PHÂN LOẠI TIÊU ĐỀ BÀI BÁO (HEADLINES) THEO CHỦ ĐỀ (TOPICS)

*Tập chủ đề : { Chính Trị (1) - Kinh Doanh (2) - Đời Sống (3) - Pháp Luật (4) - Thể Thao (5) - Khoa học & Công Nghệ (6) - Giáo Dục (7) - Giải Trí (8) }

- **Input: Tiêu đề bài báo** - ví dụ: "Trường đại học Công Nghệ Thông Tin: Những bước đi tiên phong - sáng tạo"
- **Output: Chủ đề bài báo (1-8)**: Giáo dục(7)

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors.classification import KNeighborsClassifier
from sklearn.linear_model.stochastic_gradient import SGDClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC
import pandas as pd

import json
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/utils/deprecation.py:144: FutureWarning:
  warnings.warn(message, FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/utils/deprecation.py:144: FutureWarning:
  warnings.warn(message, FutureWarning)
```

```
# To see full pandas row
pd.set_option('max_rows', 99999)
pd.set_option('max_colwidth', 400)
pd.describe_option('max_colwidth')
```

```
display.max_colwidth : int or None
    The maximum width in characters of a column in the repr of
    a pandas data structure. When the column overflows, a "..."
    placeholder is embedded in the output. A 'None' value means unlimited.
    [default: 50] [currently: 400]
```

Thu thập dữ liệu

Nguồn dữ liệu cung cấp cho train và test để xây dựng và chọn lựa model được lấy từ 2 trang web báo uy tín là vnexpress và vietnamnet. Model tốt nhất được chọn lựa sau quá trình train and test sẽ được demo phân loại tiêu đề của các trang báo khác (báo phụ nữ, báo thanh niên, báo tuổi trẻ).

CODE for Data Crawler can be found here, on my Github Repo:

- Các function: https://github.com/JunHill/CS114.K21.KHTN/blob/master/crawler.py
- Crawl dữ liệu từ VN-Express: https://github.com/JunHill/CS114.K21.KHTN/blob/master/vn_express_crawler.py
- Crawl dữ liệu từ VietnamNet: https://github.com/JunHill/CS114.K21.KHTN/blob/master/vietnamnet_crawler.py
- Training and Validating Data source: https://vnexpress.net/ https://vietnamnet.vn/vn/thoi-su/chinh-tri/trang1/
- Testing data source: https://www.phunuonline.com.vn/ https://tuoitre.vn/
- headline_data.json: https://github.com/JunHill/CS114.K21.KHTN/blob/master/headline_data.json

```
#--------------------------
# Load data from our json files
#--------------------------
data = pd.read_json('C:/backupD/headline_data.json', lines=True)
data = data.sample(frac=1)
data['topic_names'] = [None] * len(data)
data.loc[data.topic==1, 'topic_names'] = "Chính Trị & Quân Sự"
data.loc[data.topic==2, 'topic_names'] = "Kinh Doanh"
data.loc[data.topic==3, 'topic_names'] = "Đời sống"
data.loc[data.topic==4, 'topic_names'] = "Pháp Luật"
data.loc[data.topic==5, 'topic_names'] = "Thể Thao"
data.loc[data.topic==6, 'topic_names'] = "Khoa Học & Công Nghệ"
data.loc[data.topic==7, 'topic_names'] = "Giáo Dục"
data.loc[data.topic==8, 'topic_names'] = "Giải Trí"
cols = ["headline", "topic", "topic_names"]
data = data[cols].reset_index(drop=True)
data = data.drop_duplicates(subset=['headline'])
data[:10]
```

	headline	topic	topic_names
0	Những cách đơn giản giúp bạn thông minh hơn mỗi ngày	3	Đời sống
1	Vụ nổ động cơ phá hủy chiếc tiêm kích 88 triệu USD của Australia	1	Chính Trị & Quân Sự

Mô tả dữ liệu

- Số lượng: 167,322 tiêu đề
- Các tiêu đề được viết bởi tác giả của vnexpress và vietnamnet nên sẽ không có sai chính tả, sai ngữ nghĩa.
- Các từ vựng của tiêu đề thể hiện tương đối chính xác về chủ đề đã được label, tuy nhiên vẫn còn nhiều sự trùng lặp trong từ vựng của các chủ đề.
- Dữ liệu sẽ được chia 7:3 cho train và test sau khi shuffle.

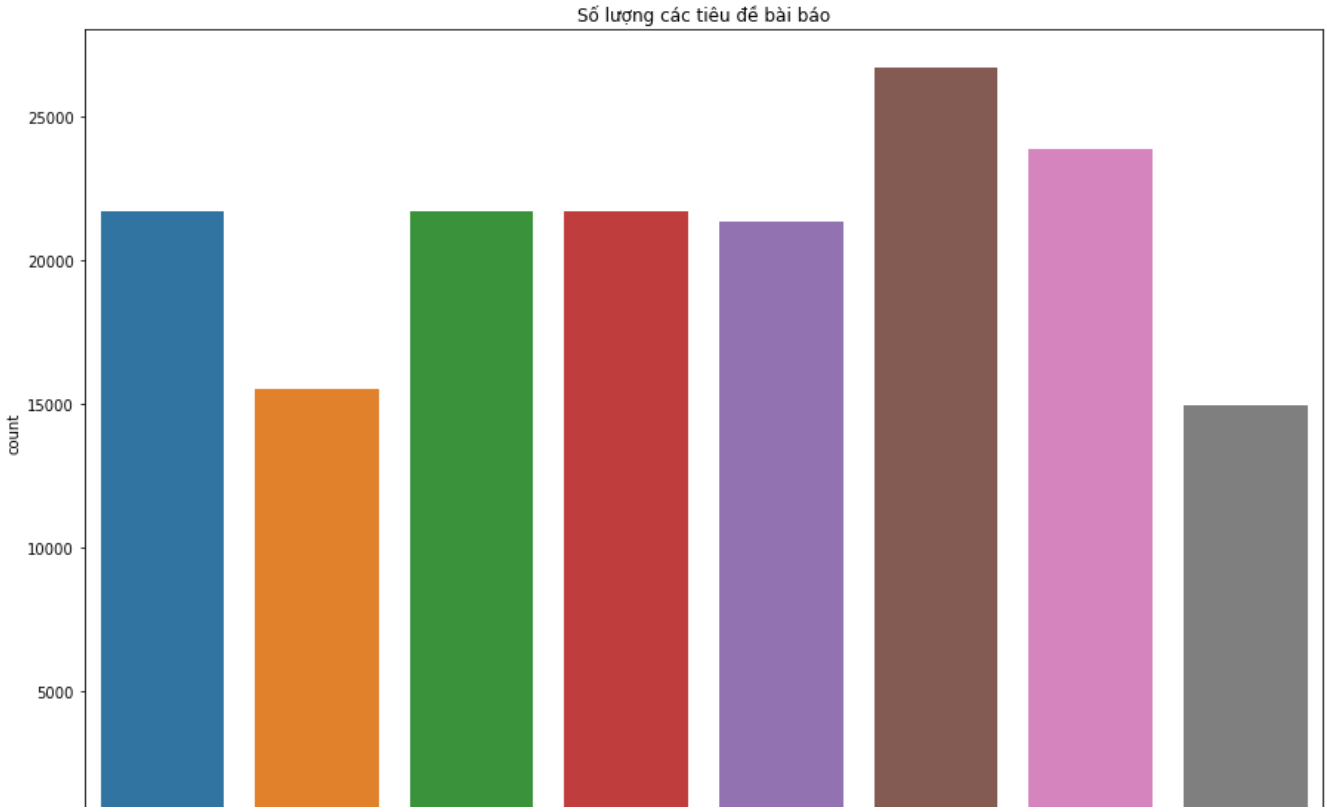
9	Vẽ kim tự tháp 3D trong vài phút	7	Giáo Dục

```
data.describe()
```

	topic
count	167322.000000
mean	4.735038
std	2.292571
min	1.000000
25%	3.000000
50%	5.000000
75%	7.000000
max	8.000000

```
import seaborn as sns
plt.figure(figsize=(15,10))
sns.countplot(data.topic_names).set_title('Số lượng các tiêu đề bài báo')
```

```
Text(0.5, 1.0, 'Số lượng các tiêu đề bài báo')
```



CLEAN DATA và TÁCH TỪ TIẾNG VIỆT

```
!pip install pyvi
from pyvi import ViTokenizer
```

```
Collecting pyvi
  Downloading https://files.pythonhosted.org/packages/10/e1/0e5bc6b5e3327b9385d6e0f1b0a
     |████████████████████████████████| 8.5MB 2.6MB/s
Collecting sklearn-crfsuite
  Downloading https://files.pythonhosted.org/packages/25/74/5b7befa513482e6dee1f3dd6817
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.6/dist-packages (
Requirement already satisfied: tabulate in /usr/local/lib/python3.6/dist-packages (from
Requirement already satisfied: tqdm>=2.0 in /usr/local/lib/python3.6/dist-packages (fro
Collecting python-crfsuite>=0.8.3
  Downloading https://files.pythonhosted.org/packages/95/99/869dde6dbf3e0d07a013c8eebfb
     |████████████████████████████████| 747kB 44.4MB/s
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from skle
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.6/dist-packages (
Requirement already satisfied: numpy>=1.11.0 in /usr/local/lib/python3.6/dist-packages
Requirement already satisfied: scipy>=0.17.0 in /usr/local/lib/python3.6/dist-packages
Installing collected packages: python-crfsuite, sklearn-crfsuite, pyvi
Successfully installed python-crfsuite-0.9.7 pyvi-0.1 sklearn-crfsuite-0.3.6
```

```
#Xoá bớt những kí tự không cần thiết như - / ? .
def clean(pd_series):
    pd_series = pd_series.str.lower()
    pd_series = pd_series.str.replace(r'[-\'.,?":/!@#$%^&*()]', '')
    pd_series = pd_series.str.replace(r'\d+', '')
```

```
    return pd_series

#Tách từ tiếng việt, sử dụng thư viện ViTokenizer: https://pypi.org/project/pyvi/  https://gi
# 'cô dâu đồng tính là á hậu' ---> 'cô_dâu đồng_tính là á_hậu'
def tokenize(text):
    return ViTokenizer.tokenize(text)


data['headline'] = clean(data['headline'])
data['headline'] = data['headline'].apply(tokenize)
data[:10]
```

	headline	topic	topic_names
0	những cách đơn_giản giúp bạn thông_minh hơn mỗi ngày	3	Đời sống
1	vụ nổ động_cơ phá_hủy chiếc tiêm_kích triệu usd của australia	1	Chính Trị & Quân Sự
2	mảnh giấy lạ chỉ_dẫn nơi tìm thi_thể đôi vợ_chồng ở thanh_hoá	4	Pháp Luật
3	cách làm chè xoài mát lạnh xóa_tan nóng_nực	3	Đời sống
4	nasa lên kế_hoạch bảo_vệ trái_đất trước tiểu hành_tinh	6	Khoa Học & Công Nghệ
5	lần bán nhà nhanh và được giá của cặp vợ_chồng sài_gòn	3	Đời sống
6	giá vàng hiện_nay khác gì năm	2	Kinh Doanh
7	shophouse sài_gòn có thiết_kế thang_máy giá triệu usd	2	Kinh Doanh
8	tình_tiết khó tin vụ trộm khoắng gần chỉ vàng nhà nữ đại_gia	4	Pháp Luật
9	vẽ kim_tự_tháp d trong vài phút	7	Giáo Dục

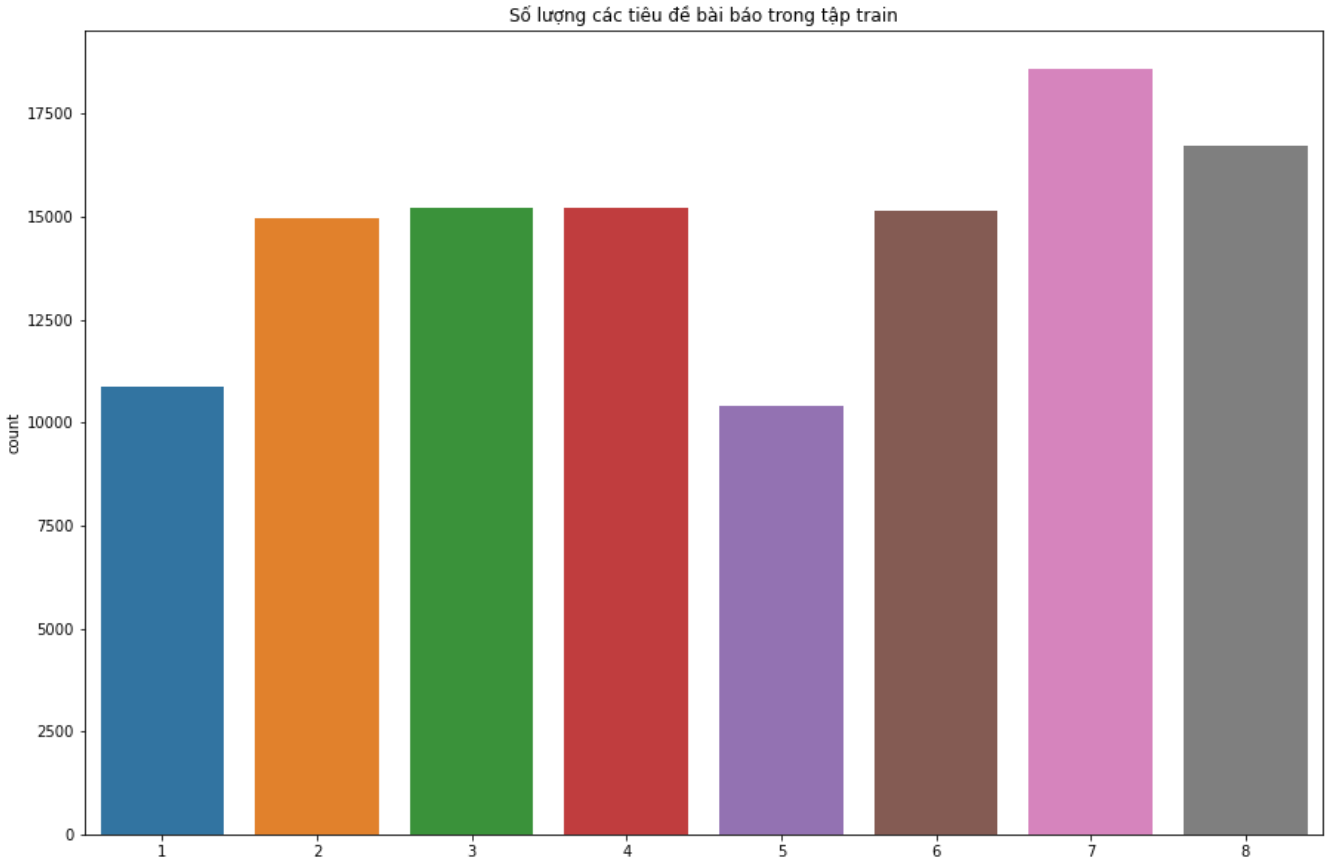
```
# Split train (70%) - test(30%)
X_train, X_test, y_train, y_test = train_test_split(data['headline'].tolist(), data['topic'].
print(X_train[:5])
print(y_train[:5])
```

