Distinguishing Between Cordcutters and Piracy Subreddits

The Opener:

In recent years traditional cable and satellite TV have been in decline as viewers transition to consuming content from Subscription Video on Demand (SVoD) services like Netflix and Hulu. Along with this rise in cord-cutters, there has also been an increase in subscriptions to illegal streaming services. Worldwide, users made a total of 300 billion visits to internet piracy sites in 2017.

The Need:

Given the shift in consumers viewing habits and the proliferation of pirated streaming content, a study and comparison of the piracy and cordcutters subreddits could provide SVoD companies insight into the type of content that is being pirated from their services. An evaluation of the intersecting and divergent elements between the two subreddits could potentially influence future content delivery methods and programming decisions these media companies.

I will build a machine learning model predicting whether a given post comes from either the piracy or cordcutter subreddit.
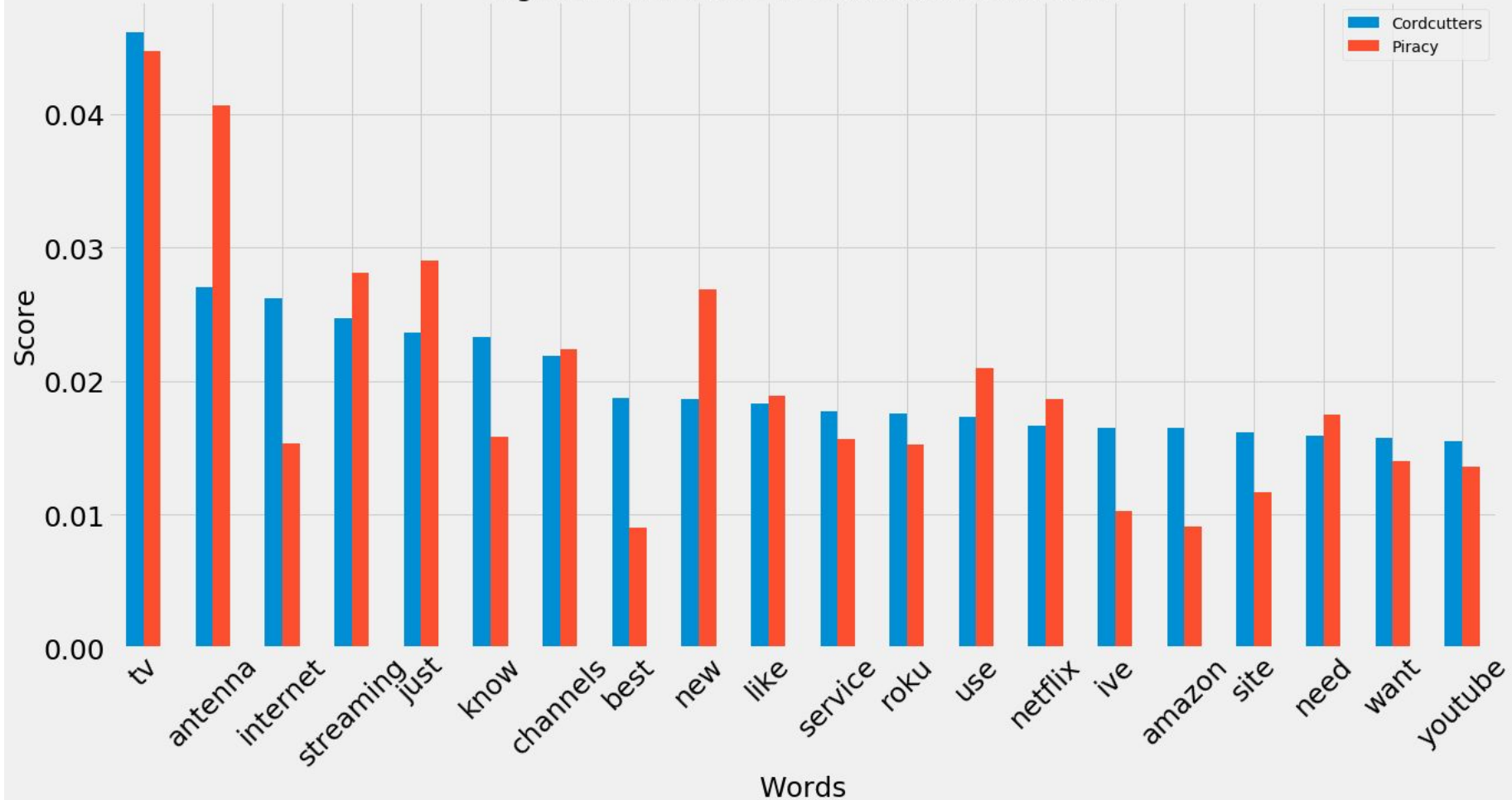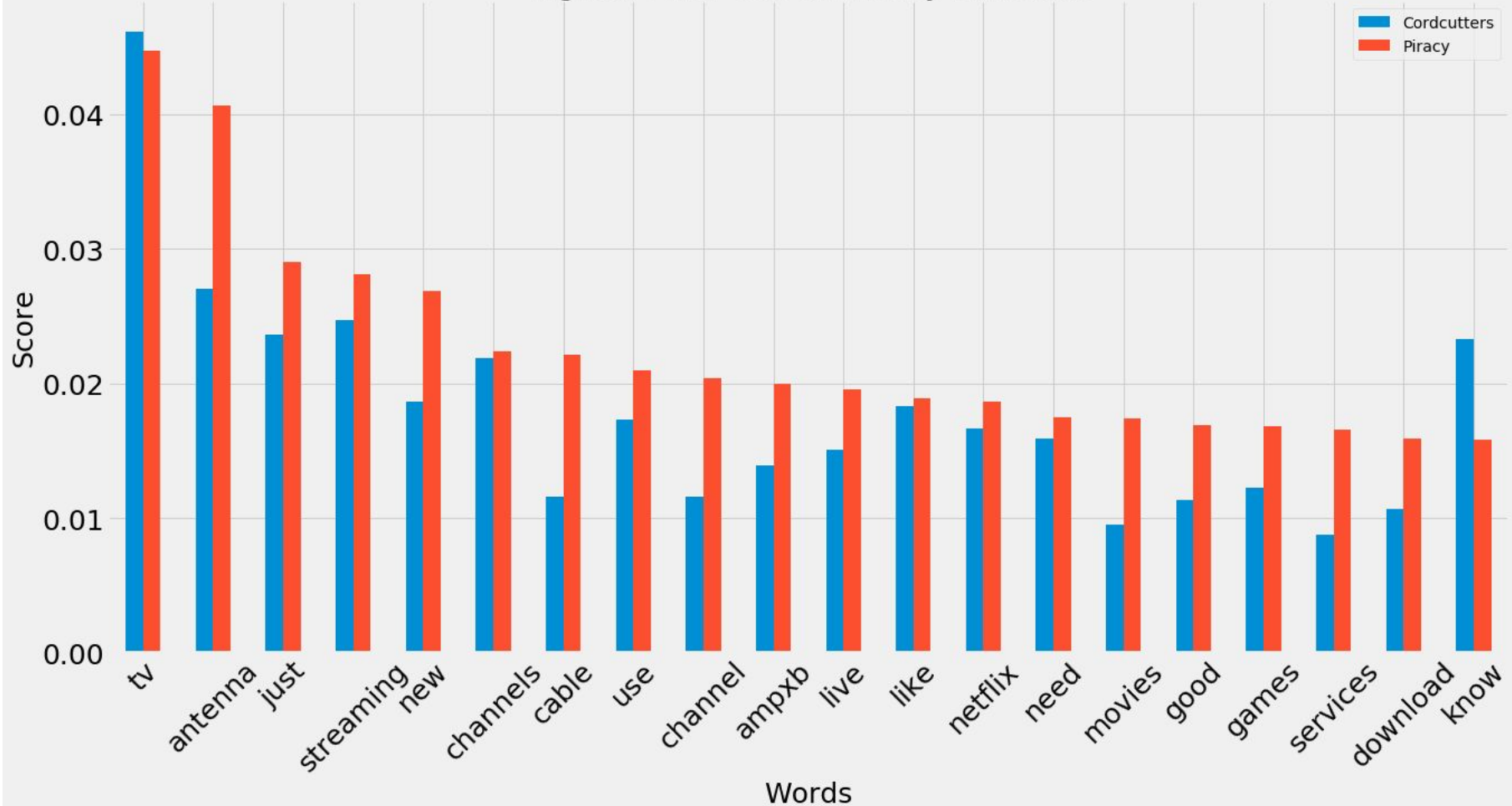
Top Words in Cordcutters
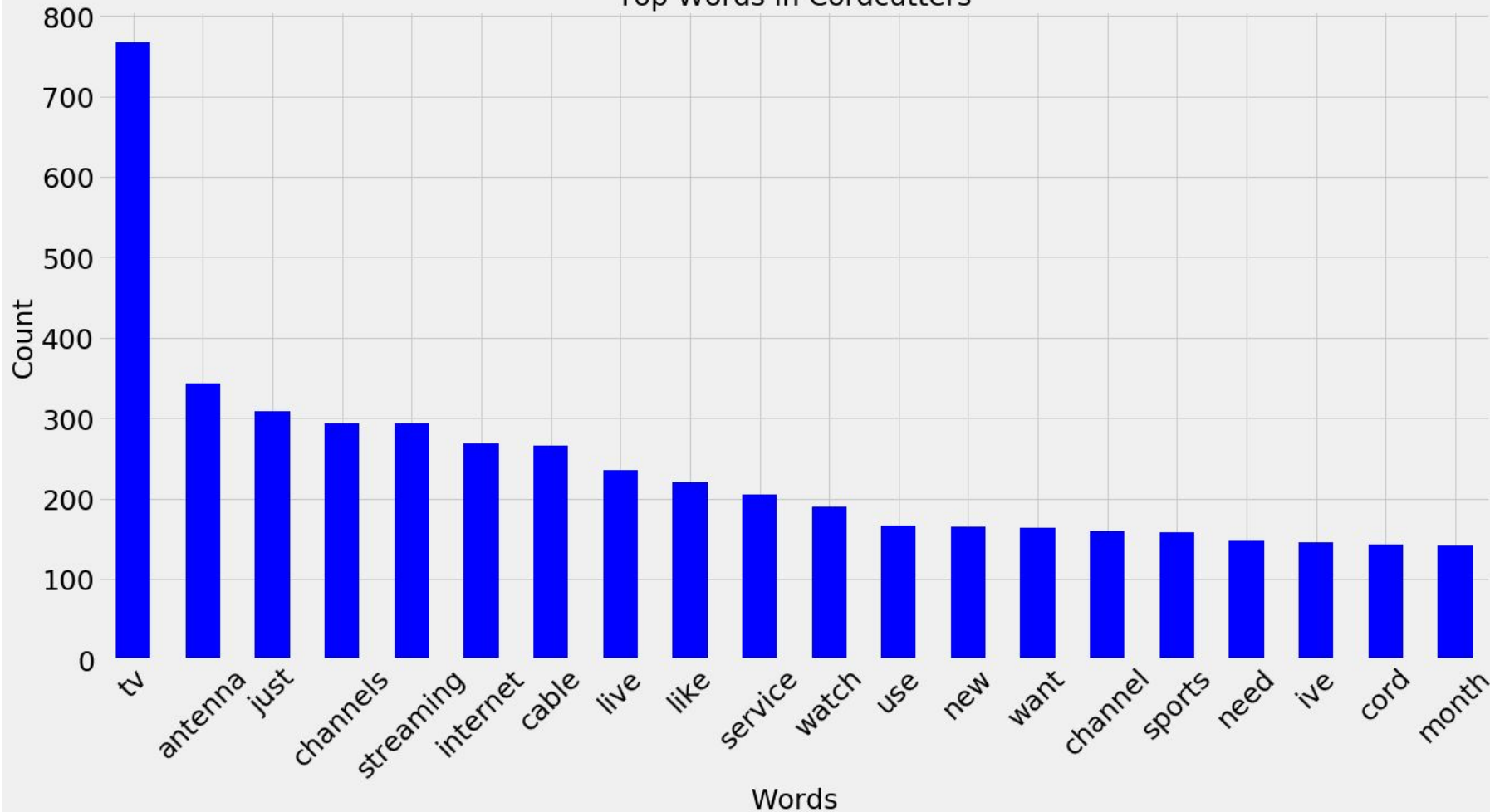
