

A close-up shot of Ewan McGregor with a surprised expression, looking directly at the camera. He has a light beard and is wearing a brown robe over a dark shirt. In the background, several droids are visible, including a large black one with a red light on its chest and several smaller tan ones. The scene is dimly lit, suggesting an indoor setting like a workshop or a base.

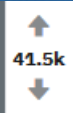
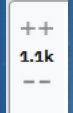





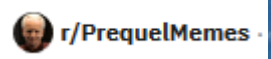

HELLO THERE.

Outline

- Background
 - Terminology
 - What is Reddit?
- Corpus
 - Where and Overall Size
 - Data Structure
- Machine Learning
 - CRC and Models Tried
 - Features

Terminology



- Subreddit – Basically a subforum or community – shortened = “sub”
- Karma – Score or Upvotes/Downvotes – will use terms interchangeably  41.5k  1.1k  3.8k
- Gild – A silver/gold/platinum awarded to a post by another user
 - This funds reddit  5  
- Linking – user profiles - /u/<username> 
- Subreddit link – r/<subreddit> 
- Cakeday – anniversary of when a user signed up 

Terminology

- Submission – the “original” post so to speak
- Comments – discussion/commenting on a submission
- Flair – equivalent to a title or status, specific to each subreddit

[u/ducktor-strange](#) Hello there!

What is Reddit?



KEEP
CALM
THERE'S
A SUBREDDIT
FOR THAT



- Created June 23, 2005
- Social Media Platform
 - Forum broken up into millions of sub-forums or rather “subreddits”
 - Want to discuss some particular topic?
- Did you know it can also produce coffee on the spot?
 - No? Good.
 - Because it doesn't
- Is Reddit Beautiful or is it Ugly?

YES

What is Reddit?

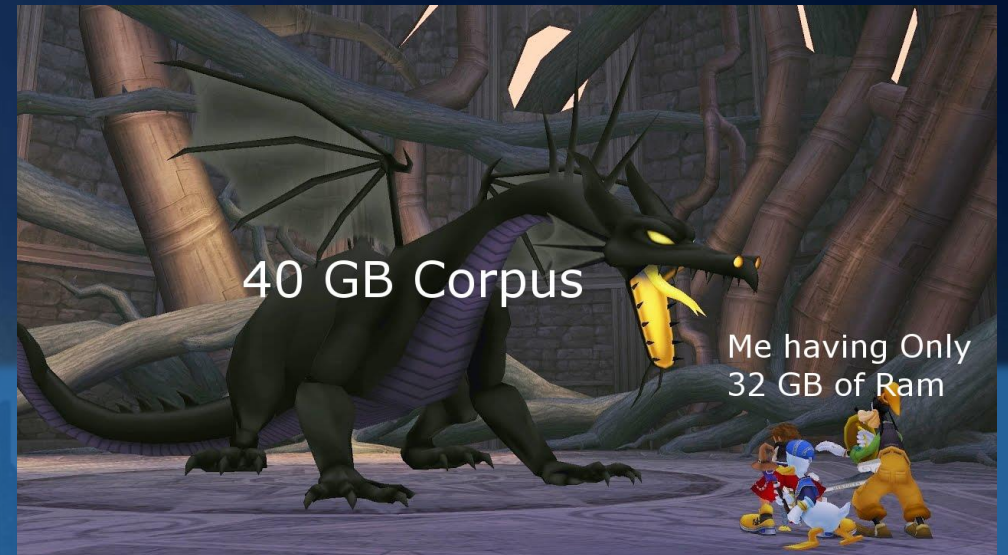
- Let's Dive straight in Shall we?
- But wait which format to use?
- Old Reddit
- <https://old.reddit.com/>
- Old Reddit
 - Better for Discussion/Text based Subs
 - Thus, better frontpage
- New Reddit
 - Better for Picture Based Subs

