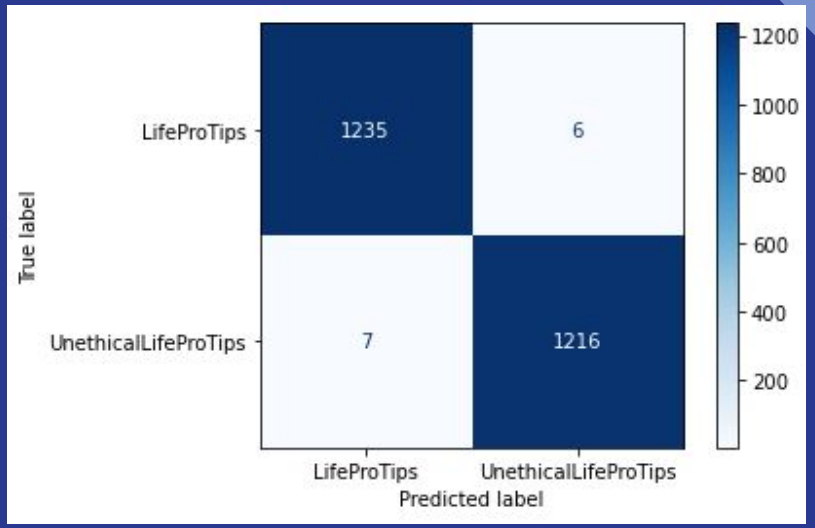
Binary Classification of Reddit Posts

/r/LifeProTips
/r/UnethicalLifeProTips

No Stop Words

Trivial problem:

Base parameters achieve .9946 accuracy



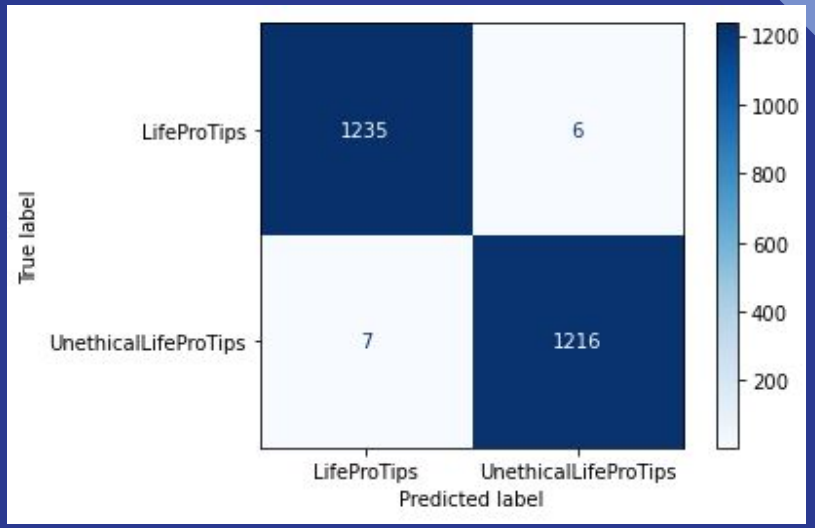
No Stop Words

Trivial problem:

Base parameters achieve .9946 accuracy

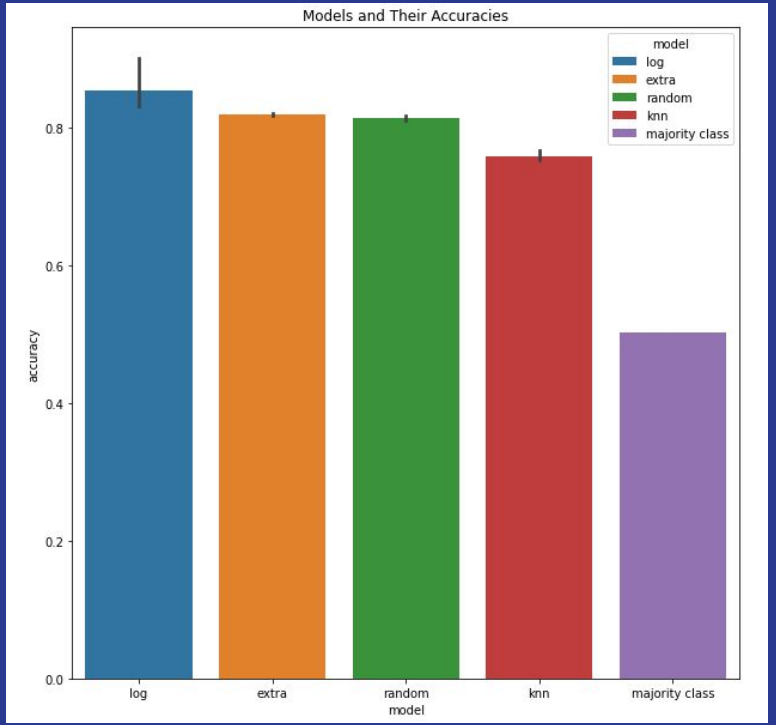
Top 5 words by count:

