Natural Language Processing

Sentiment Emojizer Phase 1

Gathering data

The source of data is twitter. When you run the gathering data script, the program gets the most recent tweets from twitter.

To get the new data, we should run the bash script located in the run folder named gather_raw_data.bash. You can add arguments to this bash to change default values. First arg is the number of tweets per class and you should enter it when running the script. Second argument is the output directory and you can specify that or leave it as the default value (data/raw). I ran this command to gather 8000 tweets per class:

```
bash run/gather_raw_data.bash 8000
```

After running this bash, you can see the results of getting tweets in console. Note that you should specify the classes details in data/classification_data/classes_data.csv, to get data correctly. Our program starts to get the emojis specified in the file above for each class and send a query to twitter api to get the tweets containing the given emoji. Then save the id and the text of the tweets to the output directory with tsv (tab separated values) format starting with the name of the class specified in the classes information csv. For more info check the “data/raw” folder in the project.

Preprocessing

To run preprocessing you should run the bash script located in run folder with name preprocess_raw_dta.bash. I ran the preprocessing with default arguments:

```
bash run/preprocess_raw_data.bash
```

You can simply run the script with default arguments. After running the script you can see the count of data before preprocessing and after preprocessing in console logs.

- “--ids'' the classes id you want to preprocess. Should be separated by ‘,’ without space. Default value is “1,2,3,4,5,6,7,8”.
- “--out” the output directory. Default value is PREPROCESSED_BASE_DIR in constants
- “--flags” preprocessing methods you want to run on raw data. Accepted values are characters ‘a’ to ‘l’. Note that after part ‘f’ I tokenized the tweets with TwitterTokenizer in NLTK module. I didn’t use sentence tokenizer, because I think this level of tokenizing is not needed in this case. Before removing emojis (part k) we should identify the label of the tweets.
    - ‘a’ : remove incomplete tweets (some tweets from twitter api are not complete and we should remove them)
    - ‘b’: remove retweets
    - ‘c’: remove mentions
    - ‘d’: remove hashtags
    - ‘e’: remove punctuations
    - ‘f’: replace newlines with empty string
    - ‘g’: stemmize tokens (this options is off by default)
    - ‘h’: remove the tweets that have more than 40% emojis in tokens.
    - ‘i’: remove emojis which are not defined for our system. Defined emojis are available in class_data.csv file.
    - ‘j’: remove consecutive repeating emojis.
    - ‘k’: remove emojis
    - ‘l’ : remove non ascii tokens
- “--input” input directory of raw data containing tsv files with name {class_name}_raw_text.tsv for each id available in ids argument.

Labeling

After preprocess data till part k, we should identify the labels. Each tweet can has multiple labels based on the emojis in the tweet. Note that in part i we removed undefined emojis. So this part and part j are important in the labeling process. After removing undefined and consecutive repeated emojis, we can count emojis used in tweet and identify its label. Then the probability that this tweet belongs to a label is the number of emojis associated with that label, divided by the total emojis count in the tweet.

After running the preprocess script, labels will be created and saved in jsons in the labels folder. Each tweet gets an index same as the file name in the labels folder. We can access the label of each tweet with the associated index.

Statistics

To show the statistics details you can run bash script and show the results.

```
bash run/show_statistics.bash
```

Arguments are:

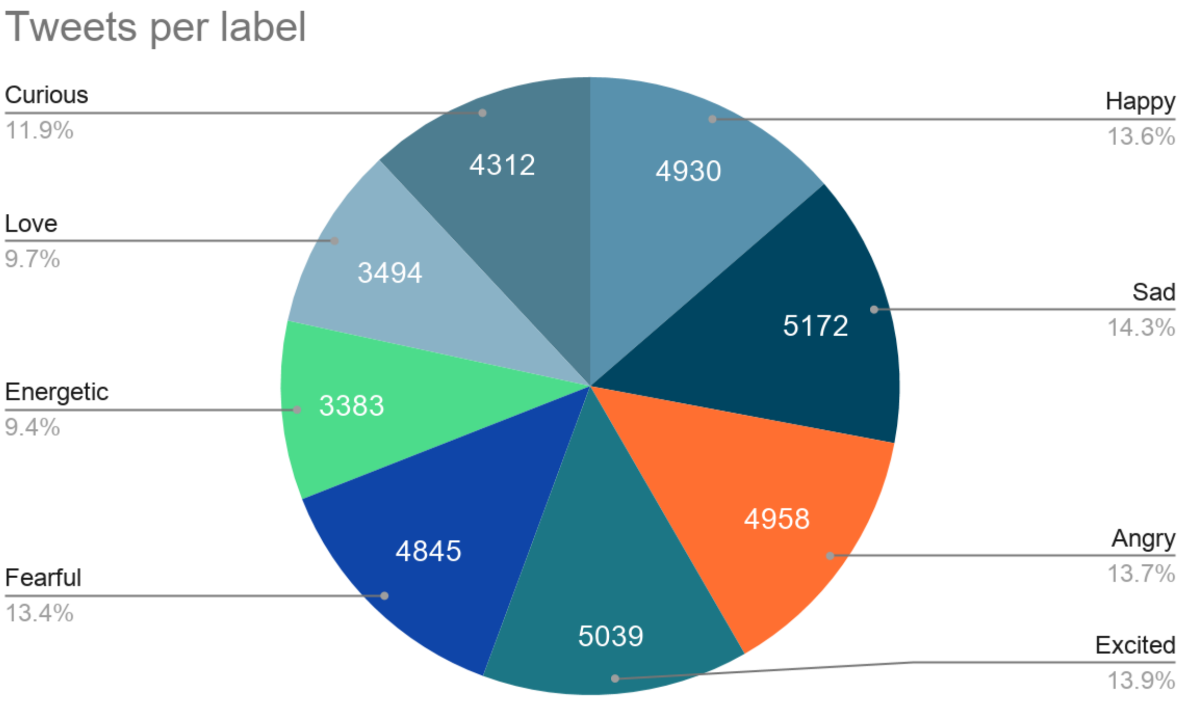
- “--flags” type of statistics you want to show
- “--input” the input directory. Default value is PREPROCESSED_BASE_DIR in constants
- “--out” by default this one is Null. You can save the csv files of statistics by giving the base directory relative to the root dir of the program. ex

```
bash run/show_statistics.bash --out=”statistics”
```

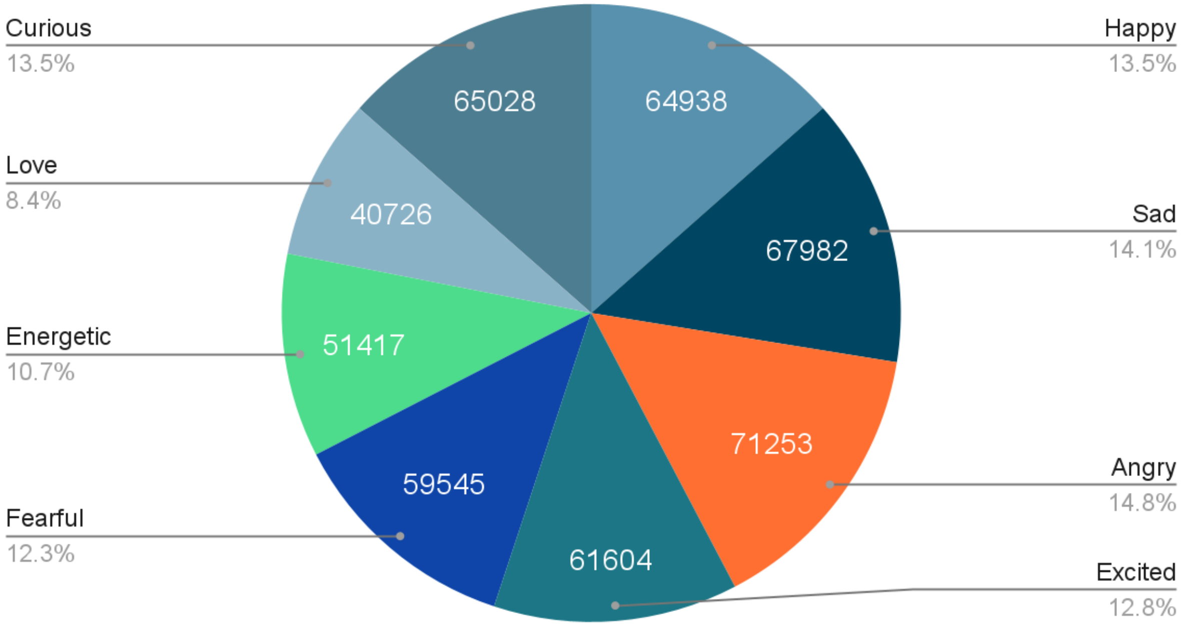
Statistics of current data

Current preprocessed data available in “data/preprocessed”.

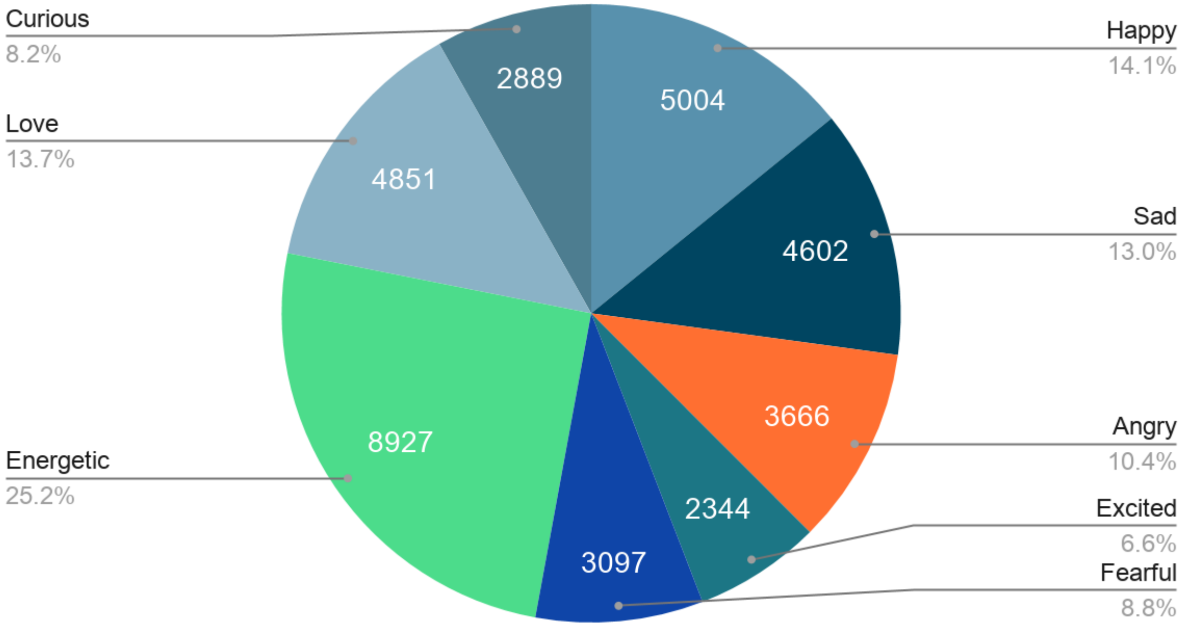
Label	Data count	Tokens Count	Unique Tokens
Happy	4930	64938	5004
Sad	5172	67982	4602
Angry	4958	71253	3666
Excited	5039	61604	2344
Fearful	4845	59545	3097
Energetic	3383	51417	8927
Love	3494	40726	4851
Curious	4312	65028	2889



