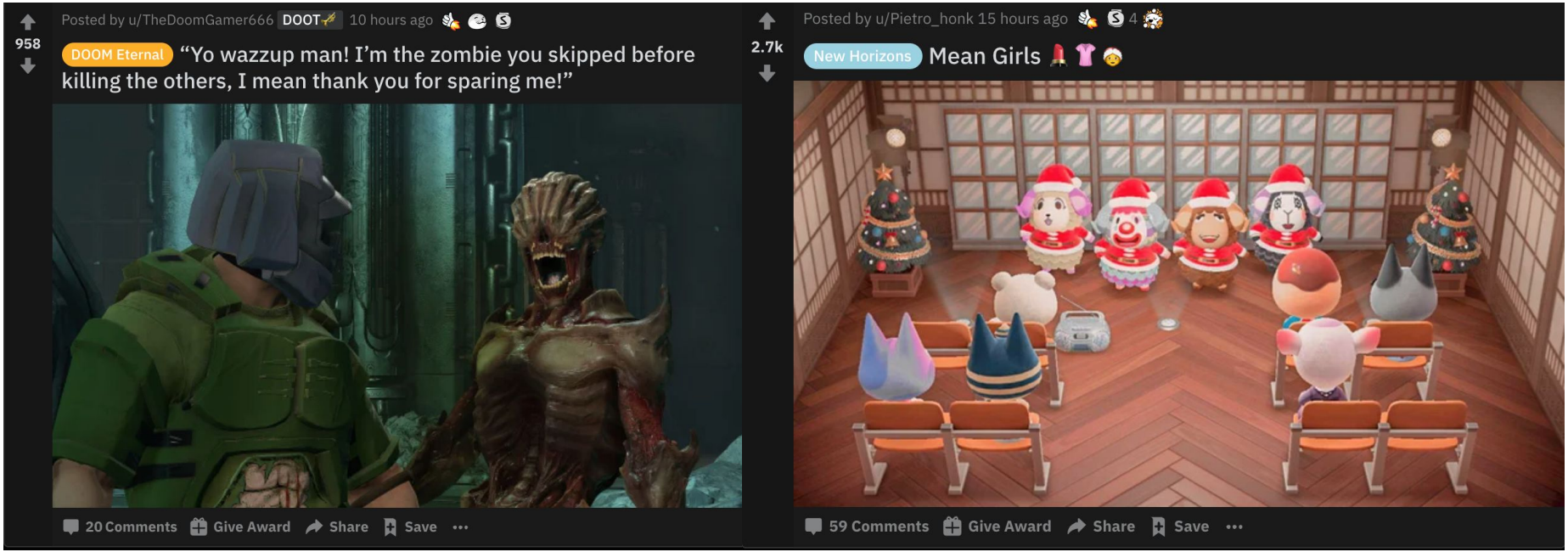
Is it Doom? Or Animal Crossing

How do we know?

Typical posts from r/Subreddit Doom, Animal Crossing



What is most important to classifying these posts?
Are the features very different? How many important features do they have in common?