1. to - 5266
2. you - 4706
3. the - 3947
4. lpt - 3689
5. ulpt - 3557



Explored Models

1. Logistic Regression (.832792)
2. Extra Trees Classifier (.821834)
3. Random Forest Classifier (.819508)
4. KNN (.767857)

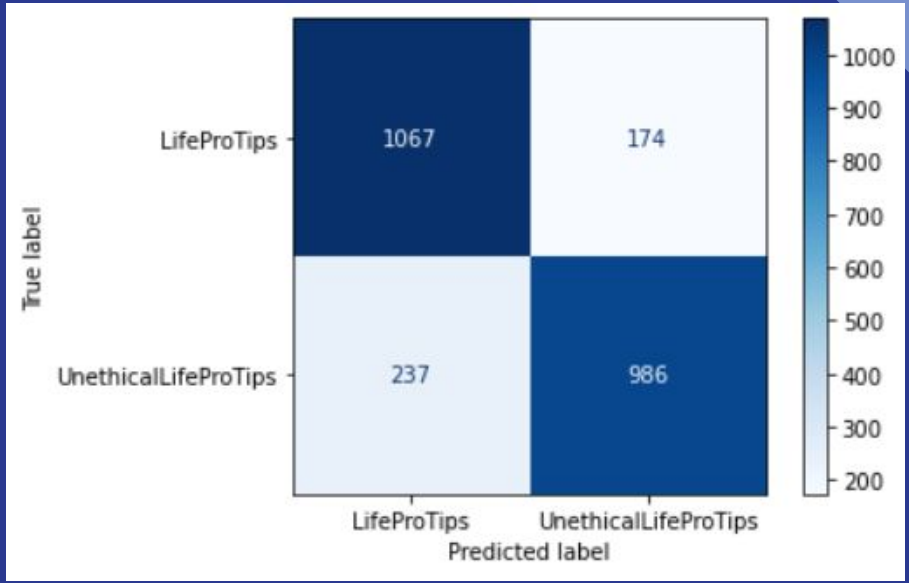


Best Model:

Train (0.9953)
Test (0.8332)

- Logistic Regression
- Lemmatized
- Count Vectorized
  - ngram_range (1, 2)
  - stop_words [‘lpt’, ‘ulpt’]

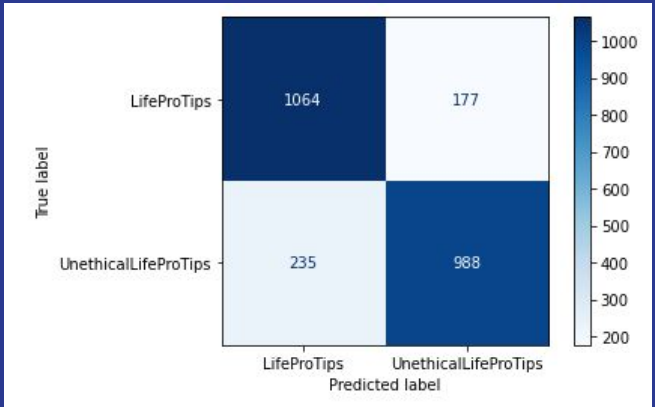
- Additional Features
  - Title word count
  - Avg word length in title



