Which Reddit Does This Post Come From?

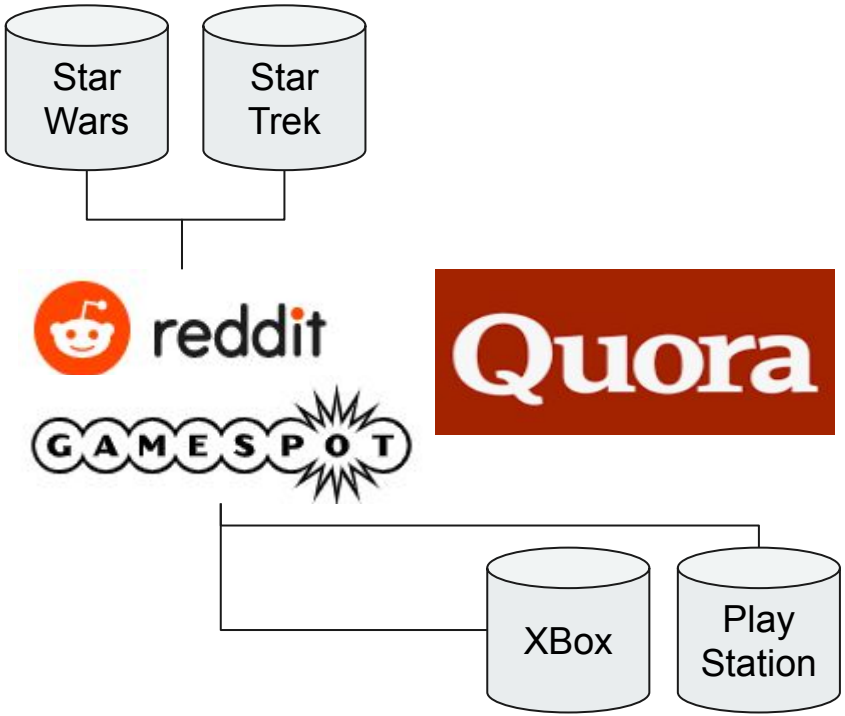
A Web-Scraping, Language Processing & Classification Project

reddit

By Howie, Wee Siang & Aishah

Business Problem

- Data engineers in a digital marketing company
- Provide social media intelligence
- Scrapping posts from a diverse network of sources
- Project 3 as a proof-of-concept



