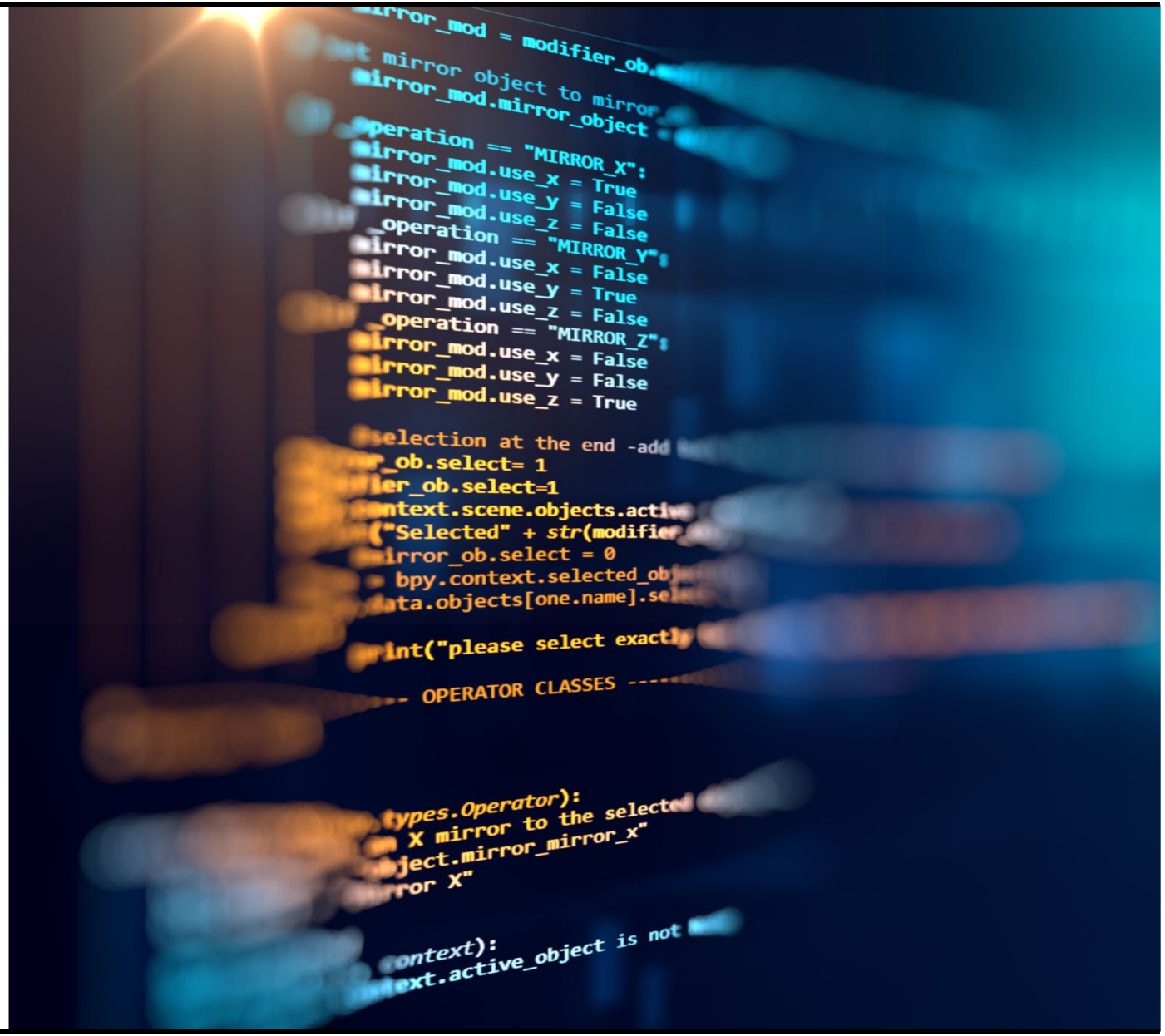
Project Three

Web API and Natural Language Processing

Basazin Belhu

Oct 08, 2021



Contents:

Problem Statement

Can you automatically classify the different reviews to targeted products for the R&D team?