New Reddit

<https://www.reddit.com/>

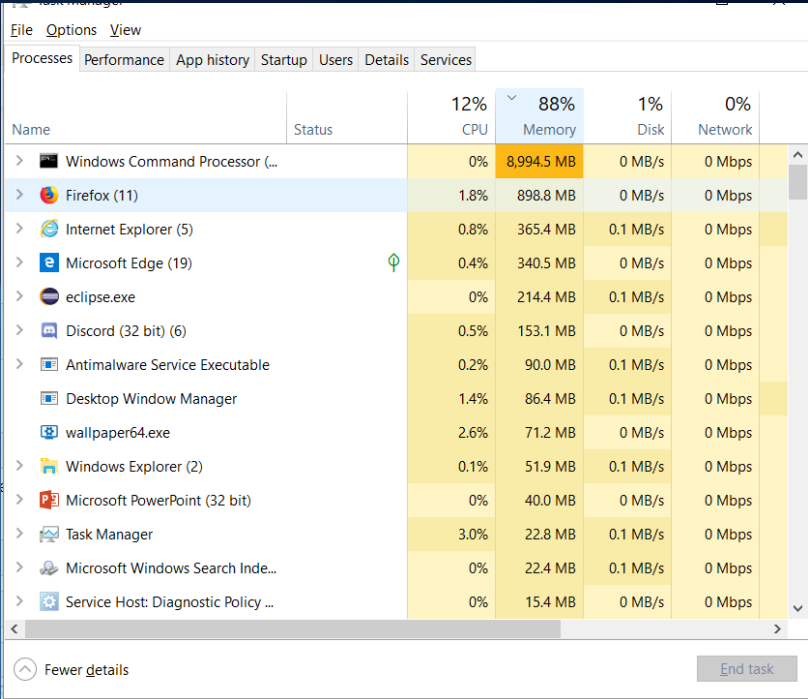
Corpus

- Found ... on a [Reddit Post](#)
- [Actual Corpus Site](#)
- Initially tried one month – 3GB compressed, 40GB uncompressed
- Then Moved onto one-day – 259 MB compressed, 2 GB uncompressed
- One day corpus Size:
 - Almost 3 million comments (2,929,215 to be exact)
 - 105,876,770 words
 - 38.7 Words per comment average



Corpus

- Why use only one day?
- Hazards of such
 - New meme formats/ideas
 - Holidays! (This particular corpus December 21)
 - Seasonal Variations
- Ideal Data
 - Multiple Months spread throughout the year
 - Multiple Years



The screenshot shows the Windows Task Manager Performance tab. The 'Memory' section is expanded, showing 88% usage (8,994.5 MB). The 'Processes' tab is also visible, showing a list of running applications and their resource usage. The table below represents the data shown in the 'Processes' tab.

Name	Status	12% CPU	88% Memory	1% Disk	0% Network
Windows Command Processor (...)		0%	8,994.5 MB	0 MB/s	0 Mbps
Firefox (11)		1.8%	898.8 MB	0 MB/s	0 Mbps
Internet Explorer (5)		0.8%	365.4 MB	0.1 MB/s	0 Mbps
Microsoft Edge (19)		0.4%	340.5 MB	0 MB/s	0 Mbps
eclipse.exe		0%	214.4 MB	0.1 MB/s	0 Mbps
Discord (32 bit) (6)		0.5%	153.1 MB	0 MB/s	0 Mbps
Antimalware Service Executable		0.2%	90.0 MB	0.1 MB/s	0 Mbps
Desktop Window Manager		1.4%	86.4 MB	0.1 MB/s	0 Mbps
wallpaper64.exe		2.6%	71.2 MB	0 MB/s	0 Mbps
Windows Explorer (2)		0.1%	51.9 MB	0.1 MB/s	0 Mbps
Microsoft PowerPoint (32 bit)		0%	40.0 MB	0 MB/s	0 Mbps
Task Manager		3.0%	22.8 MB	0.1 MB/s	0 Mbps
Microsoft Windows Search Inde...		0%	22.4 MB	0.1 MB/s	0 Mbps
Service Host: Diagnostic Policy ...		0%	15.4 MB	0 MB/s	0 Mbps

```
: redditframe.head()
```