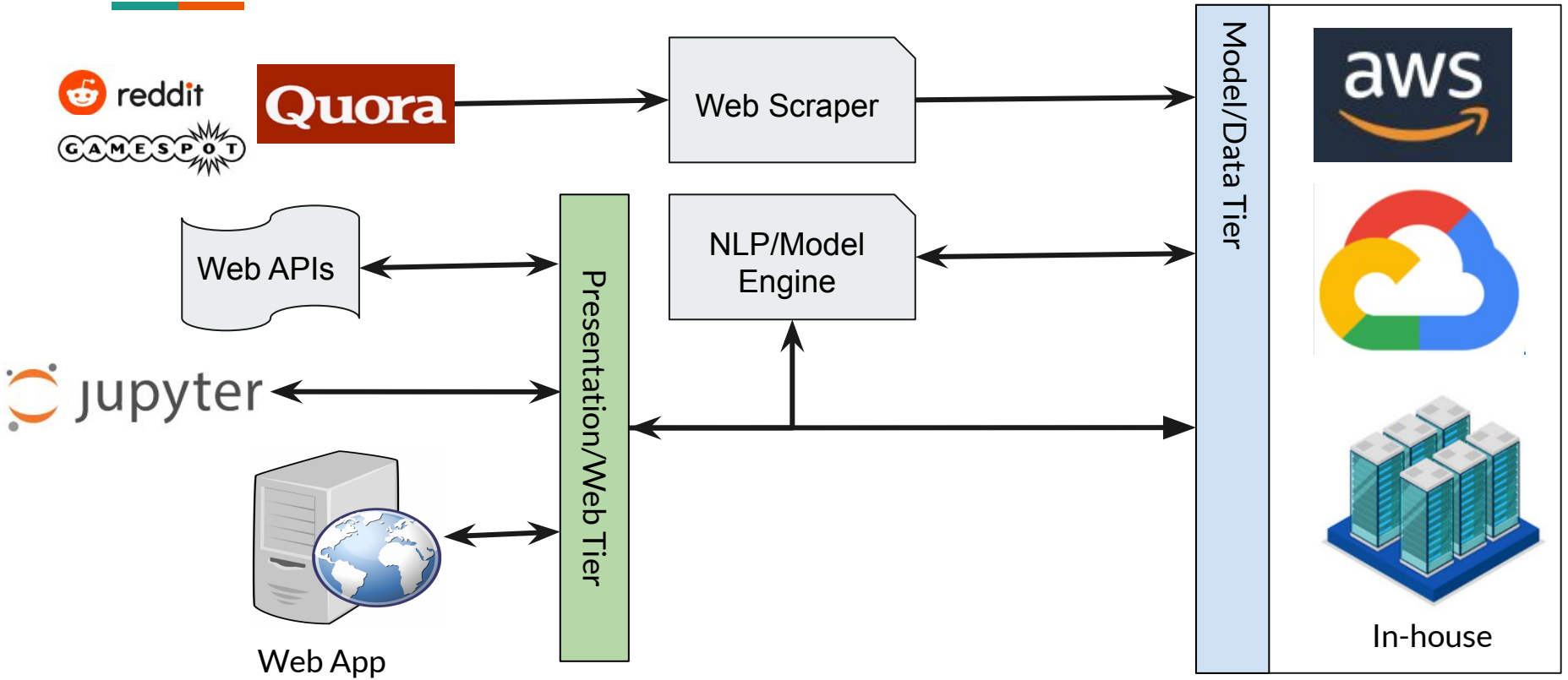
Star Wars vs Star Trek

- There have always been fans in either of these 2 camps
- Seemingly similar but not the same
- If model is good, should be able to pick up ‘keywords’ relevant to either Star Wars or Star Trek
- And classify unseen posts well
- We want the **best** model that does this (highest accuracy score)

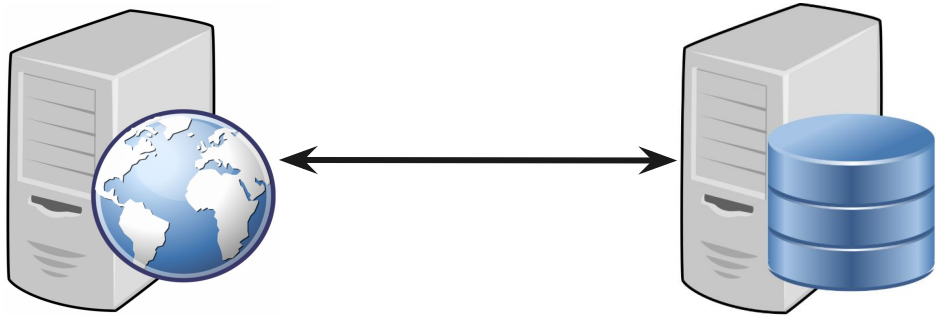
The Tech Work

1. Information Retrieval
2. Information Extraction
3. Text Preprocessing
4. More Models…
5. Hyperparameter Tuning
6. Score Comparisons
7. Top Features