Areas for Improvement

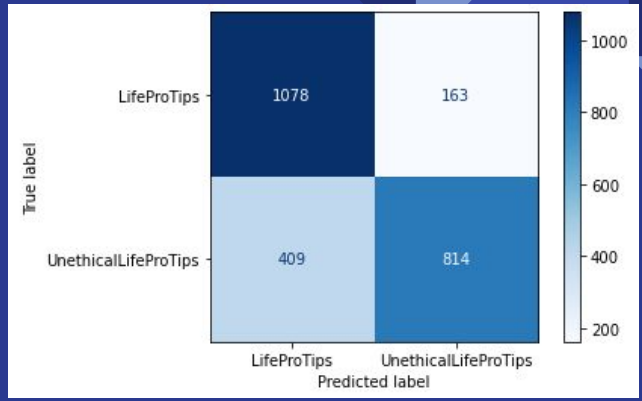
1. Using the body of the text or even comments could be considered as in some cases titles are very short or simple and the body contains the majority of relevant text. And in the case of ‘request’ type posts the comments are an important part of the post as a whole.
2. Sentiment analysis could prove to be beneficial given the nature of these subreddits.

Logistic Regression



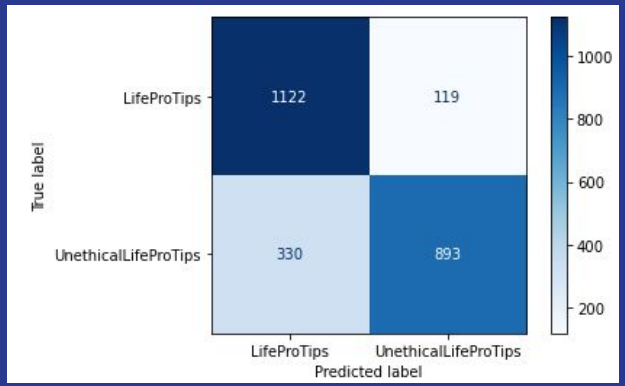
Total Misclassifications: 1894

K Nearest Neighbors



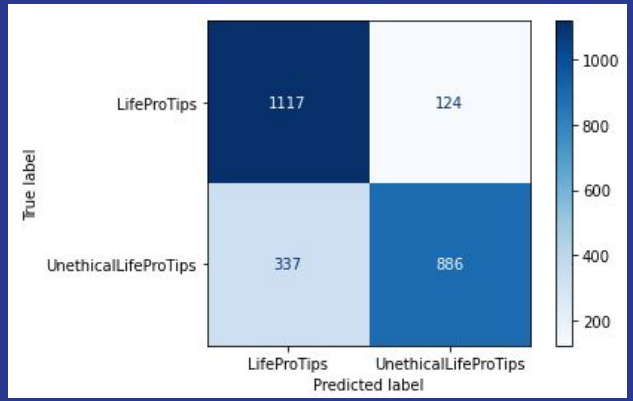
Number of Misclassifications Without Duplicates: 715

