6]:

	text	label
0	tuyệt_vời trải_nghiệm tuyệt_vời ghé tươi thích...	like
1	đồng_tiền chuyên_nghiệp hơi thích_hợp chống tr...	like

In [ ]:
```
# kiểm tra dữ liệu na/null
df_sub.isna().sum()
```

Out[7]:
```
text     0
label    0
dtype: int64
```

In [ ]:
```
df_sub.isnull().sum()
```

Out[8]:
```
text     0
label    0
dtype: int64
```

In [ ]:
```
# xóa dữ liệu trùng
df_sub = df_sub.drop_duplicates()
```

In [ ]:
```
df_sub.shape
```

Out[10]:
```
(78183, 2)
```

In [ ]:
```
# không có dữ liệu na/null
# có dữ liệu trùng
```

In [ ]:
```
df_sub.label.value_counts()
```

Out[12]:
```
like        66848
not_like    11335
Name: label, dtype: int64
```

In [ ]:
```
# Tỉ lệ like vs not_like: 6:1
```

In [ ]:
```
y_class = {'like':1, 'not_like':0}
df_sub['y'] = [y_class[i] for i in df_sub.label]
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopy
Warning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
```

In [ ]:
```
df_sub.tail(10)
```

Out[15]:

	text	label	y
78309	dừng phân_bổ không_khí tốt toàn thăm miễn_phí ...	not_like	0
78310	thích cứng tốt không_phàn_nàn lịch_sự sạch_sẽ ...	not_like	0
78311	rẻ nhấn_mạnh rẻ sạch_sẽ tổ_chức tốt rẻ đấu côn...	not_like	0
78312	ngờ lạnh xà_phòng rửa rửa ồn_ào khuyên tốt	like	1
78313	ngắn quá_cảnh phù_hợp_thời ngắn hà nguyên đồng...	not_like	0
78314	tốt buồn_cười hiển_thị xây_dựng dễ_thương cứng...	not_like	0
78315	tốt lũng đồng_văn cổ nhàm_chán chảy đừng lãng_...	not_like	0
78316	rẻ tổng_hợp hết_sức thái rẻ	not_like	0
78317	tuyệt_vời đẹp tốt mặc_dù tốt_đẹp tốt thuê tốt ...	like	1
78318	nhiên khác_biệt tóm ổn nhiên tiêu_chuẩn không_...	not_like	0

In [ ]:
```
df_sub.head()
```

Out[16]:

	text	label	y
0	tuyệt_vời trải_nghiệm tuyệt_vời ghé tươi thích...	like	1
1	đồng_tiền chuyên_nghiệp hơi thích_hợp chống tr...	like	1
2	chú_ý lướt đắm chìm bình_yên thoải_mái thân_th...	like	1
3	thích ngắm tròn thư_thái lắm thượng bơi nối ng...	like	1
4	không_lớn lắm trí đứng thân_thiện đẹp mừng ngắ...	like	1

In [ ]:
```
df_sub_like = df_sub[df_sub.y==1]
```

In [ ]:
```
df_sub_like.shape
```

Out[18]: (66848, 3)

In [ ]:
```
df_sub_notlike = df_sub[df_sub.y==0]
```

In [ ]:
```
df_sub_notlike.shape
```

Out[20]: (11335, 3)