Solution Architecture



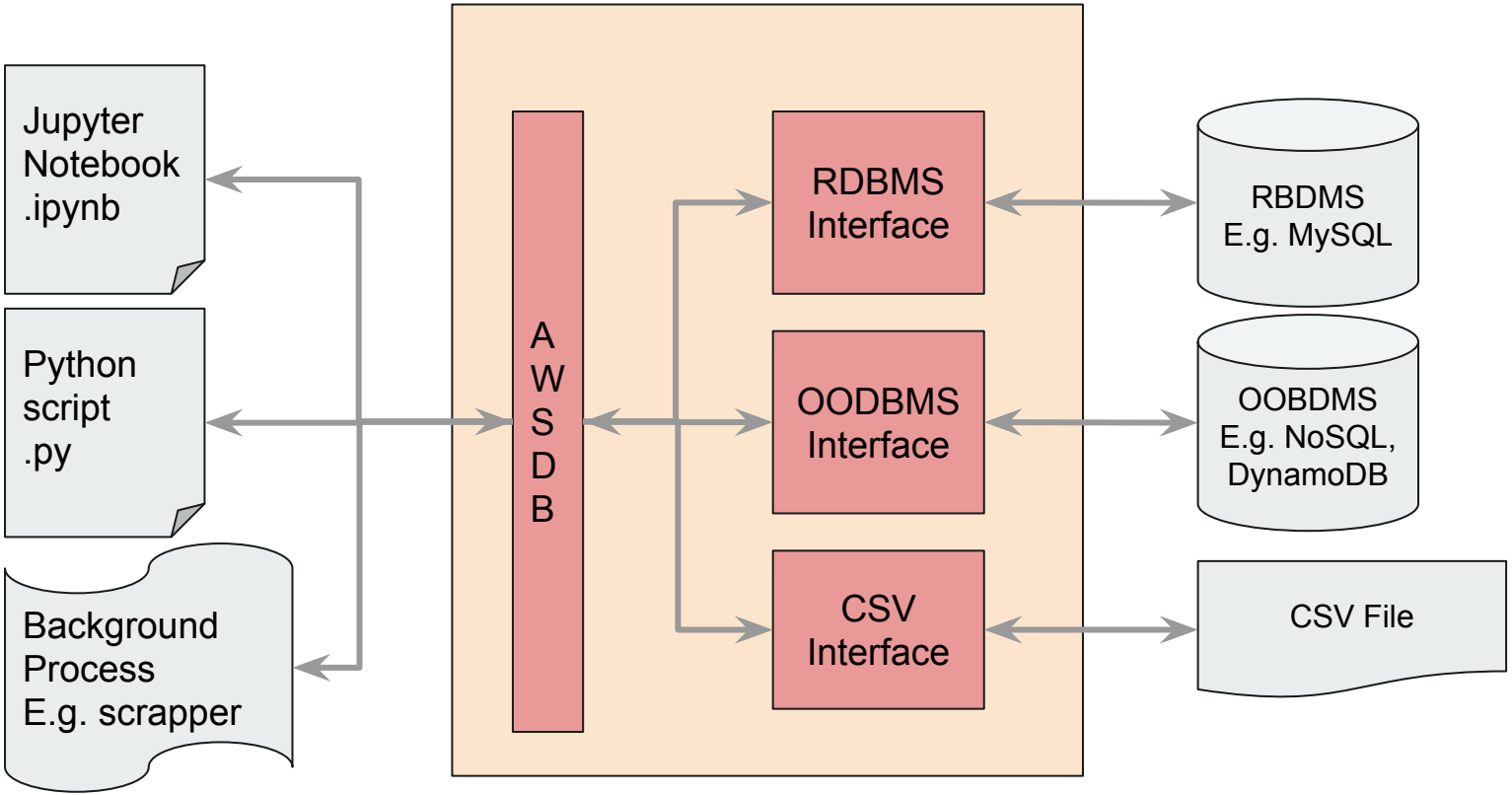
Presentation Tier



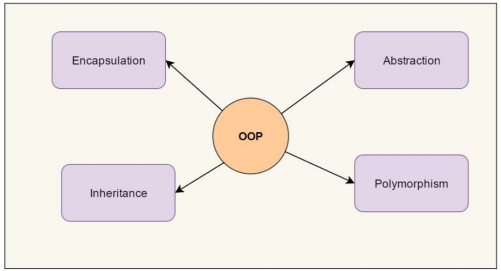
Flask Server

Database Server

Data Tier


Jupyter
Notebook
.ipynb
Python
script
.py
Background
Process
E.g. scrapper
AWSDB
RDBMS
Interface
OODBMS
Interface
CSV
Interface
RBDMS
E.g. MySQL
OOBDMS
E.g. NoSQL,
DynamoDB
CSV File