- **NumPy,** is a Seattle, WA based gaming technology startups and knows this problem well. They'd like to classify the reviews from reddit according to product, but this is tough because their customers did random comment and discussion about different games NumPy has launched.
- **NumPy** just hired me after GA. DSI, and challenged me to **build an algorithm that automatically classifies customer reddit reviews to the right product** . They provided me with a Push Shift link where most of their customers actively engage about their products where user-inputted self text, comments and tile descriptions of products.

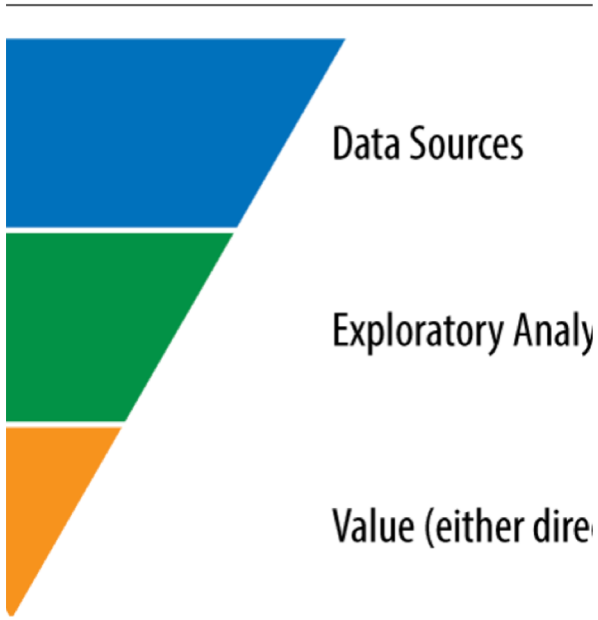
Data wrangling /gathering /acquisition

Web APIs

- API stands for Application Programming Interface
- The way of one **Software** application communication with another **Software**

Data wrangling /gathering /acquisition

- Web APIs
- Json files manipulation
- Data frame generation and manipulation



API Request → Data

```
508]: url = "https://api.pushshift.io/reddit/search/submission"
      para = "boardgames"
      response = requests.get(url, para)

      response.json()
```

Request & Response

```
    'permalink': '/r/AnythingGoesNews/comments/q3tmxt/iran_smuggling_hightech_drones_to_militant_allies/',
    'pinned': False,
    'pwls': 7,
    'retrieved_on': 1633682150,
    'score': 1,
    'selftext': '  \n\nIran's theocratic regime has ramped up its drone manufacturing operation in recent yea
rs and is now smuggling an increasingly sophisticated slate of the weaponized remote control aircraft to all
ied militant groups around the Middle East\n\nThe UAV program of the Iranian regime is the primary weapon us
ed for terrorism and warmongering and destabilizing the region, and certainly this is supplying proxies in t
he region with those UAVs',
    'send_replies': True,
    'spoiler': False,
```

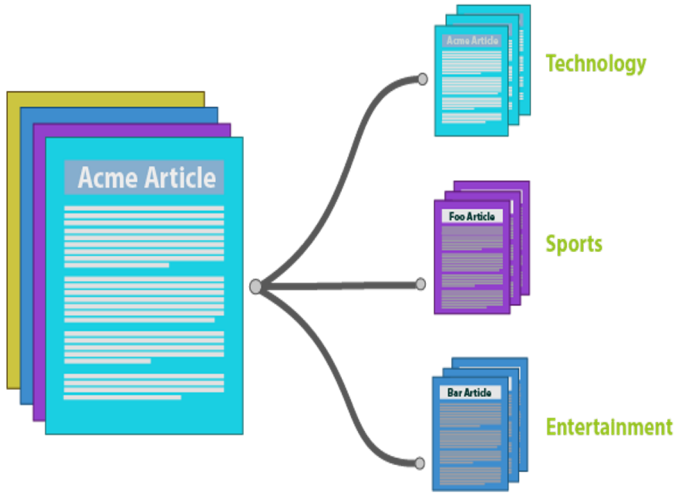
Json

What is Natural Language Processing

In short:

- It is the computer program capability of understand human language
- NLP is all about analyzing and understanding human language with computer programing and stat.