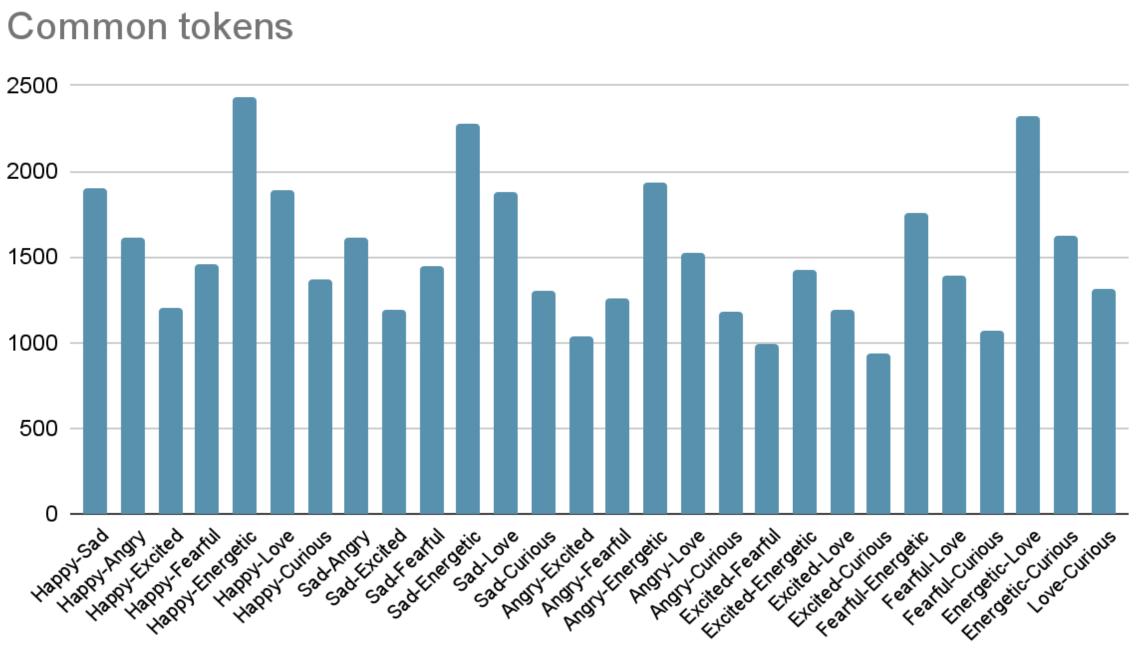
Tokens per label



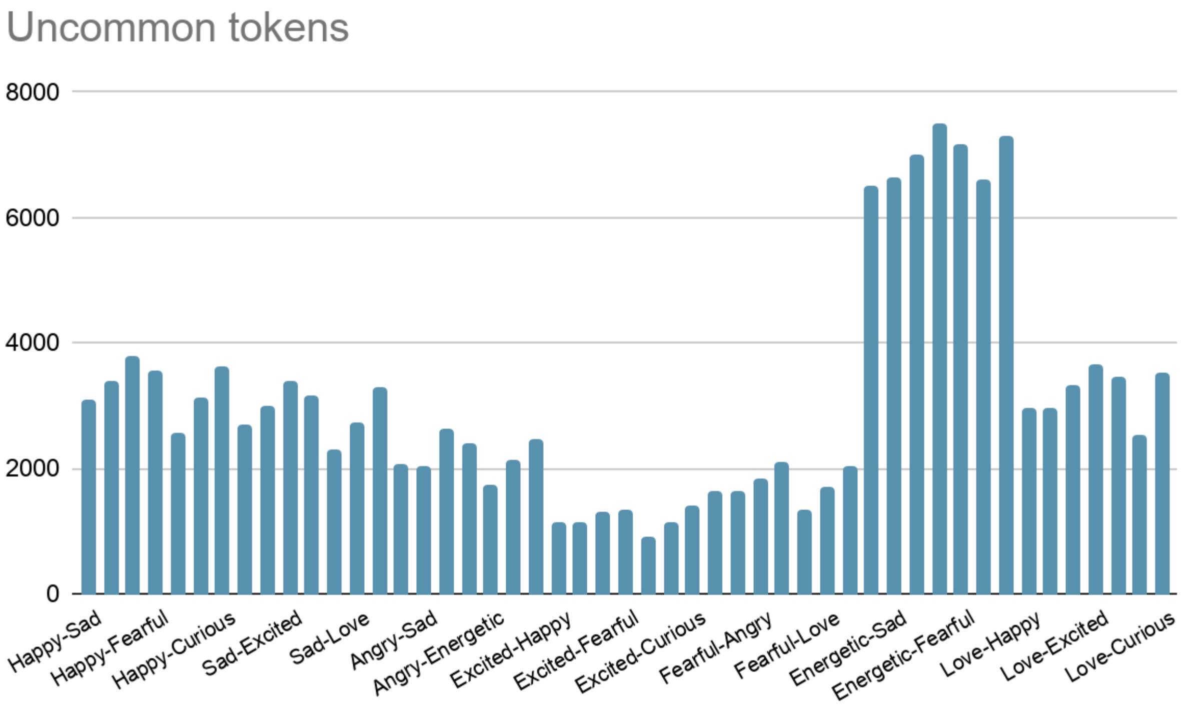
Unique tokens per label

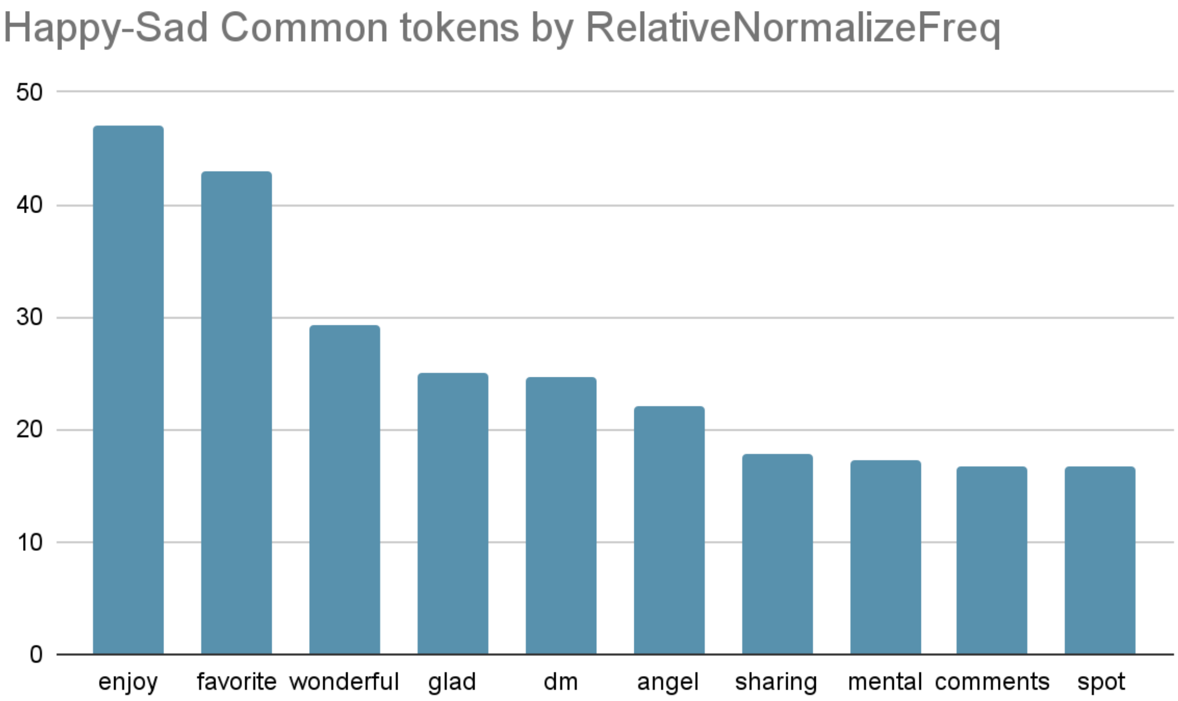


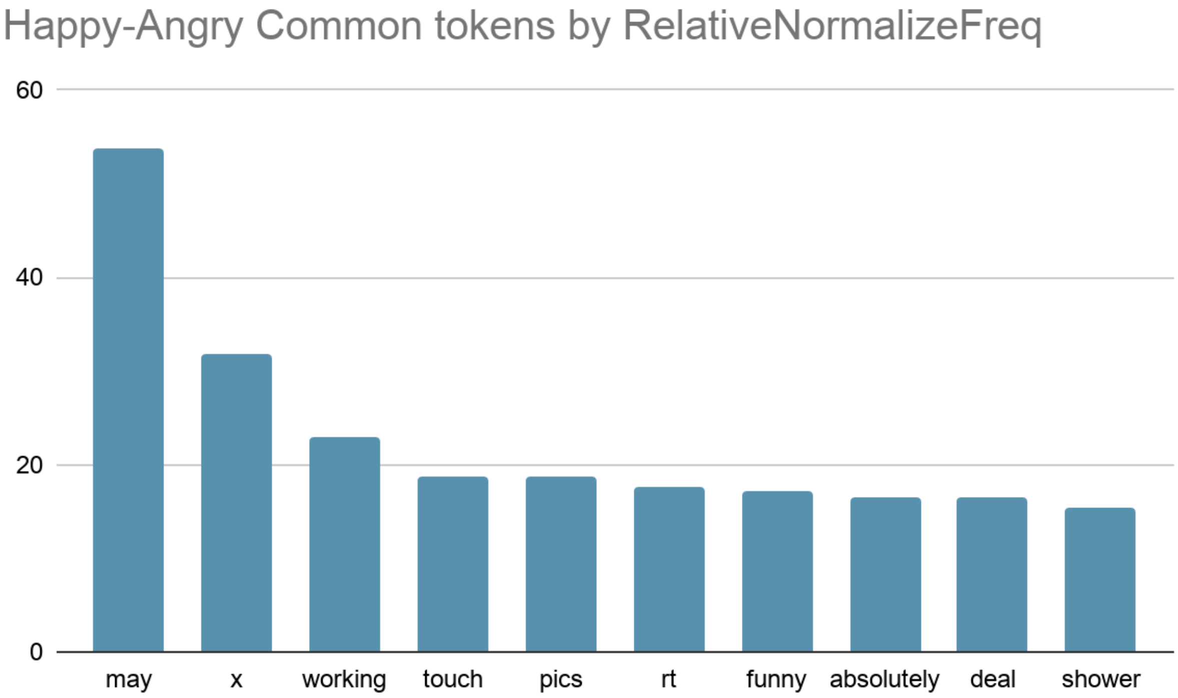
Labels	Common Tokens	Labels	Common Tokens
Happy-Sad	1896	Angry-Fearful	1261
Happy-Angry	1609	Angry-Energetic	1939
Happy-Excited	1207	Angry-Love	1526
Happy-Fearful	1455	Angry-Curious	1182
Happy-Energetic	2428	Excited-Fearful	998
Happy-Love	1886	Excited-Energetic	1428
Happy-Curious	1371	Excited-Love	1188
Sad-Angry	1615	Excited-Curious	942
Sad-Excited	1193	Fearful-Energetic	1754
Sad-Fearful	1449	Fearful-Love	1390
Sad-Energetic	2283	Fearful-Curious	1073
Sad-Love	1881	Energetic-Love	2327
Sad-Curious	1306	Energetic-Curious	1623
Angry-Excited	1037	Love-Curious	1311

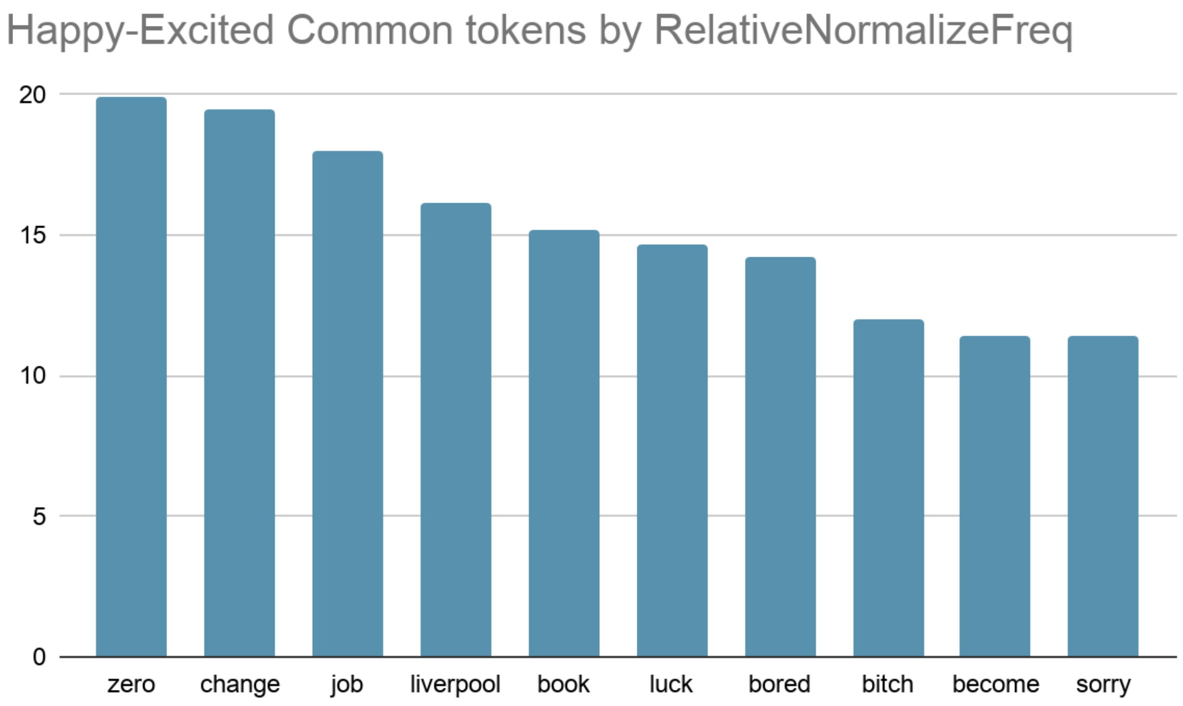


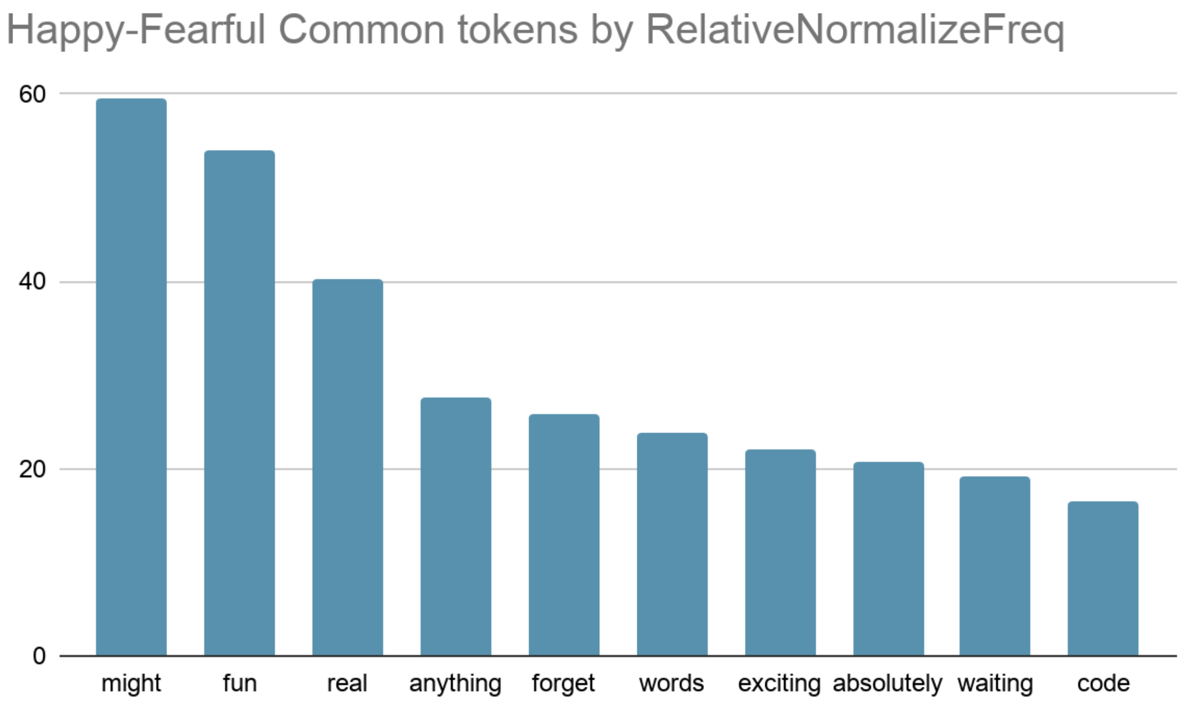
Labels	Uncommon Tokens	Labels	Uncommon Tokens	Labels	Uncommon Tokens
Happy-Sad	3108	Angry-Love	2140	Energetic-Fearful	7173
Happy-Angry	3395	Angry-Curious	2484	Energetic-Love	6600
Happy-Excited	3797	Excited-Happy	1137	Energetic-Curious	7304
Happy-Fearful	3549	Excited-Sad	1151	Love-Happy	2965
Happy-Energetic	2576	Excited-Angry	1307	Love-Sad	2970
Happy-Love	3118	Excited-Fearful	1346	Love-Angry	3325
Happy-Curious	3633	Excited-Energetic	916	Love-Excited	3663
Sad-Happy	2706	Excited-Love	1156	Love-Fearful	3461
Sad-Angry	2987	Excited-Curious	1402	Love-Energetic	2524
Sad-Excited	3409	Fearful-Happy	1642	Love-Curious	3540
Sad-Fearful	3153	Fearful-Sad	1648	Curious-Happy	1518
Sad-Energetic	2319	Fearful-Angry	1836	Curious-Sad	1583
Sad-Love	2721	Fearful-Excited	2099	Curious-Angry	1707
Sad-Curious	3296	Fearful-Energetic	1343	Curious-Excited	1947
Angry-Happy	2057	Fearful-Love	1707	Curious-Fearful	1816
Angry-Sad	2051	Fearful-Curious	2024	Curious-Energetic	1266
Angry-Excited	2629	Energetic-Happy	6499	Curious-Love	1578
Angry-Fearful	2405	Energetic-Sad	6644	Energetic-Fearful	7173
Angry-Energetic	1727	Energetic-Angry	6988	Energetic-Love	6600