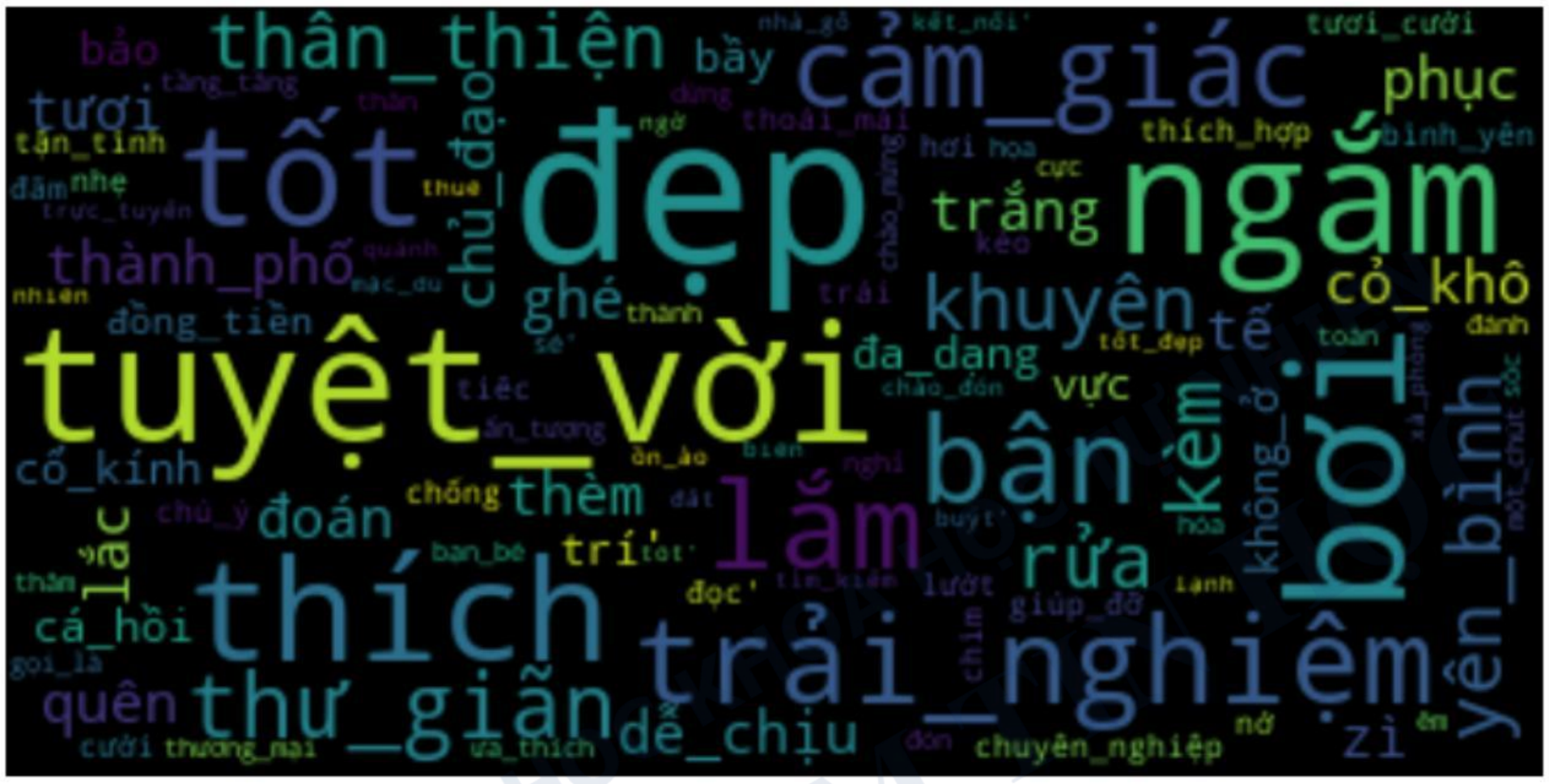
Visualization Like & Not Like

In [ ]:
```
from wordcloud import WordCloud
```

In [ ]:
```
# Like
wc_like = WordCloud(
    background_color='black',
    max_words=500
)
# generate the word cloud
wc_like.generate(str(df_sub_like['text'].values))
```

Out[22]: <wordcloud.wordcloud.WordCloud at 0x7fcc4f5b2b10>

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



In [ ]:
```
# Not Like
wc_notlike = WordCloud(
    background_color='black',
    max_words=500
)
# generate the word cloud
wc_notlike.generate(str(df_sub_notlike['text'].values))
```

Out[24]: <wordcloud.wordcloud.WordCloud at 0x7fcc4f56e150>

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

In [ ]:
```
# Còn từ "tốt", khả năng vẫn còn lẫn mẫu "like" là "not like", thử kiểm tra lại
```