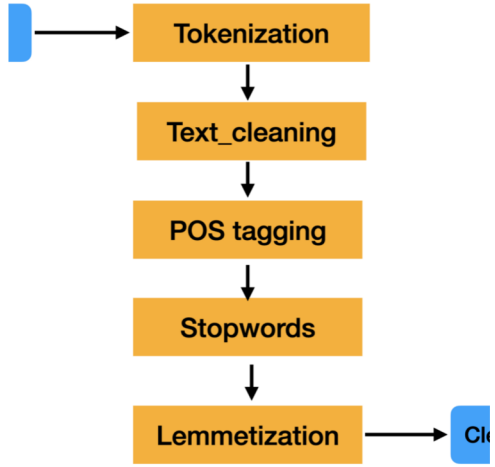
NLP Basic Layout

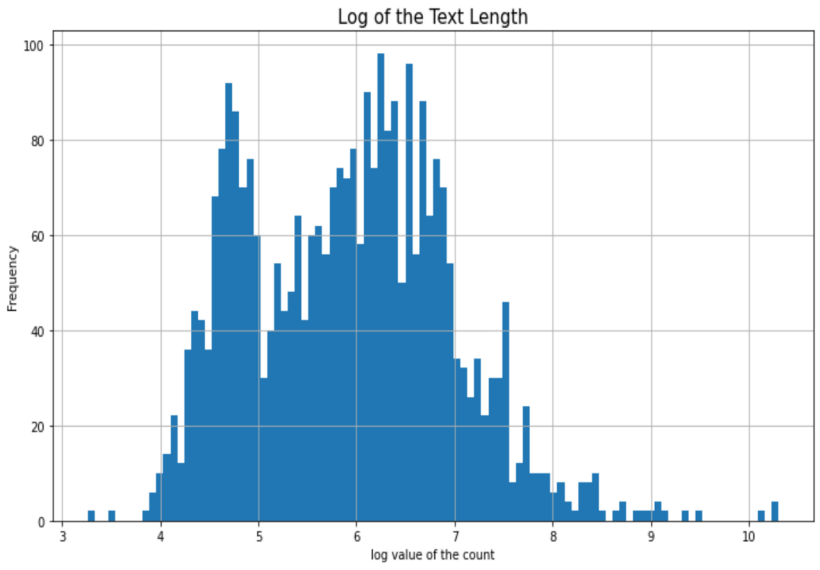


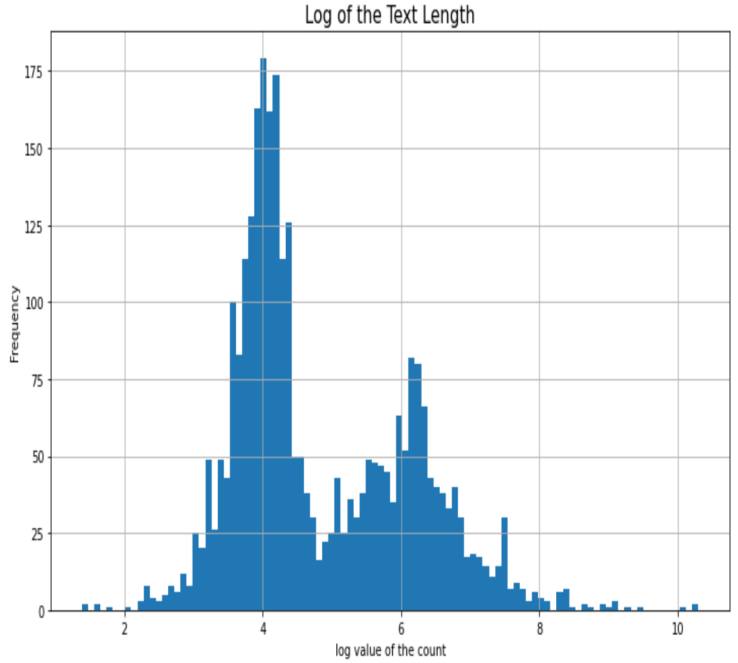
Classify the different reviews:

