ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÁY HỌC

ĐỀ TÀI: PHÂN LOẠI CHỦ ĐỀ CỦA BÀI BÁO TIẾNG ANH

Lớp: CS114.K21

18521157 – Lê Trần Phúc Nguyên

18521093 – Đoàn Xuân Minh

18521377 – Phan Quang Tấn

TPHCM, ngày 14 tháng 8 năm 2020

MỤC LỤC

1 Tổng quan

1.1 Thông tin nhóm

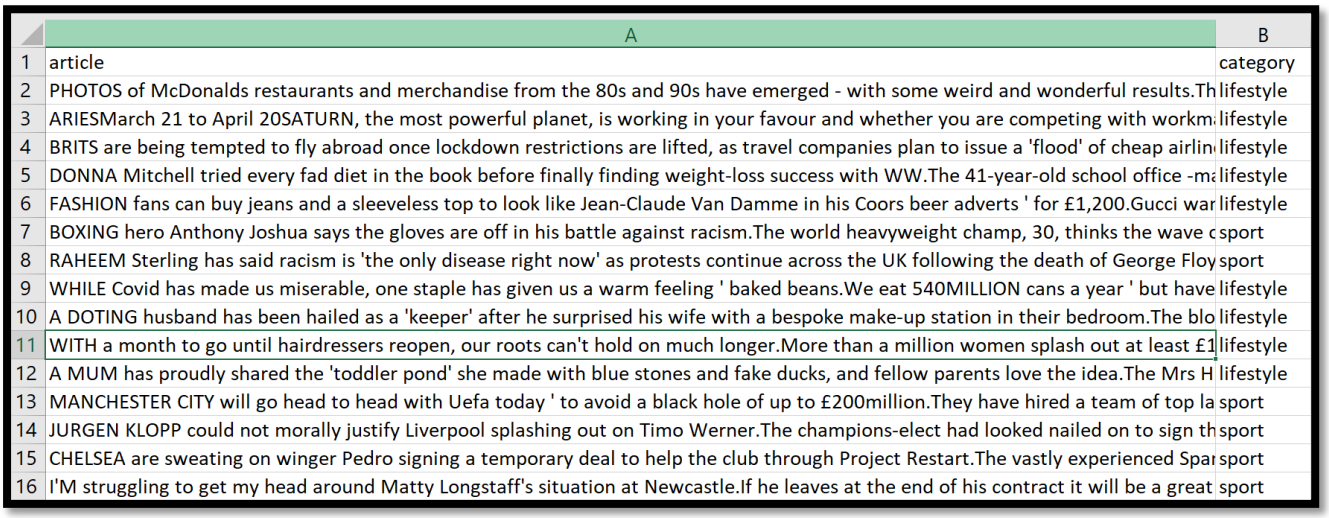
Mã số sinh viên	Họ và tên	Vai trò
18521093	Đoàn Xuân Minh	Thành viên
18521157	Lê Trần Phúc Nguyên	Thành viên
18521377	Phan Quang Tấn	Thành viên

1.2 Giới thiệu đề tài:

- Đặt vấn đề: hiện nay, việc cập nhật thông tin cũng như học tập thông qua phương thức đọc báo tiếng Anh trên các trang mạng xã hội, các trang báo điện tử khá phổ biến. Tuy nhiên với số lượng bài viết lớn cùng với nhiều chủ đề khác nhau sẽ làm tốn nhiều thời gian khi thực hiện phân loại bài báo bằng tay.
- Cách giải quyết: với vấn đề trên, cùng với việc xây dựng hệ thống phân loại chủ đề của bài báo sẽ giúp tiết kiệm thời gian phân loại.

1.3 Dataset:

- Dataset được thu thập từ: Thực hiện cào dữ liệu (crawl data) từ các trang báo tiếng Anh nổi tiếng như : The Sun, Dailymail, Telepraph.
- Tập dataset sau khi crawl gồm có hơn 2800 bài báo với 8 chủ đề khác nhau : femail, films, health, lifestyle, sport, technology, travel, tvshowbiz. Mỗi chủ đề gồm có khoảng 350 bài báo.
- Dữ liệu được chia thành 2 phần : training 80% và test 20%