Object Oriented Programming



Four Pillars of Object Oriented Programming

In [2]:

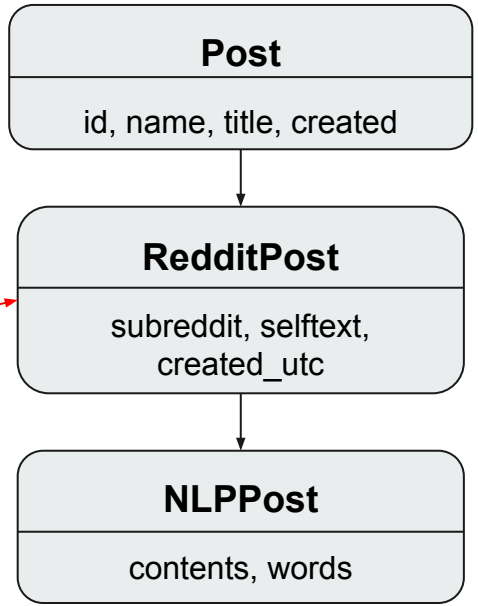
```
posts = AWSDB.get_reddit_posts_for_nlp()

df = pd.DataFrame([vars(post) for post in posts])

df.drop(columns=['batch_id', 'created_utc'], inplace=True)

df.sort_values(by='created').head()
```

Out[2]:

	created	name	subreddit	text	title
0	2019-12-05 23:58:39	t3_e6q4c5	StarWars		Baby Yoda in real life
1	2019-12-06 00:01:05	t3_e6q5fw	StarWars	Who would be the best and who would be the wor...	Best and Worst of Baby Yoda's Voice
2	2019-12-06 00:02:34	t3_e6q64h	StarWars	Hello there. I'm getting more hyped for Ep IX ...	Casual fan in need of some background info
3	2019-12-06 00:12:15	t3_e6qaxu	StarWars		Ready to lay low and stretch your legs for a c...
4	2019-12-06 00:16:04	t3_e6qcqq	StarWars		3D printed Baby-Yoda!

Encapsulation

Inheritance



r/StarWars

Search r/StarWars

155

Crossposted by u/Alextr91 2 hours ago

Fan Creations

This can be beautifully done!

'title'

i.imgur.com/SPLqSm...

r/KeanuBeingAwesome · Posted by u/Master1718 6 hours ago

Yes please

r/startrek

Search r/startrek

31

Posted by u/Redpythongoon 6 hours ago

Ode to Odo (original art)

'title'

Rene was amazing, and made Odo one of the most recognizable characters in all of Star Trek.

'self_text'

https://imgur.com/B91HnbH

8 Comments Share Save ...

26

Posted by u/TheBigSmol 7 hours ago

What is the most immoral or irresponsible action that a Starship captain has done?

91 Comments Share Save ...

13

Posted by u/FupaKoopa 3 hours ago

Someone drew an Incredible Tales (from "In the Pale Moonlight") tribute to DS9

mudron.bigcartel.com/produc...

2 Comments Share Save ...

11

Posted by u/SJFree 2 hours ago

So I started watching Enterprise last week...