- Boardgames
- Videogames

NLP Basic Processes







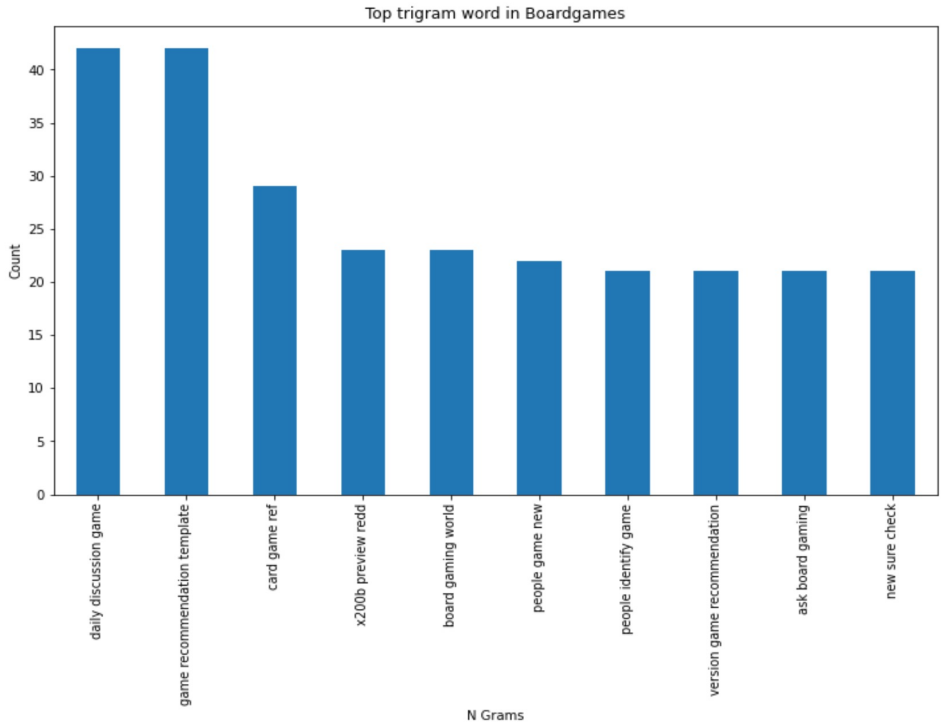
`tfidf`

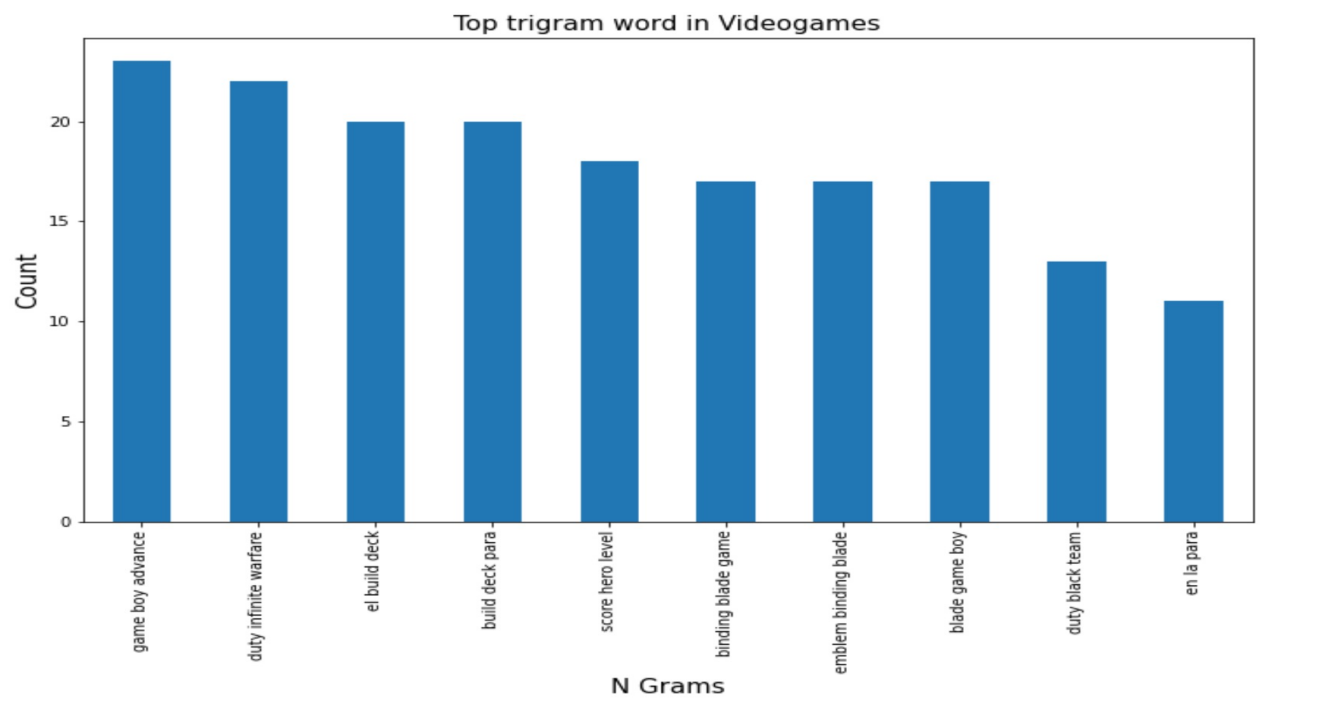
	tfidf
abductor	7.957022
ability	6.011111
able	5.066650
abound	7.957022
absent	7.957022
...	...
zodiac	7.551556
zombie	6.704259
zone	7.957022
zoo	7.551556
zoom	7.957022