```
data.shape
```

```
(167322, 3)
```

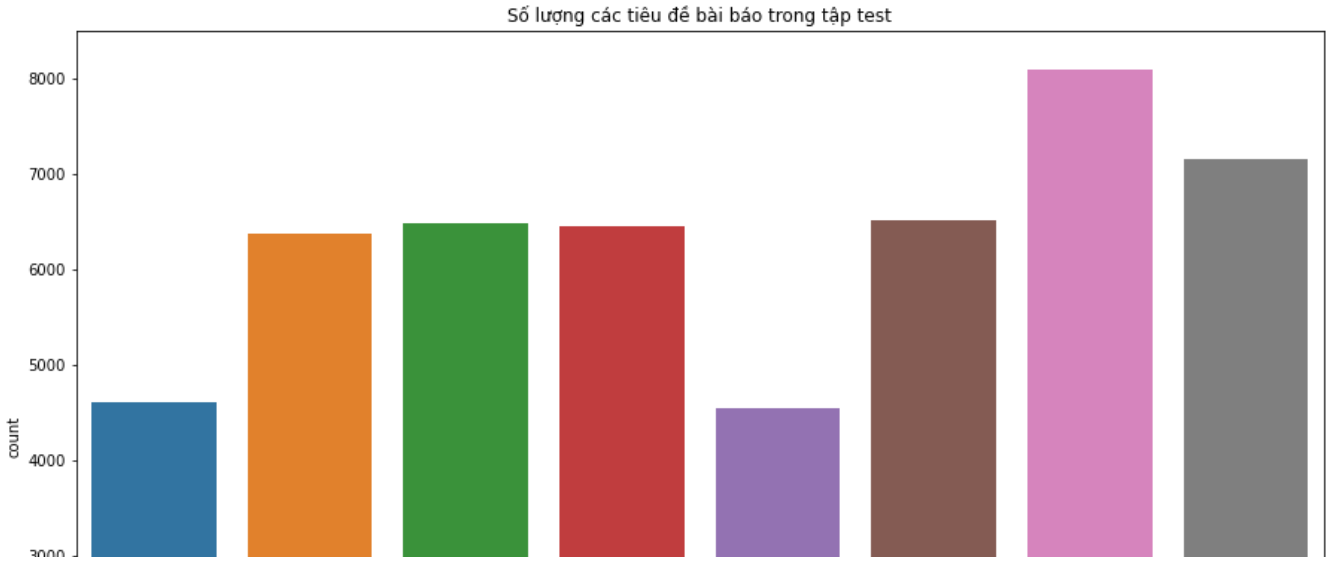
```
import seaborn as sns
print(f"training shape: {len(X_train)}")
plt.figure(figsize=(15,10))
sns.countplot(y_train).set_title('Số lượng các tiêu đề bài báo trong tập train')
```

```
training shape: 117125
Text(0.5, 1.0, 'Số lượng các tiêu đề bài báo trong tập train')
```




```
print(f"testing shape: {len(X_test)}")
plt.figure(figsize=(15,10))
sns.countplot(y_test).set_title('Số lượng các tiêu đề bài báo trong tập test')
```

```
testing shape: 50197
Text(0.5, 1.0, 'Số lượng các tiêu đề bài báo trong tập test')
```



Feature extractor: Tensor flow inverse document frequency

- Giá trị của một word trong corpus có mối quan hệ sau: quan hệ thuận với tần số xuất hiện trong 1 câu, nhưng tỉ lệ nghịch với tần số xuất hiện trong corpus

```
#dictionary dùng để lưu các classifier sau khi train xong
tfidf_model = {}
#Hàm dùng để tạo inverse document frequency từ tập tiêu đề
#Chọn ngram_range là (1,2) để bao gồm các từ đơn và từ ghép đôi
def create_idf(message_list, stopwords):
    vectorizer = TfidfVectorizer(ngram_range=(1,2), stop_words=stopwords)
    doc = vectorizer.fit_transform(message_list)
    return doc, vectorizer
```

Stop word

Do tiêu đề đã được viết theo phong cách trang trọng hoặc rất ngắn ngọn nên không cần thiết sử dụng tập stop words cho Tiếng Việt được nói hằng ngày. Chỉ cần lọc ra một số từ mà em nghĩ sẽ không cần thiết cho model.

```
stop_words = ['có_thể', 'để', 'những', 'ngày', 'làm', 'vì_sao', 'ngày_càng', 'vì', 'cùng', 't
X_train_idf, vectorizer = create_idf(X_train, stop_words)
X_test_idf = vectorizer.transform(X_test)
print(f'train shape: {X_train_idf.shape}')
print(f'test shape: {X_test_idf.shape}')
```

```
train shape: (117125, 387729)
```

```
vectorizer.vocabulary_
```

```
{'người': 193231,
 'việt': 329277,
 'tốn': 319550,
 'gần': 101540,
 'tỷ': 323657,
 'usd': 324960,
 'uống': 325894,
 'bia': 6975,
 'năm': 211204,
 'người việt': 194384,
 'việt tốn': 329839,
 'tốn gần': 319555,
 'gần tỷ': 101801,
 'tỷ usd': 323995,
 'usd uống': 325332,
 'uống bia': 325896,
 'bia năm': 7015,
 'phân_biệt': 223184,
 'see': 246186,
 'look': 150827,
 'watch': 344236,
 'view': 327736,
 'phân_biệt see': 223308,
 'see look': 246187,
 'look watch': 150831,
 'watch view': 344257,
 'cựu': 72351,
 'hoa_hậu': 107749,
 'hòa': 115709,
 'bình': 14561,
 'loại': 151143,
 'khỏi': 138001,
 'miss': 168618,
 'universe': 324853,
 'australia': 4250,
 'cựu hoa_hậu': 72376,
 'hoa_hậu hòa': 107801,
 'hòa bình': 115715,
 'bình loại': 14632,
 'loại khỏi': 151233,
 'khỏi miss': 138160,
 'miss universe': 168639,
 'universe australia': 324854,
 'nghiên_cứu': 187625,
 'mới': 181215,
 'tiết_lộ': 285882,
 'bộ_phận': 25966,
 'gợi_cảm': 102908,
 'nhất': 204181,
 'trên_cơ_thể': 295504,
 'phụ_nữ': 228528,
 'nghiên_cứu mới': 187680,
 'mới tiết_lộ': 181654,
 'tiết_lộ bộ_phận': 285897,
 'bộ_phận gợi_cảm': 25973,
 'gợi_cảm nhất': 102931,
 'nhất trên_cơ_thể': 204543,
```

```
 'trên_cơ_thể phụ_nữ': 295509,
 'quy_trình': 232182,
 'huấn_luyện': 110732,
 'lính': 156422,
 'bắn': 20979,
 'tỉa': 318552,
 'trinh_sát': 291132,
 'mỹ': 182791,
 'quy_trình huấn_luyện': 232196,
 'huấn_luyện lính': 110758,
 'lính bắn': 156428,
 'bắn tỉa': 21100,
 'tỉa trinh_sát': 318577,
 'trinh_sát mỹ': 291146,
 'xây_dựng': 349502,
 'sân_bay': 253487,
 'lên': 155885,
 'sa': 243891,
 'pa': 218620,
 'dễ': 82553,
 'như': 202663,
 'ra': 237896,
 'ngoại_thành': 190179,
 'xây_dựng sân_bay': 349624,
 'sân_bay gần': 253526,
 'tỷ lên': 323824,
 'lên sa': 156126,
 'sa pa': 243907,
 'pa dễ': 218627,
 'dễ như': 82637,
 'như ra': 203019,
 'ra ngoại_thành': 238141,
 'mẹ': 178670,
 'neymar': 186237,
 'chia_tay': 30532,
 'tình': 309479,
 'trẻ': 299801,
 'mẹ neymar': 178928,
 'neymar chia_tay': 186241,
 'chia_tay tình': 30643,
 'tình trẻ': 309652,
 'sinh_viên': 247959,
 'đeo_đuổi': 358595,
 'phim': 220228,
 'về': 339563,
 'không_chiến': 136568,
 'hàm_rồng': 112502,
 'sinh_viên năm': 248125,
 'năm đeo_đuổi': 212105,
 'đeo_đuổi phim': 358597,
 'phim về': 220594,
 'về không_chiến': 340079,
 'không_chiến hàm_rồng': 136572,
 'rapper': 239029,
 'gây': 100142,
 'bão': 13635,
 'mạng': 175751,
 'bài': 11013,
```

```
'tôi': 311344,
'kiện': 141679,
'hệ_thống': 120124,
'giáo_dục': 92087,
'rapper mỹ': 239034,
'mỹ gây': 183098,
'gây bão': 100146,
'bão mạng': 13694,
'mạng bài': 175755,
'bài tôi': 11124,
'tôi kiện': 311567,
'kiện hệ_thống': 141741,
'hệ_thống giáo_dục': 120165,
'chip': 31023,
'gián_điệp': 91964,
'trung': 292590,
'quốc': 235928,
'được': 370096,
'cài': 52375,
'sẵn': 256833,
'máy_chủ': 172878,
'của': 69094,
'amazon': 1256,
'và': 331488,
'apple': 3315,
'chip gián_điệp': 31036,
'gián_điệp trung': 91976,
'trung quốc': 292680,
'quốc được': 236858,
'được cài': 370265,
'cài sẵn': 52387,
'sẵn máy_chủ': 256848,
'máy_chủ của': 172881,
'của amazon': 69117,
'amazon và': 1322,
'và apple': 331524,
'thủ_tướng': 281137,
'động_viên': 383878,
'đoàn': 361552,
'xe': 345583,
'xuất_hành': 348308,
'tiêu_thụ': 284683,
'vải_thiều': 338114,
'bắc_giang': 20948,
'thủ_tướng động_viên': 281538,
'động_viên đoàn': 383909,
'đoàn xe': 361619,
'xe xuất_hành': 345961,
'xuất_hành tiêu_thụ': 348309,
'tiêu_thụ vải_thiều': 284698,
'vải_thiều bắc_giang': 338115,
'mẫu': 177240,
'trực_thăng': 302696,
'vũ_trang': 337132,
'cho': 32349,
'lục_quân': 164491,
'mẫu trực_thăng': 177346,
'trực_thăng trinh_sát': 302756
```

'trực_thăng trinh_sát': 302756,
'trinh_sát vũ_trang': 291154,
'vũ_trang mới': 337141,
'mới cho': 181262,
'cho lục_quân': 32924,
'lục_quân mỹ': 164495,
'vinamilk': 327852,
'sở_hữu': 258715,
'cổ_phần': 68199,
'công_ty_mẹ': 59701,
'sữa': 259931,
'mộc': 180374,
'châu': 36925,
'vinamilk sở_hữu': 327878,
'sở_hữu cổ_phần': 258740,
'cổ_phần công_ty_mẹ': 68209,
'công_ty_mẹ của': 59702,
'của sữa': 70699,
'sữa mộc': 259987,
'mộc châu': 180375,
'tượng_đài': 314611,
'vinh_danh': 328134,
'phi_công': 219656,
'liên': 149357,
'xô': 350394,
'sắp': 256456,
'xây': 349322,
'tại': 314660,
'tượng_đài vinh_danh': 314616,
'vinh_danh phi_công': 328167,
'phi_công liên': 219679,
'liên xô': 149386,
'xô sắp': 350429,
'sắp được': 256708,
'được xây': 371186,
'xây tại': 349466,
'tại mỹ': 315016,
'nai': 184475,
'sừng': 259536,
'tấm': 316151,
'quyết_chiến': 232562,
'giành': 90821,
'bạn': 17436,
'giữa': 97393,
'đường_cao_tốc': 369796,
'nai sừng': 184509,
'sừng tấm': 259552,
'tấm quyết_chiến': 316180,
'quyết_chiến giành': 232565,
'giành bạn': 90825,
'bạn tình': 17813,
'tình giữa': 309533,
'giữa đường_cao_tốc': 97711,
'tình_huống': 309886,
'bi': 6912,
'hài': 112310,
'tình_huống bi': 309889,
'bi hài': 6917,

