

# BÀI TUẦN 11: NATURAL LANGUAGE PROCESSING

## 1. Thông tin sinh viên

- **Họ và Tên: Dương Minh Lượng**
- **MSSV: 18521071**

## 2. Source

```
1. import pandas as pd
2. dataset=pd.read_csv("Restaurant_Reviews.tsv",
    delimiter="\t")
3. import nltk
4. import re
5. nltk.download("stopwords")
6. from nltk.corpus import stopwords
7. from nltk.stem.porter import PorterStemmer
8. corpus = []
9. for review in dataset.values[:, 0]:
10.     review = re.sub("[^a-zA-Z]", " ", review)
11.     review = review.lower()
12.     review = review.split()
13.     ps = PorterStemmer()
14.     review = [ps.stem(word) for word in review
    if word not in stopwords.words("english")]
15.     review = " ".join(review)
16.     corpus.append(review)
17.     # Creating the Bag of Words model
18.     from sklearn.feature_extraction.text import
    CountVectorizer
19.     cv = CountVectorizer(max_features = 1500)
20.     X = cv.fit_transform(corpus).toarray()
21.     y = dataset.iloc[:, 1].values
22.     print(X.shape[1])
23.     from sklearn.model_selection import
    train_test_split
24.     X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size = 0.2, random_state
    = 0)
25.     from sklearn.naive_bayes import GaussianNB
26.     classifier = GaussianNB()
27.     classifier.fit(X_train, y_train)
28.     y_pred = classifier.predict(X_test)
29.     print(classifier.score(X_train, y_train))
30.     print(classifier.score(X_test, y_test))
```

```

31.     from sklearn.metrics import confusion_matrix
32.     cm = confusion_matrix(y_test, y_pred)
33.     print(cm) (model, X_):
34.         X1 = X[:, 0]
35.

```

### 3. Kết quả

#### Confusion\_matrix test

|   | 0  | 1  |
|---|----|----|
| 0 | 55 | 42 |
| 1 | 12 | 91 |

Nhận xét:

|                     | Dự đoán là tiêu cực | Dự đoán là tích cực |
|---------------------|---------------------|---------------------|
| Thực sự là tiêu cực | 55                  | 42                  |
| Thực sự là tích cực | 12                  | 91                  |

- Độ chính xác của mô hình trên tập train và test: 0.92125(train) và 0.73(test)-> độ chính xác của tập test không cao so với train.
- Xác suất dự đoán sai là  $(42+12)/200=0,27\%$ .
- Các bình luận thực sự là tiêu cực nhưng dự đoán là tích cực tương đối nhiều -> cần xem lại.