	author	author_cakeday	author_flair_css_class	author_flair_text	body	can_gild	controversiality	distinguished	gilded	id	is_submitter	
0	StrayYoshi	0.0	asc-hierophant	Hierophant	Have to kill him in 1 portal is the way I do i...	True	0	none	0	drjocb3	False	t3_
1	vgeh	0.0	none	none	3 season- Mostly thrunite ti3. \n\n\n3 season ...	True	0	none	0	drjocb5	False	t:
2	Litbus_TJ	0.0	none	none	Beautiful, isn't it?	True	0	none	0	drjocb6	False	t3

link_id	parent_id	permalink	score	stickied	subreddit	subreddit_id	subreddit_type
t3_7l0zv1	t3_7l0zv1	/r/pathofexile/comments/7l0zv1/ggg_please_fix_...	1	False	pathofexile	t5_2sf6m	public
t3_7klrjl	t1_drjkcuz	/r/Ultralight/comments/7klrjl/rultralight_disc...	3	False	Ultralight	t5_2s7p2	public

Machine Learning

- Chose 14 subreddits to fine tune my Models
 - relationships, aww, nfl, PrequelMemes, gaming, mildlyinteresting, politics, Showerthoughts, worldnews, gifs, StarWars, funny
 - Filtered by total karma score of 50 or greater
- Largest Data Portion – Baseline Percent
 - Politics – 22%

Comment Totals

```
In [8]: redditframe.subreddit.value_counts()
```

```
Out[8]: AskReddit      186327
         politics      57114
         nba           34359
         The_Donald    33815
         news          31364
         worldnews     30099
         StarWars      25322
         survivor      22102
         Bitcoin       21427
         nfl           19521
         movies        17501
         DestinyTheGame 17374
         PUBBATTLEGROUND 16511
         videos        16101
         gaming        15940
         pics          15838
         leagueoflegends 15377
         funny         15216
         Showerthoughts 14788
         CryptoCurrency 14217
         todayilearned  13474
         CFB           12190
         fantasyfootball 11753
         soccer        11514
         teenagers     11252
         SquaredCircle 11028
         RocketLeagueExchange 10505
         btc           10340
         DBZDokkanBattle 9608
         europe        9526
```

Comment Totals (filtered)

```
In [41]: redditframe[redditframe.score >= 50].subreddit.value_counts()
```

```
Out[41]: AskReddit      6735
         nba            1813
         politics       1786
         The_Donald    1435
         nfl           1019
         worldnews      995
         StarWars       851
         videos         766
         movies         760
         soccer         698
         gaming         691
         news           689
         survivor       664
         todayilearned  653
         relationships  631
         CFB            621
         leagueoflegends 596
         funny          580
         pics           553
         SquaredCircle  536
         Showerthoughts 479
         hiphopheads    441
         BlackPeopleTwitter 404
         aww            362
         baseball       332
         gifs           324
         europe         323
         DestinyTheGame 321
         WTF            321
         Overwatch      302
```


Most negative comments

```
In [5]: redditframe[redditframe.score < 0].subreddit.value_counts()
```

```
Out[5]: politics      4723
AskReddit      4509
worldnews      3250
news           2709
nba            2238
StarWars       1365
videos         1180
nfl            1130
movies         1110
leagueoflegends 1045
PUBATTLEGROUNDS 1015
soccer         1014
DestinyTheGame  993
europe         973
pics           795
gaming         794
Bitcoin        774
conspiracy     761
survivor       757
todayilearned  735
SquaredCircle  706
funny          660
PoliticalHumor  651
DotA2          629
Showerthoughts 613
canada         554
Cryptocurrency 553
technology     498
Games          469
australia      429
```

Most Positive comments

```
In [9]: redditframe[redditframe.score > 0].subreddit.value_counts()
```

```
Out[9]: AskReddit      176476
politics      49591
The_Donald    33080
nba           30907
news          26479
worldnews     24598
StarWars      22675
survivor      20599
Bitcoin       19473
nfl           17656
movies        15498
DestinyTheGame 15290
gaming        14506
PUBATTLEGROUNDS 14499
pics          14410
videos        14051
funny         13947
Showerthoughts 13369
leagueoflegends 13241
Cryptocurrency 13061
todayilearned 12078
CFB           11539
fantasyfootball 11252
teenagers     10987
RocketLeagueExchange 10095
soccer        10034
SquaredCircle  9863
btc           9436
DBZDokkanBattle 9004
FIFA          8824
```