Extra Trees Classifier

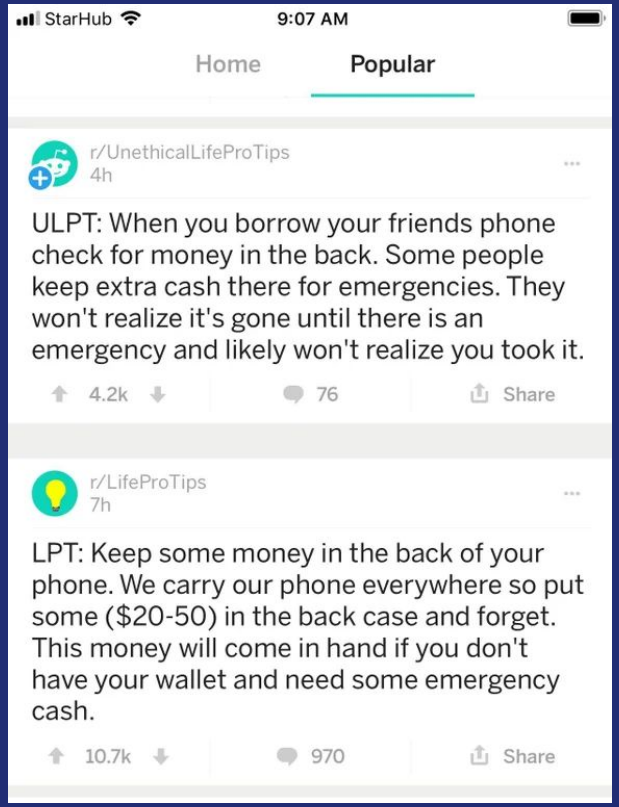


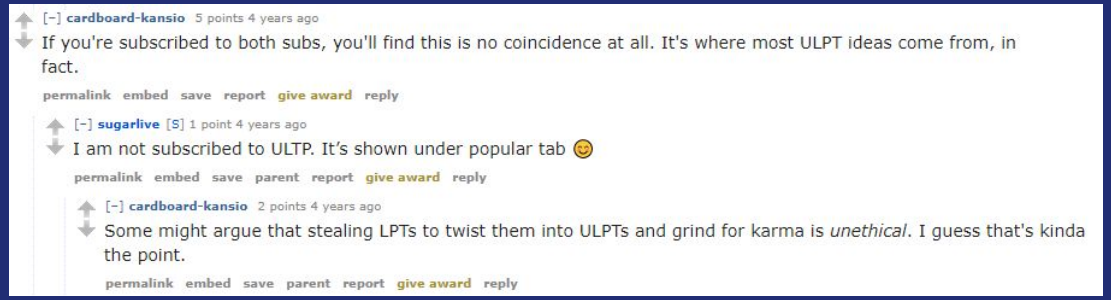
Averages: 2.65

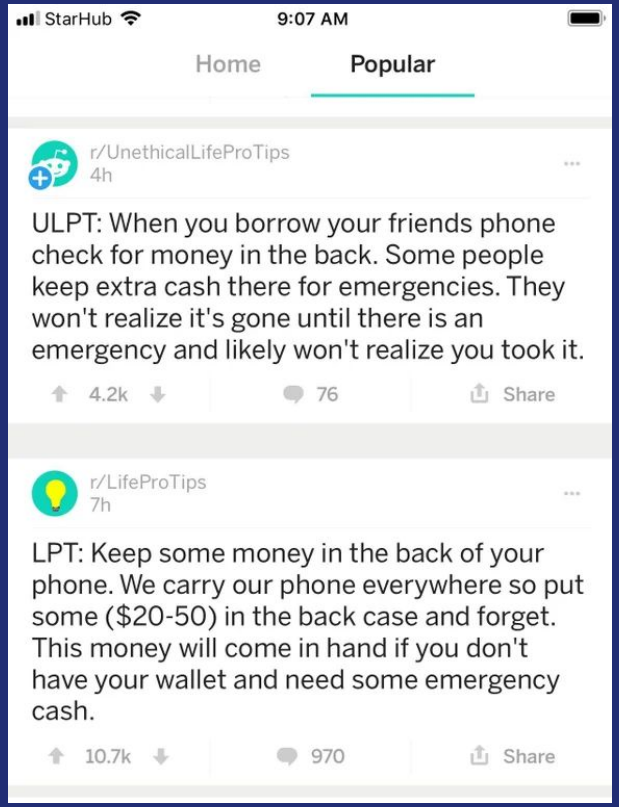
Random Forest Classifier

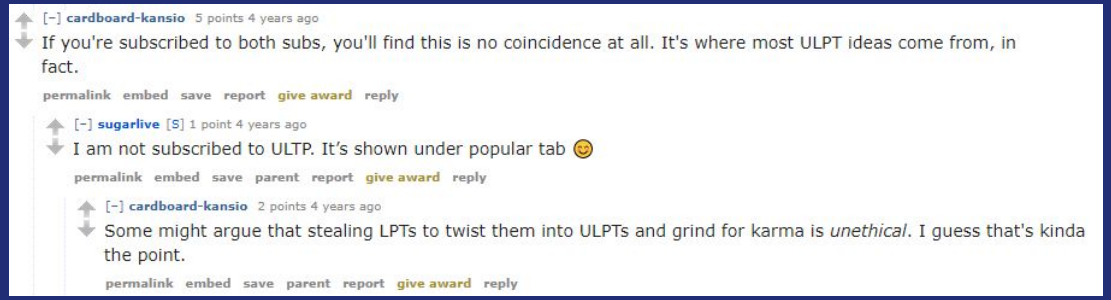



Reddit feed coincidence ?! (i.redd.it)
submitted 4 years ago by sugarlive
8 comments share save hide give award


