Chapter 7 - Ex2: NLP Thời Trang Nam - Comments (Shopee)

In [ ]:
```
# !pip install underthesea
```

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
from underthesea import word_tokenize, pos_tag, sent_tokenize
import regex
import string
from Viet_lib import *
```

* Dữ liệu đọc ra từ file 'Products_ThoiTrangNam_comments_20K.csv'
* Bạn hãy làm phần tiền xử lý liệt kê dưới đây:

1. Đọc dữ liệu -> dataframe
2. Từ dataframe vừa đọc hãy lọc ra những dữ liệu có số lượng từ trong comment >=7 từ
3. Với kết quả câu trên -> Tạo bộ dữ mới df_sub chỉ gồm 2 cột là 'comment' và 'rating' từ dữ liệu
4. Xử lý dữ liệu thiếu, dữ liệu trùng trong df_sub
5. Trong df_sub, từ cột 'rating' => tạo cột 'label' theo tiêu chí >=4: 1 (like), <4: 0 (not like)/ hoặc theo tiêu chí: <=2: 3 (not_like), 3: 2 (neutral), >=4: 1 (like)
6. Trong df_sub, từ cột comment -> tạo cột **comment_new** theo các bước đã được hướng dẫn trong phần **Tiền xử lý dữ liệu tiếng Việt** (có thể bổ sung, hiệu chỉnh cho phù hợp với bộ dữ liệu này) để có dữ liệu xử lý.
7. Dùng wordcloud để trực quan hóa dữ liệu 'comment_new' theo từng loại (like/not_like...)
8. Lưu dữ liệu df_sub vào tập tin (ví dụ: "Products_ThoiTrangNam_comments_20K_pre.csv") để thực hiện build model ở phần sau.

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 [ ]:
```
##LOAD EMOJICON
file = open('files/emojicon.txt', 'r', encoding="utf8")
emoji_lst = file.read().split('\n')
emoji_dict = {}
for line in emoji_lst:
    key, value = line.split('\t')
    emoji_dict[key] = str(value)
#print(teen_dict)
file.close()
#################
#LOAD TEENCODE
file = open('files/teencode.txt', 'r', encoding="utf8")
teen_lst = file.read().split('\n')
teen_dict = {}
for line in teen_lst:
    key, value = line.split('\t')
    teen_dict[key] = str(value)
#print(teen_dict)
file.close()
###############
#LOAD TRANSLATE ENGLISH -> VNMESE
file = open('files/english-vnmese.txt', 'r', encoding="utf8")
englist_lst = file.read().split('\n')
for line in englist_lst:
    key, value = line.split('\t')
    teen_dict[key] = str(value)
#print(teen_dict)
file.close()
###############
#LOAD wrong words
file = open('files/wrong-word.txt', 'r', encoding="utf8")
wrong_lst = file.read().split('\n')
file.close()
#################
#LOAD STOPWORDS
file = open('files/vietnamese-stopwords.txt', 'r', encoding="utf8")
stopwords_lst = file.read().split('\n')
file.close()
```

In [ ]:
```
df = pd.read_csv('Products_ThoiTrangNam_comments_20K.csv')
```

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

Out[159]:
```
(20000, 6)
```

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

Out[160]:

	product_id	category	sub_category	user	rating	comment
0	588	Thời Trang Nam	Quần jeans	quyenanh99x	5	chất lượng sản phẩm tốt
1	1333	Thời Trang Nam	Đồ lót	hoai_anh2992	5	, Chất lượng sản phẩm tuyệt vời, Đóng gói sản ...
2	1671	Thời Trang Nam	Đồ Bộ	r*****5	5	, Chất lượng sản phẩm tuyệt vời
3	320	Thời Trang Nam	Áo	thanhvui.mt	5	Đóng gói giao hàng nhanh. Chất lượng tốt trong...