Word Vectorizations

	abductor demand like	ability alter gravity	ability alter personal	ability considering open	ability determine necessary	ability half think	ability learn game	ability let trigger	ability magic attack	ability missile bomb	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..
	...	...	...	...	...	...	...	...	...	...	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..
	0	0	0	0	0	0	0	0	0	0	..

Word Frequency in the corpus





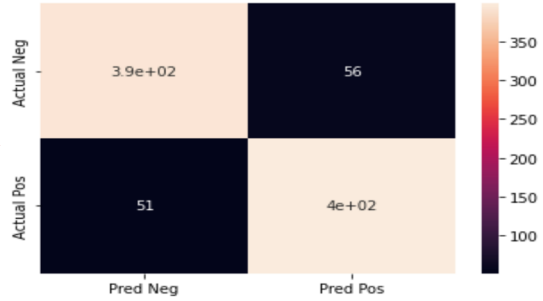
Classification Model Evals:

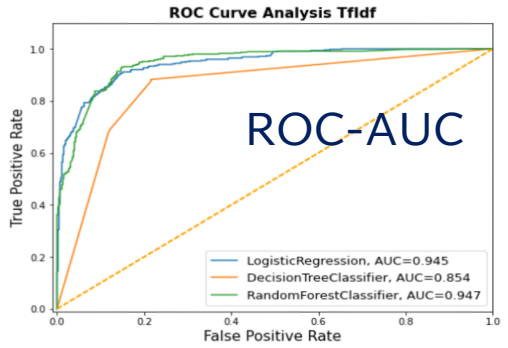
	Model	Accuracy
0	Log_tfidf_train	0.947619
1	Log_tfidf_test	0.881111
2	Rnd_tfidf_train	0.991429
3	Rnd_tfidf_test	0.875556
4	Log_vc_train	0.965238
5	Log_vc_test	0.881111
6	Rnd_vc_train	0.991429
7	Rnd_vc_test	0.884444

Accuracy

```
Lg_Training Accuray: 0.9476190476190476
Lg_Testing Accuracy:  0.8811111111111111
```

Confusion Matrix





Interpretation:

- Random forest has better traing accuracy
- Logistic Reg. is better because less overfit
- ”Probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance AUC”

Conclusion & Recommendation

All the Data Science steps executed we have discovered pattern from the messy data

We have developed an logarithms that able classify the reviews or tweets in 0.94 on training and .88 on testing accuracy:

Future Work:

- Tune some the hyperparameter and more train the model
- Deploy this model to the production

Thank You!