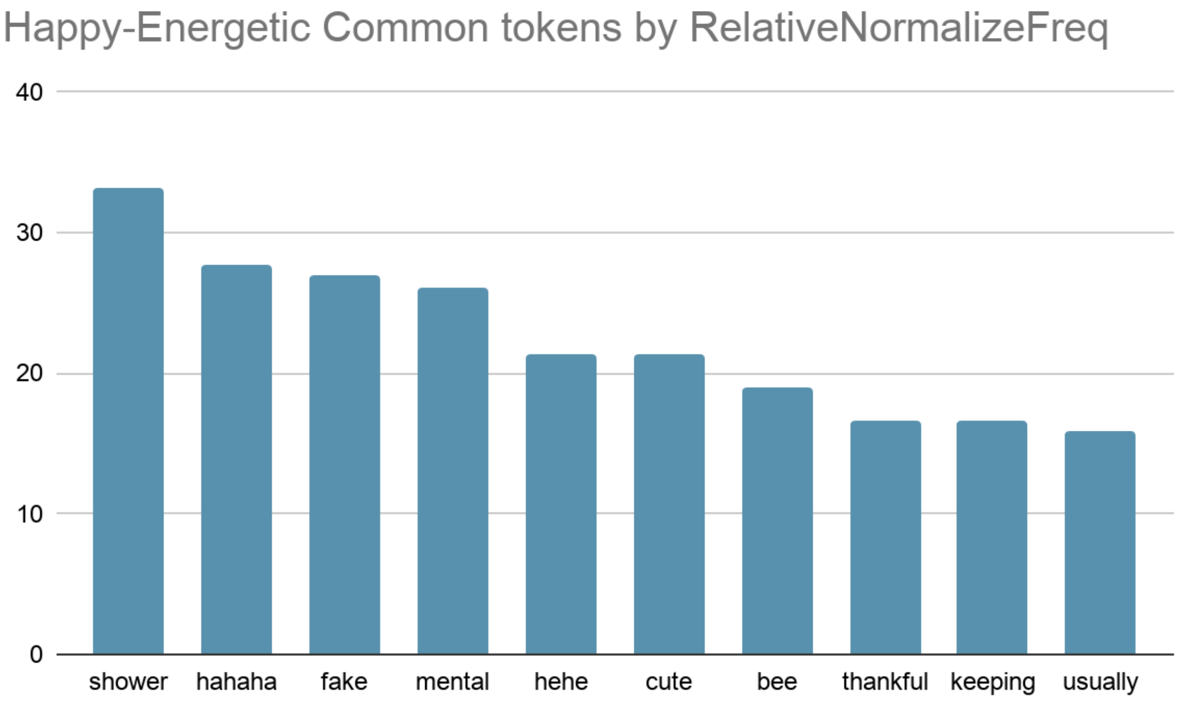

Uncommon tokens
8000
6000
4000
2000
0
Happy-Sad
Happy-Fearful
Happy-Curious
Sad-Excited
Sad-Love
Angry-Sad
Angry-Energetic
Excited-Happy
Excited-Fearful
Excited-Curious
Fearful-Angry
Fearful-Love
Energetic-Sad
Energetic-Fearful
Love-Happy
Love-Excited
Love-Curious

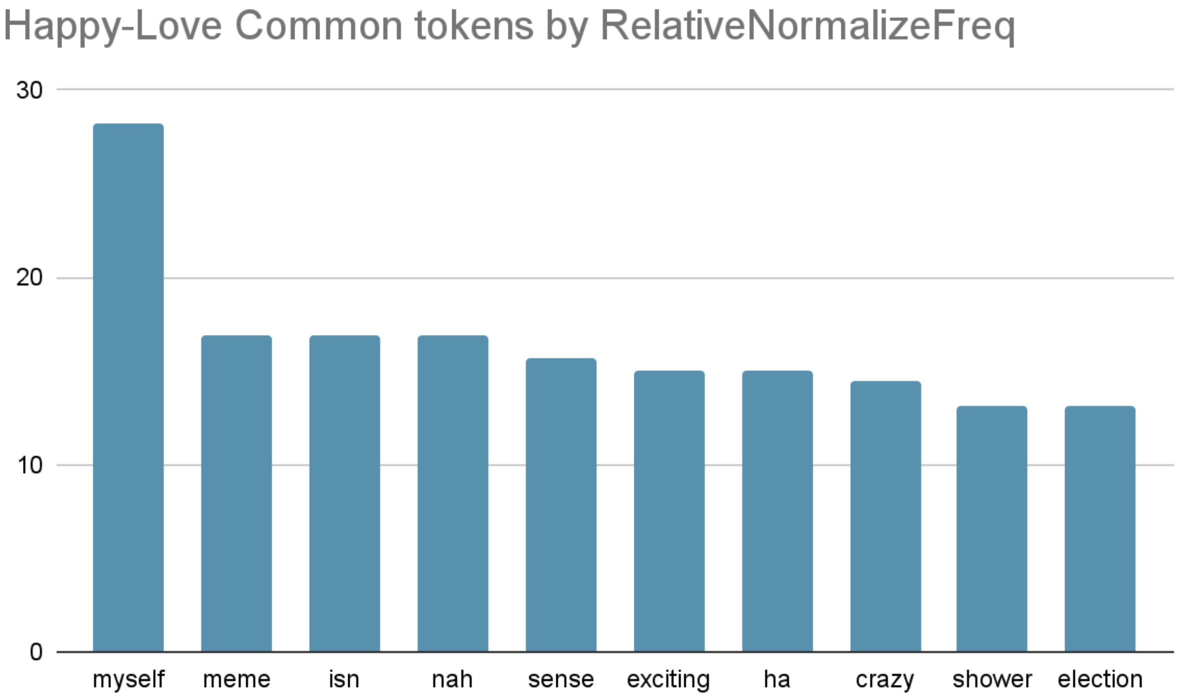

Happy-Sad Common tokens by RelativeNormalizeFreq

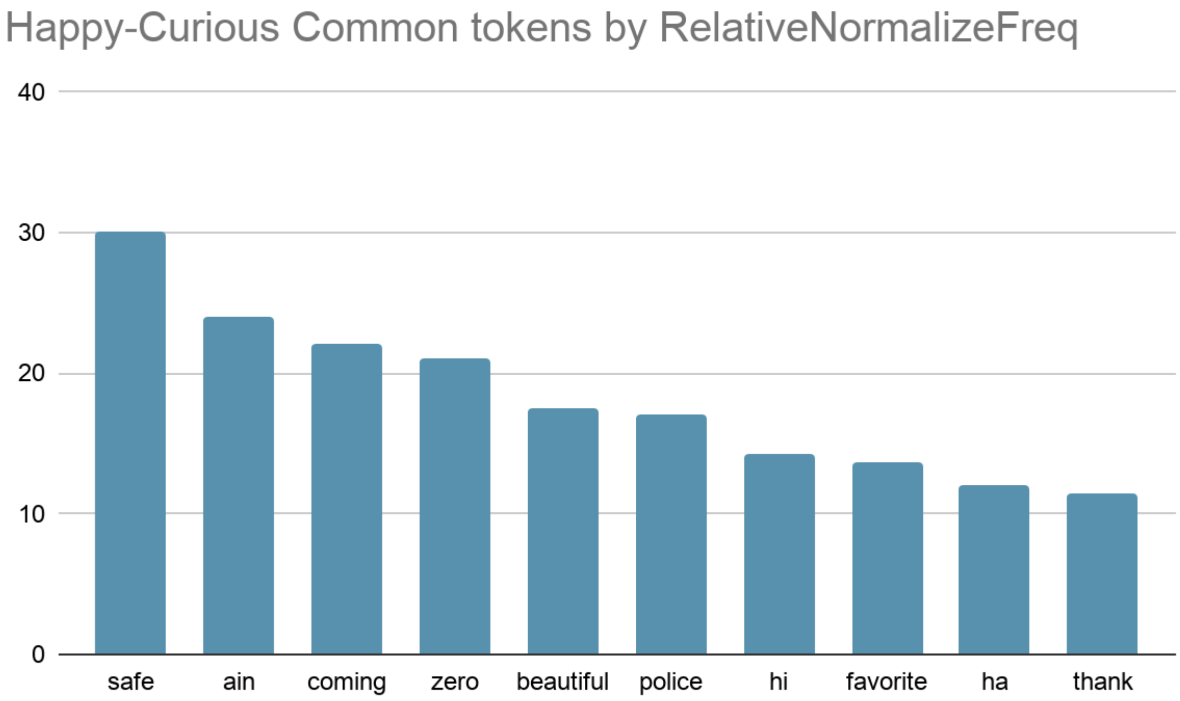

Happy-Angry Common tokens by RelativeNormalizeFreq

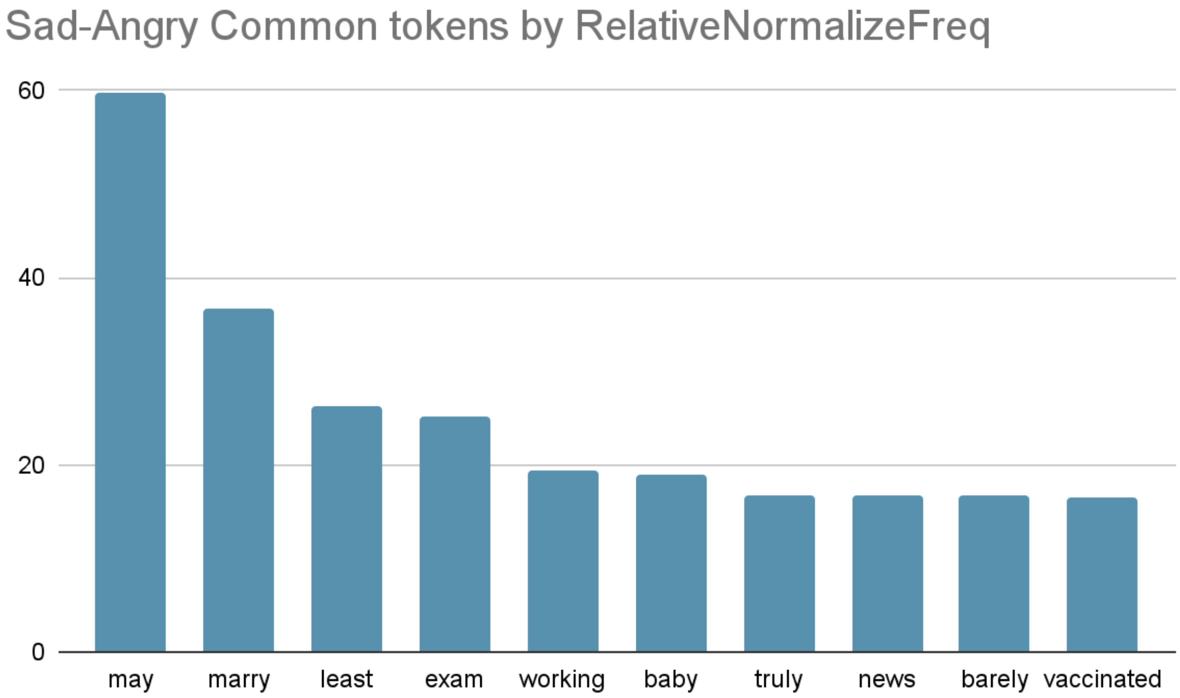

Happy-Excited Common tokens by RelativeNormalizeFreq
20
15
10
5
0
zero
change
job
liverpool
book
luck
bored
bitch
become
sorry

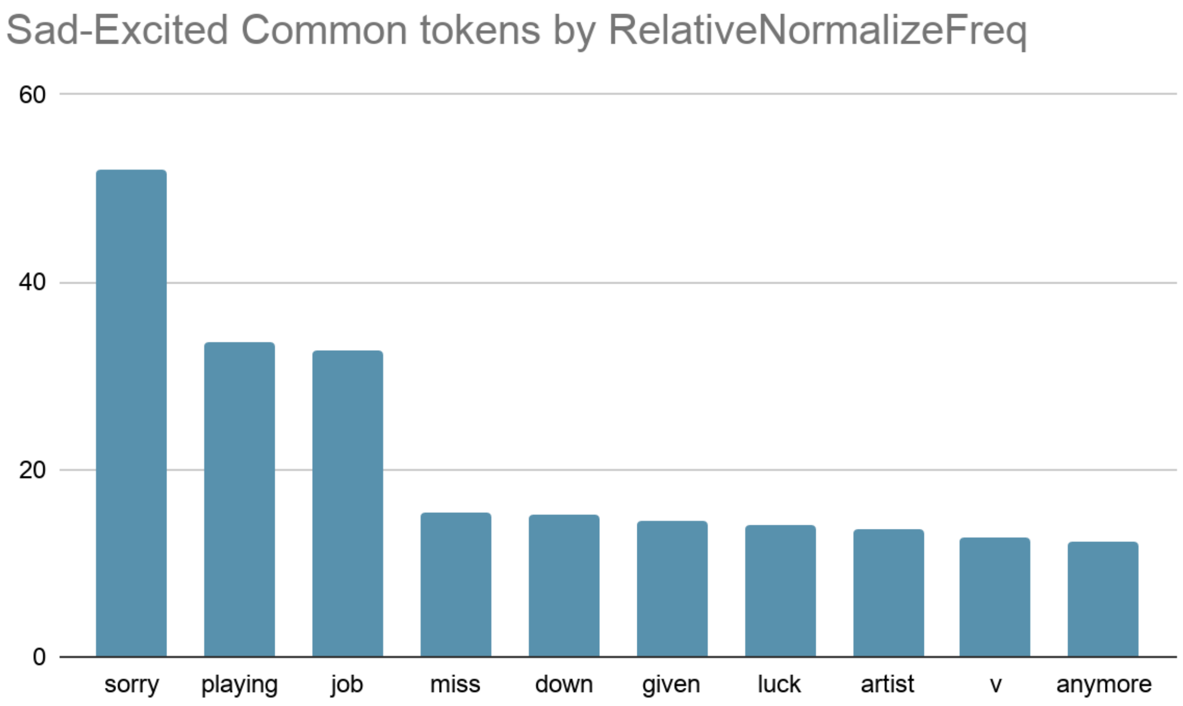

Happy-Fearful Common tokens by RelativeNormalizeFreq
60
40
20
0
might
fun
real
anything
forget
words
exciting
absolutely
waiting
code