	A	B
1	article	category
2	PHOTOS of McDonalds restaurants and merchandise from the 80s and 90s have emerged - with some weird and wonderful results.Th	lifestyle
3	ARIESMarch 21 to April 20SATURN, the most powerful planet, is working in your favour and whether you are competing with workm	lifestyle
4	BRITS are being tempted to fly abroad once lockdown restrictions are lifted, as travel companies plan to issue a 'flood' of cheap airlin	lifestyle
5	DONNA Mitchell tried every fad diet in the book before finally finding weight-loss success with WW.The 41-year-old school office -m	lifestyle
6	FASHION fans can buy jeans and a sleeveless top to look like Jean-Claude Van Damme in his Coors beer adverts ' for £1,200.Gucci war	lifestyle
7	BOXING hero Anthony Joshua says the gloves are off in his battle against racism.The world heavyweight champ, 30, thinks the wave c	sport
8	RAHEEM Sterling has said racism is 'the only disease right now' as protests continue across the UK following the death of George Floy	sport
9	WHILE Covid has made us miserable, one staple has given us a warm feeling ' baked beans.We eat 540MILLION cans a year ' but have	lifestyle
10	A DOTING husband has been hailed as a 'keeper' after he surprised his wife with a bespoke make-up station in their bedroom.The blo	lifestyle
11	WITH a month to go until hairdressers reopen, our roots can't hold on much longer.More than a million women splash out at least £1	lifestyle
12	A MUM has proudly shared the 'toddler pond' she made with blue stones and fake ducks, and fellow parents love the idea.The Mrs H	lifestyle
13	MANCHESTER CITY will go head to head with Uefa today ' to avoid a black hole of up to £200million.They have hired a team of top la	sport
14	JURGEN KLOPP could not morally justify Liverpool splashing out on Timo Werner.The champions-elect had looked nailed on to sign th	sport
15	CHELSEA are sweating on winger Pedro signing a temporary deal to help the club through Project Restart.The vastly experienced Spar	sport
16	I'M struggling to get my head around Matty Longstaff's situation at Newcastle.If he leaves at the end of his contract it will be a great	sport

1.4 Mô tả bài toán:

- Input: Nội dung của bài báo - đoạn văn bản của một bài báo.
- Output: Chủ đề của bài báo.

2 Nội dung đồ án

2.1 Chuẩn bị dữ liệu (Prepare Data)

- Sử dụng thư viện BeautifulSoup để cào dữ liệu từ các trang báo điện tử nổi tiếng: The sun, Dailymail, Telegraph.
- Code Crawl dữ liệu: https://github.com/tankien76/CS114.K21/blob/master/CrawlerTest.ipynb

2.2 Xử lý dữ liệu (Data Preprocessing)

- Với dữ liệu văn bản trong bài báo, chúng ta sử dụng nltk trong thư viện scikit-learn để xử lý ngôn ngữ tự nhiên bao gồm các thư viện và các hàm:
    - Các thư viện được sử dụng trong đó gồm: thư viện nltk để xử lý các stopword, tách từ trong dữ liệu văn bản:

```
[ ]  from nltk.corpus import stopwords
     from nltk.tokenize import word_tokenize
```

    - Loại bỏ các ký tự đặc biệt, dấu câu, số:

```
[ ]  features = features.apply(lambda s : re.sub('[^a-zA-Z]', ' ', s))
```

    - Loại bỏ các stopword – các từ không ảnh hưởng đến nghĩa của từ trong tiếng Anh nhưng xuất hiện nhiều lần (vd: a, the, an, …)

```
[ ]  stop_words = set(stopwords.words('english'))
     for i in range (len(features)):
       word_tokens = word_tokenize(features[i])

       filtered_sentence = [w for w in word_tokens if not w in stop_words]

       filtered_sentence = []

       for w in word_tokens:
           if w not in stop_words:
               filtered_sentence.append(w)
```

2.3 Trích xuất đặc trưng (Feature Engineering)

- Sử dụng TF-IDF trong thư viện sklearn để trích xuất đặc trưng của bài báo – các từ thường xuyên xuất hiện trong mỗi chủ đề:

```
[ ]  from sklearn.feature_extraction.text import TfidfVectorizer
     tfIdfVec = TfidfVectorizer()
     tfIdfVec.fit(features)
```

- Hàm .fit để học các từ thường xuyên xuất hiện trong features.