StarHub
9:07 AM
Home
Popular
r/UnethicalLifeProTips
4h
ULPT: When you borrow your friends phone check for money in the back. Some people keep extra cash there for emergencies. They won't realize it's gone until there is an emergency and likely won't realize you took it.
4.2k
76
Share
r/LifeProTips
7h
LPT: Keep some money in the back of your phone. We carry our phone everywhere so put some ($20-50) in the back case and forget. This money will come in hand if you don't have your wallet and need some emergency cash.
10.7k
970
Share


[–] cardboard-kansio 5 points 4 years ago
If you're subscribed to both subs, you'll find this is no coincidence at all. It's where most ULPT ideas come from, in fact.
permalink embed save report give award reply
[–] sugarlive [S] 1 point 4 years ago
I am not subscribed to ULTP. It's shown under popular tab
permalink embed save parent report give award reply
[–] cardboard-kansio 2 points 4 years ago
Some might argue that stealing LPTs to twist them into ULPTs and grind for karma is *unethical*. I guess that's kinda the point.
permalink embed save parent report give award reply



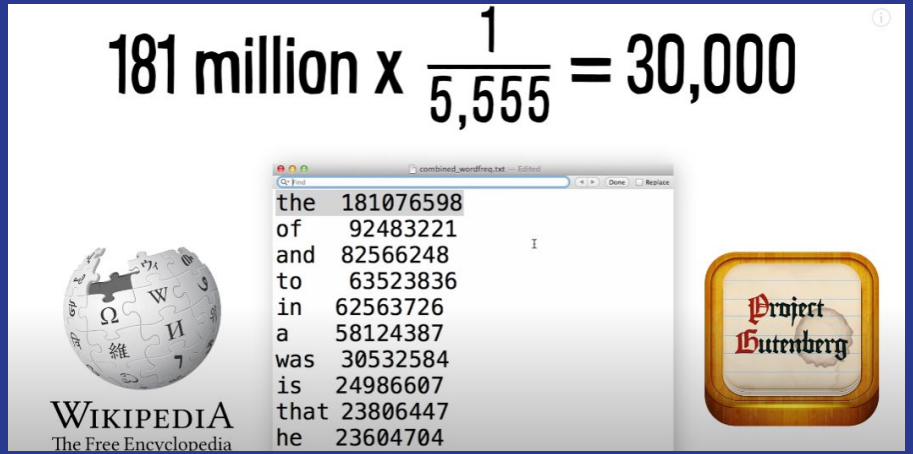


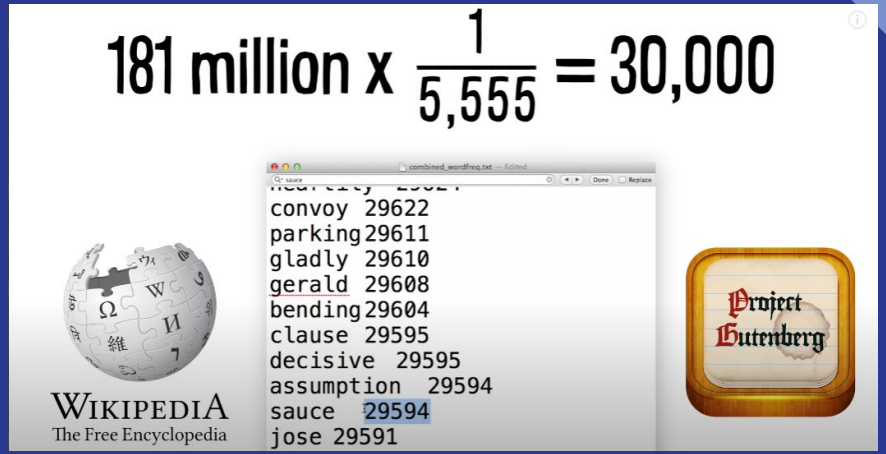
Actual	Predicted	Predict Probabilities (LPT, ULPT)
ULPT	LPT	(0.5816, 0.4184)
LPT	LPT	(0.5816, 0.4184)