Most neutral comments

```
In [10]: redditframe[redditframe.score == 0].subreddit.value_counts()
```

```
Out[10]: AskReddit          5342
         politics          2800
         worldnews         2251
         news             2176
         StarWars         1282
         nba              1214
         Bitcoin          1180
         leagueoflegends  1091
         DestinyTheGame   1091
         PUBATTLEGROUNDS   997
         movies           893
         videos           870
         Showerthoughts    806
         survivor          746
         nfl              735
         todayilearned     661
         gaming           640
         pics             633
         funny            609
         conspiracy        606
         Cryptocurrency    603
         btc              504
         technology        498
         Dota2            486
         europe           476
         soccer           466
         SquaredCircle     459
         The_Donald        454
         CringeAnarchy     433
         gifs             418
```



```
above50t.groupby("subreddit").score.describe()
```

Out[36]:

	count	mean	std	min	25%	50%	75%	max
subreddit								
PrequelMemes	130.0	150.469231	163.691437	50.0	63.25	90.0	149.50	1164.0
Showerthoughts	479.0	400.661795	1040.023066	50.0	73.50	131.0	277.00	13328.0
StarWars	851.0	149.481786	222.027572	50.0	64.00	90.0	156.00	3692.0
aww	362.0	376.994475	1135.255738	50.0	69.00	110.0	247.00	16122.0
funny	580.0	450.589655	1222.346296	50.0	72.00	117.5	270.25	11239.0
gaming	691.0	419.299566	1256.652680	50.0	70.00	119.0	275.50	22738.0
gifs	324.0	757.317901	2844.763898	50.0	72.75	122.0	289.25	28613.0
mildlyinteresting	206.0	492.194175	1340.742110	50.0	71.00	115.5	356.00	14349.0
nfl	1019.0	175.197252	352.575079	50.0	65.00	91.0	164.00	8178.0
politics	1786.0	228.754759	498.984808	50.0	65.00	101.0	192.00	7727.0
relationships	631.0	154.890650	183.116709	50.0	64.00	94.0	167.00	2157.0
worldnews	995.0	399.067337	939.977101	50.0	72.50	129.0	328.50	11045.0

```
In [33]: pd.set_option('display.max_colwidth', -1)
         above50t[above50t.score > 25000].permalink
```

Out[33]: 2679664 /r/gifs/comments/7l1dc7d/its_all_downhill_from_here/dr1e9vt/
Name: permalink, dtype: object

```
In [34]: above50t[above50t.score > 25000].body
```

Out[34]: 2679664 Well at least he is on the right side to go back up and give it a second try.
Name: body, dtype: object

Machine Learning

- Unfortunately, nothing happened at all. Still at 22%



Machine Learning – The Real Truth

- No, in all seriousness, I've made plenty of progress on the this front
- Currently using two models : MultinomialNB, Support Vectors
- Hope to use more if possible
- With minimum setting adjustment –
- Naïve Bayes 36%
- Support Vectors 48%
- 14% Increase from Baseline

