Trường Đại học Công nghệ Thông tin
Khoa Khoa học máy tính

TEXT CLASSIFICATION

2-Aug-20

Phân loại đánh giá trong lĩnh vực khách sạn

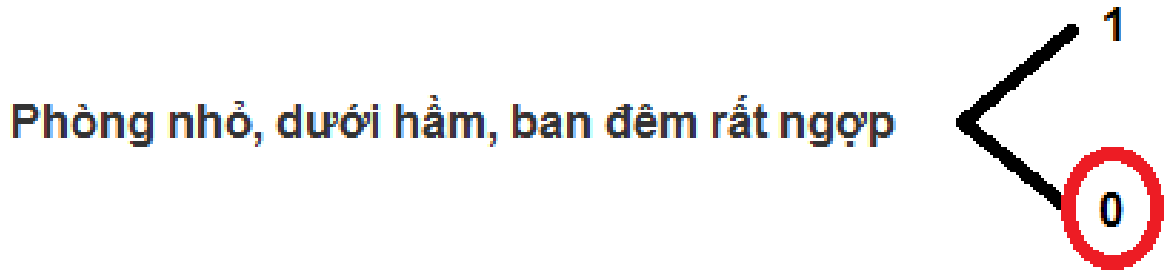
Giáo viên HD: Lê Đình Duy
Phạm Nguyễn Trường An

Sinh viên thực hiện:

Nguyễn Duy Nhật	18520118
Đỗ Nguyễn Thuận Phong	18520126
Nguyễn Dương Trúc Phương	18520133

BÀI TOÁN

- Bài toán phân loại đánh giá trong lĩnh vực khách sạn
- Input: Một câu đánh giá về khách sạn
- Output: Nhãn tương ứng (Đánh giá tiêu cực (0) hay tích cực(1)).



ỨNG DỤNG

- Dựa vào đánh giá có thể đưa ra kết luận về chất lượng khách sạn (phòng ở, thái độ nhân viên, …).
- Xây dựng hệ thống Recommended Systems.
- Là cơ sở để ứng dụng vào các bài toán cao hơn

DATA

- Sử dụng công cụ ***Data Miner*** để lấy dữ liệu từ trang https://www.agoda.com/vi-vn/
- Data sau khi thu thập về lưu dưới file (.xlsx)

DATA

	Column 1
750	Nhiều muỗi quá, nhưng có lẽ khách sạn nào cũng vậy ..... Phòng khá mới và sạch sẽ, nhân viên nhiệt tình!
751	Resort đẹp, điểm trừ hơi xa trung tâm
752	Ngoài việc đi lại vào trung tâm thì xa còn lại tất cả mọi điểm tuyệt vời
753	Đồ ăn ngon tuy nhiên chưa phong phú, không gian đẹp, nhân viên thân thiện
754	Resort đẹp, dịch vụ tuyệt vời tuy nhiên hơi xa trung tâm thành phố
755	Đáng để quay lại
756	Resort khá đẹp, nhân viên chuyên nghiệp, vị trí thuânj lợi
757	Ưu điểm: không gian đẹp, phục vụ tốt Nhược điểm: đi qua nhiều mồ mả.
758	Dịch vụ spa chưa trau chuốt
759	Voi nhung gi hien tai ma The Shells da mang den cho toi trong 3ngay nghi thi qua tuyet, cach don tiep, thuc an, phuc vu thi k con gi bang. Toi se quay lai The Shells lan sa
760	khung cảnh đẹp, tiện nghi, đồ ăn rất ok. Mỗi tội đi qua nhiều mồ mả .
761	Cần thuê một đầu bếp giỏi về, với lại phải để sẵn khăn giấy trên bàn ăn.
762	Mình có bị huỷ 1 phòng do 2 bạn đi cùng thời tiết xấu bị lỡ chuyến bay. Bên khách sạn có đổi phòng huỷ thêm 1 đêm cho 2 vợ chồng. Rất nhiệt tình giúp đo
763	Phòng ốc sạch, tiện nghi,nhà hàng nấu ăn ngon. Vị trí tương đối thuận lợi.
764	View đẹp, phòng tiện nghi và sạch sẽ, nhân viên thân thiện và chuyên nghiệp. Nhà hàng đồ ăn ngon tuy nhiên phục vụ chậm khách sạn cần khắc phục điều này để hoàn h
765	Ocean Bungalow ở khu vực riêng biệt, ngay biển, có hồ bơi riêng, xung quanh sạch sẽ, cây xanh còn ít. Thực đơn nhà hàng ít món, kì nghỉ 3 ngày trở đi thật sự sẽ trở nên nhàm chán.
766	Khách sạn đẹp, không gian rộng rãi, thoải mái, nhân viên rất nhiệt tình
767	Khong gian sach. Rong. Thoang mat. Ho Boi dep
768	Phòng đẹp, view đẹp. Tuy nhiên ăn sáng quá đơn điệu.
769	Nhân viên nhiệt tình, thân thiện.