Bonus: Zipf

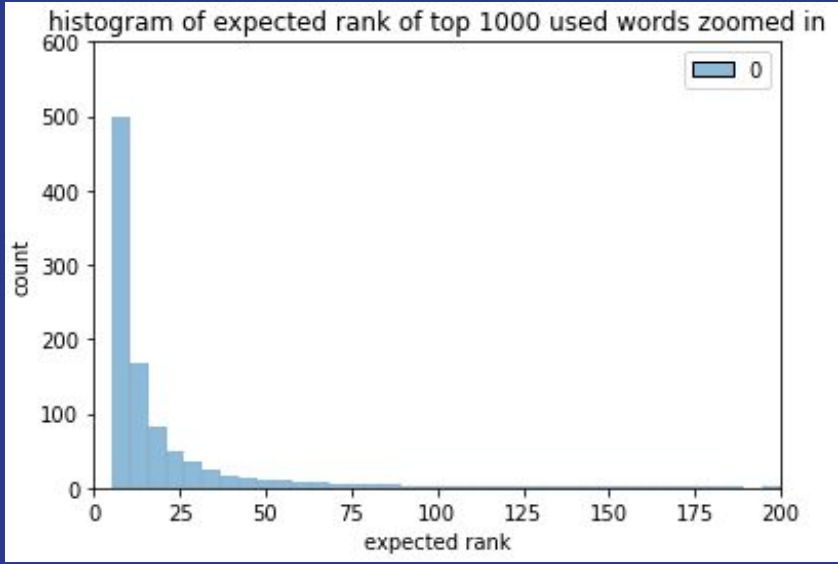
Vsauce: The Zipf Mystery

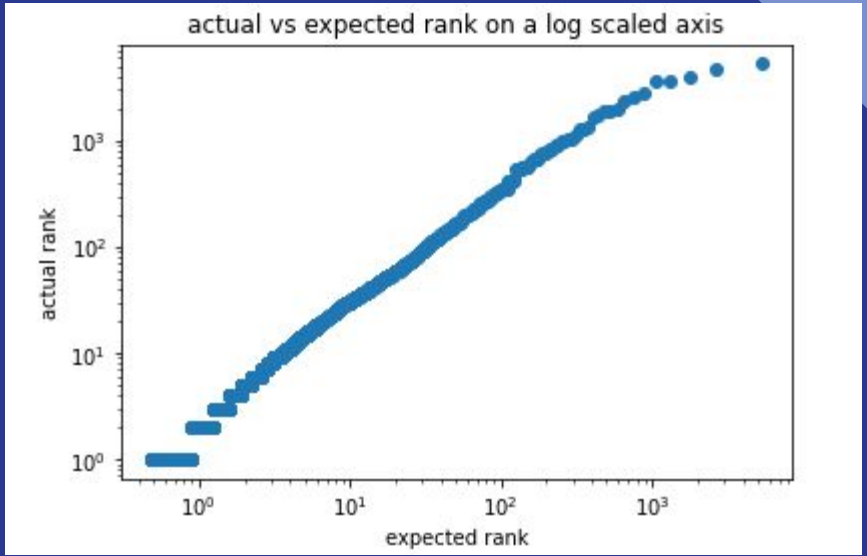
Zipf's law (/zɪf/, not /tsɪpf/ as in German) is **an empirical law formulated using mathematical statistics that refers to the fact that for many types of data studied in the physical and social sciences, the rank-frequency distribution is an inverse relation**.





Bonus: Zipf


histogram of expected rank of top 1000 used words zoomed in
0
count
expected rank
0
25
50
75
100
125
150
175
200
100
200
300
400
500
600


actual vs expected rank on a log scaled axis
actual rank
expected rank
10^0
10^1
10^2
10^3