CRC

- Could handle data, but CRC to speed up Grid Search
- Support Vectors... giving the most trouble....

```
print("Multinomial Best Parameters")
print("-----")
model = Pipeline(steps=[('Tfidf', TfidfVectorizer(min_df=2)), ('MNB', MultinomialNB)])
param_grid = {
    "Tfidf__max_features": [None, 1500, 3000, 5000, 7500],
    "Tfidf__min_df": [1, 2, 3, 4, 5],
    "Tfidf__norm": ['l1', 'l2', None],
    "Tfidf__ngram_range": [(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)],
    "MNB__alpha": [0.01, 0.25, 0.5, 0.75, 1.00]
}
grid = GridSearchCV(model, param_grid, n_jobs=4, cv=5)
grid.fit(above50t["body"], above50t["subreddit"])
print("Best Parameters :")
print(grid.best_params_)

print("\n")
print("Support Vector Parameters")
print("-----")
model = Pipeline(steps=[('Tfidf', TfidfVectorizer(min_df=2)), ('SVC', SVC(C=1E5))])
param_grid = {
    "Tfidf__max_features": [None, 1500, 3000, 5000, 7500],
    "Tfidf__min_df": [1, 2, 3, 4, 5],
    "Tfidf__norm": ['l1', 'l2', None],
    "Tfidf__ngram_range": [(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)],
    "SVC__kernel": ["linear", "poly", "rbf", "sigmoid"],
    "SVC__gamma": ['auto', 'scale']
}
grid = GridSearchCV(model, param_grid, n_jobs=10, cv=5)
grid.fit(above50t["body"], above50t["subreddit"])
print("Best Parameters :")
print(grid.best_params_)
```


Support Vector Parameters

multiprocessing.pool.RemoteTraceback:

"""