- Tổng số lượng data thu thập : 2503 câu

DATA

- Hướng dẫn gán nhãn:

-Tách câu

-Sửa các chữ viết tắt (Vd: Ko, Dc, Ks, …)

-Xoá các dấu câu đặc biệt (!, ?, :, ...)

-Đọc bình luận và gán nhãn.

*Đối với những bình luận không rõ về nghĩa, mơ hồ hoặc là tường thuật -> Xóa

DATA

	text	label
0	Không có chào đón như trong giới thiệu	0
1	Lựa chọn tốt cho du lịch nhóm ,gia đình	1
2	phòng yên tĩnh dù sát mặt đường lớn , sạch sẽ.	1
3	phòng sạch đẹp	1
4	Khách sạn sạch sẽ, gần biển, gần siêu thị Lotte gần quán cafe sân vườn rộng đẹp.	1
5	Khách sạn mới, sạch và đẹp	1
6	Nhà hàng hơi chật	0
7	Tốt, tôi sẽ trở lại đây khi đến Vũng Tàu	1
8	Khách sạn ngay trung tâm, tránh được khu nhà nghỉ khách sạn đông đúc	1
9	Tốt	1
10	hệ thống nước ko tốt nên nước nóng ko ổn định, lúc nóng lúc lạnh ko đều	0
11	Khách sạn đẹp vừa túi tiền, gần lottle có cho mượn xe đạp miễn phí.	1
12	Khách sạn mới, đẹp, rất sạch, yên tĩnh, tôi đã có những ngày nghỉ rất dễ chịu ở đây. .	1
13	Nếu là người Việt bạn không nên ở khách sạn này mọi thứ đều tệ nhất là bữa sáng và cách phụ	0
14	Khách sạn sạch sẽ, không có hồ bơi chỉ có 3 bồn matxa nhỏ, có bãi xe nhưng nằm trong hẽm gầ	1
15	Sàn gỗ dơ, buffet tạm	0