...and just got to S1:E19, "Oasis". Imagine my shock when René shows up. Gave me some

Info Retrieval & Extraction


```
{'kind': 't3',
 'data': {'approved_at_utc': None,
  'subreddit': 'StarWars',
  'selftext': '',
  'author_fullname': 't2_fx04m',
  'saved': False,
  'mod_reason_title': None,
  'gilded': 0,
  'clicked': False,
  'title': "Rewatching the prequels and it's heartbreaking seeing Qui Gon Jinn acting as a father figure to Anakin. Giving the boy who endured so much, a dad-like figure, even if it was only for a short time. Qui Gon was the father Anakin needed.",
  'link_flair_richtext': [],
  'subreddit_name_prefixed': 'r/StarWars',
```


.to_csv()

	post	subreddit
0	Reminder: We don't allow memes or image macros	StarWars
1	The Mandalorian - Discussion Thread - S1E4	StarWars
2	Dying Star Wars Fan Sees The Rise Of Skywalker...	StarWars
3	Our 3d printed Baby Yoda Christmas tree topper!	StarWars
4	I repainted some cheap Droid toys and made the...	StarWars

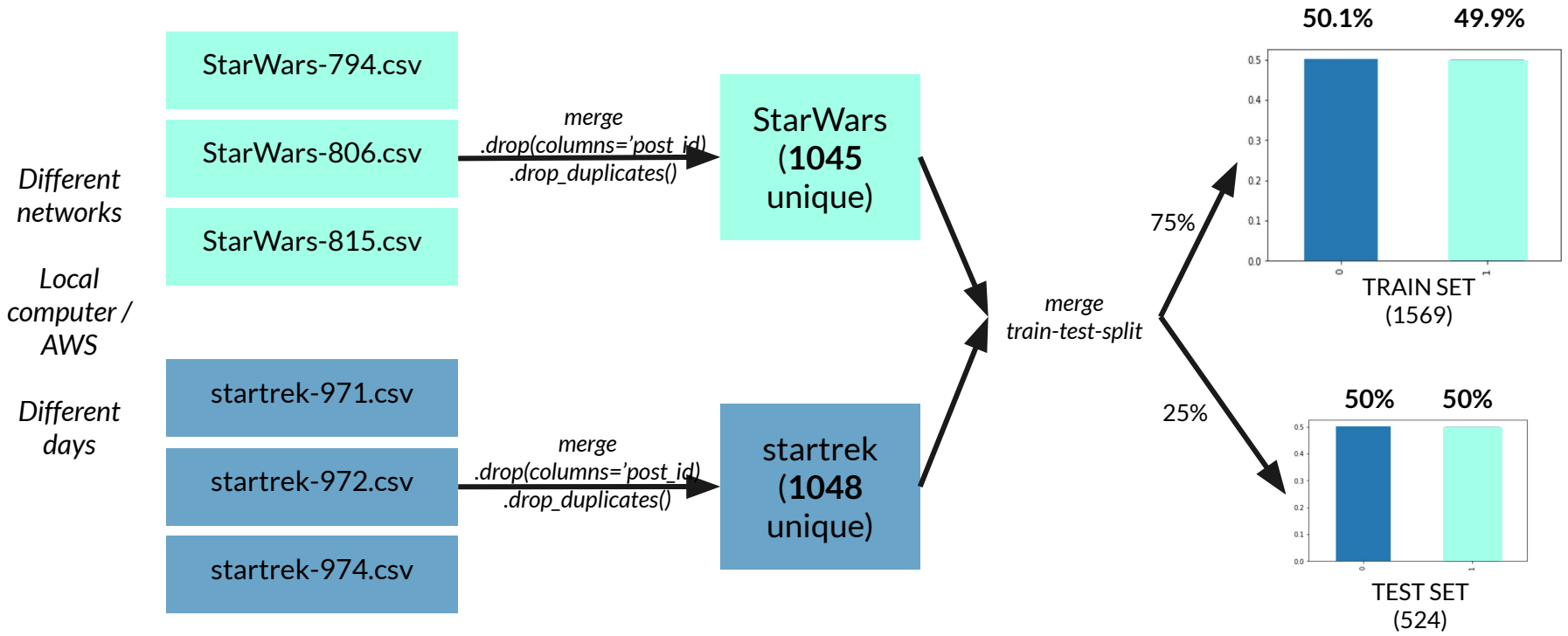

```
{'kind': 't3',
 'data': {'approved_at_utc': None,
  'subreddit': 'startrek',
  'selftext': '',
  'author_fullname': 't2_46d9z1gz',
  'saved': False,
  'mod_reason_title': None,
  'gilded': 0,
  'clicked': False,
  'title': 'Will this English theme park have a Star Trek land? It has a partnership with Paramount and one planned area is described as a "bustling 23rd century landing zone." Hmmm',
  'link_flair_richtext': [],
  'subreddit_name_prefixed': 'r/startrek',
```

```
'downs': 0,
'thumbnail_height': 73,
'hide_score': False,
'name': 't3_e6pbnp',
'quarantine': False,
'link_flair_text_color': 'dark',
```