2.4 Xây dựng và huấn luyện Model

- Tiến hành train và test với 3 model:

Model	LinearSVC	MultinomialNB	LogisticRegression
Accuracy	87,15%	78,47%	85,06%

- Qua kết quả thu được từ tập test, các Model đều có độ chính xác tương đối tốt, độ chính xác từ 78% trở và Model LinearSVC có độ chính xác cao nhất trong các Model với 87,15%.

2.5 Đánh giá Model

- Tuy nhiên, để đánh giá performance của các model thì cách tính bằng accuracy không phản ánh chính xác dữ liệu được phân loại như thê nào vì vậy chúng ta cần 1 cách tính độ chính xác khác đó là **F1 Score.**
- Sử dụng Confusion Matrix của thư viện sklearn để trực quan hóa từng class được phân loại như thế nào.

2.5.1 Model LinearSVC

```
Confusion Matrix
[[56  5  6  7  0  0  4  8]
 [ 2 68  1  0  0  0  1  4]
 [ 4  0 47  3  1  2  1  3]
 [ 4  1  5 58  1  1  1  0]
 [ 0  0  0  0 54  0  0  0]
 [ 0  0  1  1  0 76  0  0]
 [ 3  1  1  0  0  0 76  0]
 [ 0  1  0  0  1  0  0 67]]
```

	precision	recall	f1-score	support
femail	0.81	0.65	0.72	86
films	0.89	0.89	0.89	76
health	0.77	0.77	0.77	61
lifestyle	0.84	0.82	0.83	71
sport	0.95	1.00	0.97	54
technology	0.96	0.97	0.97	78
travel	0.92	0.94	0.93	81
tvshowbiz	0.82	0.97	0.89	69
accuracy			0.87	576
macro avg	0.87	0.88	0.87	576
weighted avg	0.87	0.87	0.87	576

2.5.2 Model Multinomial Naive Bayes

```
Confusion Matrix
[[32  2  0 42  0  0  1  9]
 [ 4 63  1  1  0  1  1  5]
 [ 1  0 26 29  0  1  2  2]
 [ 1  0  1 68  0  1  0  0]
 [ 0  0  0  1 53  0  0  0]
 [ 0  0  2  4  0 72  0  0]
 [ 0  1  0  8  0  0 72  0]
 [ 0  1  0  1  0  0  1 66]]
```

	precision	recall	f1-score	support
femail	0.84	0.37	0.52	86
films	0.94	0.83	0.88	76
health	0.87	0.43	0.57	61
lifestyle	0.44	0.96	0.60	71
sport	1.00	0.98	0.99	54
technology	0.96	0.92	0.94	78
travel	0.94	0.89	0.91	81
tvshowbiz	0.80	0.96	0.87	69
accuracy			0.78	576
macro avg	0.85	0.79	0.79	576
weighted avg	0.85	0.78	0.78	576

2.5.3 Model Logistic Regression

Confusion Matrix

```
[[51  6  8  8  0  1  4  8]
 [ 3 69  1  0  0  0  1  2]
 [ 6  0 46  3  1  1  1  3]
 [ 3  0  7 56  1  2  1  1]
 [ 0  1  0  0 53  0  0  0]
 [ 0  0  1  1  0 76  0  0]
 [ 1  3  1  2  0  0 74  0]
 [ 0  2  0  1  0  0  1 65]]
```

	precision	recall	f1-score	support
femail	0.84	0.37	0.52	86
films	0.94	0.83	0.88	76
health	0.87	0.43	0.57	61
lifestyle	0.44	0.96	0.60	71
sport	1.00	0.98	0.99	54
technology	0.96	0.92	0.94	78
travel	0.94	0.89	0.91	81
tvshowbiz	0.80	0.96	0.87	69
accuracy			0.78	576
macro avg	0.85	0.79	0.79	576
weighted avg	0.85	0.78	0.78	576

2.6 Kết luận

2.6.1 Nhận xét:

- Các Model có độ chính xác tương đối tốt vì:
  - Số lượng tập test ít dẫn đến độ chính xác của model cao.
  - Xử lý dữ liệu kĩ càng
- Tuy độ chính xác của các model tối nhưng khi sử dụng các model để phân loại các bài báo mới thì kết quả giảm xuống vì xuất hiện các từ mới chưa xuất hiện trong data đã train.

