BÀI TUẦN 11: NATURAL LANGUAGE PROCESSING

1. Thông tin sinh viên

Họ và Tên: Dương Minh Lượng

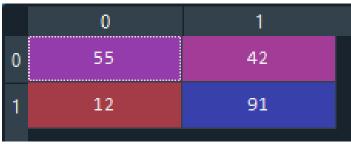
- MSSV: 18521071

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]:
           review = re.sub("[^a-zA-Z]"," ", review)
11.
           review = review.lower()
12.
           review = review.split()
13.
           ps = PorterStemmer()
           review = [ps.stem(word) for word in review
 if word not in stopwords.words("english")]
           review =" ".join(review)
15.
            corpus.append(review)
      # Creating the Bag of Words model
17.
       from sklearn.feature extraction.text import
 CountVectorizer
19. cv = CountVectorizer (max features = 1500)
      X = cv.fit_transform(corpus).toarray()
21.
      y = dataset.iloc[:, 1].values
       print(X.shape[1])
        from sklearn.model selection import
 train test split
       X train, X test, y train, y test =
  train test split(X, Y, test size = 0.2, random state
25.
        from sklearn.naive bayes import GaussianNB
        classifier = GaussianNB()
26.
27.
        classifier.fit(X train, y train)
        y pred = classifier.predict(X test)
28.
29.
        print(classifier.score(X train, y train))
        print(classifier.score(X test, y test))
30.
```

3. Kết quả

Confusion_matrix test



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 xuất 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.