Traceback (most recent call last):

```
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 3:
    return self.func(*args, **kwargs)
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__
    return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in <listcomp>
    return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/model_selection/_validation.py", line 458, in _fit_estimator
    estimator.fit(X_train, y_train, **fit_params)
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/pipeline.py", line 250, in fit
    self._final_estimator.fit(Xt, y, **fit_params)
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/svm/base.py", line 187, in fit
    fit(X, y, sample_weight, solver_type, kernel, random_seed=seed)
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/svm/base.py", line 276, in _sparse_fit
    random_seed)
File "sklearn/svm/libsvm_sparse.pyx", line 75, in sklearn.svm.libsvm_sparse.libsvm_sparse_train
TypeError: must be real number, not str
```

During handling of the above exception, another exception occurred:

Traceback (most recent call last):

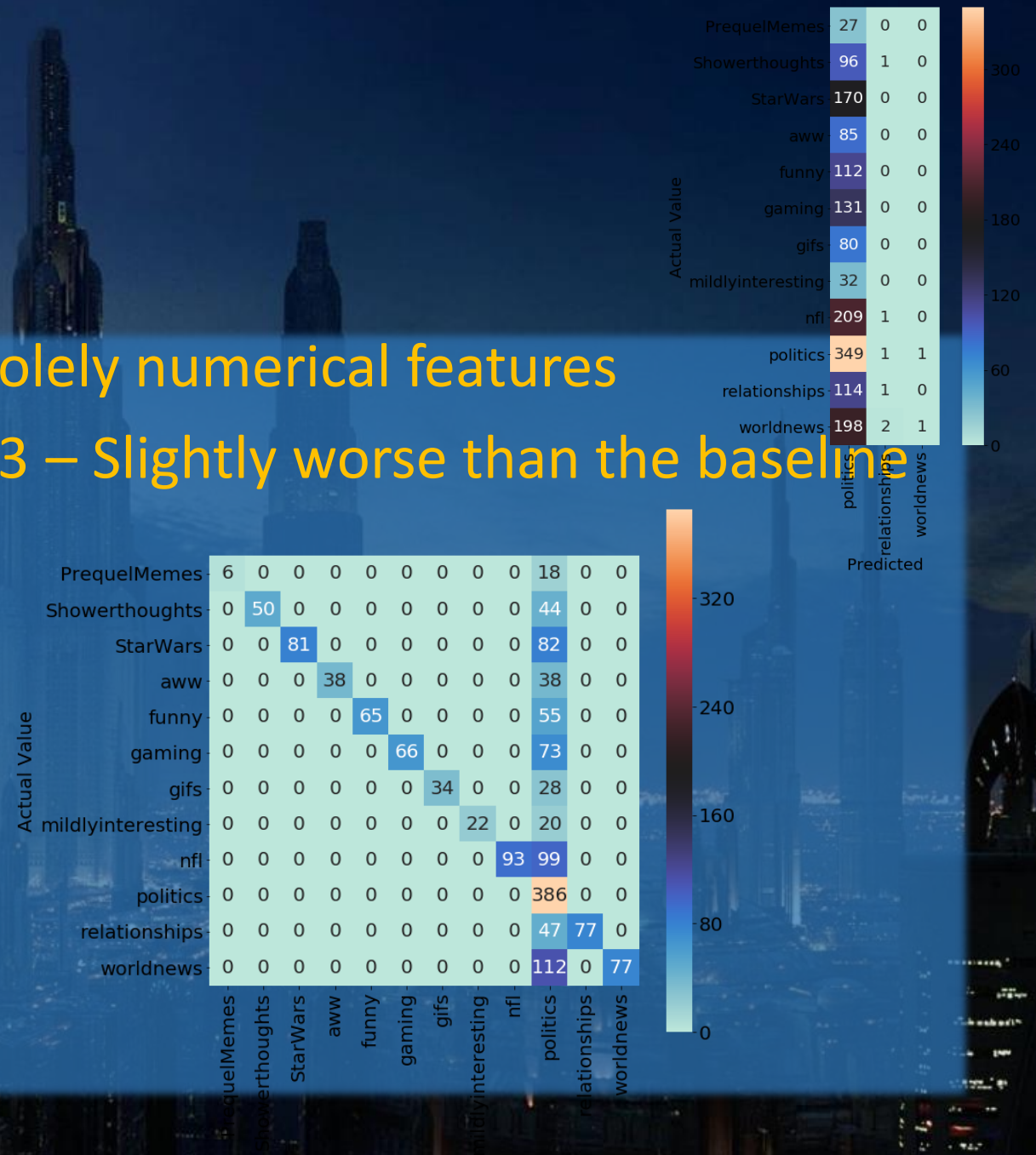
```
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/multiprocessing/pool.py", line 121, in worker
    result = (True, func(*args, **kwds))
File "/ihome/crc/install/python/miniconda3-3.7/lib/python3.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 3:
    raise TransportableException(text, e.type)
```