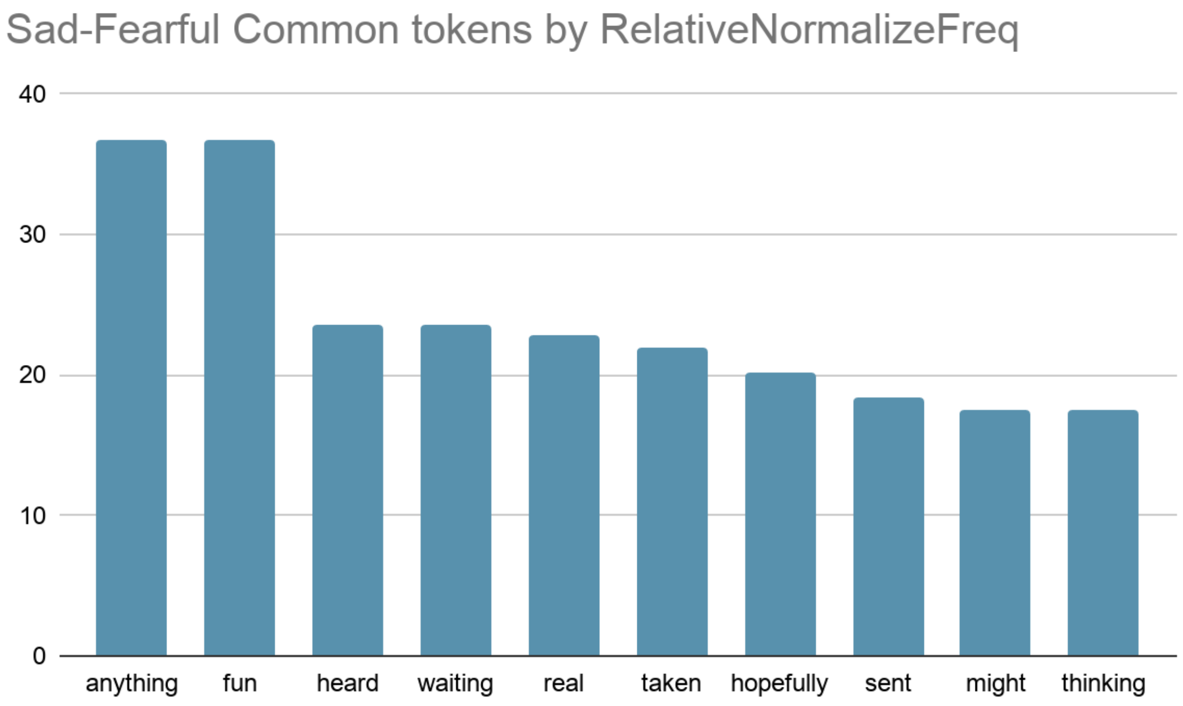

Happy-Energetic Common tokens by RelativeNormalizeFreq


Happy-Love Common tokens by RelativeNormalizeFreq

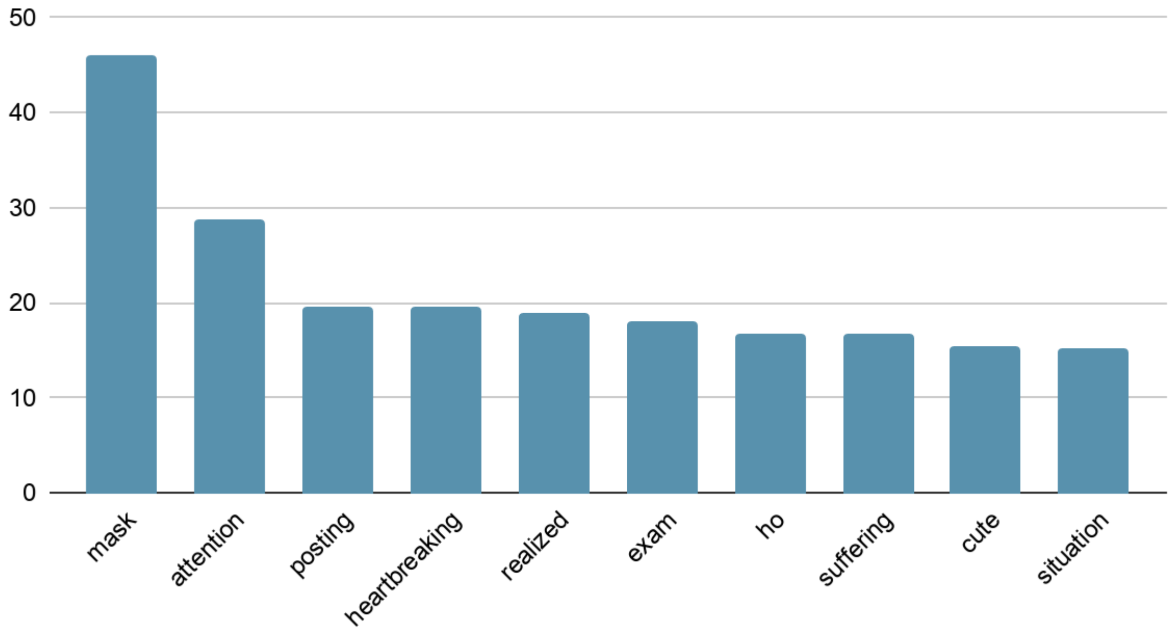

Happy-Curious Common tokens by RelativeNormalizeFreq


Sad-Angry Common tokens by RelativeNormalizeFreq

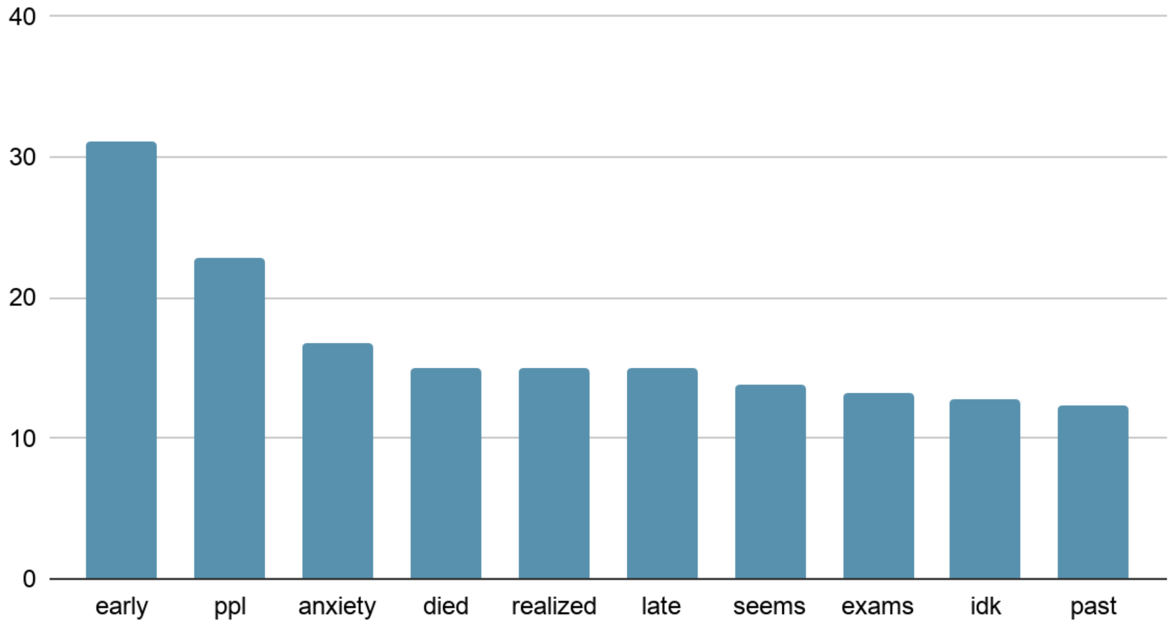

Sad-Excited Common tokens by RelativeNormalizeFreq


Sad-Fearful Common tokens by RelativeNormalizeFreq

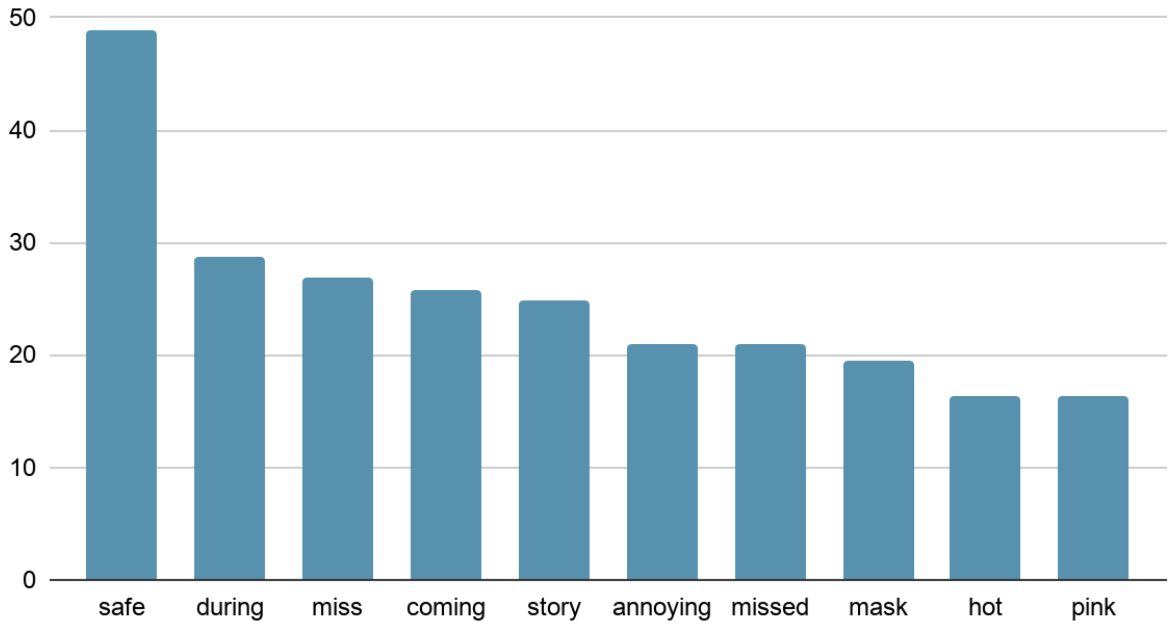
Sad-Energetic Common tokens by RelativeNormalizeFreq



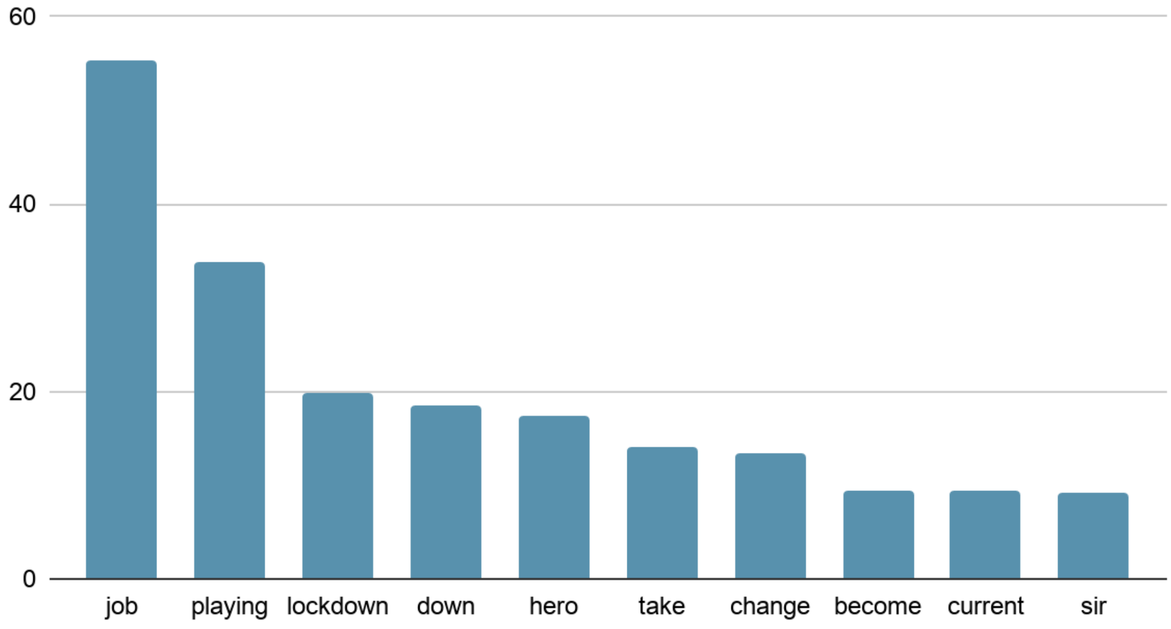
Sad-Love Common tokens by RelativeNormalizeFreq