```
 'hài của': 112326,
 'của người': 70294,
 'người mới': 193859,
 'mới mẹ': 181493,
 'thủ_môn': 280998,
 'navas': 185813,
 'đòi': 366564,
 'rời': 243339,
 'real': 239247,
 'madrid': 165619,
 'thủ_môn navas': 281029,
 'navas đòi': 185816,
 'đòi rời': 366676,
 'rời real': 243412,
 'real madrid': 239316,
 'clb': 47722,
 'bến': 22799,
 'tre': 290848,
 'lại': 159203,
 'xin': 346862,
 'bỏ': 23721,
 'giải': 94189,
 'hạng': 118116,
 'nhì': 201844,
 'clb bến': 47725,
 'bến tre': 22813,
 'tre lại': 290878,
 'lại xin': 159841,
 'xin bỏ': 346872,
 'bỏ giải': 23799,
 'giải hạng': 94306,
 'hạng nhì': 118134,
 'chị_em': 43439,
 'ruột': 240878,
 'trốn': 301028,
 'nã': 209624,
 'án': 353632,
 'ma_túy': 165351,
 'bắt': 21188,
 'giáp': 92798,
 'tết': 318189,
 'chị_em ruột': 43491,
 'ruột trốn': 240923,
 'trốn nã': 301096,
 'nã án': 209636,
 'án ma_túy': 353774,
 'ma_túy bắt': 165359,
 'bắt giáp': 21318,
 'giáp tết': 92817,
 'trao': 290726,
 'quyết_định': 232658,
 'chủ_tịch': 46404,
 'nước': 212728,
 'thiếu_tướng': 266676,
 'tạ': 314619,
 'quang': 231626,
 'khải': 137396,
 'trao quyết_định': 290767,
```

```
 'quyết_định của': 232688,
 'của chủ_tịch': 69400,
 'chủ_tịch nước': 46504,
 'nước cho': 212768,
 'cho thiếu_tướng': 33296,
 'thiếu_tướng tạ': 266699,
 'tạ quang': 314644,
 'quang khải': 231655,
 'acb': 85,
 'muốn': 170698,
 'thoái': 266923,
 'sạch': 255594,
 'vốn': 341539,
 'kem': 128423,
 'thuỷ_tạ': 268876,
 'acb muốn': 90,
 'muốn thoái': 171073,
 'thoái sạch': 266927,
 'sạch vốn': 255690,
 'vốn tại': 341644,
 'tại kem': 314942,
 'kem thuỷ_tạ': 128443,
 'chi': 30047,
 'triệu': 291661,
 'mua': 169869,
 'lá_chắn': 154483,
 'bảo_vệ': 19182,
 'xe_tăng': 346305,
 'trophy': 292315,
 'israel': 126797,
 'mỹ chi': 182883,
 'chi triệu': 30145,
 'triệu usd': 291984,
 'usd mua': 325153,
 'mua lá_chắn': 170063,
 'lá_chắn bảo_vệ': 154485,
 'bảo_vệ xe_tăng': 19387,
 'xe_tăng trophy': 346371,
 'trophy của': 292316,
 'của israel': 69910,
 'cảm_hứng': 63824,
 'hai': 103499,
 'mặt': 178122,
 'xã_hội': 349814,
 'us': 324937,
 'cảm_hứng về': 63885,
 'về hai': 339953,
 'hai mặt': 103794,
 'mặt xã_hội': 178324,
 'xã_hội mỹ': 349880,
 'mỹ phim': 183456,
 'phim us': 220580,
 'guardiola': 98790,
 'mừng': 182597,
 'man': 166130,
 'city': 47489,
 'thoát': 266955,
 'guardiola mừng': 98844,
```

```
 'guardiola mừng': 98844,
'mừng man': 182641,
'man city': 166135,
'city thoát': 47600,
'thoát án': 267035,
'vật_thể': 339132,
'nhô': 202410,
'cao': 27822,
'đầu': 374236,
'giống': 96478,
'quái_vật': 233514,
'hồ': 122547,
'loch': 150557,
'ness': 186107,
'albania': 951,
'vật_thể nhô': 339144,
'nhô cao': 202411,
'cao đầu': 28129,
'đầu giống': 374372,
'giống quái_vật': 96581,
'quái_vật hồ': 233528,
'hồ loch': 122602,
'loch ness': 150558,
'ness albania': 186108,
'kỷ_niệm': 146273,
'đặc_biệt': 376274,
'bác_sĩ': 11923,
'thú_y': 273735,
'hà_nội': 111578,
'đại_sứ': 372526,
'kỷ_niệm đặc_biệt': 146318,
'đặc_biệt của': 376294,
'của bác_sĩ': 69231,
'bác_sĩ thú_y': 12027,
'thú_y hà_nội': 273738,
'hà_nội và': 112094,
'và đại_sứ': 333402,
'đại_sứ mỹ': 372547,
'cựu_phó': 72603,
'thống_đốc': 279869,
'nhnn': 198596,
'đặng': 376605,
'thanh_bình': 262913,
'nghẹn_ngào': 188629,
'nói': 209971,
'lời': 163737,
'sau_cùng': 245733,
'cựu_phó thống_đốc': 72610,
'thống_đốc nhnn': 279879,
'nhnn đặng': 198599,
'đặng thanh_bình': 376616,
'thanh_bình nghẹn_ngào': 262921,
'nghẹn_ngào nói': 188636,
'nói lời': 210107,
'lời sau_cùng': 163847,
'tình_yêu': 310160,
'sét': 253941,
'đánh': 364337,
```

```
 'đôi': 367585,
 'uyên_ương': 325821,
 'đều': 378630,
 'nặng': 214410,
 'trên': 294919,
 'rưỡi': 242041,
 'tình_yêu sét': 310222,
 'sét đánh': 253963,
 'đánh của': 364395,
 'của đôi': 71213,
 'đôi uyên_ương': 367808,
 'uyên_ương đều': 325847,
 'đều nặng': 378673,
 'nặng trên': 214493,
 'trên tạ': 295382,
 'tạ rưỡi': 314645,
 'thời_lượng': 280343,
 'pin': 229470,
 'smartphone': 248996,
 'đời': 384251,
 'càng': 52419,
 'kém': 142766,
 'thời_lượng pin': 280346,
 'pin smartphone': 229506,
 'smartphone đời': 249111,
 'đời càng': 384278,
 'càng mới': 52467,
 'mới càng': 181294,
 'càng kém': 52454,
 'bv': 10757,
 'sản': 255750,
 'nhi': 197623,
 'quảng_ninh': 235408,
 'không': 135061,
 'sử_dụng': 259587,
 'bệnh_án': 23307,
 'giấy': 95822,
 'bv sản': 10760,
 'sản nhi': 255753,
 'nhi quảng_ninh': 197658,
 'quảng_ninh không': 235436,
 'không sử_dụng': 136029,
 'sử_dụng bệnh_án': 259602,
 'bệnh_án giấy': 23311,
 'ngành': 191597,
 'sóc': 254003,
 'trăng': 296783,
 'mát_tay': 172361,
 'đồng': 380919,
 'ngành giáo_dục': 191642,
 'giáo_dục sóc': 92278,
 'sóc trăng': 254020,
 'trăng chi': 296786,
 'chi mát_tay': 30109,
 'mát_tay tỷ': 172362,
 'tỷ đồng': 324051,
 'nữ_sinh': 217343,
 'quốc gia': 236933,
```

```
'môn': 174577,
'địa_lý': 378937,
'nữ_sinh giành': 217402,
'giành giải': 90844,
'giải nhất': 94377,
'nhất quốc_gia': 204459,
'quốc_gia môn': 237008,
'môn địa_lý': 174644,
'kiatisuk': 139501,
'cầu_thủ': 66444,
'nam': 184550,
'kỹ_thuật': 146518,
'bằng': 22006,
'thái_lan': 270520,
'kiatisuk cầu_thủ': 139502,
'cầu_thủ việt': 66617,
'việt nam': 329589,
'nam không': 184855,
'không kỹ_thuật': 135626,
'kỹ_thuật bằng': 146523,
'bằng thái_lan': 22465,
'khởi_tố': 138991,
'vụ': 343141,
'lớp': 163366,
'mang': 166179,
'bầu': 20275,
'rước': 242013,
'dâu': 78953,
'đêm': 366087,
'rồi': 242946,
'mất_tích': 176993,
'khởi_tố vụ': 139109,
'vụ nữ_sinh': 343390,
'nữ_sinh lớp': 217454,
'lớp mang': 163499,
'mang bầu': 166194,
'bầu được': 20375,
'được rước': 370828,
'rước dâu': 242018,
'dâu đêm': 79052,
'đêm rồi': 366221,
'rồi mất_tích': 243017,
'sai_lầm': 244325,
'khi': 129316,
'dùng': 79760,
'máy_giặt': 172892,
'khiến': 130874,
'máy': 172455,
'hỏng': 122275,
'nhanh': 197121,
'mất': 176554,
'tiền': 286097,
'oan': 217723,
'sai_lầm khi': 244352,
'khi dùng': 129630,
'dùng máy_giặt': 80047,
'máy_giặt khiến': 172896,
'khiến máy': 131056
```

```
 'khiên máy': 131056,
'máy hỏng': 172508,
'hỏng nhanh': 122306,
'nhanh dễ': 197164,
'dễ mất': 82627,
'mất tiền': 176784,
'tiền oan': 286474,
'chịu': 43532,
'khoanh_tay': 131716,
'trước': 296994,
'tên_lửa': 308851,
'syria': 251290,
'israel không': 126825,
'không chịu': 135239,
'chịu khoanh_tay': 43558,
'khoanh_tay trước': 131717,
'trước tên_lửa': 297369,
'tên_lửa syria': 309018,
'con': 48261,
'đường_cao': 369792,
'mét': 173494,
'xuyên': 347883,
'qua': 230716,
'hẻm': 119803,
'núi': 210970,
'con đường_cao': 49140,
'đường_cao mét': 369793,
'mét xuyên': 173593,
'xuyên qua': 347916,
'qua hẻm': 230847,
'hẻm núi': 119811,
'chủ': 45894,
'nhàn': 200869,
'tênh': 309138,
'nhờ': 207081,
'cún': 60170,
'cưng': 62157,
'người chủ': 193393,
'chủ nhàn': 46022,
'nhàn tênh': 200879,
'tênh nhờ': 309141,
'nhờ cún': 207168,
'cún cưng': 60172,
'ong': 218172,
'bắp_cày': 21158,
'chích': 37699,
'nọc': 215102,
'tha': 262016,
'nhện': 206508,
'tổ': 319987,
'ong bắp_cày': 218175,
'bắp_cày chích': 21161,
'chích nọc': 37727,
'nọc tha': 215107,
'tha nhện': 262023,
'nhện về': 206564,
'về tổ': 340747,
'cá': 52563,
```

