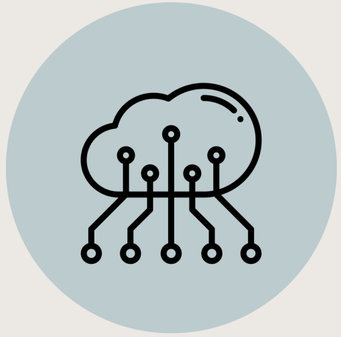
SENTIMENT ANALYSIS

BY LE HONG QUANG

OUTLINE

Data Prepare

Prepare and preprocess data ready for training



Training model

Use TF-IDF fit with Linear SVC

Improve model

Hyperparameter tuning to consider model changes

DATA PREPARE

IMDB Dataset

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

DATA PREPROCESS

Implementation steps

1. Delete content HTML & punctuation and special characters
2. Switch to regular words and split them
3. Use stopwords to eliminate words that do not affect the meaning
4. Stemming and Lemmatization to grouping similar words and easier to analyze and understand
5. Join words and save them in the corpus

TRAINING MODEL

Vectorized using TF-IDF technique

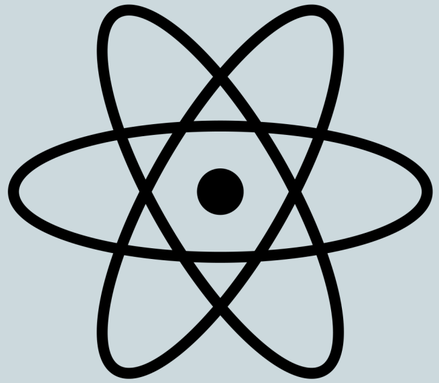
- The tf–idf is the product of two statistics: term frequency and inverse document frequency
- Calculation formula:

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

TRAINING MODEL

Code



Vectorize TF-IDF

```
tfidf_vecto = TfidfVectorizer(ngram_range=(1, 3))
tfidf_vecto_train = tfidf_vecto.fit_transform(corpus_train)
tfidf_vecto_test = tfidf_vecto.transform(corpus_test)
```

TF-IDF fit with Linear SVC

```
LSVC = LinearSVC(C=0.5, random_state=42)
LSVC.fit(tfidf_vecto_train, y_train)
du_doan = LSVC.predict(tfidf_vecto_test)
```

TRAINING MODEL

Report of model

```
Classification report of Linear SVC:
              precision    recall  f1-score   support

    Positive       0.91      0.89      0.90      6157
    Negative       0.89      0.92      0.91      6343

    accuracy                           0.90     12500
   macro avg       0.90      0.90      0.90     12500
weighted avg       0.90      0.90      0.90     12500

Confusion Matrix of Linear SVC:
 [[5467  690]
 [ 524 5819]]
```

IMPROVE MODEL

Hyperparameter tuning

- Hyperparameter tuning is the process of finding and selecting optimal values for hyperparameters
- Adjust and find the best hyperparameter C for the model

```
LSVC = LinearSVC(C=1, random_state=42)
LSVC = LinearSVC(C=0.1, random_state=42)
LSVC = LinearSVC(C=10, random_state=42)
```

IMPROVE MODEL

Result of Hyperparameter tuning

C = 0.5

Accuracy = 90.29%

C = 1

Accuracy = 90.59%

C = 0.1

Accuracy = 88.63%

C = 10

Accuracy = 90.79%

Thank you