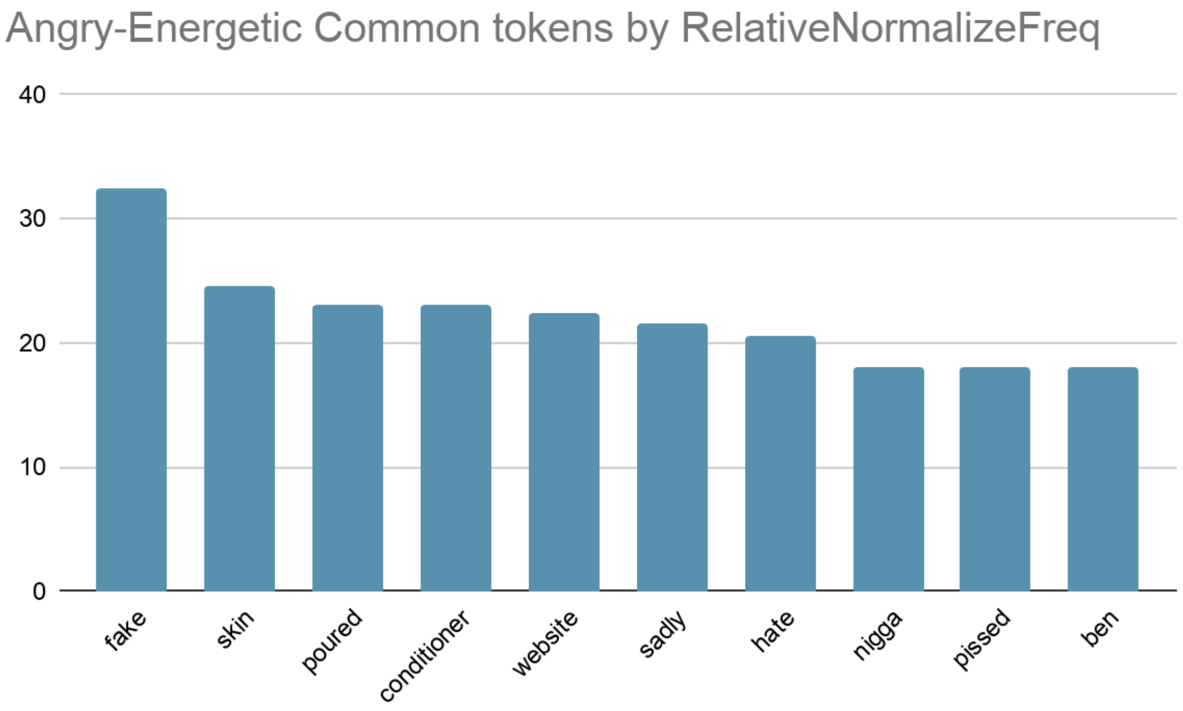


Sad-Curious Common tokens by RelativeNormalizeFreq

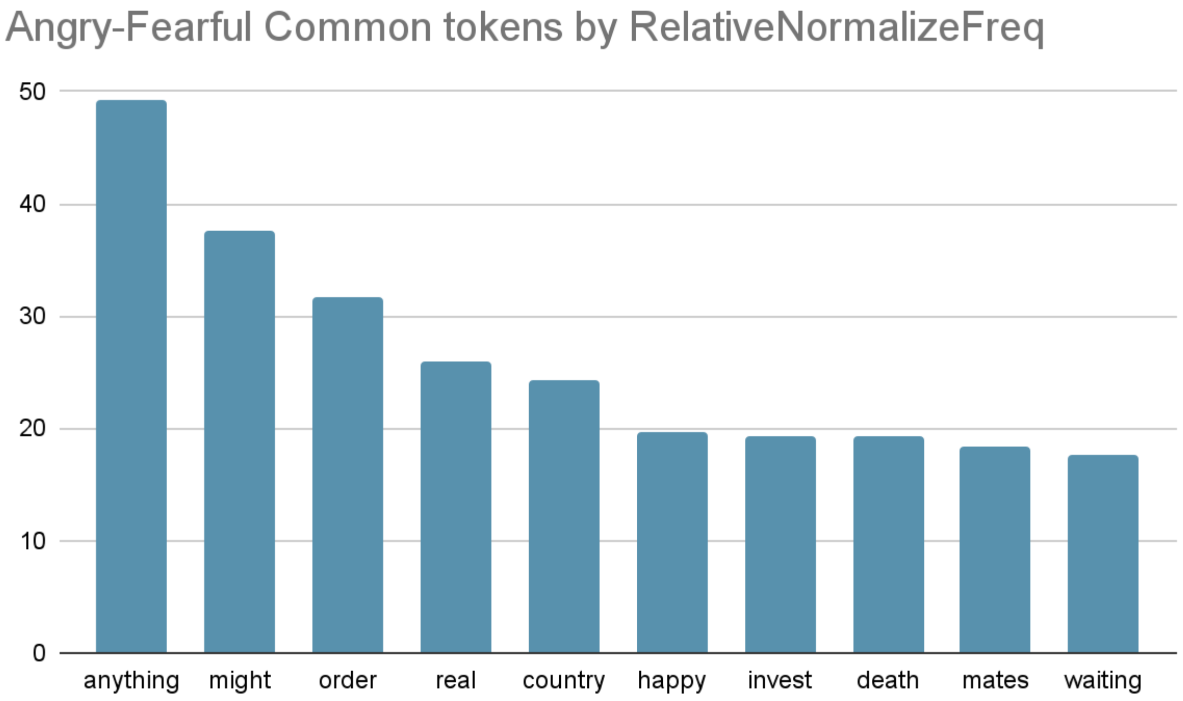


Angry-Excited Common tokens by RelativeNormalizeFreq

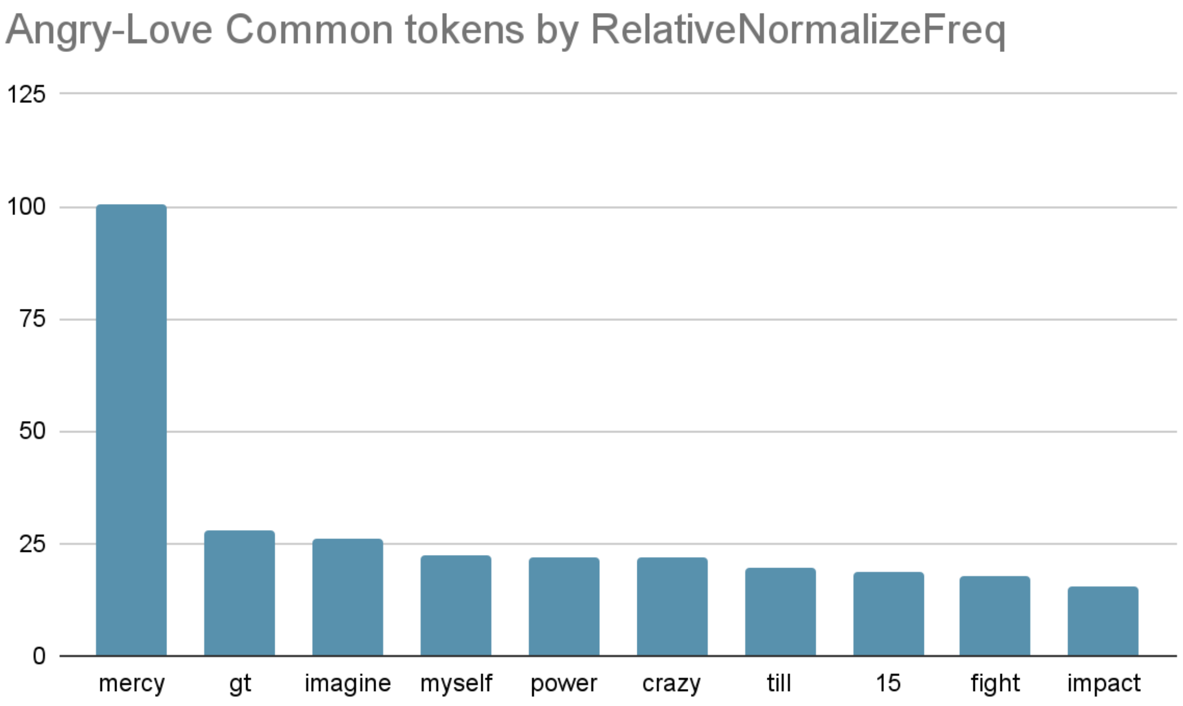


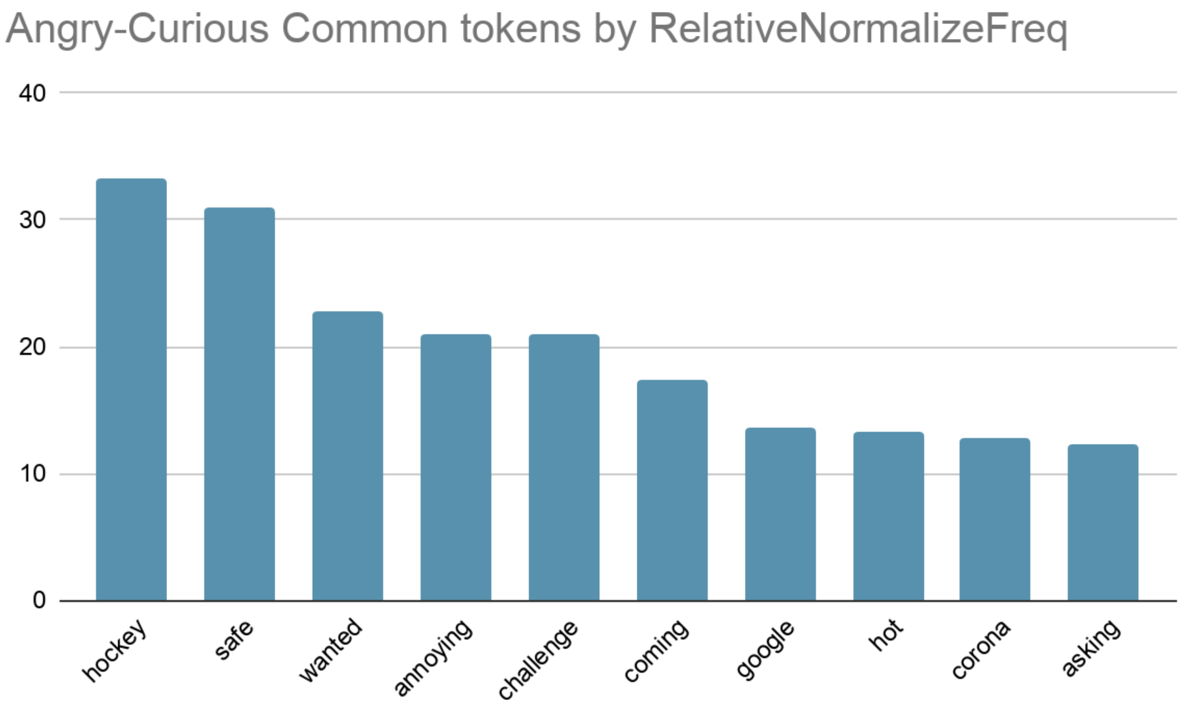


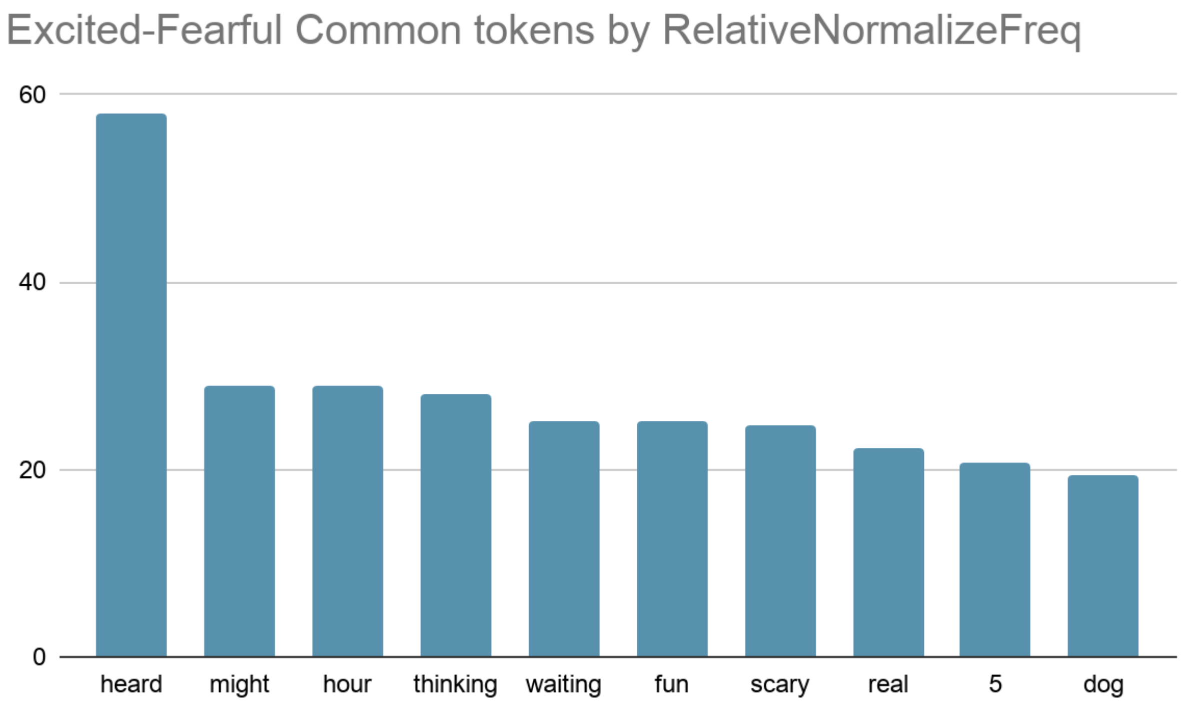
Angry-Energetic Common tokens by RelativeNormalizeFreq



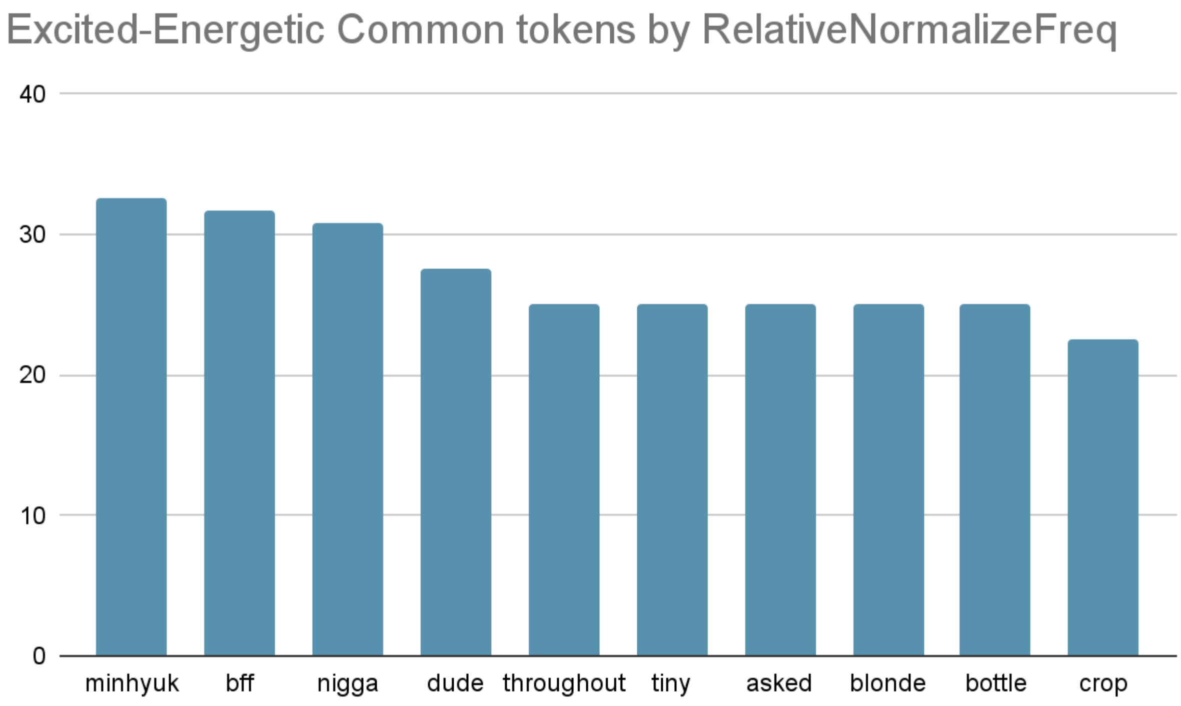
Angry-Fearful Common tokens by RelativeNormalizeFreq