.to_csv()

1043	I make Star Trek pet portraits. By Furburn Franks	startrek
1044	In what circumstances will a higher ranking of...	startrek
1045	Guenine question : Why is the subreddit icon r...	startrek
1046	Trying to draw one Star Trek character every d...	startrek
1047	Just in time for the holidays and to add to th...	startrek

Info Extraction -> Data Sets



Text Preprocessing - Document Cleaning

nltk

```
from nltk.corpus import stopwords

nltk_stops = stopwords.words('english')
len(nltk_stops)
```

```
179
```

sklearn

```
from sklearn.feature_extraction import stop_words

skl_stops = list(stop_words.ENGLISH_STOP_WORDS)
len(skl_stops)
```

```
318
```

Initial stopwords 'library'

```
skl_stops.extend(nltk_stops)
stopwords = set(skl_stops)
len(set(stopwords))
```

```
378
```

	post	subreddit
1566	When the mushrooms wear off and you realize...	1
1567	Star Trek writer D.C. Fontana has passed away	0
1568	Mount Rushmore of Droids	1

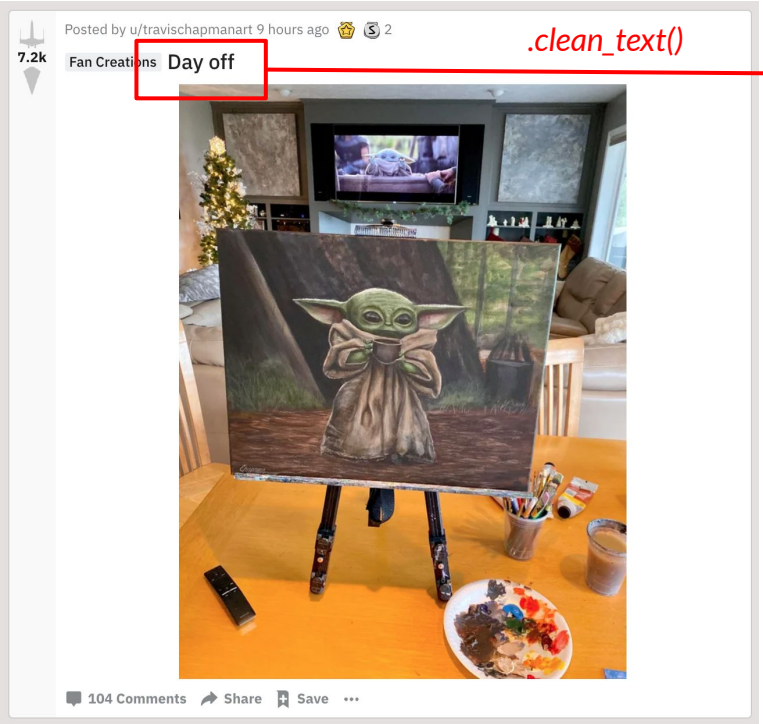
1 | StarWars
0 | startrek

def clean_text(doc):
1. *Tokenize*
2. *Lower*
3. *Remove stops*
4. *Lemmatize*
5. *Re-string*

	post	subreddit
1566	mushroom wear realize	1
1567	star trek writer c fontana passed away	0
1568	mount rushmore droids	1

to penalize common word ‘star’