Tokenize/TF-IDF the words and Clustering the colors

KMeans clustering

Find RGB Values for each pixel for each document-round to nearest 5

Find Count Each Cluster RGB

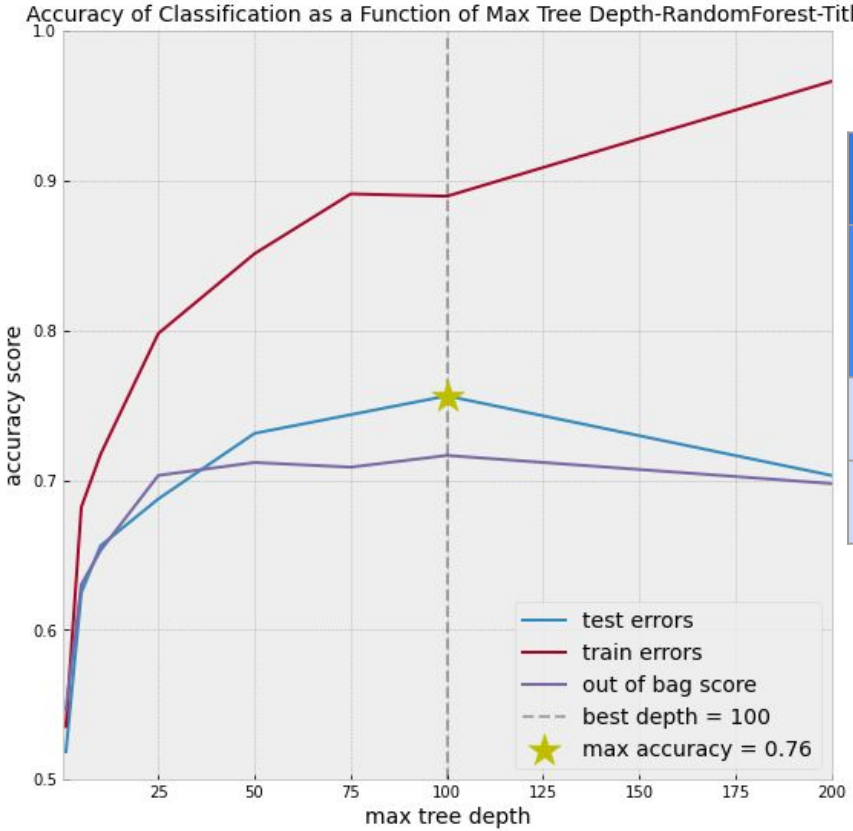
Order by count, then split into count, red, green and blue

	filename	100.0, 125.0, 30.0	100.0, 175.0, 105.0	115.0, 145.0, 45.0	125.0, 115.0, 85.0	140.0, 165.0, 75.0	140.0, 215.0, 240.0
0	yjrafuspulu41.png	7526.0	2186.0	9001.0	2743.0	8390.0	2098.0
1	2ehmq0fjpj051.png	0.0	0.0	0.0	0.0	0.0	0.0
2	a2vvzng7g7v41.png	0.0	0.0	0.0	0.0	0.0	0.0
3	t88z1sn736x41.png	0.0	0.0	0.0	0.0	0.0	0.0
4	176wf6duetq41.png	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
752	23tuxt9bkev41.png	0.0	0.0	0.0	0.0	0.0	0.0
753	7o0coqjaqfo41.png	0.0	0.0	0.0	0.0	0.0	0.0
754	w0ddyr9atvu41.png	0.0	0.0	0.0	0.0	0.0	0.0
755	2v9ke0u7u3251.png	0.0	0.0	0.0	0.0	0.0	0.0
756	stvociaym2o41.png	0.0	0.0		0.0	0.0	0.0

757 rows × 11245 columns

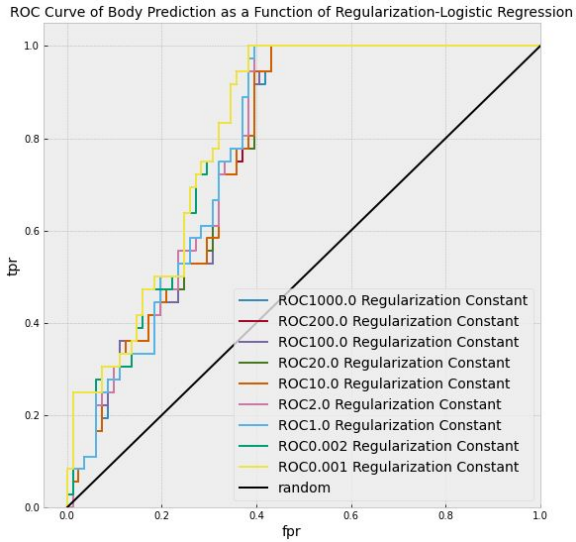
7g	7r	8b	8count	8g	8r	9b
70.0	80.0	205.0	1983.0	200.0	210.0	90.0
65.0	110.0	110.0	2963.0	100.0	120.0	95.0
250.0	155.0	220.0	3253.0	255.0	195.0	170.0
175.0	175.0	200.0	839.0	200.0	200.0	120.0
100.0	120.0	20.0	1436.0	30.0	185.0	95.0
...	...	...	...	...	...	...
85.0	120.0	135.0	3736.0	175.0	10.0	80.0
215.0	240.0	90.0	2511.0	65.0	100.0	195.0
15.0	45.0	90.0	3823.0	50.0	130.0	175.0
0.0	50.0	145.0	1951.0	90.0	195.0	105.0
95.0	95.0	135.0	723.0	135.0	135.0	150.0