```
 'hút': 116667,
 'máu': 172367,
 'tàn_sát_sinh_vật': 306535,
 'bản_địa': 18541,
 'vùng': 335808,
 'cá hút': 52625,
 'hút máu': 116714,
 'máu tàn_sát_sinh_vật': 172436,
 'tàn_sát_sinh_vật bản_địa': 306536,
 'bản_địa vùng': 18545,
 'vùng hồ': 335836,
 'hồ mỹ': 122617,
 'mở': 181980,
 'nắp': 214272,
 'chai': 29514,
 'mà': 171588,
 'cần': 65665,
 'dụng_cụ': 83637,
 'chuyên_biệt': 34513,
 'mở nắp': 182060,
 'nắp chai': 214277,
 'chai mà': 29528,
 'mà không': 171639,
 'không cần': 135323,
 'cần dụng_cụ': 65763,
 'dụng_cụ chuyên_biệt': 83639,
 'thứ_trưởng': 282159,
 'công_thương': 59257,
 'quản_lý': 234984,
 'thị_trường': 279152,
 'kiểm_tra': 141314,
 'phiền_nhiễu': 220907,
 'thứ_trưởng công_thương': 282166,
 'công_thương quản_lý': 59303,
 'quản_lý thị_trường': 235080,
 'thị_trường kiểm_tra': 279216,
 'kiểm_tra con': 141325,
 'con cưng': 48421,
 'cưng không': 62174,
 'không gây': 135440,
 'gây phiền_nhiễu': 100228,
 'chuyện': 35146,
 'lạ': 158817,
 'nuôi': 208560,
 'cỏ': 67498,
 'um_tùm': 324690,
 'vườn': 337543,
 'chanh': 29634,
 'ha': 103268,
 'chả': 40669,
 'lo': 150116,
 'chống': 44057,
 'hạn': 118021,
 'chuyện lạ': 35290,
 'lạ nuôi': 158936,
 'nuôi cỏ': 208608,
 'cỏ um_tùm': 67541,
 'um_tùm vườn': 324692,
```

```
'vườn chanh': 337553,
'chanh ha': 29638,
'ha chả': 103270,
'chả lo': 40680,
'lo chống': 150139,
'chống hạn': 44103,
'căn_hộ': 60341,
'ốc_đảo': 386875,
'cây_xanh': 56394,
'phố': 227891,
'căn_hộ như': 60427,
'như ốc_đảo': 203251,
'ốc_đảo cây_xanh': 386877,
'cây_xanh giữa': 56397,
'giữa phố': 97578,
'rút': 241643,
'giấy_phép': 95917,
'kinh_doanh': 139983,
'cửa_hàng_hiệu': 72102,
'thuốc': 268596,
'tăng_giá': 313350,
'khẩu_trang': 137577,
'rút giấy_phép': 241657,
'giấy_phép kinh_doanh': 95929,
'kinh_doanh cửa_hàng_hiệu': 140008,
'cửa_hàng_hiệu thuốc': 72103,
'thuốc tăng_giá': 268676,
'tăng_giá khẩu_trang': 313370,
'cách': 54054,
'kiểm_soát': 141168,
'công_nghệ': 58643,
'cách trung': 54627,
'quốc kiểm_soát': 236287,
'kiểm_soát xã_hội': 141242,
'xã_hội bằng': 349824,
'bằng công_nghệ': 22106,
'linh': 148606,
'hát': 114509,
'đam_mê': 357441,
'mv': 171345,
'mỹ linh': 183264,
'linh hát': 148664,
'hát về': 114731,
'về đam_mê': 340867,
'đam_mê của': 357457,
'người trẻ': 194267,
'trẻ mv': 300055,
'hlv': 107193,
'tp': 288905,
'hcm': 105212,
'thiên_vị': 265688,
'công': 57517,
'phượng': 225782,
'hlv tp': 107369,
'tp hcm': 288915,
'hcm tôi': 105498,
'tôi không': 311556,
```

```
 'không thiên_vị': 136059,
 'thiên_vị công': 265691,
 'công phượng': 57571,
 'anh': 1890,
 'thư': 274366,
 'đưa': 368960,
 'con_trai': 49834,
 'sang': 244853,
 'las': 147800,
 'vegas': 326673,
 'xem': 346537,
 'show': 246949,
 'anh thư': 2571,
 'thư đưa': 274470,
 'đưa con_trai': 369009,
 'con_trai sang': 49992,
 'sang las': 244962,
 'las vegas': 147801,
 'vegas xem': 326680,
 'xem show': 346653,
 'học_sinh': 121146,
 'là': 152624,
 'văn_hóa': 336293,
 'đọc': 379416,
 'thủ_đô': 281616,
 'học_sinh là': 121324,
 'là đại_sứ': 153483,
 'đại_sứ văn_hóa': 372568,
 'văn_hóa đọc': 336372,
 'đọc thủ_đô': 379465,
 'võ_sĩ': 335697,
 'quyền': 232789,
 'đấm': 373465,
 'thua': 267724,
 'trận': 299234,
 'võ_sĩ quyền': 335717,
 'quyền anh': 232790,
 'anh đấm': 2827,
 'đấm hlv': 373482,
 'hlv khi': 107278,
 'khi thua': 130288,
 'thua trận': 267860,
 'cậu': 66956,
 'bé': 13794,
 'quyên_góp': 232429,
 'nghìn': 188211,
 'đôla': 367875,
 'mộ': 180283,
 'ung_thư': 324732,
 'cậu bé': 66958,
 'bé quyên_góp': 14003,
 'quyên_góp nghìn': 232437,
 'nghìn đôla': 188352,
 'đôla bia': 367877,
 'bia mộ': 7009,
 'mộ cho': 180292,
 'cho bạn': 32423,
 'bạn mất': 17639,
```

```
 'mất ung_thư': 176816,
 'tuyển': 303642,
 'trễ': 300576,
 'tập': 317205,
 'một': 180444,
 'tiếng': 284983,
 'mưa': 175341,
 'lớn': 163031,
 'myanmar': 171515,
 'tuyển việt': 303758,
 'nam trễ': 185207,
 'trễ tập': 300586,
 'tập một': 317266,
 'một tiếng': 180887,
 'tiếng rưỡi': 285092,
 'rưỡi mưa': 242044,
 'mưa lớn': 175387,
 'lớn myanmar': 163161,
 'bộ_trưởng': 25989,
 'chúng_tôi': 39215,
 'con_trẻ': 50085,
 'thí_nghiệm': 272068,
 'bộ_trưởng giáo_dục': 26010,
 'giáo_dục chúng_tôi': 92114,
 'chúng_tôi không': 39241,
 'không mang': 135706,
 'mang con_trẻ': 166218,
 'con_trẻ ra': 50090,
 'ra thí_nghiệm': 238241,
 'cựu_binh': 72465,
 'si_tình': 247113,
 'cô': 56956,
 'gái': 99292,
 'tuổi': 304509,
 'vẫn': 338252,
 'đi': 358713,
 'tìm': 309148,
 'cựu_binh mỹ': 72472,
 'mỹ si_tình': 183545,
 'si_tình cô': 247116,
 'cô gái': 56993,
 'gái việt': 99947,
 'việt tuổi': 329810,
 'tuổi năm': 304925,
 'năm vẫn': 212045,
 'vẫn đi': 338657,
 'đi tìm': 359182,
 'đạo_diễn': 372885,
 'last': 147848,
 'tango': 261073,
 'in': 125567,
 'paris': 218830,
 'qua_đời': 231200,
 'đạo_diễn last': 372920,
 'last tango': 147850,
 'tango in': 261074,
 'in paris': 125619,
 'paris qua_đời': 218851,
```

```
 'paris qua_doi': 218891,
 'gia_đình': 89230,
 'dừng': 83756,
 'việc': 328770,
 'kiếm': 140638,
 'đônăm': 368207,
 'học': 120402,
 'gia_đình dừng': 89297,
 'dừng việc': 83863,
 'việc kiếm': 328928,
 'kiếm triệu': 140705,
 'triệu đônăm': 292026,
 'đônăm con': 368208,
 'con được': 49144,
 'được học': 370463,
 'học tiếng': 120751,
 'tiếng việt': 285117,
 'nguyên_nhân': 190914,
 'con_người': 49583,
 'nói_lắp': 210376,
 'nguyên_nhân khiến': 190941,
 'khiến con_người': 130930,
 'con_người nói_lắp': 49644,
 'ông': 355775,
 'park': 218861,
 'thất_vọng': 276301,
 'tuột': 305351,
 'chiến_thắng': 31923,
 'malaysia': 165934,
 'ông park': 355912,
 'park thất_vọng': 218939,
 'thất_vọng khi': 276313,
 'khi tuột': 130443,
 'tuột chiến_thắng': 305353,
 'chiến_thắng trước': 31999,
 'trước malaysia': 297197,
 'báo': 13176,
 'philippines': 220048,
 'còn': 56514,
 'phép': 223545,
 'màu': 171970,
 'đình': 366327,
 'báo philippines': 13256,
 'philippines không': 220104,
 'không còn': 135282,
 'còn phép': 56658,
 'phép màu': 223599,
 'màu mỹ': 172024,
 'mỹ đình': 183950,
 'các': 53428,
 'nàng': 208887,
 'hàn': 112512,
 'khổ': 138587,
 'chồng': 44333,
 'quá': 233119,
 'chiều': 32259,
 'các nàng': 53762,
 'nàng dâu': 208900,
```

```
 'dau nan': 78978,
 'hàn khổ': 112581,
 'khổ mẹ': 138603,
 'mẹ chồng': 178737,
 'chồng quá': 44796,
 'quá chiều': 233129,
 'chiều con_trai': 32266,
 'dân_chơi': 78683,
 'tiệc': 287325,
 'dân_chơi mở': 78693,
 'mở tiệc': 182090,
 'tiệc ma_túy': 287361,
 'tháng': 270609,
 'kết_luận': 144931,
 'thanh_tra': 263487,
 'xuất_khẩu': 348312,
 'gạo': 101237,
 'giữa tháng': 97623,
 'tháng kết_luận': 270706,
 'kết_luận thanh_tra': 144945,
 'thanh_tra xuất_khẩu': 263539,
 'xuất_khẩu gạo': 348331,
 'học_bổng': 120879,
 'tổng_chi_phí': 320433,
 'the': 264091,
 'read': 239237,
 'school': 246020,
 'học_bổng tổng_chi_phí': 120999,
 'tổng_chi_phí tại': 320434,
 'tại the': 315178,
 'the read': 264161,
 'read school': 239238,
 'school anh': 246021,
 'babylift': 5112,
 'gửi': 103138,
 'người phụ_nữ': 194026,
 'phụ_nữ babylift': 228530,
 'babylift tìm': 5114,
 'tìm được': 309378,
 'được mẹ': 370635,
 'mẹ việt': 179133,
 'việt năm': 329644,
 'năm gửi': 211462,
 'gửi đi': 103254,
 'đi mỹ': 359006,
 'sáng_chế': 252503,
 'túi': 312741,
 'hỗ_trợ': 123385,
 'phân_loại': 223434,
 'rác': 241058,
 'học_sinh sáng_chế': 121462,
 'sáng_chế túi': 252535,
 'túi hỗ_trợ': 312782,
 'hỗ_trợ phân_loại': 123495,
 'phân_loại rác': 223449,
 'hóa': 115985,
 'trang': 289839,
 'halloween': 104172,
```

```
 'kẹt': 143945,
 'răng': 241795,
 'giả': 93827,
 'miệng': 169029,
 'hóa trang': 116021,
 'trang halloween': 289879,
 'halloween người': 104177,
 'phụ_nữ kẹt': 228678,
 'kẹt răng': 143969,
 'răng giả': 241807,
 'giả miệng': 93940,
 'thông': 272906,
 'thật': 277351,
 'thông thật': 272928,
```

Data visualize

- Giảm chiều dữ liệu (truncated SVD) của tập train còn 2 components. Edit lại hàm plot_LSA từ link github để plot giá trị ngữ nghĩa của đặc headline-idf giữa các topics. Đồ thị bên dưới cho thấy sử dụng đặc trưng idf, giá trị idf của từ vựng thuộc các topics có sự phân biệt nhưng không quá rõ ràng, vẫn còn nhiều giá trị gần giống nhau giữa các chủ đề.
- Đối với chủ đề Giáo Dục (chủ đề 7), do headline thu thập về từ vnexpress có một số bài học tiếng anh ("Ý nghĩa thành ngữ 'apple of my eye'", "Trắc nghiệm phân biệt 'see', 'look', 'watch'",...) nên có rất nhiều giá trị từ vựng khác với những chủ đề còn lại.