Bình luận tích cực: >2400 câu

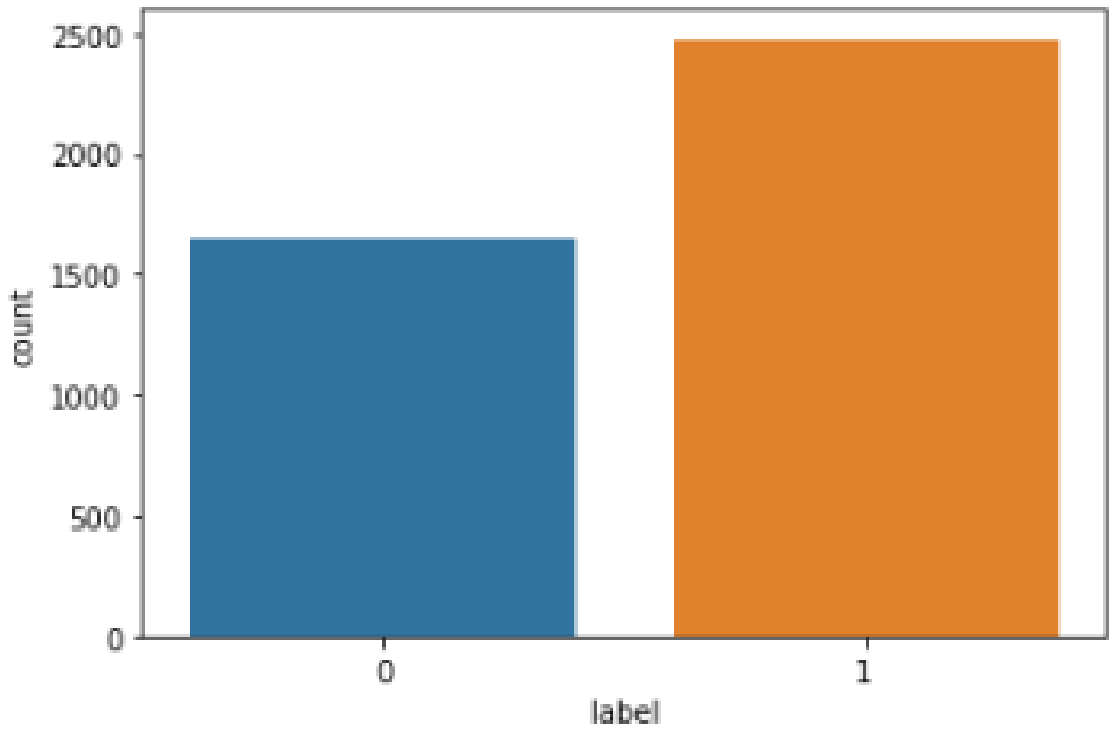
Bình luận tiêu cực: >900 câu

DATA

- Tăng thêm data: thu thập thêm khoảng 800 bình luận tiêu cực và tiến hành gán nhãn

```
[8]  1 sns.countplot(df['label'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0f21c5a0f0>
```



DATA

- CountVectorizer được hỗ trợ bởi Scikit Learn

```
>>> from sklearn.feature_extraction.text import CountVectorizer
>>> corpus = [
...     'This is the first document.',
...     'This document is the second document.',
...     'And this is the third one.',
...     'Is this the first document?',
... ]
>>> vectorizer = CountVectorizer()
>>> X = vectorizer.fit_transform(corpus)
>>> print(vectorizer.get_feature_names())
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
>>> print(X.toarray())
[[0 1 1 1 0 0 1 0 1]
 [0 2 0 1 0 1 1 0 1]
 [1 0 0 1 1 0 1 1 1]
 [0 1 1 1 0 0 1 0 1]]
```

DATA

- Tf-idf được hỗ trợ bởi Scikit Learn

```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2 corpus = [
3     'This is the first document.',
4     'This document is the second document.',
5     'And this is the third one.',
6     'Is this the first document?']
7 tf=TfidfVectorizer()
8 tf.fit(corpus)
9 print(tf.get_feature_names())
10 corpus_new=tf.transform(corpus).toarray()
11 print(corpus_new)
```

```
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
[[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]
 [0.         0.6876236  0.         0.28108867 0.         0.53864762
  0.28108867 0.         0.28108867]
 [0.51184851 0.         0.         0.26710379 0.51184851 0.
  0.26710379 0.51184851 0.26710379]
 [0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]]
```

DATA

- Tách từ: Pyvi


```
[76]  1 from pyvi import ViTokenizer
      2 def Token(t):
      3   return ViTokenizer.tokenize(t)
      4
      5 print(Token('Khách sạn nằm tại trung tâm thành phố'))
```

```
Khách_sạn nằm tại trung_tâm thành_phố
```

MACHINE LEARNING

Tách từ theo dấu cách:

Model	Bernouli NB	Logistic Regression	SVC	Decision Trees	RandomForest
CountVectorizer	88,97	93,57	93,69	88,73	94,18
TF-IDF Vectorizer	88,96	94,67	95,27	87,15	93,21

MACHINE LEARNING

Tách từ theo nghĩa (sử dụng Pyvi):

Model	Bernouli NB	Logistic Regression	SVC	Decision Trees	RandomForest
CountVectorizer	88,85	94,54	94,06	89,82	94,54
TF-IDF Vectorizer	88,85	94,67	95,51	89,45	94,67

DEMO


THANK
YOU