Introduction to Natural Language Processing

By Bhupesh Kumar

What is Natural language processing?

It's a form of artificial intelligence that focus on analyzing the human language to draw insights. It is the art and science which helps us extract information from text and use it in our computations and algorithms.

Step by step example

Approach

1. COLLECT

Redditlist : 2 subreddits

- **Iphone (0)**
  - 1230 posts
  - 903 Unique posts
- **Google Pixel (1)**
  - 1235 posts
  - 908 Unique posts

Data shape (2465 , 2)

2. CLEAN

- Change data to pandas DataFrame
- Delete the null values
- Removed duplicate posts from data collection

Data shape (1807 , 2)

3. MODEL

- Bayes Classifier

  -Multinomial classifier

- Logistic Regression

Before Modeling

Countvectorizer

It provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words.

Simply, It converts text documents to a matrix

Date shape after splitting

75 /25(Training/Testing) and then Countvectorizer

X_train_cv = (1355 , 2238)

X_test_cv = (452 , 2238)

Modeling Result

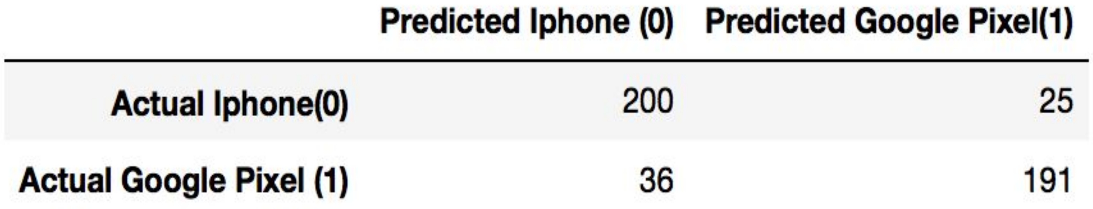
Logistic Regression

Confusion Matrix

Accuracy :

Training - 98.6 %

Testing - 86.5 %



	Predicted Iphone (0)	Predicted Google Pixel(1)
Actual Iphone(0)	200	25
Actual Google Pixel (1)	36	191

True Positive	191
True Negative	200
False Positive	25
False Negative	36

Modeling Result

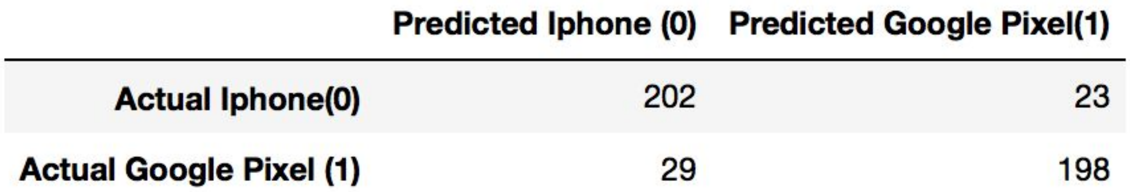
Multinomial classifier

Confusion Matrix

Accuracy :

Training - 95.6 %

Testing - 88.45 %



	Predicted Iphone (0)	Predicted Google Pixel(1)
Actual Iphone(0)	202	23
Actual Google Pixel (1)	29	198

True Positive	198
True Negative	202
False Positive	23
False Negative	29

Conclusion

- Both of my models were overfitting, Naive classifier by 7 % and logistic regression by 12% .

Ways to fix it:

- Change the number of features
- Change different parameter of the model
- Still overfitting, then try different model.