```
 'bực_mình trên': 18015,
```

```
#github: https://github.com/hundredblocks/concrete_NLP_tutorial/blob/master/NLP_notebook.ipyr
from sklearn.decomposition import TruncatedSVD
import matplotlib
import matplotlib.patches as mpatches


def plot_LSA(test_data, test_labels, savepath="PCA_demo.csv", plot=True):
        lsa = TruncatedSVD(n_components=2)
        lsa.fit(test_data)
        lsa_scores = lsa.transform(test_data)
        color_mapper = {label:idx for idx,label in enumerate(set(test_labels))}
        color_column = [color_mapper[label] for label in test_labels]
        colors = ['orange','blue','black', 'red', 'green', 'pink', 'purple', 'yellow']
        if plot:
            plt.scatter(lsa_scores[:,0], lsa_scores[:,1], s=8, alpha=.6, c=test_labels, cmap=
            orange_patch = mpatches.Patch(color='orange', label='1')
            blue_patch = mpatches.Patch(color='blue', label='2')
            black_patch = mpatches.Patch(color='black', label='3')
            red_patch = mpatches.Patch(color='red', label='4')
            green_patch = mpatches.Patch(color='green', label='5')
            pink_patch = mpatches.Patch(color='pink', label='6')
            purple_patch = mpatches.Patch(color='purple', label='7')
            yellow_patch = mpatches.Patch(color='yellow', label = '8')
            plt.legend(handles=[orange_patch, blue_patch, black_patch, red_patch, green_patch
```

```
fig = plt.figure(figsize=(16, 16))
plot_LSA(X_train_idf[:20000], y_train[:20000])
plt.show()
```

Train models and Evaluate test result

- Chọn ra 5 classifier từ sklearn: LinearSVC, MultinomialNB, KNN, Logistic Regression và SGD classifier.
- Score trong classification thường là mean accuracy: số lần đoán trúng / tổng số lần đoán.
- Lưu các model đã train vào dictionary.

```
clf = make_pipeline(StandardScaler(with_mean=False),LinearSVC(max_iter=2000)).fit(X_train_idf
tfidf_model['LinearSVC'] = clf
print("Performance on train set:{}\nPerformance on test set:{}".format(clf.score(X_train_idf,
```

```
Performance on train set:0.999982924226254
Performance on test set:0.8718050879534633
c:\users\jundevice\appdata\local\programs\python\python37\lib\site-packages\sklearn\svm
  "the number of iterations.", ConvergenceWarning)
```

```
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(X_train_idf, y_train)
tfidf_model['MultinomialNB'] = clf
print("Performance on train set:{}\nPerformance on test set:{}".format(clf.score(X_train_idf,
```

```
Performance on train set:0.935752401280683
Performance on test set:0.8709883060740682
```

```
clf = KNeighborsClassifier(n_neighbors=15).fit(X_train_idf, y_train)
tfidf_model['KNN'] = clf
print("Performance on train set:{}\nPerformance on test set:{}".format(clf.score(X_train_idf,
```

```
Performance on train set:0.8465656350053362
Performance on test set:0.8200689284220173
```

```
clf = make_pipeline(StandardScaler(with_mean=False),SGDClassifier(loss='perceptron',max_iter=
tfidf_model['SGDClassifier'] = clf
print("Performance on train set:{}\nPerformance on test set:{}".format(clf.score(X_train_idf,
```

```
Performance on train set:0.999974386339381
Performance on test set:0.7839113891268402
```

```
clf = make_pipeline(StandardScaler(with_mean=False), LogisticRegression(max_iter=2000)).fit(X
tfidf_model['LogisticRegression'] = clf
print("Performance on train set:{}\nPerformance on test set:{}".format(clf.score(X_train_idf,
```

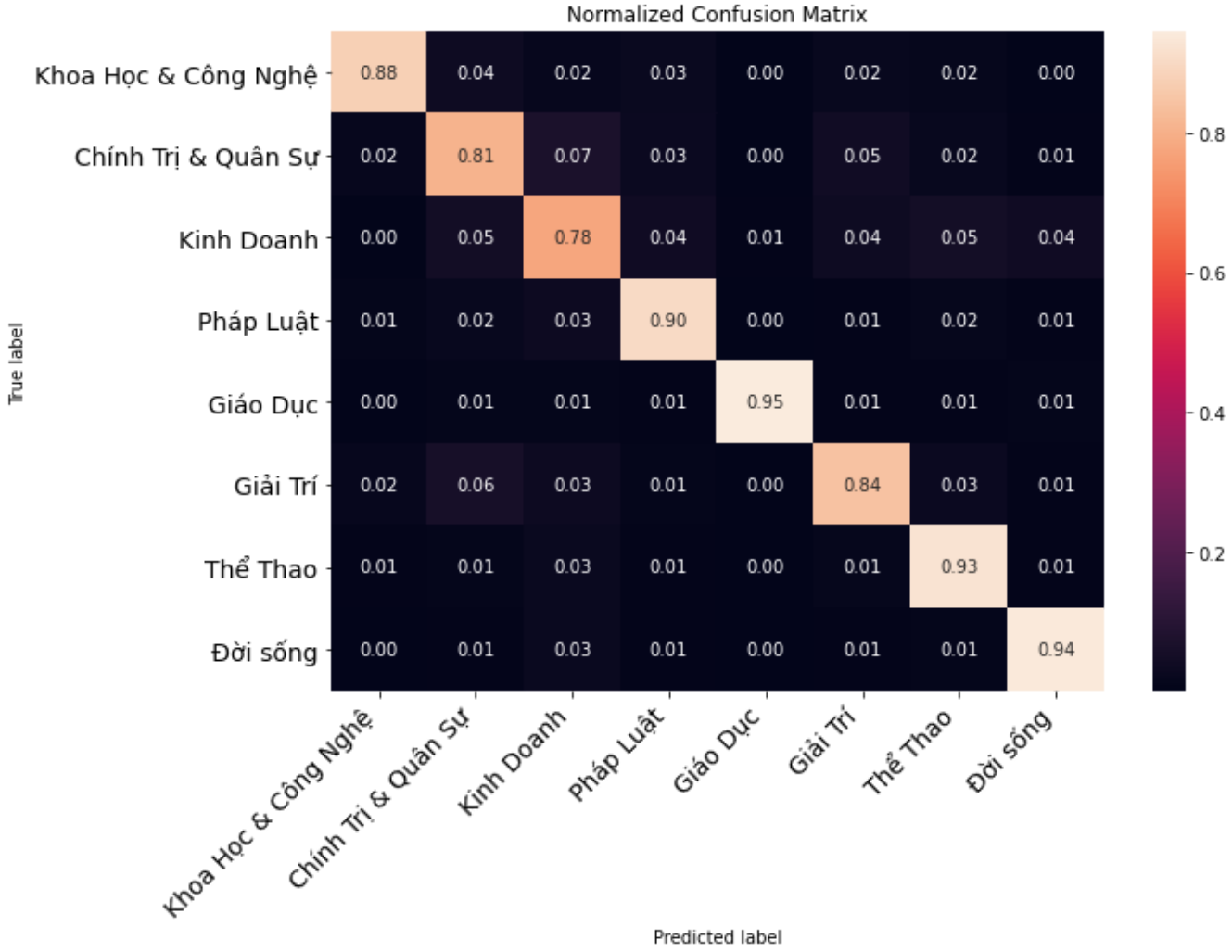
```
Performance on train set:0.999982924226254
```

Model	Multi. NB	Logictic Regr.	SGD Clas.	Linear SVC	KNN
Score	0.8709	**0.877**	0.783	0.8718	0.820

Ta thấy Logistic Regression cho ra score cao nhất (0.877). Đây không phải một kết quả cao và có thể được cải thiện nhờ nếu tìm được các parameter phù hợp hơn. Bởi vì chưa có nhiều phân biệt trong giá trị ngữ nghĩa giữa các từ vựng thuộc các chủ đề khác nhau nên model đoán nhầm trong một số trường hợp.

Confusion matrix thể hiện tỉ lệ đoán nhầm giữa các chủ đề với nhau. Hàm in confusion matrix được tham khảo và edit từ link github bên dưới.

```
#Based on shaypal5's gist: https://gist.github.com/shaypal5/94c53d765083101efc0240d776a23823
from sklearn.metrics import confusion_matrix
import numpy as np

def print_confusion_matrix(confusion_matrix, class_names, figsize = (10,7), fontsize=14, norm
    if normalize:
        confusion_matrix = confusion_matrix.astype('float') / confusion_matrix.sum(axis=1)[:,
        fmt = '.2f'
        title = 'Normalized Confusion Matrix'
    else:
        fmt = 'd'
        title = 'Confusion Matrix'

    df_cm = pd.DataFrame(confusion_matrix, index=class_names, columns=class_names)
    fig = plt.figure(figsize=figsize)
    heatmap = sns.heatmap(df_cm, annot=True, fmt=fmt)

    heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', font
    heatmap.xaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=45, ha='right', for
    heatmap.set_ylabel('True label')
    heatmap.set_xlabel('Predicted label')
    heatmap.set_title(title)
    return fig

conf_mat = confusion_matrix(y_test, clf.predict(X_test_idf))

# Get some readable labels
labels = [x for x in set(data['topic_names'])]
ax = print_confusion_matrix(conf_mat, labels, normalize=True)
plt.show()
```



Hyper parameter tuning

Sử dụng thuật toán GridSearch (giống như brute-force search qua các tập parameter định sẵn để tìm ra bộ parameter tốt nhất) vào 3 model có kết quả tốt nhất. Tuy nhiên, thời gian thực thi quá lâu nên tạm thời chưa thể đưa ra những models với score tốt hơn.

```
from sklearn.model_selection import GridSearchCV
tuned_parameters = [
  {'kernel': ['rbf','sigmoid', 'linear'], 'gamma': [1e-3, 1e-4, 1e-2], 'C': [1, 10]}]

# Objective metrics
scores = ['precision']


clf = GridSearchCV(SVC(), tuned_parameters, cv=3, scoring='%s_macro' % scores[0], n_jobs=-1,
clf.fit(X_train_idf, y_train)

print("Best Hyperparameters found are:")
print(clf.best_params_)
```

```
print("Grid scores are:")

means = clf.cv_results_['mean_test_score']
for mean,params in zip(means, clf.cv_results_['params']):
  print("%0.3f for %r" % (mean, params))
```

```
Fitting 3 folds for each of 18 candidates, totalling 54 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
```

```
tuned_parameters = {
     'n_neighbors': [5,11,15],
     'weights': ['uniform', 'distance'],
     'metric': ['euclidean', 'manhattan']
}

# Objective metrics
scores = ['precision']

clf = GridSearchCV(
    KNeighborsClassifier(),
    tuned_parameters,
    verbose = 10,
    cv = 2,
    n_jobs = -1,
    scoring = scoring='%s_macro' % scores[0]
)

clf.fit(X_train_idf, y_train)

print("Best Hyperparameters found are:")
print(clf.best_params_)

print("Grid scores are:")

means = clf.cv_results_['mean_test_score']
for mean,params in zip(means, clf.cv_results_['params']):
  print("%0.3f for %r" % (mean, params))


tuned_parameters = {"C":np.logspace(-3,3,7), "penalty":["l1","l2"]}
# Objective metrics
scores = ['precision']

clf = GridSearchCV(
    LogisticRegression(),
    tuned_parameters,
    verbose = 2,
    cv = 3,
    n_jobs = -1,
    scoring = scoring='%s_macro' % scores[0]
)
```

```
clf.fit(X_train_idf, y_train)

print("Best Hyperparameters found are:")
print(clf.best_params_)

print("Grid scores are:")

means = clf.cv_results_['mean_test_score']
for mean,params in zip(means, clf.cv_results_['params']):
  print("%0.3f for %r" % (mean, params))
```

Lưu lại các models để thực hiện demo phân loại.