Text Preprocessing - TFIDF Vectorization -> 3080 features



.clean_text()

	post	subreddit
1375	NaN	1

1 present in train set

impute

	post	subreddit
1375	a	1

unimportant word

learn!

model

predict

1? ***0?***

Model Selection

Dataset Review

- Categorical Target
- ‘Fat’ & ‘Short’ Dataset
- Sparse Matrix
- May be prone to noise (e.g. spurious postings that have no relevancy or just confusing postings)

Outcome : To get the highest accuracy score given dataset, time and computational constraint.

Machine Learning Model Selection

- Without any feature reductions, the ML that works best in ‘Fat’ & ‘Short’ dataset is SVM and Random Forest
- With feature reductions, dataset can become relatively ‘Thin’ & ‘Tall’. ML model that works best is Logistic Regression.
- With no constraint in time/computer resources, boosting trees ML may be considered.

Hyperparameter Tuning

Why hyperparameter tuning is needed

- Presence of ‘Noisy’ data points
- Prone to overfitting to such noise
- Potential non-linearity in dataset(?)
- Collinearity(?)

Machine Learning Model Hyperparameter Tuning

- Logistic Regression : regularization parameter (C)
- SVM : Kernel, regularization parameter (C)
- Random Forest : max depth, number of trees.
- Gradient Boosting : Learning rate, max depth, number of boosting stages

Score Comparisons

Advice from Andrew Ng for such a dataset like ours

Greater dimension than samples

Model	Accuracy Score	Interpretability	Speed	Notes
NB	87.9%	Medium	High	Non-parametric method.
LR	*88.5%*	*High*	*High*	*Regularization*
Linear SVM	*88.9%*	*Medium*	*Medium*	*Linear Kernel hyperparameter surprisingly work best, Work best when noise-to-signal ratio is high.*
RF	88.4%	Medium	Medium	Work best when dataset is ‘fat’ and ‘short’.
Boosting	85.6%	Medium	Low	Very slow and computationally intensive to tune for best performance. May work best if dataset turns unbalanced.

Feature Importances (via linear SVM and LR)

Linear SVM

features	svmcoef
war	3.203302
mandalorian	3.191092
yoda	2.146545
lightsaber	1.624892
jedi	1.606748
...	...
enterprise	-2.051433
picard	-2.272456
ds9	-2.408547
tng	-2.678605
trek	-4.913250

similar

same!

LogReg

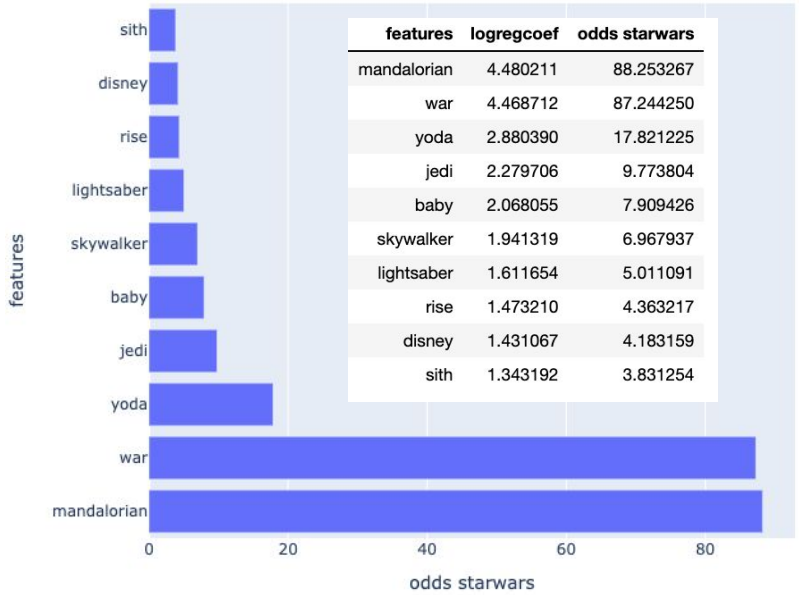
features	logregcoef
mandalorian	4.480211
war	4.468712
yoda	2.880390
jedi	2.279706
baby	2.068055
...	...
enterprise	-2.445224
picard	-2.640324
ds9	-2.700804
tng	-2.871139
trek	-6.220168

but...