4	871	Thời Trang Nam	Đồ lót	t*****8	5	, Đóng gói sản phẩm rất đẹp và chắc chắn

In [ ]:
```
df.rating.value_counts()
```

Out[161]:
```
5    10000
4     4000
3     2000
2     2000
1     2000
Name: rating, dtype: int64
```

In [ ]:
```
df["words"] = [len(x.split(" ")) for x in df['comment']]
```

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

Out[163]:

	product_id	category	sub_category	user	rating	comment	words
0	588	Thời Trang Nam	Quần jeans	quyenanh99x	5	chất lượng sản phẩm tốt	6
1	1333	Thời Trang Nam	Đồ lót	hoai_anh2992	5	, Chất lượng sản phẩm tuyệt vời, Đóng gói sản ...	30
2	1671	Thời Trang Nam	Đồ Bộ	r*****5	5	, Chất lượng sản phẩm tuyệt vời	7
3	320	Thời Trang Nam	Áo	thanhvui.mt	5	Đóng gói giao hàng nhanh. Chất lượng tốt trong...	13
4	871	Thời Trang Nam	Đồ lót	t*****8	5	, Đóng gói sản phẩm rất đẹp và chắc chắn	10

In [ ]:
```
df = df[df["words"]>=7]
df.shape
```

Out[164]:
```
(16378, 7)
```

In [ ]:
```
df.rating.value_counts()
```

Out[165]:
```
5    8911
4    2891
3    1622
1    1605
2    1349
Name: rating, dtype: int64
```

In [ ]:
```
# Datasub
df_sub = df[['comment', 'rating']]
```

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

Out[167]:

	comment	rating
1	, Chất lượng sản phẩm tuyệt vời, Đóng gói sản ...	5
2	, Chất lượng sản phẩm tuyệt vời	5

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

Out[168]:
```
comment    0
rating     0
dtype: int64
```

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

Out[169]:
```
comment    0
rating     0
dtype: int64
```

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

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

Out[171]:
```
(11915, 2)
```

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

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

Out[173]:
```
5    5741
4    2451
2    1283
3    1229
1    1211
Name: rating, dtype: int64
```

In [ ]:
```
df_sub['label'] = [1 if x>=4 else 0 for x in df_sub.rating]
```

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

Out[175]:
```
1    8192
0    3723
Name: label, dtype: int64
```

In [ ]:
```
# Tỉ lệ like vs not_like: 2:1
```

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

Out[178]:

	comment	rating	label
19993	phiếu thì là 5 sản phẩm nhưng mà là 4 cái kẹo ...	1	0
19994	Shop phục vụ rất kém Chất lượng sản phẩm rất kém	1	0
19996	Bọn giao hàng tự lấy hàng của khách tự ...	1	0
19997	Shop không hề che tên sản phẩm dù đó là 1 yêu ...	1	0
19998	Tiền nào của nấy nên k bàn về chất lượng. Nhưn...	1	0

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

Out[179]:

	comment	rating	label
1	, Chất lượng sản phẩm tuyệt vời, Đóng gói sản ...	5	1
2	, Chất lượng sản phẩm tuyệt vời	5	1
3	Đóng gói giao hàng nhanh. Chất lượng tốt trong...	5	1
4	, Đóng gói sản phẩm rất đẹp và chắc chắn	5	1
5	Mình mua bị chật nhưng shop nhiệt tình đồng ý ...	5	1

In [ ]:
```
# Tiền xử lý dữ liệu tiếng Việt
df_sub['comment_new'] = df_sub['comment'].apply(
                lambda x: process_text(str(x), emoji_dict,
                teen_dict, wrong_lst))
```

In [ ]:
```
df_sub['comment_new'] = df_sub['comment_new'].apply(
                lambda x: covert_unicode(str(x)))
```

In [ ]:
```
df_sub['comment_new'] = df_sub['comment_new'].apply(
                lambda x: process_postag_thesea(str(x)))
```

In [ ]:
```
df_sub['comment_new'] = df_sub['comment_new'].apply(
                lambda x: remove_stopword(str(x), stopwords_lst))
```

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

Out[184]:

	comment	rating	label	comment_new
5818	Sản phẩm chỉ nhận xu thôi nhé ok brooo vopppppp...	5	1	sản_phẩm đồng_ý brooo
7150	Đặt hàng dùm bạn bạn khen xinh nên em đánh giá...	5	1	đặt_hàng dùm ngợi ố yêu
15654	Áo thì giống mẫu ,vãi đẹp nhưng bị rách 1 lỗ ...	3	0	áo mẫu vãi đẹp rách lỗ tương_đối mông cửa tiệm...