```
import joblib
classifiers = ['KNN', 'LinearSVC', 'LogisticRegression', 'SGDClassifier', 'MultinomialNB']
for clf_ in classifiers:
    joblib.dump(tfidf_model[clf_], clf_+'.sav')
```

Lưu lại tập corpus

```
import pickle
with open('vectorizer.pickle', 'wb') as handle:
    pickle.dump(vectorizer, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

Demonstration (test with other sources)

- Other sources bao gồm genk, techz, báo phụ nữ và báo thanh niên
- Tải thêm 994 tiêu đề bài báo để phân loại với model đã chọn

```
import requests
from lxml import html
import json
import joblib
import pickle

#structure for JSON files
def createData():
    data = {}
    data["topic"] = ""
    data["headline"] = ""
    return data

#Append topic + headline at the end of the file
def writeJSON(data, articles,topics, writeFile):
```

```
    for i in range(len(articles)):
        data["topic"] = topics
        data["headline"] = articles[i]
        with open(writeFile, "a+", encoding='utf8') as outfile:
            json.dump(data, outfile,ensure_ascii=False)
            outfile.write("\n")
        outfile.close()
    print('Done writing!')

#Send GET HTTP requests to server and receive [response 200]
#Parse html tree to get headline data
def getHeadline(link, headline_Xpath):
    h = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like
    response = requests.get(link, headers=h)
    print(f'{link}: {response}')
    tree = html.fromstring(response.content)
    return tree.xpath(headline_Xpath), response.status_code



for i in range(1,10):
    thanhnien, _ = getHeadline(f'https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-{i}.htm
                               '//*[@class="title"]/@title')
    thanhnien_data = createData()
    writeJSON(thanhnien_data, thanhnien, 5, 'input.json')

    phunu_headlines, _ = getHeadline(f"https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuye
                                     '//*[@class="b-content"]/a/@title')
    phunu_data = createData()
    writeJSON(phunu_data, phunu_headlines, 3, 'input.json')

    thanhnien_headlines, _ = getHeadline(f"https://thanhnien.vn/thoi-su/chinh-tri/trang-{i}.h
                                     '//*[@class="relative"]/article/h2/a/@title')
    thanhnien_data = createData()
    writeJSON(thanhnien_data, thanhnien_headlines, 1, 'input.json')

    phunu_headlines, _ = getHeadline(f"https://www.phunuonline.com.vn/giai-tri/?p={i}",
                                     '//*[@class="b-content"]/a/@title')
    phunu_data = createData()
    writeJSON(phunu_data, phunu_headlines, 8, 'input.json')

    phunu_headlines, _ = getHeadline(f"https://www.phunuonline.com.vn/giao-duc/?p={i}",
                                     '//*[@class="b-content"]/a/@title')
    phunu_data = createData()
    writeJSON(phunu_data, phunu_headlines, 7, 'input.json')

    tuoitre_headlines, _ = getHeadline(f"https://tuoitre.vn/phap-luat/trang-{i}.htm",
                                               '//*[@class="title-news"]/a/@title')
    tuoitre_data = createData()
    writeJSON(tuoitre_data, tuoitre_headlines, 4, 'input.json')
```

```
    tuoitre_headlines, _ = getHeadline(f"https://tuoitre.vn/kinh-doanh/trang-{i}.htm",
                                                '//*[@class="title-news"]/a/@title')
    tuoitre_data = createData()
    writeJSON(tuoitre_data, tuoitre_headlines, 2, 'input.json')


tuoitre_headlines, _ = getHeadline(f"https://congnghe.tuoitre.vn/",
                                                '//*[@class="title-news"]/a/@title')
tuoitre_data = createData()
writeJSON(tuoitre_data, tuoitre_headlines, 6, 'input.json')

genk_headlines, _ = getHeadline(f"https://genk.vn/",
                                                '//*[@class="knswli-title"]/a/@title')
genk_data = createData()
writeJSON(genk_data, genk_headlines, 6, 'input.json')
```

```
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-1.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=1: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-1.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=1: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=1: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-1.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-1.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-2.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=2: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-2.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=2: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=2: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-2.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-2.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-3.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=3: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-3.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=3: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=3: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-3.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-3.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-4.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=4: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-4.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=4: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=4: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-4.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-4.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-5.html: <Response [200]>
```

```
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=5: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-5.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=5: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=5: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-5.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-5.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-6.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=6: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-6.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=6: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=6: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-6.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-6.htm: <Response [200]>
Done writing!
https://thanhnien.vn/the-thao/bong-da-viet-nam/trang-7.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/hon-nhan-gia-dinh/chuyen-nha/?p=7: <Response [200]>
Done writing!
https://thanhnien.vn/thoi-su/chinh-tri/trang-7.html: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giai-tri/?p=7: <Response [200]>
Done writing!
https://www.phunuonline.com.vn/giao-duc/?p=7: <Response [200]>
Done writing!
https://tuoitre.vn/phap-luat/trang-7.htm: <Response [200]>
Done writing!
https://tuoitre.vn/kinh-doanh/trang-7.htm: <Response [200]>
```

```
input = pd.read_json('input.json', lines=True)
input = input.sample(frac=1)
input['topic_names'] = [None] * len(input)
input.loc[input.topic==1, 'topic_names'] = "Chính Trị & Quân Sự"
input.loc[input.topic==2, 'topic_names'] = "Kinh Doanh"
input.loc[input.topic==3, 'topic_names'] = "Đời sống"
input.loc[input.topic==4, 'topic_names'] = "Pháp Luật"
input.loc[input.topic==5, 'topic_names'] = "Thể Thao"
input.loc[input.topic==6, 'topic_names'] = "Khoa Học & Công Nghệ"
input.loc[input.topic==7, 'topic_names'] = "Giáo Dục"
input.loc[input.topic==8, 'topic_names'] = "Giải Trí"
cols = ["headline", "topic", "topic_names"]
input = input[cols].reset_index(drop=True)
input = input.drop_duplicates(subset=['headline'])
input[:10]
```

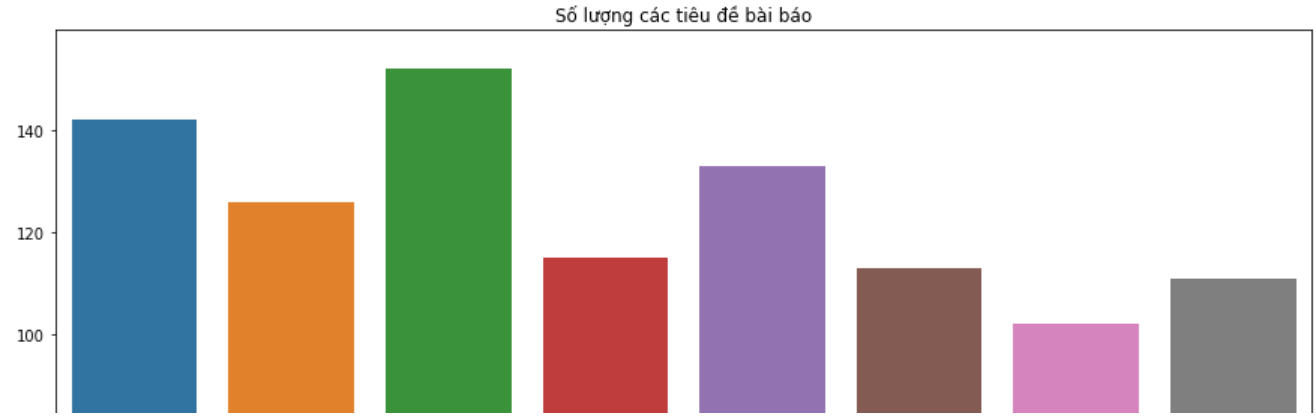
	headline	topic	topic_names
0	Không để tham nhũng, lãng phí đất nông nghiệp	1	Chính Trị & Quân Sự
1	Truyền hình báo Thanh Niên trực tiếp chương trình bình luận “Tiêu điểm bóng đá”	5	Thể Thao
2	Masan chờ "át chủ bài" The CrownX ngăn VinCommerce bớt lỗ	2	Kinh Doanh
3	HLV Hứa Hiền Vinh và học trò bị cấm 2 trận, Võ Văn Huy thoát án phạt	5	Thể Thao
4	Con virus dạy chúng ta điều gì?	3	Đời sống
5	Hủy tạm giam cựu bí thư Bến Cát	4	Pháp Luật
6	Điều tra “Ellen show” sau tố cáo phân biệt chủng tộc	8	Giải Trí
7	Quảng Nam: Ông Bùi Ngọc Ảnh giữ chức Chủ tịch UBND TP.Tam Kỳ	1	Chính Trị & Quân Sự

```
input['headline'] = clean(input['headline'])
input['headline'] = input['headline'].apply(tokenize)
input.shape
```

```
(994, 3)
```

```
import seaborn as sns
plt.figure(figsize=(15,10))
sns.countplot(input.topic_names).set_title('Số lượng các tiêu đề bài báo')
```

```
Text(0.5, 1.0, 'Số lượng các tiêu đề bài báo')
```




```
def __predict(input, model='SGDClassifier.sav', vectorizer='vectorizer.pickle'):
    classifier = joblib.load(model)
    with open(vectorizer, 'rb') as handle:
      vocab = pickle.load(handle)
    input = vocab.transform(input)
    return classifier.predict(input)
```

Đánh giá chung

- Khác với khi validate trên tập test cùng một source với tập train, đổi source để tải tiêu đề thì accuracy giảm đi khá nhiều.
- Quan sát confusion matrix bên dưới, ta thấy 3 chủ đề Giải trí, đời sống và giáo dục có accuracy thấp nhất.
- Để hiểu rõ hơn, khi nhìn vào danh sách những tiêu đề bị phân loại sai, có rất nhiều từ vựng của chủ đề này có thể đúng với chủ đề khác và còn có những từ vựng không có trong corpus được train.
- Ví dụ:

404	" tình_yêu và tham_vọng " càng dài càng chán	Giải Trí	8	3
18	tvb có người mắc covid	Giải Trí	8	3

```
prediction = __predict(list(input['headline']), 'LinearSVC.sav')
num_correct = 0
for i in range(len(prediction)):
    if prediction[i] == list(input['topic'])[i]:
        num_correct += 1

print(f'\n\n----> GET {num_correct}/{len(input)} ({(num_correct/len(input))*100}%) CORRECT!\r
```

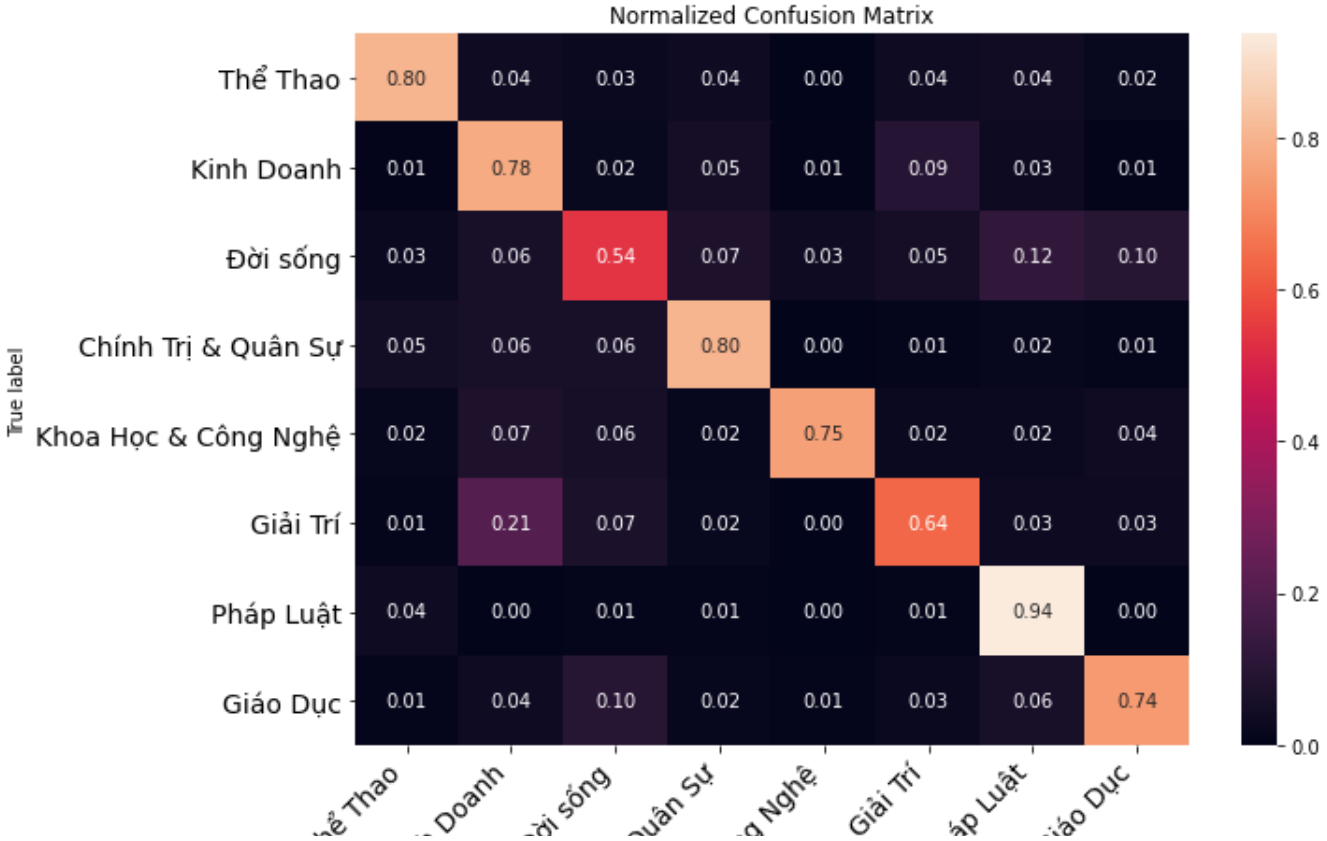
```
      /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
        UserWarning)
      /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
        UserWarning)
      /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
        UserWarning)
      /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
        UserWarning)


      ----> GET 739/994 (74.34607645875252%) CORRECT!

      /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
```

```
prediction = __predict(list(input['headline']), 'LogisticRegression.sav')
num_correct = 0
for i in range(len(prediction)):
    if prediction[i] == list(input['topic'])[i]:
        num_correct += 1

print(f'\n\n----> GET {num_correct}/{len(input)} ({(num_correct/len(input))*100}%) CORRECT!\r
```

```
 ⮡  /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
      UserWarning)
    /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
      UserWarning)
    /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
      UserWarning)
    /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
      UserWarning)


    ----> GET 750/994 (75.45271629778671%) CORRECT!


    /usr/local/lib/python3.6/dist-packages/sklearn/base.py:318: UserWarning: Trying to unpi
      UserWarning)
```

```
conf_mat = confusion_matrix(input['topic'], prediction)

# Get some readable labels
labels = [x for x in set(input['topic_names'])]
ax = print_confusion_matrix(conf_mat, labels, normalize=True)
plt.show()
```

```
 ⮡
```




```
input['prediction'] = prediction
input = input[["headline", "topic_names", "topic", "prediction"]]
input[input['topic'] != input['prediction']]
```