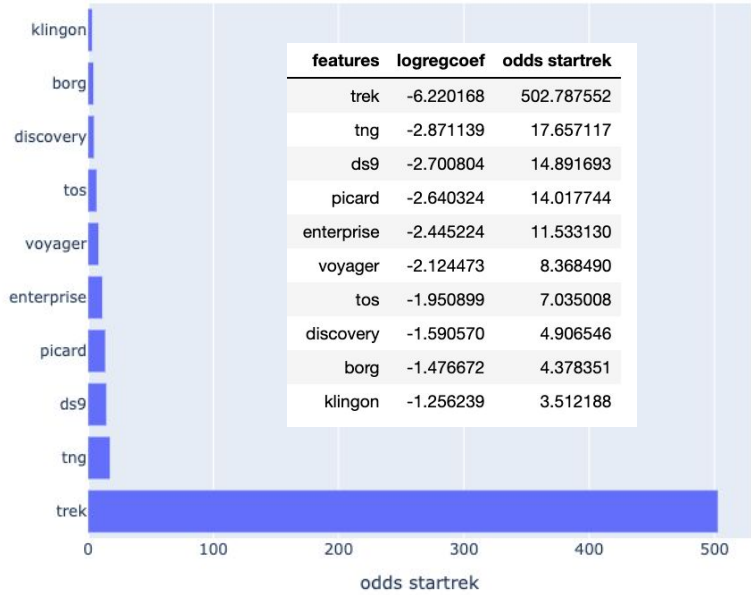
Feature Importances (via LR)

no ‘star’

$$ln\left(\frac{P(subreddit=StarWars)}{P(subreddit=startrek)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$



features	logregcoef	odds starwars
mandalorian	4.480211	88.253267
war	4.468712	87.244250
yoda	2.880390	17.821225
jedi	2.279706	9.773804
baby	2.068055	7.909426
skywalker	1.941319	6.967937
lightsaber	1.611654	5.011091
rise	1.473210	4.363217
disney	1.431067	4.183159
sith	1.343192	3.831254



features	logregcoef	odds startrek
trek	-6.220168	502.787552
tng	-2.871139	17.657117
ds9	-2.700804	14.891693
picard	-2.640324	14.017744
enterprise	-2.445224	11.533130
voyager	-2.124473	8.368490
tos	-1.950899	7.035008
discovery	-1.590570	4.906546
borg	-1.476672	4.378351
klingon	-1.256239	3.512188

Feature Importances (via RF and Gradient Boosting)

Random Forest

Stemmed Word	Feature Importance
enterpris	0.140824
trek	0.099557
season	0.02637
war	0.023613
archer	0.022653
star	0.019689
mandalorian	0.014065
theme	0.013378
song	0.012995
nx01	0.012459

Gradient Boosting

Stemmed Word	Feature Importance
enterpris	0.302887
trek	0.121199
theme	0.036631
archer	0.031019
season	0.021645
seri	0.016578
nx01	0.015864
catch	0.015013
episod	0.013313
tpol	0.011854

Feature Importance are ranked by how well they improve the purity of the node (Gini Purity)

Conclusion & Recommendation

- Provide social media intelligence for e-commerce websites, for eg.

- If features related to ‘topic’ is important
  → Logistic Regression
  → accurate and fast

- If labelling is unimportant
  → Random Forest
  → reduces collinearity
  → generalize better for any 2 to classify