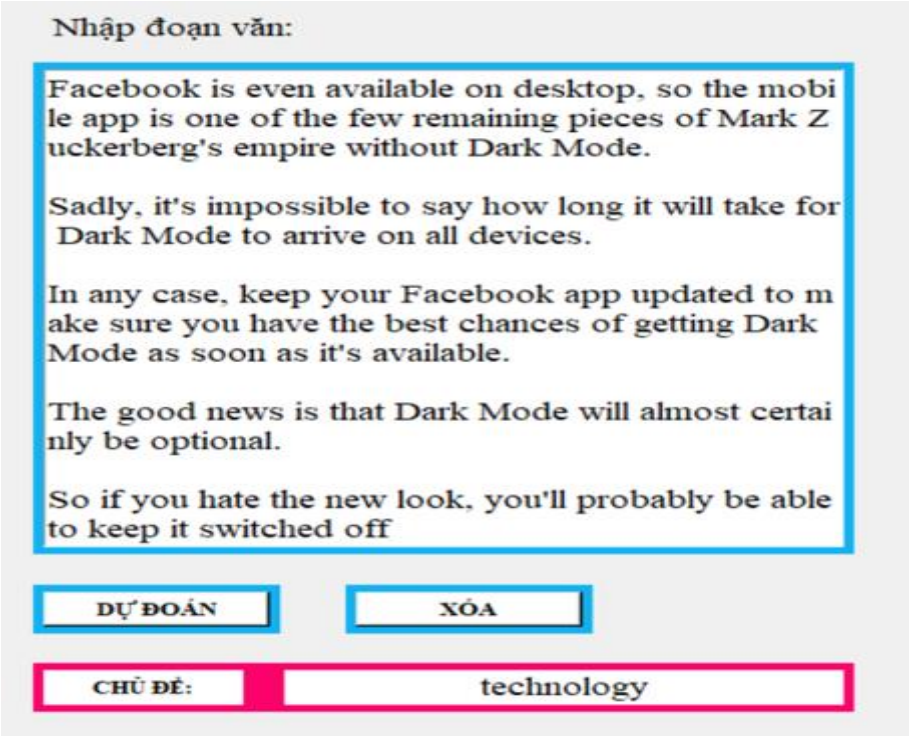
2.6.2 Các khó khăn:

- Số lượng dataset ít dẫn đến việc hạn chế về từ vựng, số lượng chủ đề.
- Thông tin cập nhật mỗi ngày, dẫn đến các từ mới xuất hiện gần đây làm kết quả dự đoán có thể sai. VD: Covid-19

2.6.3 Hướng phát triển:

- Tăng thêm số lượng dataset.
- Cập nhật các từ vựng mới xuất hiện.
- Điều chỉnh các hyperparameter hoặc sử dụng Deep Learning để xử lý ngôn ngữ tự nhiên.

2.7 Một số kết quả dự đoán:



Nhập đoạn văn:

More and more people are considering cycling to w ork, with many eager to avoid public transport as a result of coronavirus and others keen to embrace th e health benefits of cycling. Mark Bailey investigate d whether arriving by bike should become our prim ary method of getting to work.

However, the predicted ten-fold surge in bike journe ys presents some challenges, argued Tom Welsh. N amely, the lack of a compulsory proficiency test an d any form of identification for cyclists. The idea of a compulsory cycling licence was starting to look m ore persuasive, he argued.

Telegraph readers debated whether cycling to work will become the norm as well as the idea of a comp

DỰ ĐOÁN XÓA

CHỦ ĐỀ: health

Nhập đoạn văn:

The 23-year-old was solo and rocked a very chic an d stylish look for her errands run to Blue Bottle Caf e in Beverly Hills on Sunday.

Hailey slicked her blonde hair back into a ballerina b un and appeared to be makeup free underneath her Drew mask.

Her appearance comes after she took to social medi a on Saturday to post a series of photos from her tri p to Utah with Justin.

The pair appeared to have had some pool time as w ell as moments in the outdoor sightseeing.

Their trip comes after Justin announced his decision

DỰ ĐOÁN XÓA

CHỦ ĐỀ: tvshowbiz

3 Tài liệu tham khảo

https://towardsdatascience.com/sarcasm-detection-step-towards-sentiment-analysis-84cb013bb6db

https://pythonspot.com/nltk-stop-words/

https://towardsdatascience.com/how-sklearns-tf-idf-is-different-from-the-standard-tf-idf-275fa582e73d

https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826