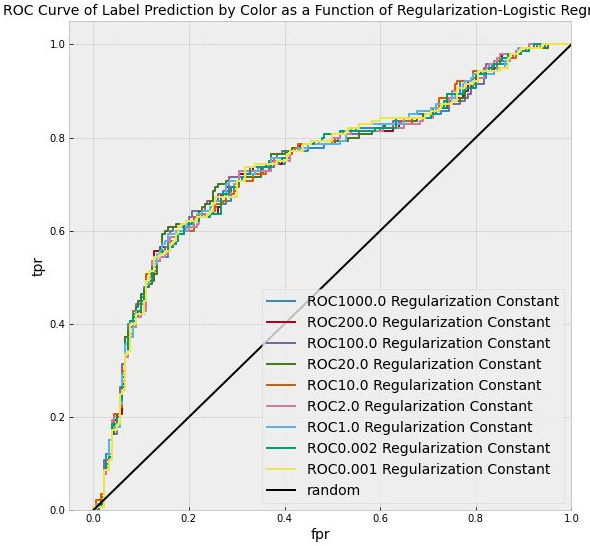
Render unto animal crossing that which is animal crossing's

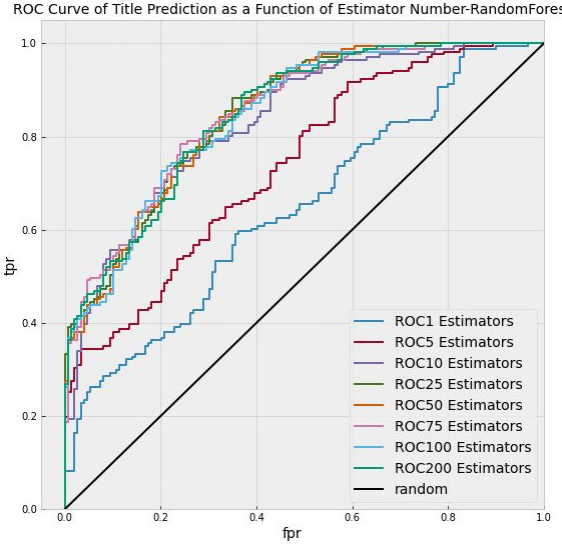


Accuracy Data for Model Types				
	Logistic Regression (color)	Random Forest (color)	Gradient Boost (color)	Logistic Regression (body)
Colors/Body	0.68	0.75	0.76	0.88
Title	0.71	0.71	0.73	0.81

Where do we go wrong?



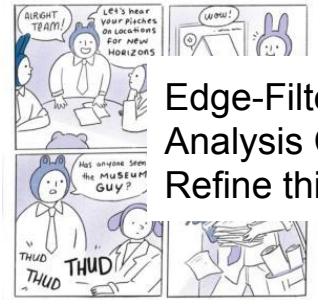




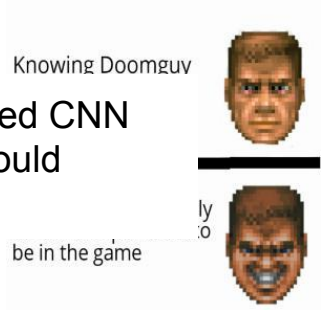
Predicted AC: ‘Mine was 16 and I got it on Arc Complex’

Sentiment Analysis Could Refine this

Pre
care : easy, how is everyone else doing this so well? Am I being stupid or is their a method to doing this?"



Edge-Filtered CNN Analysis Could Refine this



Predicted AC: 'A shot I took in Camera Mode. Really like how it looks.'