CRC Results

- To compensate for the error, I used the Tfidf features found for the Naïve Bayes, ran grid search locally on the SVC unique parameters
- “None” was found to be the best for max features, but after some discussion, caps are more practical, so I slapped a 15000 max features limit
- NB – 58% CV – 54%
- SVC – 52% CV – 48%

Features

- What about numerical features solely numerical features
- Accuracy - 0.21787709497206703 – Slightly worse than the baseline
- Score alongside Text?
 - 54% on Naïve Bayes
- What about parent ID's?
 - Basically cheating



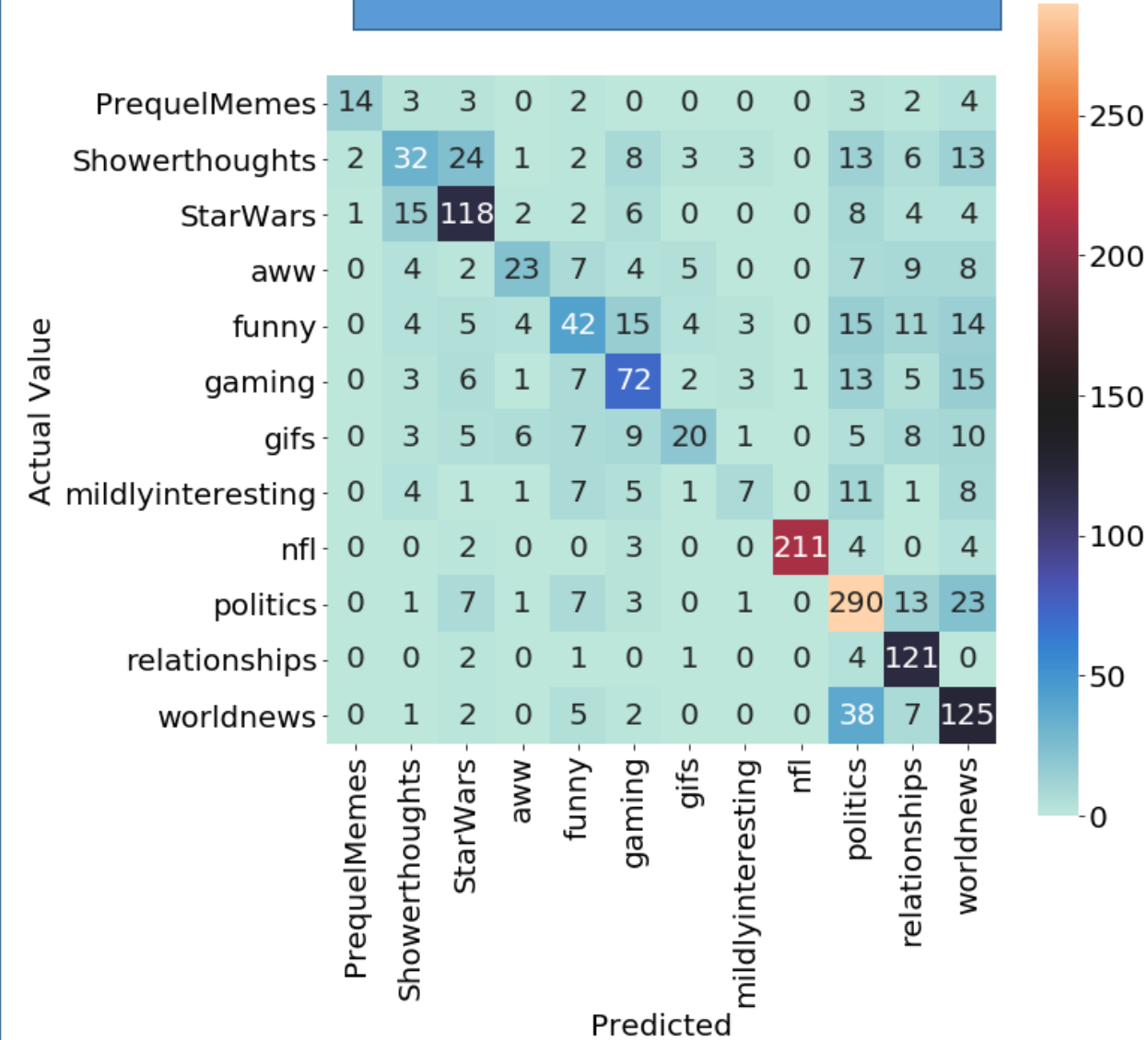
Best run – with flairs

- Highest Percentage - About 67% with Multinomial Naïve Bayes
- Support Vectors came behind at 64% - 3% difference
- 5-fold Crossvalidation :
 - 62% and 57% respectively
- Support Vectors could be tuned more

