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

THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU
BÀI TẬP 4

Nhận diện cảm xúc từ những bình luận

Họ và tên : Lưu Quang Tiến Hoàng

MSSV : 20521342

Lớp : DS103.M21

1. Tổng quan về bộ dữ liệu: số lượng dữ liệu, số nhãn.

- Số lượng dữ liệu bao gồm 206 comments

- Số cột bao gồm 2 cột (comment và label_id)

- Số nhãn bao gồm 3 nhãn :

+ 0: positive (cảm xúc tích cực)

+1: negative (cảm xúc tiêu cực)

+2: neutral (trung tính)

- Lưu file "*nguoi gan nhan 1.xlsx*" và "*nguoi gan nhan 2.xlsx*" thành 2 file "*annotator_1.csv*" và "*annotator_2.csv*"

```
: import pandas as pd
  read_1=pd.read_excel("nguoi gan nhan 1.xlsx")
  read_1.to_csv("annotator_1.csv",header=True,index=None)
  f1=pd.DataFrame(pd.read_csv("annotator_1.csv"))
```

```
: import pandas as pd
  read_2=pd.read_excel("nguoi gan nhan 2.xlsx")
  read_2.to_csv("annotator_2.csv",header=True,index=None)
  f2=pd.DataFrame(pd.read_csv("annotator_2.csv"))
```

Hình 0. Chuyển file xlsx thành file csv

	comment_text	label_id
0	Comment 1	0
1	Comment 2	0
2	Comment 3	0
3	Comment 4	0
4	Comment 5	0
...	...	...
201	Comment 202	0
202	Comment 203	2
203	Comment 204	0
204	Comment 205	2
205	Comment 206	2

206 rows × 2 columns

Hình 1. annotator_1.csv

	comment_text	label_id
0	Comment 1	0
1	Comment 2	0
2	Comment 3	0
3	Comment 4	0
4	Comment 5	0
...	...	...
201	Comment 202	0
202	Comment 203	2
203	Comment 204	0
204	Comment 205	1
205	Comment 206	1

206 rows × 2 columns

Hình 2. annotator_2.csv

2. Thống kê số lượng nhãn:

* *annotator_1.csv:*

Tổng nhãn 0: 122

Tổng nhãn 1: 23

Tổng nhãn 2: 61

* *annotator_2.csv:*

Tổng nhãn 0: 121

Tổng nhãn 1: 36

Tổng nhãn 2: 49

```
: f1['label_id'].value_counts()
: 0    122
  2     61
  1     23
  Name: label_id, dtype: int64
```

Hình 3. annotator_1.csv:

```
: f2['label_id'].value_counts()
: 0    121
  2     49
  1     36
  Name: label_id, dtype: int64
```

Hình 4. annotator_2.csv:

3. Tính toán độ đồng thuận theo công thức Cohen-Kappa. Ghi rõ số liệu tính toán

Theo công thức Cohen-Kappa

$$\mathcal{K} = \frac{p_0 - p_e}{1 - p_e}$$

ta được kết quả:

```
: from sklearn.metrics import cohen_kappa_score
  true = f1['label_id'].values.tolist()
  pred = f2['label_id'].values.tolist()
  print(cohen_kappa_score(true,pred))

  0.49054784759190173
```

Hình 5. Tính độ đồng thuận

4. Kết luận về độ đồng thuận theo thang đo của Landis và Koch (1977)

Theo thang đo của Landis và Koch (1977) thì kết quả độ đồng thuận theo công thức Cohen – Kappa của annotator_1 và annotator_2 đã tính đạt mức độ Moderate (vừa phải)**.**