	headline	topic_names	topic	prediction
1	truyền_hình báo thanh_niên trực_tiếp chương_trình bình_luận “ tiêu_điểm bóng_đá ”	Thể Thao	5	7
4	con virus dạy chúng_ta điều gì	Đời sống	3	6
8	bà rịa vũng_tàu brvt lần đầu_tiên việt nam in vé qr code chống giả	Thể Thao	5	6
11	hướng_dẫn tải video facebook từ nhóm kín	Khoa Học & Công Nghệ	6	4
17	kỷ_luật quan_chức quảng_ngãi liên_quan đến dự_án đất_đai bổ_nhiệm	Pháp Luật	4	1
18	tvb có người mắc covid	Giải Trí	8	3
20	một trọng_tài được dí thổi liền trận ép một đội bóng công_bằng ở đâu	Thể Thao	5	3
23	nghề tay_trái của sao thể_thao kiếm thủ số đông nam á bán trà sữa online	Thể Thao	5	2
29	viếng ông trần quốc hương trong niềm tiếc_thương vô_hạn	Chính Trị & Quân Sự	1	8
33	sao phải thi vậy mẹ	Đời sống	3	8
39	phan_thiết hướng đến xây_dựng thành_phố văn_minh thân_thiện nghĩa_tình	Chính Trị & Quân Sự	1	2
44	điều lẽ_ra phải làm từ lâu	Đời sống	3	5
45	filip nguyễn ‘ tôi vẫn chưa đủ điều_kiện nhập quốc_tịch việt_nam ’	Thể Thao	5	8
48	nạn_nhân vô_hình của đại_dịch	Đời sống	3	7
50	gia_tài của ba không có ánh hào_quang	Đời sống	3	7
51	trắc_trở học_hành vì covid cơ_hội chia_sẻ bất_an cùng nhau	Đời sống	3	6
63	công_an hà_nội thu_giữ thiết_bị công_nghệ_cao để gian_lận khi thi	Pháp Luật	4	2
69	những viên ngọc sáng được chờ_đợi của u việt_nam	Thể Thao	5	3
72	cậu bé tuổi chết vì bị chích kim có virút hiv	Đời sống	3	4
78	việt_nam mỗi ngày có triệu lượt mua hàng trực_tuyến	Kinh Doanh	2	4
86	cuộc_gọi dựng tóc_gáy không trả nợ sau tiếng sẽ khởi_tố	Khoa Học & Công Nghệ	6	8
89	đạo_diễn chu thiện chỉ một câu “ hết_sức hết_lòng ”	Giải Trí	8	7
107	qua những ngày biến_động tựa vào mình để an	Đời sống	3	2
109	cúp quốc_gia lại lùi lịch thêm ngày vì than quảng_ninh bị	Thể Thao	5	2

	cách_ly	Thể Thao	5	2
112	nhiều người lại bị tin giả lừa ‘ tôi không đồng_ý cho facebook chia_sẻ ảnh hoặc tin ’	Khoa Học & Công Nghệ	6	4
115	ông trump ra sắc_lệnh mới buộc bytedance rút khỏi mỹ trong ngày	Khoa Học & Công Nghệ	6	2
122	vụ đường nhuệ thủ thiêm vai_trò của đoàn đại_biểu quốc_hội ở đâu	Chính Trị & Quân Sự	1	4
128	dalatmilk tinh_khiết ‘ di_sản từ cao_nguyên ’	Kinh Doanh	2	6
131	bao_giờ những cây cầu ngàn tỉ nối nhịp cao_tốc thông xe	Kinh Doanh	2	6
132	công_an các tỉnh nghệ_an hà tĩnh quảng_bình có giám_đốc mới	Chính Trị & Quân Sự	1	7
136	atalanta sự trỗi dậy của nữ_thần	Thể Thao	5	6
138	bắt tạm giam giám_đốc người hàn lừa nhiều nhà đầu_tư hơn tỉ đồng	Kinh Doanh	2	4
144	chứng_kiến thảm_họa và ôm con chặt hơn	Đời sống	3	6
150	tháng tư về …	Đời sống	3	1
154	ba giải bóng_đá lớn nhất việt nam có nguy_cơ bị hủy hẳn trong năm nay	Thể Thao	5	2
160	tập cơ bụng đơn_giản và hiệu_quả nhất tại nhà	Khoa Học & Công Nghệ	6	3
163	báo_động hàm rồng trộm ‘ viếng ’ đại bản_doanh hagl bị rượt té chỏng_gọng	Thể Thao	5	6
171	đề_nghị ban_hành luật an_ninh kinh_tế do lo_ngại trung quốc thâu_tóm đất_đai	Chính Trị & Quân Sự	1	2
172	giữa u_ám dịch_bệnh cây táo lại nở hoa	Đời sống	3	2
173	nghỉ covid mạc hồng_quân gây choáng khi tậu biệt_thự siêu sang chục tỉ đồng	Thể Thao	5	4
174	những bộ phim “ giải_cứu ” mùa hè cho trẻ	Giải Trí	8	7
182	tôi là một người_mẫu và tôi biết rằng trí_tuệ nhân_tạo rồi sẽ lấy mất việc_làm của mình	Khoa Học & Công Nghệ	6	3
189	cổ giai “ chưa phải là hết ” “ phụ_nữ đừng tự giới_hạn bản_thân ”	Giải Trí	8	3
190	garmin mua lại firstbeat analytics	Khoa Học & Công Nghệ	6	2
194	phó_chủ_tịch nước ấn_tượng với sự phát_triển của miền tây xứ nghệ	Kinh Doanh	2	7
201	khi chị_em nói nói nhiều nhưng chưa bén	Giải Trí	8	2
203	bị ép quá mức chủ_nhân tiktok không thèm bán quyết đấu ông trump	Kinh Doanh	2	6

	quyết_đấu ông trump			
204	bí_quyết cho tổ_ấm hạnh_phúc và bình_yên	Đời sống	3	7
210	hướng_dẫn cách sử_dụng hashtag đúng và mang lại hiệu_quả cao nhất	Khoa Học & Công Nghệ	6	3
211	tháo_gỡ điểm nghẽn cho tphcm phát_triển	Chính Trị & Quân Sự	1	2
212	cách tản_nhiệt laptop đơn_giản mà hiệu_quả co không lo hại máy	Khoa Học & Công Nghệ	6	3
225	hlv lê huỳnh đức ngồi ghế một chân	Thể Thao	5	8
246	ngắm bạn gái mới của neymar đẹp tuyệt_trần không tì_vết	Thể Thao	5	3
251	chủ sở_hữu bản_quyền phát_sóng tuyển việt nam có_thể mất_trắng hàng chục triệu usd	Thể Thao	5	2
253	top bộ phim siêu anh_hùng marvel không thuộc mcu được đánh_giá cao nhất	Khoa Học & Công Nghệ	6	8
265	chiến_thắng bắt_nguồn từ lòng yêu nước nồng_nàn	Chính Trị & Quân Sự	1	3
268	đại_biểu quốc_hội đề_xuất ban_hành luật bảo_vệ người làm_việc tốt	Chính Trị & Quân Sự	1	2
273	clip ngắm trọn_vẻ đẹp iphone g đẹp không tì_vết	Khoa Học & Công Nghệ	6	3
285	“ siêu_thị má ” sau dịch	Đời sống	3	2
293	căn nhà nhỏ trong hẻm đường cộng hòa chứa toàn xì_gà cuba thuốc ba số	Pháp Luật	4	3
294	‘ ông bố quốc dân ’ quế ngọc hải lại đốn tim fan với tình_yêu cho sunny	Thể Thao	5	3
296	công_ty thu_mua máy_tính cũ tphcm giá cao thanh_lý cường phát	Khoa Học & Công Nghệ	6	2
298	làm sáng_tỏ hơn nữa lý_luận về chủ_nghĩa xã_hội	Chính Trị & Quân Sự	1	6
302	xin đừng bỏ phí tài_năng	Thể Thao	5	3
312	rộn_ràng đổi hàng bao tươi miễn_phí trên smartphone	Kinh Doanh	2	6
322	phạt nặng doanh_nghiệp dùng mủ trôm sản_xuất … nước yến	Pháp Luật	4	2
331	đi nhậu về thấy người đàn_ông đứng bên giường vợ mình nên tức_giận chém nhát	Pháp Luật	4	3
344	vleague lên các phương_án tái_xuất vào tháng	Thể Thao	5	8
345	ấm lòng những món quà gửi vùng dịch covid miền trung	Kinh Doanh	2	3
355	đang lo_lắng về lệnh cấm wechat apple bị ông trump dội một gáo nước_lạnh	Khoa Học & Công Nghệ	6	2

364	ca_sĩ sara lưu vết sẹo sinh mổ là minh chứng của hạnh_phúc	Đời sống	3	8
375	mô_hình chính_quyền đô_thị đà_nẵng kỳ phù_hợp với đô_thị nhỏ gọn	Chính Trị & Quân Sự	1	6
377	nhìn bằng trái_tim cho đời nhẹ_nhàng	Đời sống	3	7
382	bố ơi cố lên	Đời sống	3	8
393	chưa hết kỷ_luật một chuyên_viên sở gdđt tphcm được bổ_nhiệm làm hiệu_trưởng	Giáo Dục	7	1
401	hồ chí minh gương_mặt lớn của nhân_loại trong thế_kỷ xx	Chính Trị & Quân Sự	1	6
404	“ tình_yêu và tham_vọng ” càng dài càng chán	Giải Trí	8	3
407	đưa thanh_niên hàn quốc tuổi từ tphcm qua campuchia giá usd	Pháp Luật	4	3
408	duy mạnh sẽ bị mời làm_việc vì phát_ngôn lệch_lạc về chủ_quyền việt_nam	Giải Trí	8	1
441	ngân_hàng chạy_đua chuyển_đổi số trước giờ g	Kinh Doanh	2	6
446	sau microsoft twitter cũng muốn mua lại tiktok từ bytedance	Khoa Học & Công Nghệ	6	2
454	những máy_tính bảng giá rẻ bất_ngờ và đáng mua nhất năm	Khoa Học & Công Nghệ	6	2
455	cơ_hội vàng mua galaxy_note với ưu_đãi tốt nhất thị_trường từ viettel trị_giá đến triệu đồng	Khoa Học & Công Nghệ	6	2
459	tạm giữ gần triệu khẩu_trang y_tế không rõ nguồn_gốc	Kinh Doanh	2	4
487	doanh_nghiệp thái dựng trại điện gió lớn nhất ở lào để bán điện cho việt_nam	Kinh Doanh	2	6
496	đã tháo_dỡ khoảng diện_tích vi_phạm của gia trang resort	Pháp Luật	4	1
507	yêu_thương cả những người khác mình	Giải Trí	8	3
508	tphcm thay toàn_bộ giám_thị trở về từ đà_nẵng	Giáo Dục	7	3
510	nghề tay_trái của sao thể_thao đông triều với ước_mơ làm_chủ chuỗi nhà_hàng món quảng	Thể Thao	5	2
512	từ “ cát đỏ ” nhìn về phận đàn_bà	Giải Trí	8	4
520	cất_cánh sự_nghiệp sau vấp_ngã	Đời sống	3	8
535	mốc son chói_ngời trong lịch_sử nhân_loại	Chính Trị & Quân Sự	1	6
537	kiến_nghị điều_tra nhiều sai_phạm tại khu đô_thị quốc_tế đa_phước	Pháp Luật	4	2
540	làm phim thời giãn cách xã_hội	Giải Trí	8	2
546	kỷ_luật khiển_trách giám_đốc sở tư_pháp tỉnh lâm đồng	Pháp Luật	4	1