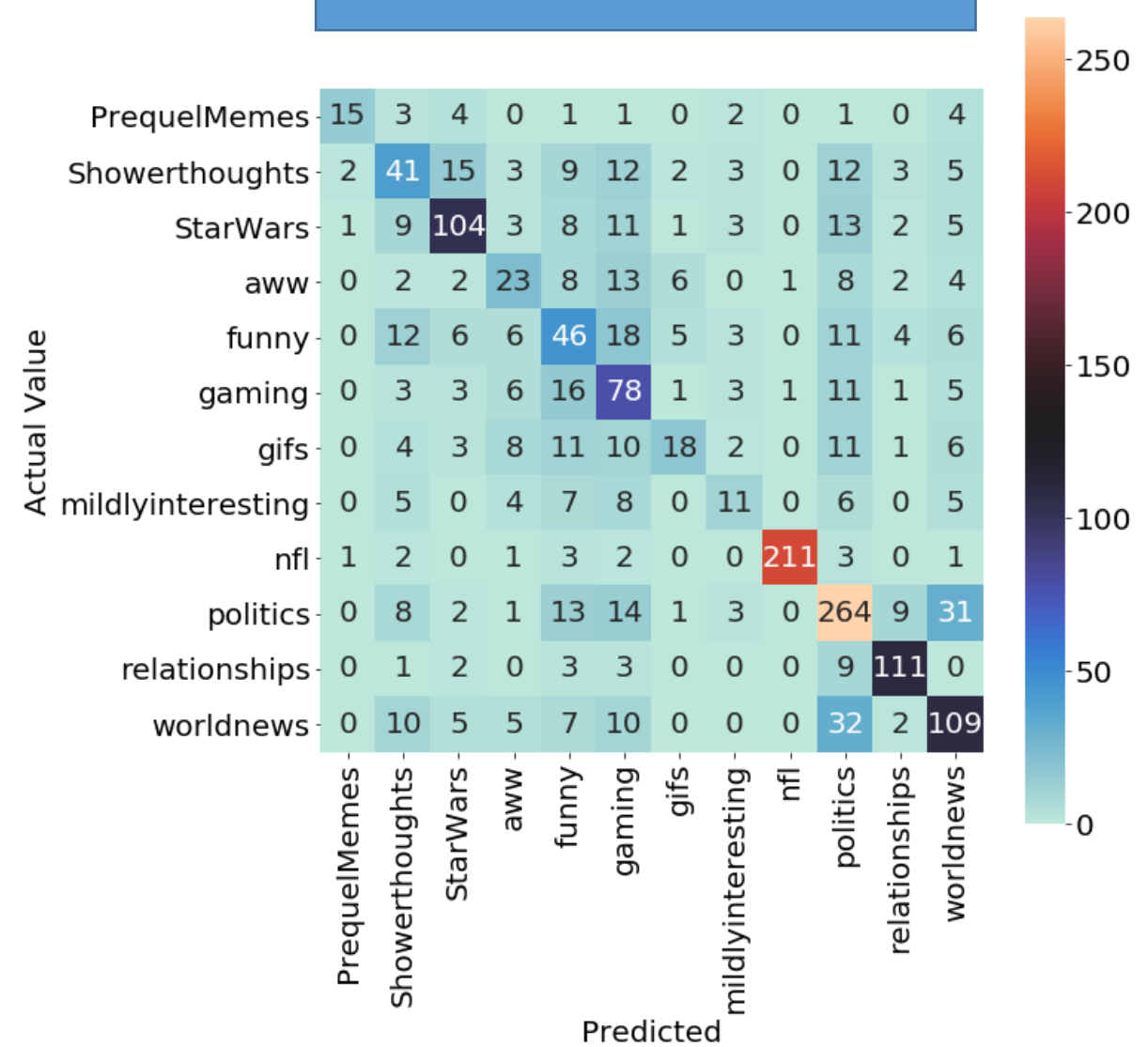


Written and Directed by
GEORGE LUCAS

Multinomial Naïve Bayes

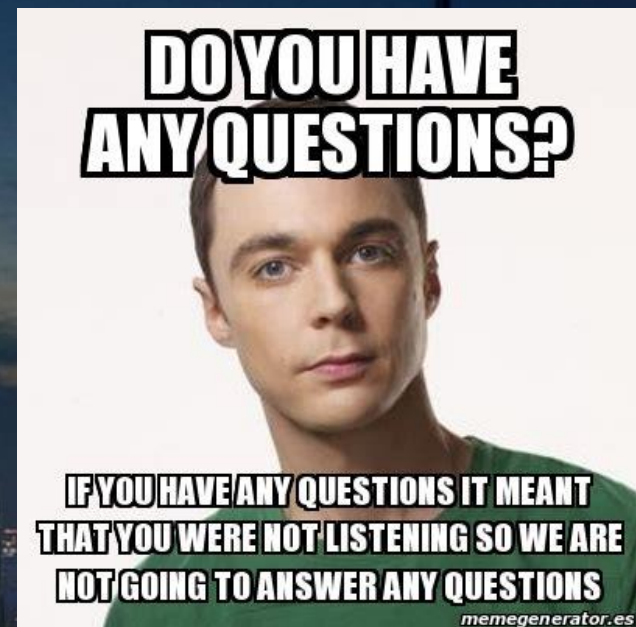


Support Vectors



Want more?

- Checkout r/SubredditSimulator
 - <https://www.reddit.com/r/SubredditSimulator/>
- Bots that generate random submissions and comments
 - Model used: Markov Chains





ONE DOES NOT SIMPLY

**SAY THANK YOU WITHOUT A
MEME**

makeameme.org