Angry-Love Common tokens by RelativeNormalizeFreq
125
100
75
50
25
0
mercy
gt
imagine
myself
power
crazy
till
15
fight
impact

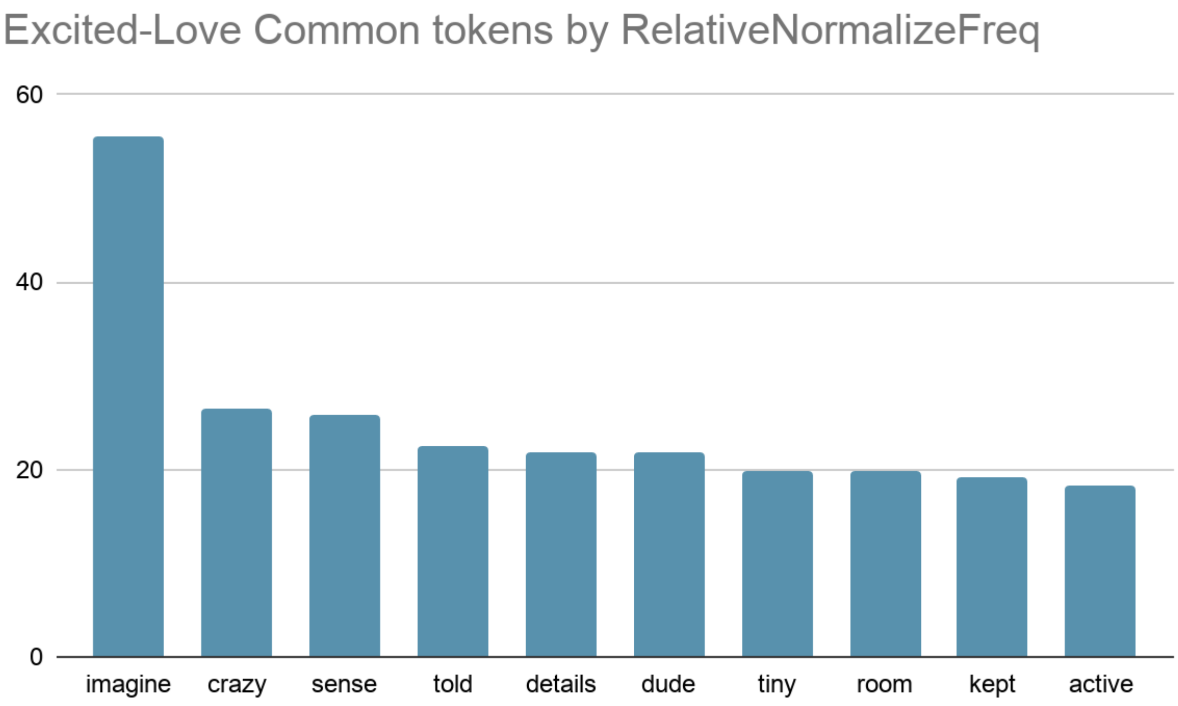

Angry-Curious Common tokens by RelativeNormalizeFreq
40
30
20
10
0
hockey
safe
wanted
annoying
challenge
coming
google
hot
corona
asking



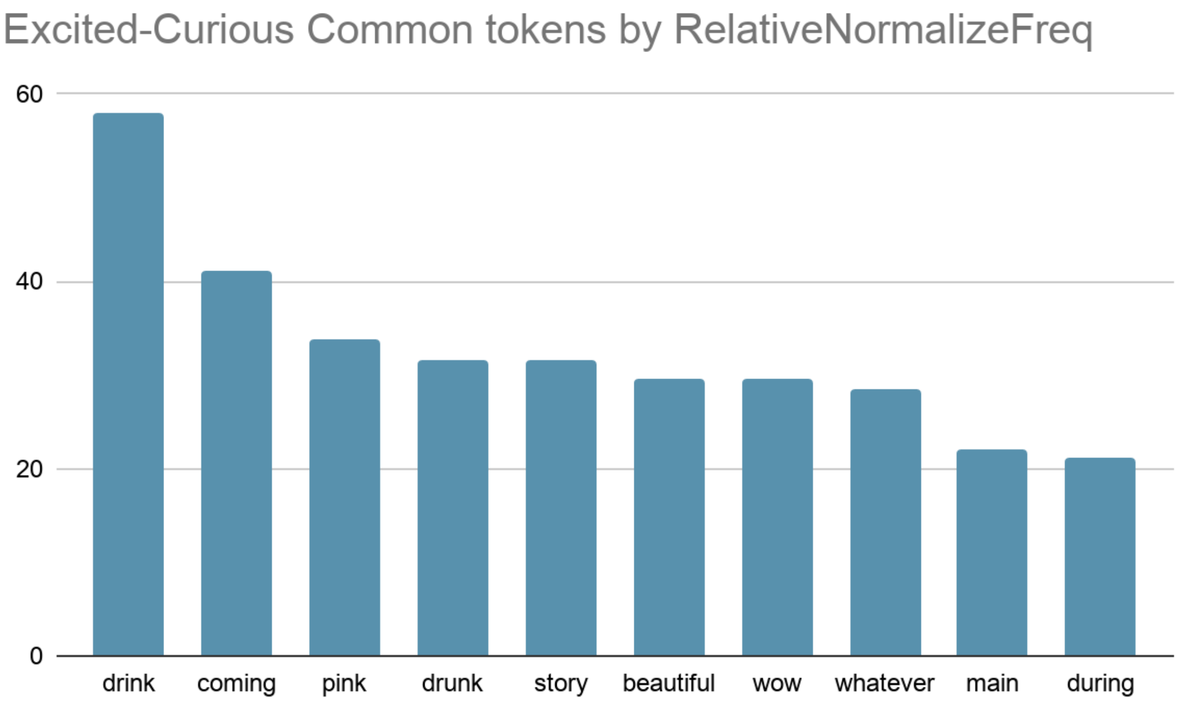
Excited-Fearful Common tokens by RelativeNormalizeFreq



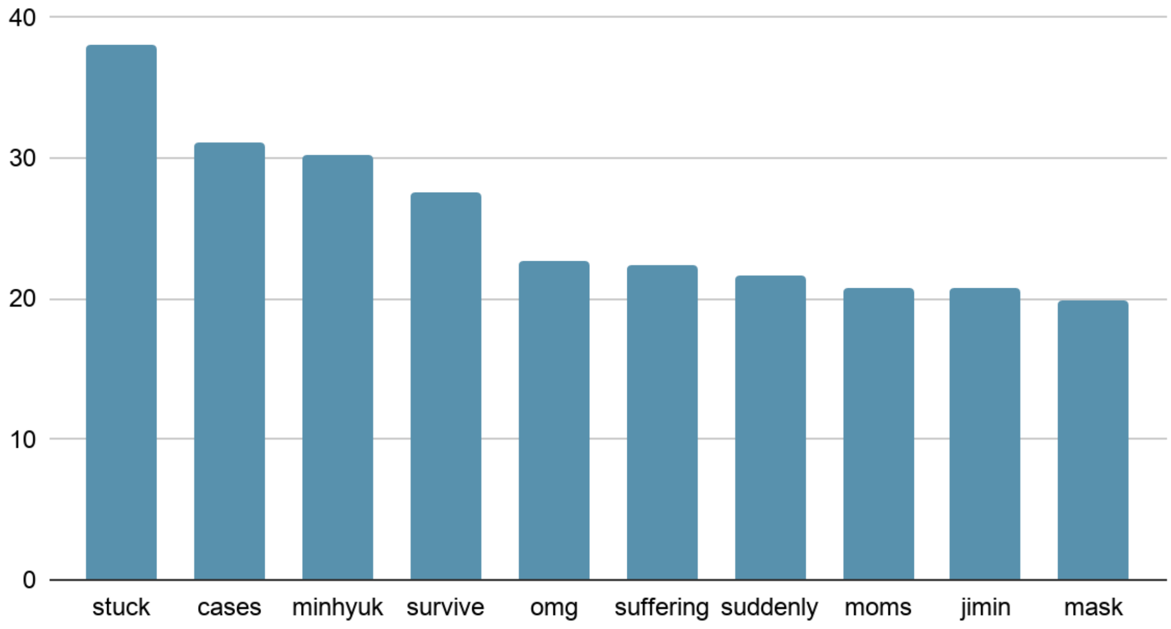
Excited-Energetic Common tokens by RelativeNormalizeFreq

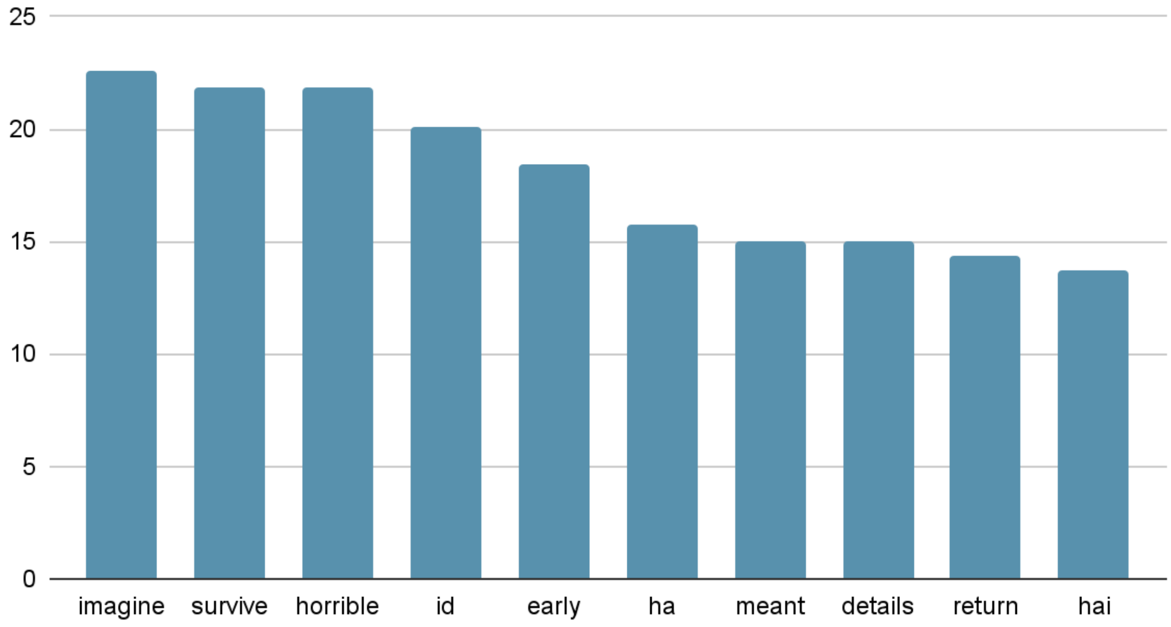


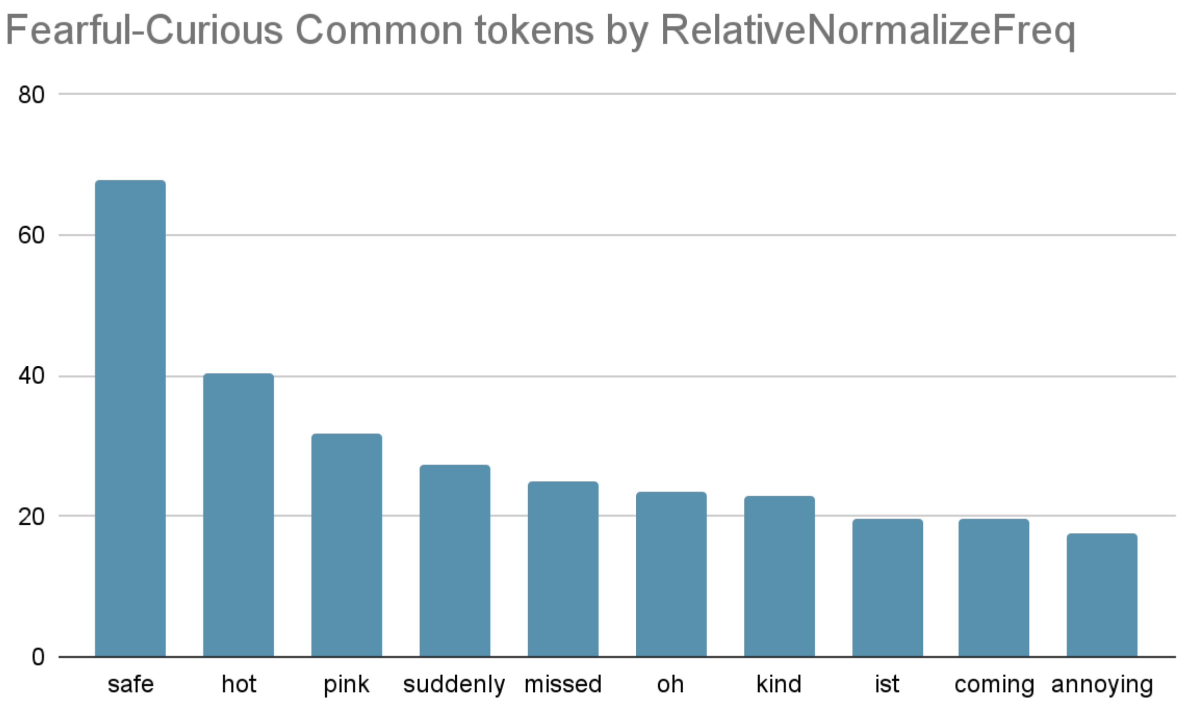
Excited-Love Common tokens by RelativeNormalizeFreq