546	ky_luật khiển_trách giám_đốc sở tư_pháp tỉnh lâm đồng	Pháp Luật	4	1
547	biết_ơn kẻ dữ	Đời sống	3	8
552	nhật_ký trong tù đã đánh_thức lương_tâm của nhiều người mỹ	Chính Trị & Quân Sự	1	4
555	phú yên cách_chức thị_ủy_viên đối_với phó_chủ_tịch txsông cầu	Chính Trị & Quân Sự	1	7
572	bác hồ vị lãnh_tụ có ảnh_hưởng lớn đối_với nhân_loại trong thế_kỷ xx	Chính Trị & Quân Sự	1	6
582	bị kiện đòi tỉ đồng evn nói_gì	Kinh Doanh	2	4
589	một ngày để yêu	Giải Trí	8	3
596	now nói gì về việc hàng trăm shipper quây kín trụ_sở công_ty phản_đối chính_sách mới	Khoa Học & Công Nghệ	6	2
597	vừa lên chức bố phan văn đức lóng_ngóng khi chăm_sóc con_gái đầu_lòng	Thể Thao	5	3
600	ngân_hàng thế_giới việt nam là ngôi_sao sáng trong bầu_trời tăm_tối	Kinh Doanh	2	5
602	những chung_cư vắng người ở campuchia	Kinh Doanh	2	3
605	nghề tay_trái sao thể_thao lê đức lương xây_dựng nhãn hàng thời_trang cho tỉnh nhà	Thể Thao	5	2
606	trục_xuất người đàn_ông trung quốc đi chui sang việt_nam để lấy vợ	Pháp Luật	4	3
625	microsoft theo_đuổi việc mua tiktok tiktok tố facebook đạo nhái	Kinh Doanh	2	6
630	back to school đã trở_lại acer giới_thiệu chương_trình ưu_đãi lớn nhất trong năm	Khoa Học & Công Nghệ	6	7
636	thấy gì từ clip cô gái xả rác trả_thù chủ_khách_sạn	Đời sống	3	4
639	xóm_làng ngỡ_ngàng vụ nữ bác_sĩ đầu_độc cháu nội bại não để giải_thoát	Pháp Luật	4	3
641	ít ai được như con dâu tui	Đời sống	3	2
650	xem miễn_phí những câu_chuyện nực_cười	Giải Trí	8	7
652	russell kirsch người phát_minh ra điểm_ảnh đặt nền_móng cho ngành nhiếp_ảnh số đã qua_đời	Khoa Học & Công Nghệ	6	7
653	là vai siêu phụ trong endgame nhưng anh_chàng này đã biết iron man sẽ hi_sinh từ năm trước cả một_số diễn_viên chính	Khoa Học & Công Nghệ	6	8
663	có ngoại_binh từng đến khám ở bệnh_viện c đội shb đà_nẵng xét_nghiệm covid	Thể Thao	5	4
665	vingroup xuất_khẩu linh_kiện máy thở đi mỹ và ireland	Khoa Học & Công Nghệ	6	2

671	vnpt đạt giải_thưởng tại stevie awards châu á thái_bình_dương	Kinh Doanh	2	6
672	sẵn_sàng đương_đầu với tai_ương	Đời sống	3	5
674	bắc kinh nói doanh_nghiệp nước_ngoài không muốn rời trung_quốc	Kinh Doanh	2	4
687	đủ chiêu trò lừa_đảo qua email	Pháp Luật	4	2
689	sẽ thông_báo thời_gian trả hành_lý khi kết_thúc chuyến bay	Kinh Doanh	2	6
702	những con_số “ biết nói ” của “ chưa phải là hết ”	Giải Trí	8	7
705	nữ nhà_văn mỹ lady borton tôi ngưỡng_mộ chủ_tịch hồ chí minh	Chính Trị & Quân Sự	1	8
709	lựa_chọn cho game thủ mobile và bạn trẻ yêu nhạc	Khoa Học & Công Nghệ	6	7
717	hồ chí minh một con_người diệu_kỳ cho mọi thời_đại	Chính Trị & Quân Sự	1	3
719	nông_nghiệp kết_hợp điện_áp mái thiệt_hại tiền tỉ vì thiếu hướng_dẫn	Kinh Doanh	2	6
738	lee hyori tiết_lộ kế_hoạch tương_lai sau khi ssak tan_rã	Giải Trí	8	7
743	bộ tttt buộc gỡ nội_dung khiêu_dâm netflix ngó lơ	Giải Trí	8	6
755	thị_trường chứng_kiến ipad pro chính hãng có mức giá thấp kỷ_lục	Khoa Học & Công Nghệ	6	2
758	ai giữ “ két sắt sức_khỏe ” trong gia_đình	Đời sống	3	4
762	không có bàn_tay mẹ mọi thứ rối nùi	Đời sống	3	7
763	năm ngày truyền_thống ngành tuyên_giáo tạo động_lực sáng_tạo cho trí_thức	Chính Trị & Quân Sự	1	7
772	bảo_hiểm_nhân_thọ triệu bệnh nằm viện không được bồi_thường vì không cần_thiết	Pháp Luật	4	2
779	tại_sao chủ_tịch ubnd tp hà_nội nguyễn đức chung bị tạm đình_chỉ công_tác	Pháp Luật	4	1
789	bệnh_viện chuẩn nhật bản trường chuẩn mỹ trong đô_thị ecopark	Kinh Doanh	2	7
794	tiktok dọa khởi_kiện ông trump bắc kinh cáo_buộc mỹ đàn_áp doanh_nghiệp trung_quốc	Khoa Học & Công Nghệ	6	1
797	bài_học từ rơm	Đời sống	3	7
816	oneplus chính_thức gia_nhập thị_trường việt_nam	Khoa Học & Công Nghệ	6	2
818	“ khu rừng bí_mật ” lên sóng với nhiều áp_lực	Giải Trí	8	3
829	giúp đại_bàng chim_sẻ cùng đi cao_tốc evfta	Kinh Doanh	2	6

833	sẽ giám_sát đặc_biệt ca_sĩ rất đáng lên_án duy mạnh	Pháp Luật	4	1
835	diệt virút … đố_kỵ	Đời sống	3	1
841	đại_tá lê hồng nam chính_thức ra_mắt công_an tphcm	Chính Trị & Quân Sự	1	4
843	công_ty lý hải khiếu_nại vụ ca_khúc gánh mẹ nói không phải bị_đơn	Pháp Luật	4	8
844	tập_đoàn hưng_thịnh đã trích tỉ đồng ủng_hộ phòng_chống dịch covid	Kinh Doanh	2	6
855	học_sinh vĩnh long nhận học_bổng ‘ ủng_hộ nông_sản việt ’	Kinh Doanh	2	7
861	tổng_lãnh_sự_quán mỹ đóng_cửa thành đô lo mất luôn đầu_tư ngoại	Kinh Doanh	2	7
862	chuyện ít biết về chiếc máy_cày bác hồ tặng cho xã vĩnh_kim	Chính Trị & Quân Sự	1	8
878	phật_giáo việt nam luôn đồng_hành cùng dân_tộc	Chính Trị & Quân Sự	1	7
883	ye ye của gia_đình phép_thuật khác lạ ngày trở_lại	Giải Trí	8	3
891	bác hồ với các nhà_báo ở paris năm	Chính Trị & Quân Sự	1	3
906	thanh_xuân của mẹ đâu rồi	Đời sống	3	8
912	vỡ_mộng “ soái ca ” “ tình_yêu cổ_tích ” vui thôi đừng tin quá	Giải Trí	8	3
935	bộ công_an có thêm thứ_trưởng	Chính Trị & Quân Sự	1	4
938	tiến_sĩ dương ngọc dũng “ chức_năng giáo_dục của gia_đình ngày_càng hạn_chế ”	Đời sống	3	7
940	bắt kho hàng nghi nhái giả thương_hiệu mỗi tháng bán hàng tỉ đồng	Pháp Luật	4	2
946	ông trump yêu_cầu bytedance thoái vốn tiktok trong ngày	Khoa Học & Công Nghệ	6	2
966	con_cái chúng_ta sẽ ra sao khi rời tổ	Đời sống	3	6
967	tphcm vẫn thực_hiện chế_độ lương thâm_niên với nhà_giáo	Giáo Dục	7	6
969	sài_gòn fc chung tay vì cộng_đồng trên nền nhạc ghen covy đá bay virus corona	Thể Thao	5	2
980	bài_toán chung haesoung và chiến_lược của tphcm	Thể Thao	5	1
989	bí_quyết thoát khỏi lo_lắng	Đời sống	3	7
1000	đăng_tải cách phòng_chống covid không đúng bị phạt triệu đồng	Pháp Luật	4	7

1001	cách_ly ngày ở khách_sạn phí trung_bình là triệu đồng	Kinh Doanh	2	7
1010	ca mắc covid ca bạch_hầu tỉnh đắk lắk rà_soát các thí_sinh thuộc diện f f	Giáo Dục	7	1
1017	làm cha tôi dạy con không được đánh phụ_nữ	Đời sống	3	4
1018	nghề tay_trái của sao thể_thao máu kinh_doanh của nguyễn tiến minh	Thể Thao	5	7
1023	đà_nẵng tạm giữ thêm khẩu_trang không rõ nguồn_gốc	Kinh Doanh	2	4
1024	hành_lý tình_yêu se duyên hay tạo “ sóng ”	Giải Trí	8	3
1029	thất_nghiệp mùa dịch cô_giáo mầm_non làm kênh youtube để rèn con	Đời sống	3	7
1030	tphcm phát_hiện triệu găng_tay đã qua sử_dụng tái_chế bán ra thị_trường	Pháp Luật	4	6
1031	thủ_tướng tiếp_tục tinh_thần chống dịch như chống giặc	Kinh Doanh	2	1
1032	mẹ trả lương cho con	Đời sống	3	7
1036	doanh_nhân không_thể buông tay trước đại_dịch	Kinh Doanh	2	3
1043	quá_độ lên chủ_nghĩa xã_hội bao_lâu có mấy chặng đường cần tiếp_tục làm rõ	Chính Trị & Quân Sự	1	4
1051	ngô_cẩn ngôn muốn thoát khỏi hào_quang cũ của nguỵ anh lạc	Giải Trí	8	3
1064	thu_giữ nhiều tai_nghe siêu nhỏ dùng gian_lận thi_cử	Pháp Luật	4	7
1078	huỳnh anh gây sốt khi ra sân làm ‘ quân xanh ’ cho quang hải	Thể Thao	5	8
1080	đà_nẵng xử_phạt trường_hợp vi_phạm_phòng_chống dịch covid	Pháp Luật	4	3
1081	cảm_ơn cha vợ	Đời sống	3	4
1082	tháng đầu năm doanh_thu nhiều thị_trường nước_ngoài của viettel tăng_trưởng con_số	Khoa Học & Công Nghệ	6	2
1088	đà nẵng tháng đầu năm kỷ_luật đảng_viên	Chính Trị & Quân Sự	1	7
1089	viettel được công_nhận là doanh_nghiệp việt nam có ảnh_hưởng lớn nhất châu á	Khoa Học & Công Nghệ	6	2
1090	về nhà sau trùng_trùng trắc_trở	Đời sống	3	8
1094	ca_sĩ tuấn hưng tuyên_bố giải_nghệ vì gia_đình lựa_chọn quá khôn_ngoan	Đời sống	3	8
1099	thế_giới bất_định_đề văn bắt người trẻ lắng_nghe	Giáo Dục	7	4
1101	kpop phục_hồi thần_tốc	Giải Trí	8	2
1103	thanh_tra mẫu_mực bóc_trần thực_trạng xã_hội	Giải Trí	8	4

1104	cô gái chạy dưới hàng cây hoa vàng	Đời sống	3	5
1106	học_sinh hôn nhau trong lễ bế_giảng trường_học không phải công_viên	Đời sống	3	7

Kết Luận

Hiệu quả:

- Chưa thể hiện được tính ổn định và hiệu quả để có thể tự động hóa một cách đáng tin cậy hoàn toàn bài toán đặt ra

Hướng phát triển:

- Chạy thêm parameter tuning cho cả model lẫn feature extraction để cải thiện accuracy
- Thay vì chỉ dùng headline, có thể lấy thêm một đoạn ngắn từ bài báo để cải thiện việc phân loại. Do tiêu đề có thể chưa thể hiện đầy đủ nội dung cần để phân loại chủ đề.
- Cải thiện bước tiền xử lí tiếng Việt
- Có thể phát triển thành bài toán tự động gắn news tag thay vì chỉ gắn chủ đề

	tuyệt vời ngôi_sao quảng_cáo thể_thao			
	[illegible]			
1202	lừa_đảo trúng thưởng chiêu cũ nhiều nạn_nhân mới	Kinh Doanh	2	4
1206	người đàn bà nơi " thành phố cách ly "	Đời sống	3	4