Pre

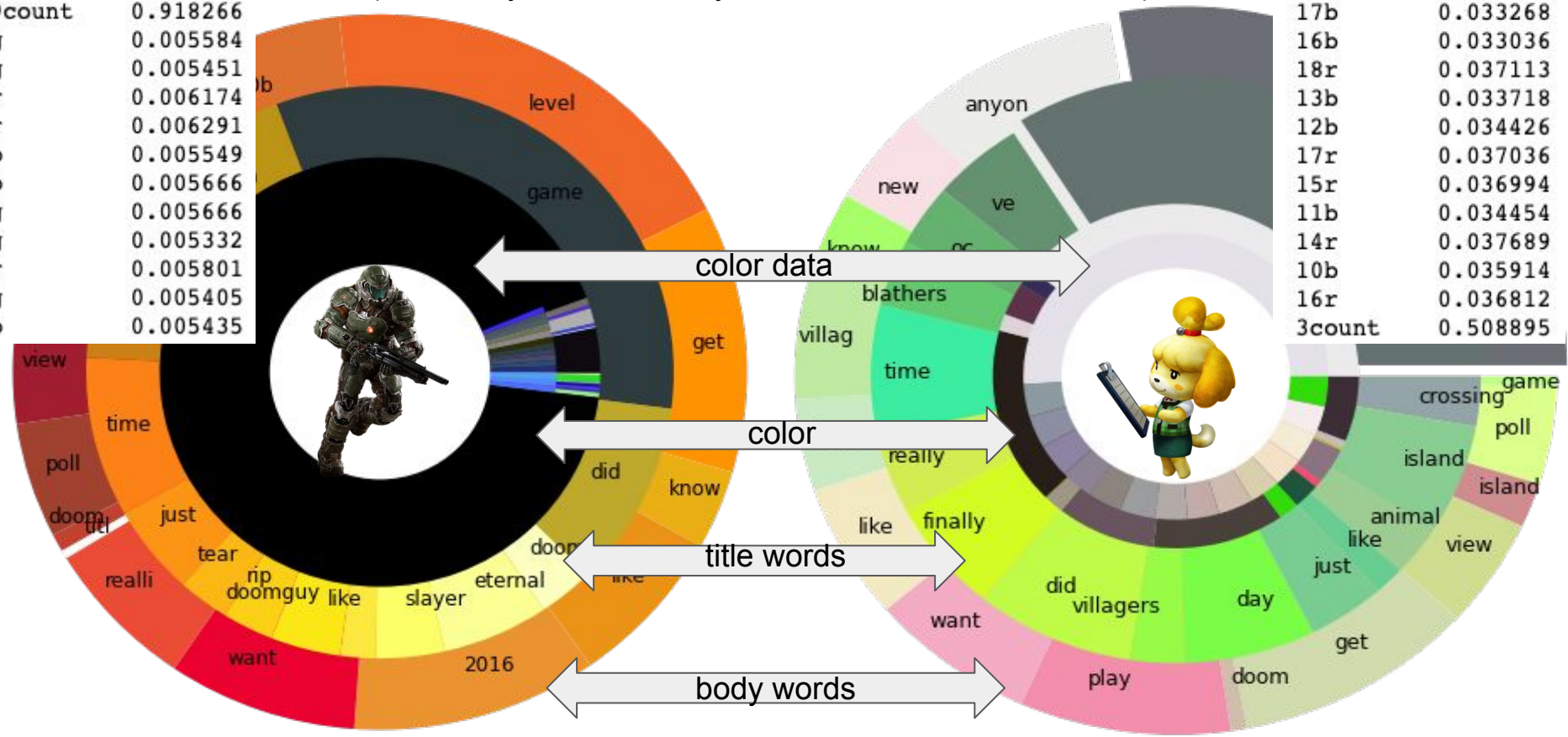
Sentiment Analysis May or May Not Refine this

Building Components to Increase Certainty

(Naive Bayes - Probability of Feature Given A Subreddit)

0r	0.006519
1r	0.006479
2r	0.006385
19count	0.918266
1g	0.005584
2g	0.005451
4r	0.006174
3r	0.006291
1b	0.005549
0b	0.005666
0g	0.005666
4g	0.005332
6r	0.005801
3g	0.005405
2b	0.005435

18b	0.034117
14b	0.033284
15b	0.033244
17b	0.033268
16b	0.033036
18r	0.037113
13b	0.033718
12b	0.034426
17r	0.037036
15r	0.036994
11b	0.034454
14r	0.037689
10b	0.035914
16r	0.036812
3count	0.508895



Next Steps: use a neural network for cumulative probability

Questions?