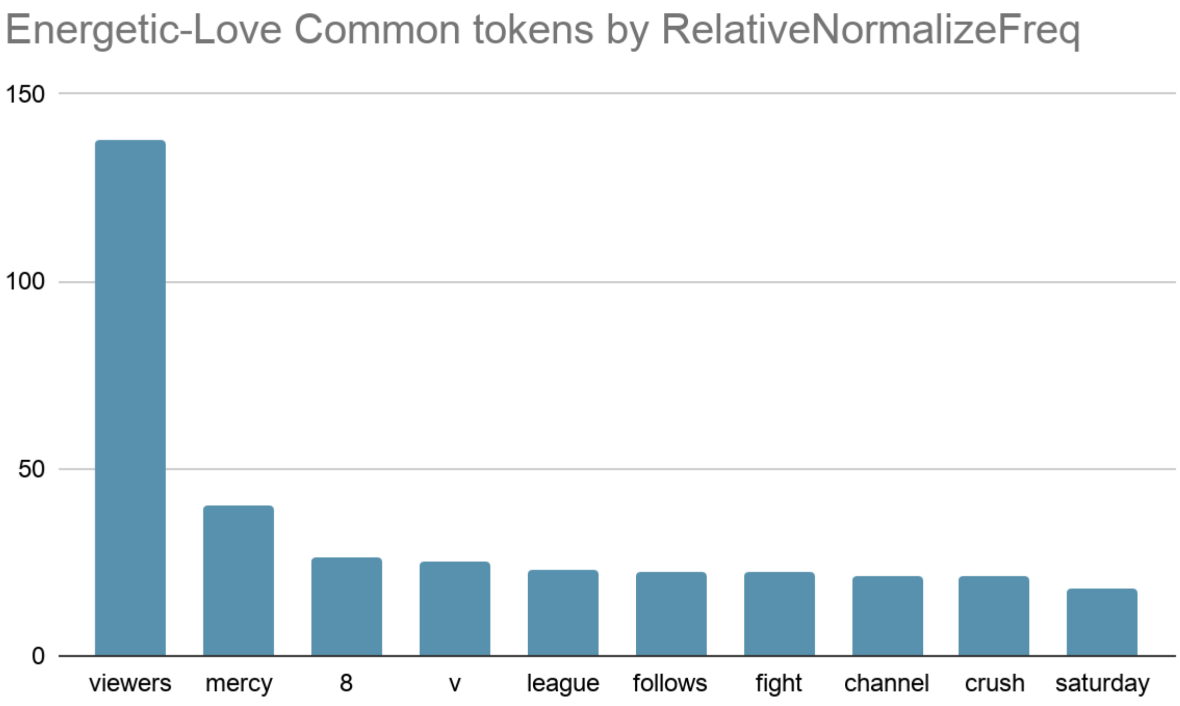


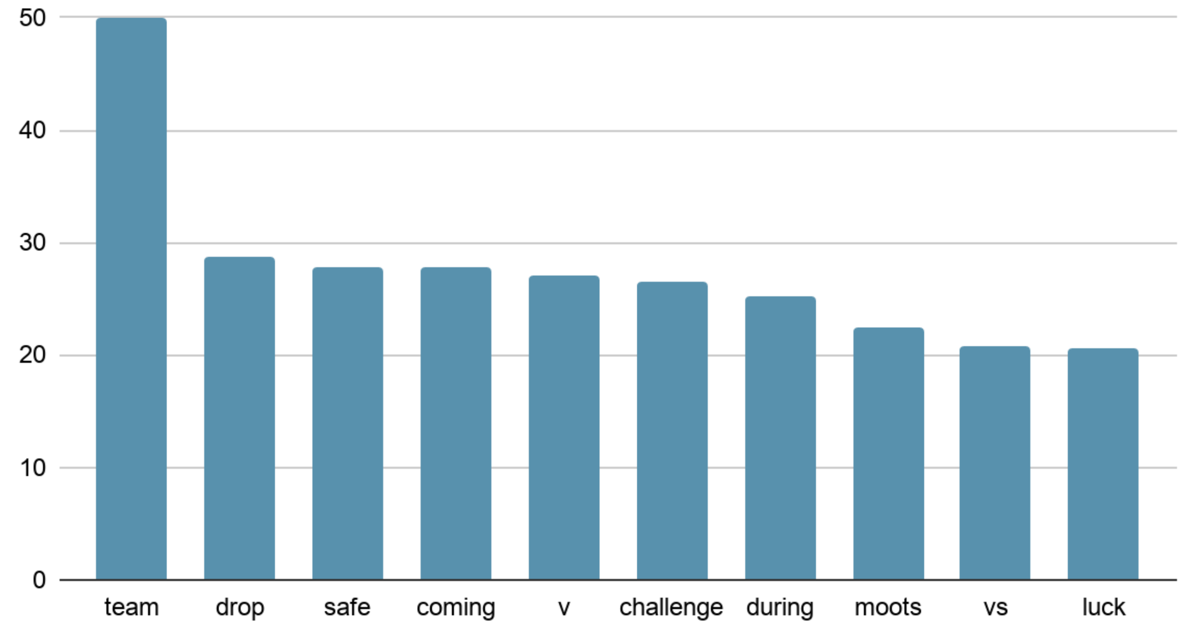
Excited-Curious Common tokens by RelativeNormalizeFreq


Fearful-Energetic Common tokens by RelativeNormalizeFreq

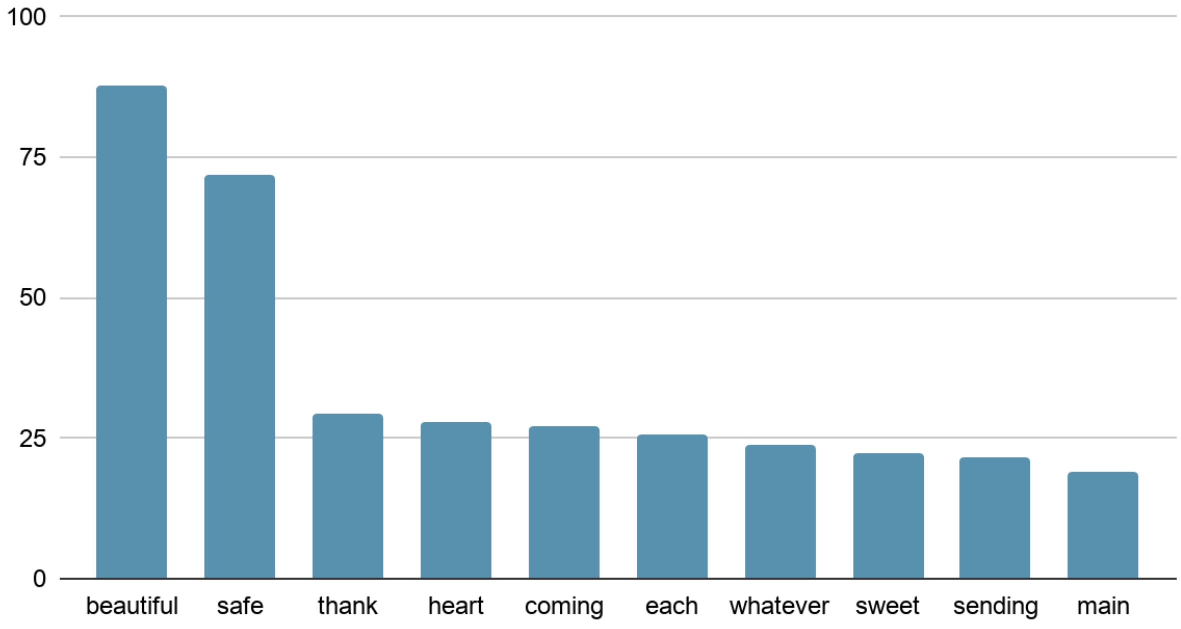

Fearful-Love Common tokens by RelativeNormalizeFreq


Fearful-Curious Common tokens by RelativeNormalizeFreq
80
60
40
20
0
safe
hot
pink
suddenly
missed
oh
kind
ist
coming
annoying


Energetic-Love Common tokens by RelativeNormalizeFreq
150
100
50
0
viewers
mercy
8
v
league
follows
fight
channel
crush
saturday

