Assignment 3

> Tuan Trinh 1001885663 tqt5664

question 1

> Answer is d_6 is classified as a N label document

- Total number of documents, $N_{\text{doc}} = 5$
- Total P documents, $N_P = 3$
- Total N documents, $N_N = 2$

$$P(P) = \frac{N_P}{N_{\text{doc}}} = \frac{3}{5} = 0.6$$
$$P(N) = \frac{N_N}{N_{\text{doc}}} = \frac{2}{5} = 0.4$$

Vocabulary $V = \{\text{'excellent', 'definitely', 'good', 'not', 'bad', 'so', 'enough'}\}$

Vocabulary Count $|V| = 7$

$$P(P) : \{'good' : 2, 'excellent' : 1, 'definitely' : 1, 'not' : 1, 'bad' : 1, 'so' : 1\}$$

$$P(N) : \{'so' : 2, 'not' : 1, 'good' : 1, 'enough' : 1, 'bad' : 1\}$$

Add-1 Smoothing

$$P(w_i|P) = \frac{\text{Count}(w_i, P) + 1}{\text{Total words in } P + |V|}$$
$$P(w_i|N) = \frac{\text{Count}(w_i, N) + 1}{\text{Total words in } N + |V|}$$

$$P(d_6|P) = P(\text{'so'}|P) \times P(\text{'so'}|P) \times P(\text{'good'}|P)$$
$$P(d_6|P) = 0.143 \times 0.143 \times 0.214$$
$$P(d_6|N) = P(\text{'so'}|N) \times P(\text{'so'}|N) \times P(\text{'good'}|N)$$
$$P(d_6|N) = 0.231 \times 0.231 \times 0.154$$

$$P(P|d_6) = P(P) \times P(d_6|P)$$
$$P(P|d_6) = 0.6 \times (0.143 \times 0.143 \times 0.214) = 0.002624$$
$$P(N|d_6) = P(N) \times P(d_6|N)$$
$$P(N|d_6) = 0.4 \times (0.231 \times 0.231 \times 0.154) = 0.003277$$

$$Normalize\ P(P|d_6) = \frac{P(P|d_6)}{P(P|d_6) + P(N|d_6)}$$
$$P(P|d_6) = \frac{0.002624}{0.005901} \approx 0.444$$

$$Normalize\ P(N|d_6) = \frac{P(N|d_6)}{P(P|d_6) + P(N|d_6)}$$
$$P(N|d_6) = \frac{0.003277}{0.005901} \approx 0.556$$

question 2

The method I used to train the data on was the Add-1 Smoothing or Laplace smoothing because the dataset that was given was relatively small. I performed really well and this was definitely noticed after normalizing the scores

sample output with metrics

```
2024-02-10 23:55:40.720687 - INFO - Processing vocabulary...
2024-02-10 23:55:40.720721 - INFO - Training model...
2024-02-10 23:55:40.720751 - INFO - Word counts: {<Label.Zero:
2024-02-10 23:55:40.720780 - INFO - Trained model: {<Label.Zer
2024-02-10 23:55:40.720814 - INFO - Initial scores: {<Label.Ze
2024-02-10 23:55:40.720827 - INFO - Normalized scores: {<Label
2024-02-10 23:55:40.720833 - INFO - Predicted label: Label.One
```