8069	Chưa bao giờ mua đc cái nào ưng như vậy luôn á...	5	1	ưng xịn quá_trời chất chê cực đẹp lunnn
12093	Màu kem ???? Nhìn cũng được rộng hơn thì đẹphs...	4	1	màu rộng
8871	Xịn xò quá mn ạ chất lượng ok mát mẻ thoải mái...	5	1	xịn xò chất_lượng đồng_ý mát_mẻ
17267	Thông số ghi sai so với ảnh ,size 29 ghi bụng ...	2	0	thông_số ảnh bụng đi_đi sửa bụng chán
1429	Quàn chất vải mát đẹp lắm mọi người nên mua nh...	5	1	quàn chất vải mát đẹp lắm nhaaaaaaaaaaaa
15570	Vải khá dày, hơi thô nên chắc mặc mùa đông thì...	3	0	vải dày hơi thô mặc mùa đông ổn hơi chát tiền
13001	Áo đẹp, khá mát..... Áo đen mỏng hơn nhiều so...	4	1	áo đẹp mát áo đen mỏng màu trắng

In [ ]:
```
df_sub.to_csv("Products_ThoiTrangNam_comments_20K_pre.csv")
```

Visualization Like & Not Like

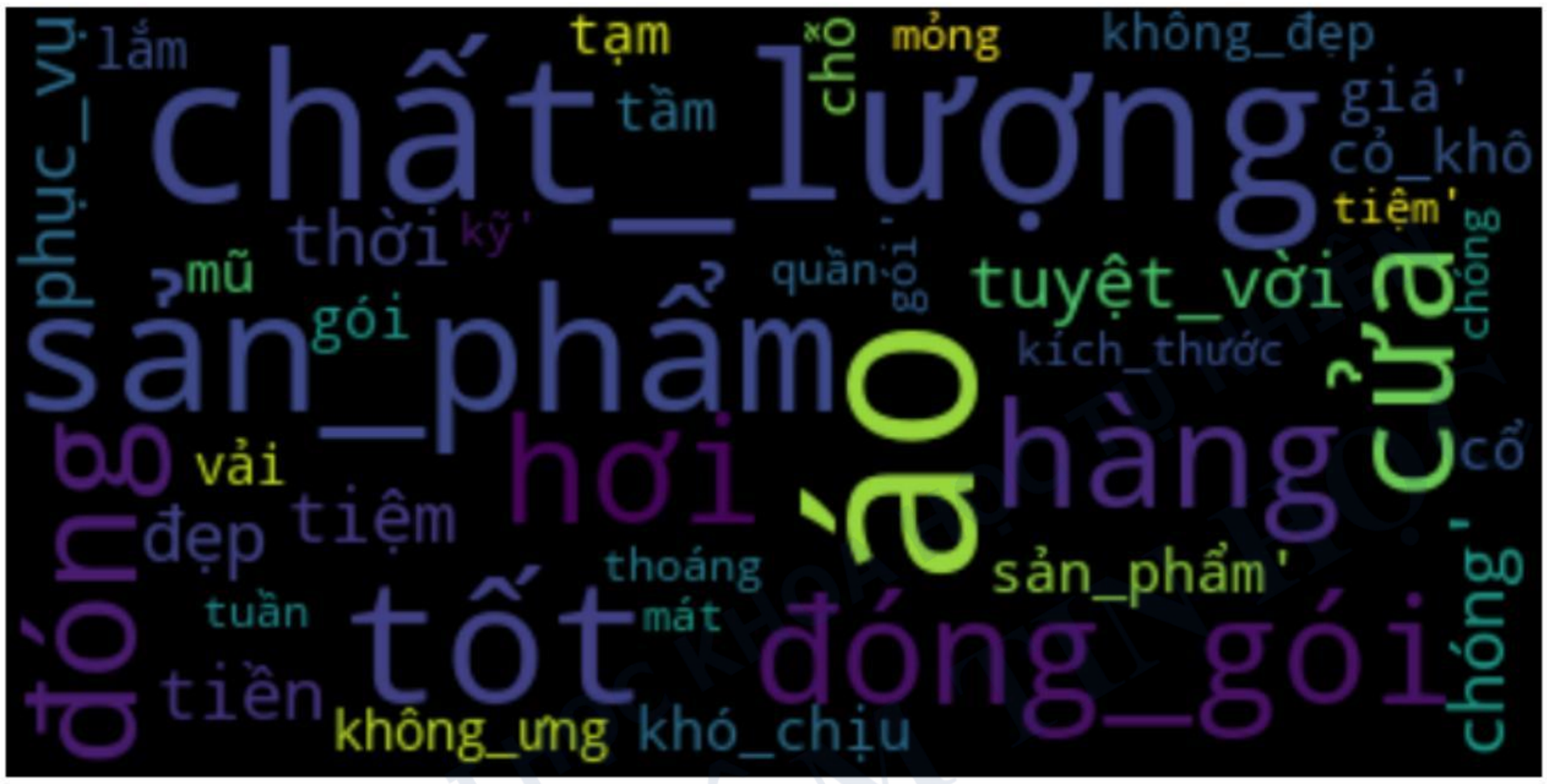
In [ ]:
```
from wordcloud import WordCloud
```

In [ ]:
```
df_sub_like = df_sub[df_sub.label ==1]
df_sub_notlike = df_sub[df_sub.label ==0]
```

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

Out[188]:
```
<wordcloud.wordcloud.WordCloud at 0x7f94ff19ab10>
```

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['comment_new'].values))
```

Out[190]: <wordcloud.wordcloud.WordCloud at 0x7f94fee74350>

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