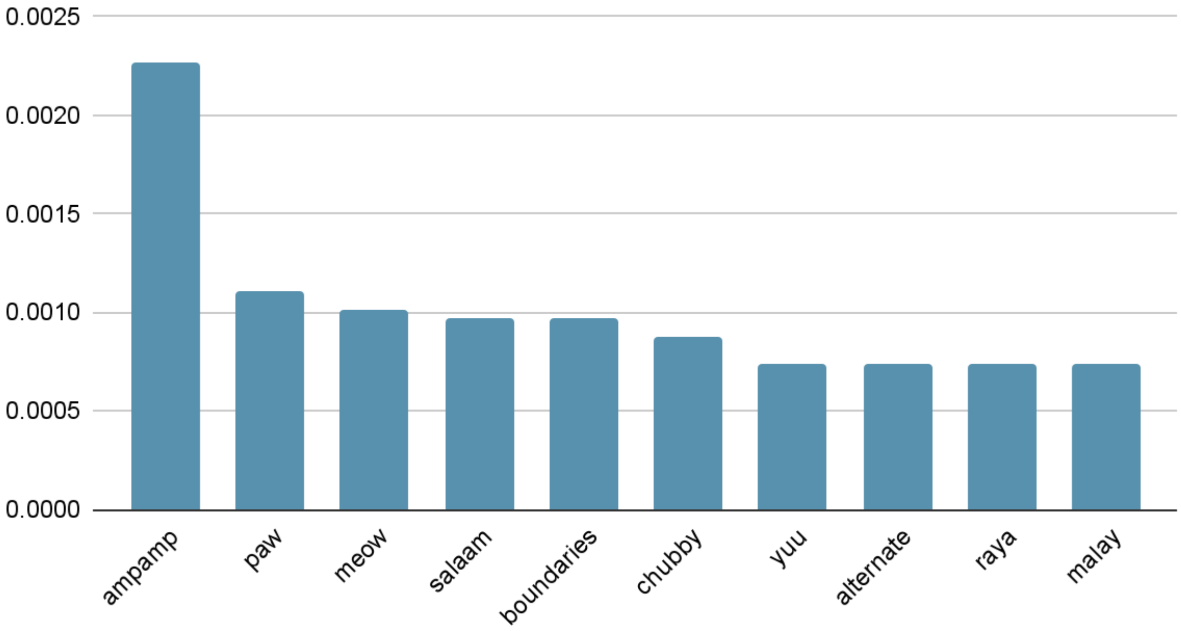


Energetic-Curious Common tokens by RelativeNormalizeFreq

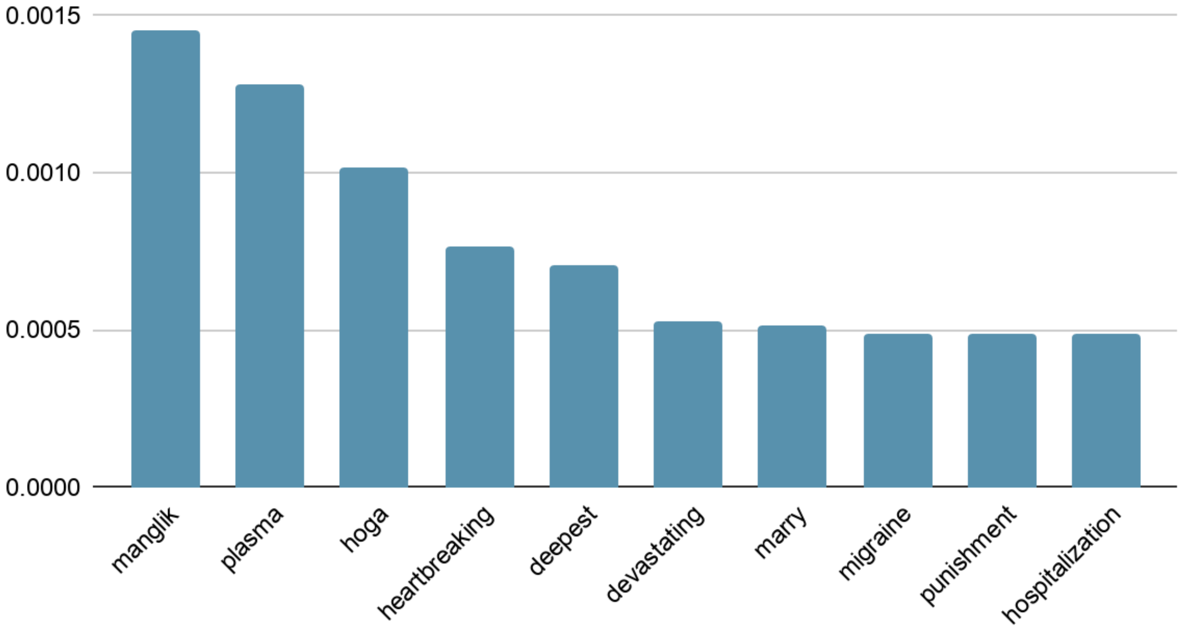


Love-Curious Common tokens by RelativeNormalizeFreq

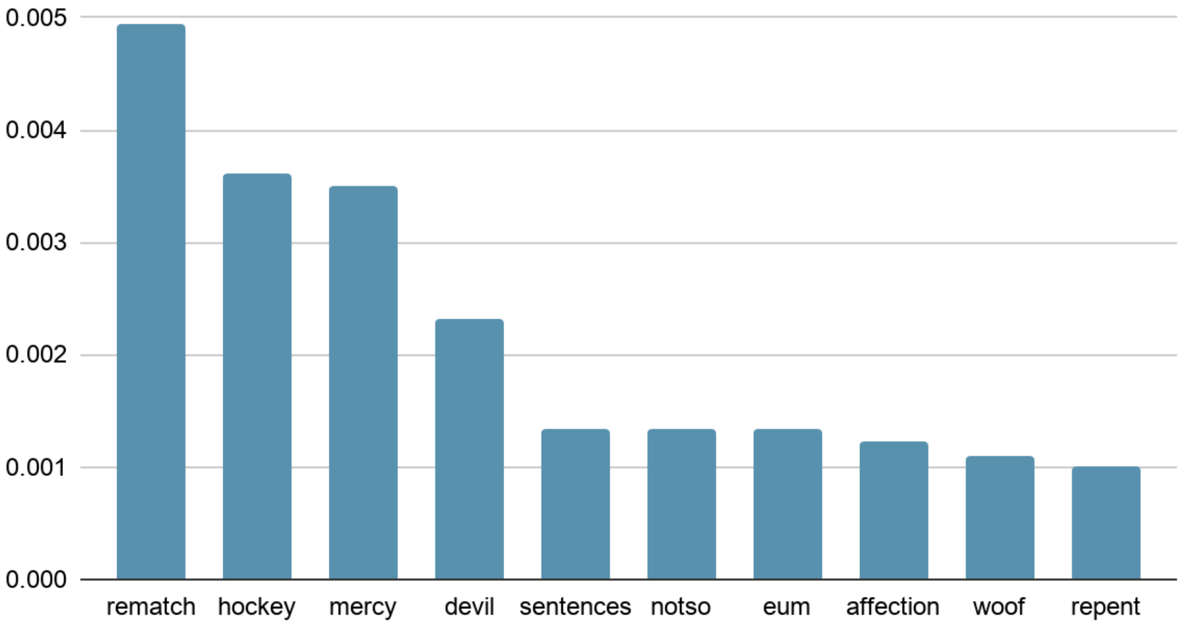
Happy Tf-Idf Most frequent tokens



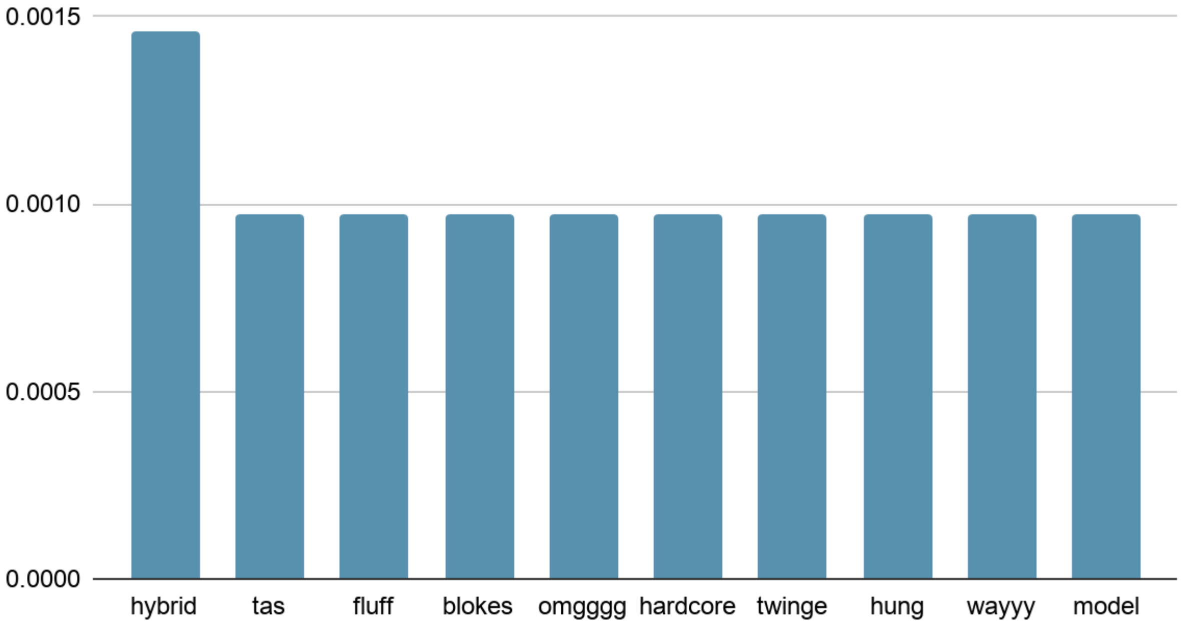
Sad Tf-Idf Most frequent tokens

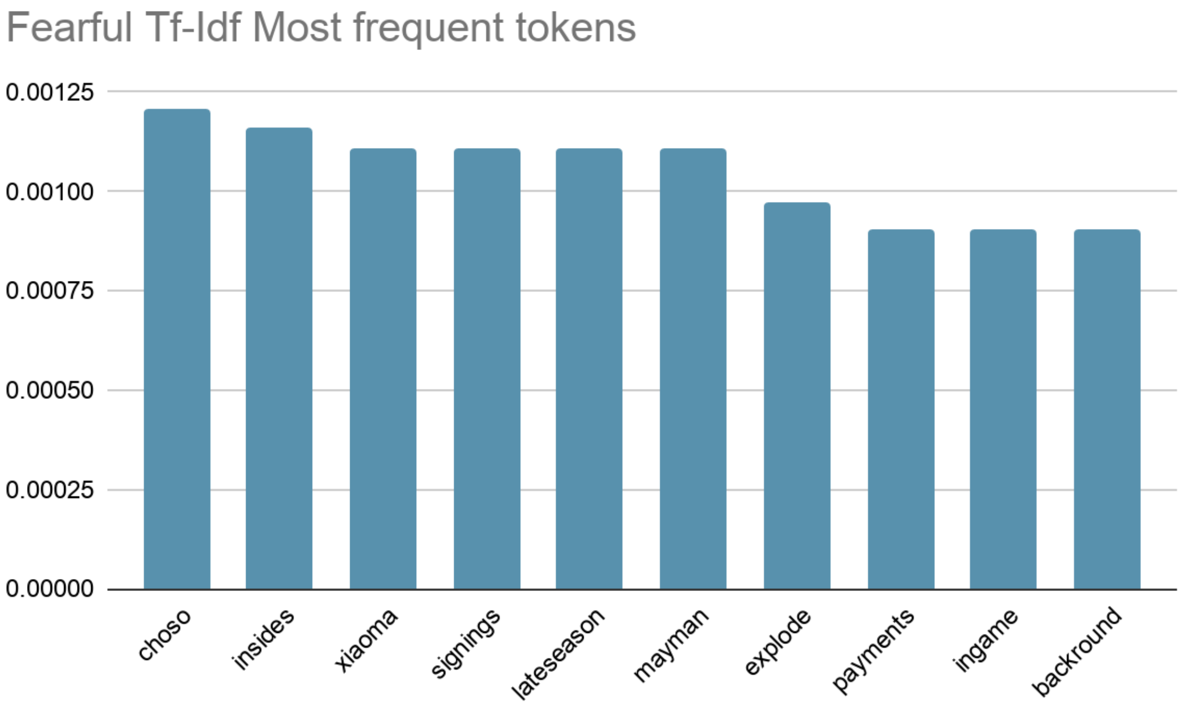


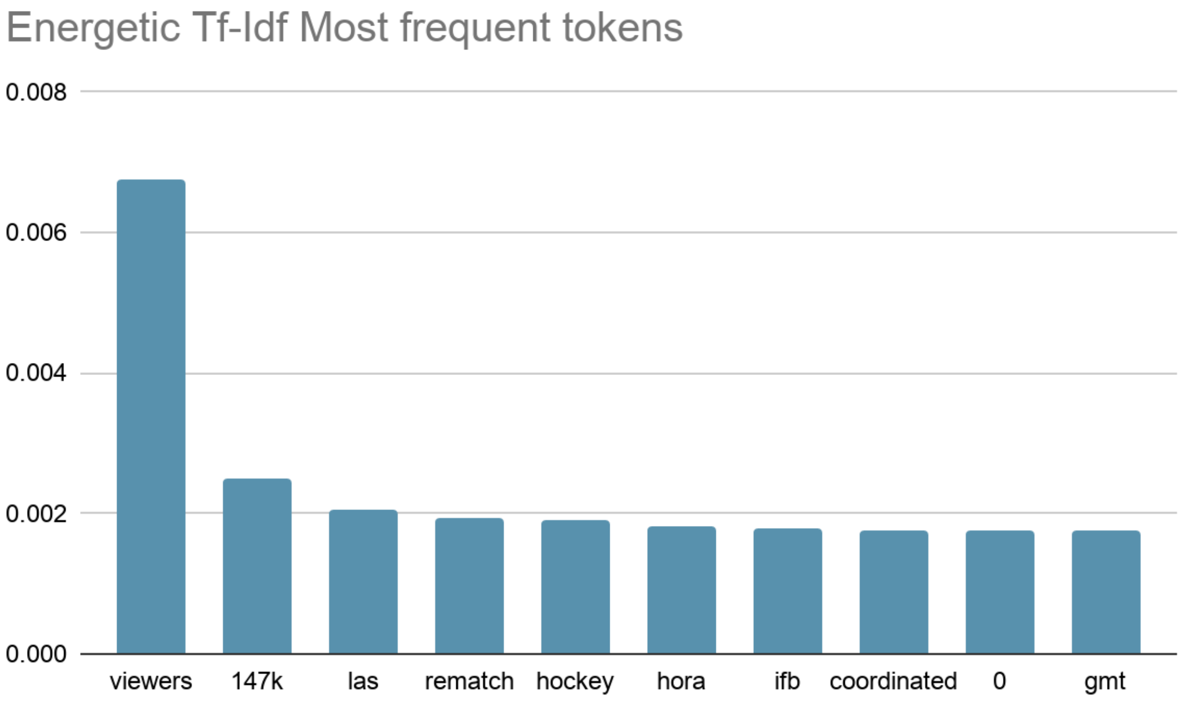
Angry Tf-Idf Most frequent tokens

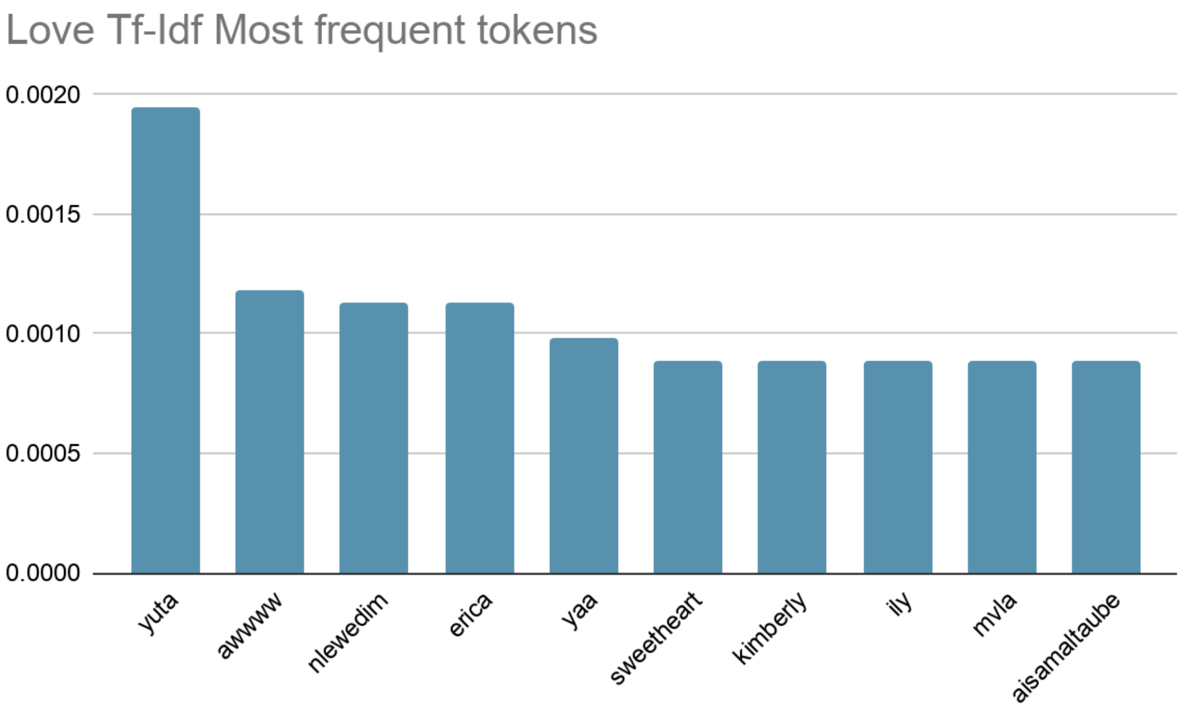


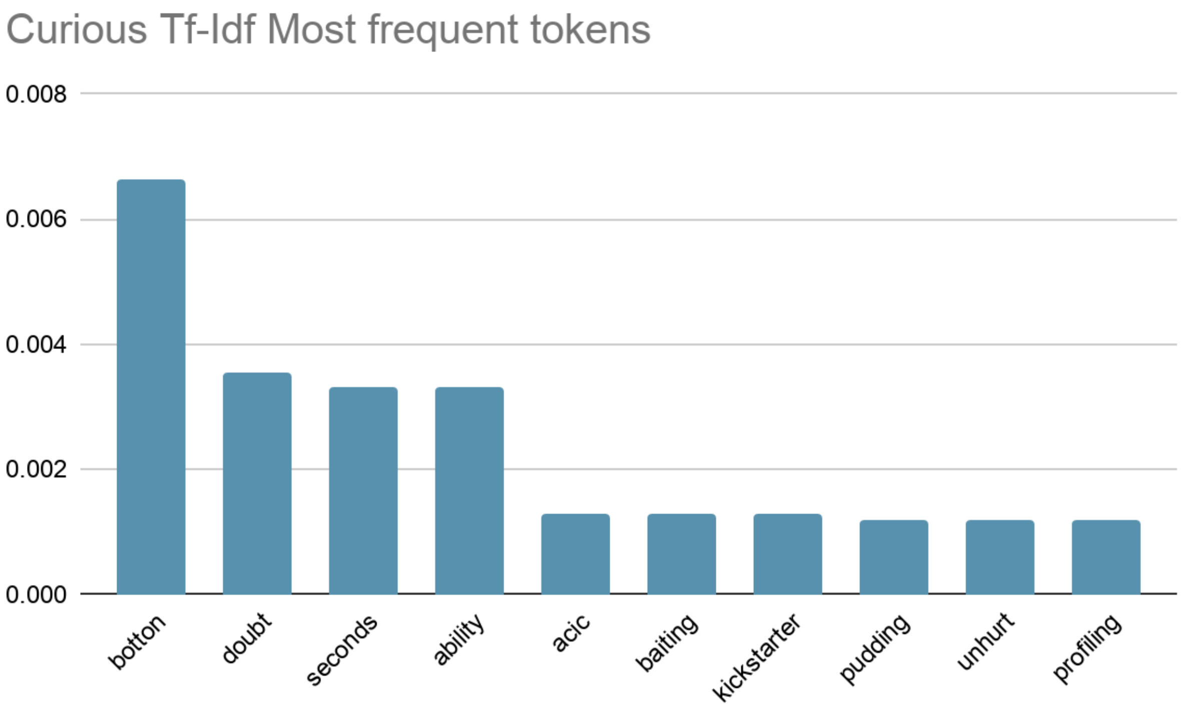
Excited Tf-Idf Most frequent tokens




Fearful Tf-Idf Most frequent tokens
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
choso
insides
xiaoma
signings
lateseason
mayman
explode
payments
ingame
backround


Energetic Tf-Idf Most frequent tokens
0.008
0.006
0.004
0.002
0.000
viewers
147k
las
rematch
hockey
hora
ifb
coordinated
0
gmt


Love Tf-Idf Most frequent tokens


Curious Tf-Idf Most frequent tokens