Top Words in Piracy
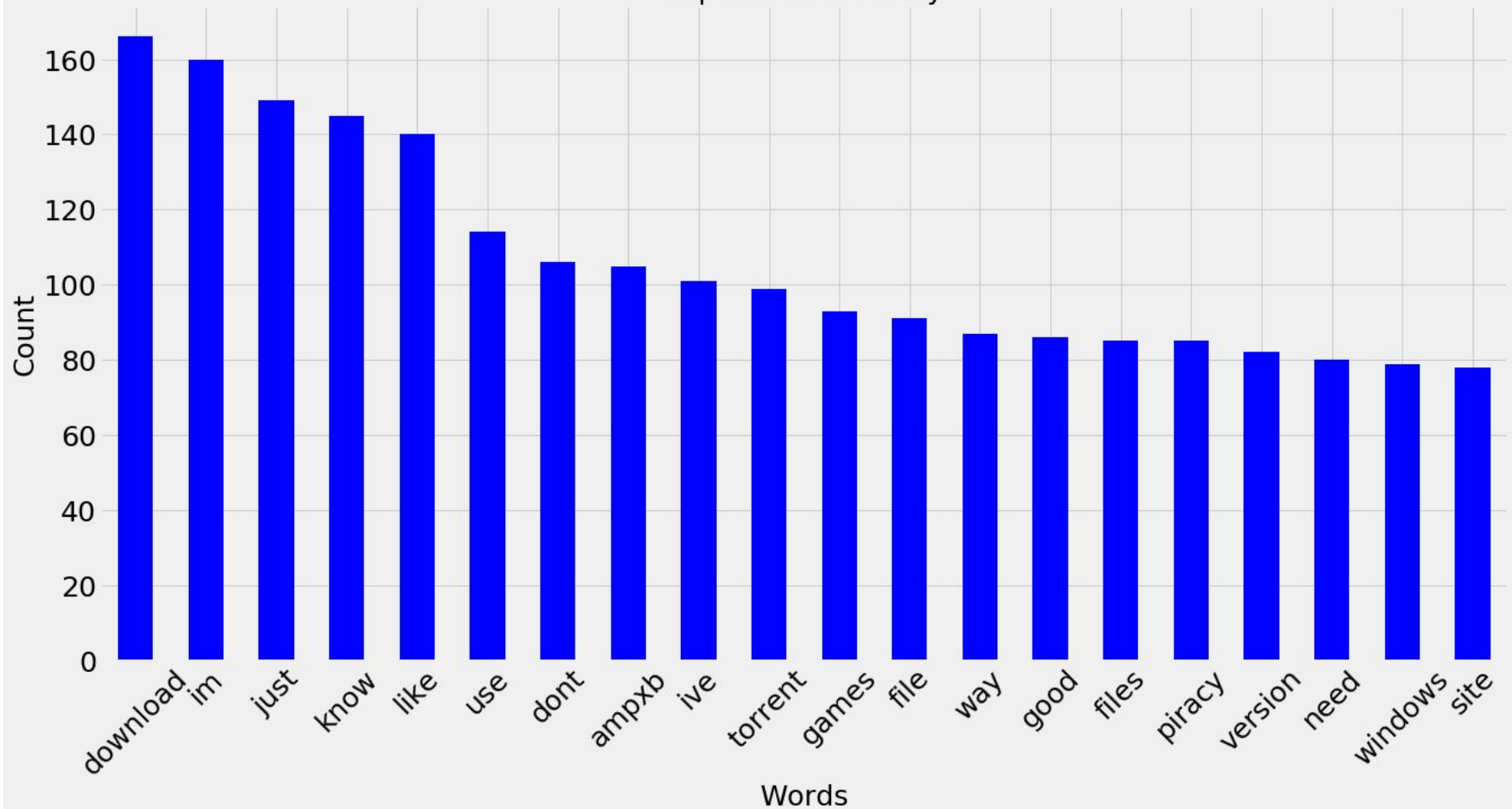
Highest TFIDF Score in Cordcutters Subreddit

Highest TFIDF Score in Piracy Subreddit

Common Words Between Cordcutters after Count Vectorization and TFIDF:

- TV
- Antenna
- Internet
- Streaming
- Just
- Channels
- New
- Like
- Service
- Use
- Ive
- Need
- Want

Common Words Between Piracy after Count Vectorization and TFIDF:

- Just
- Use
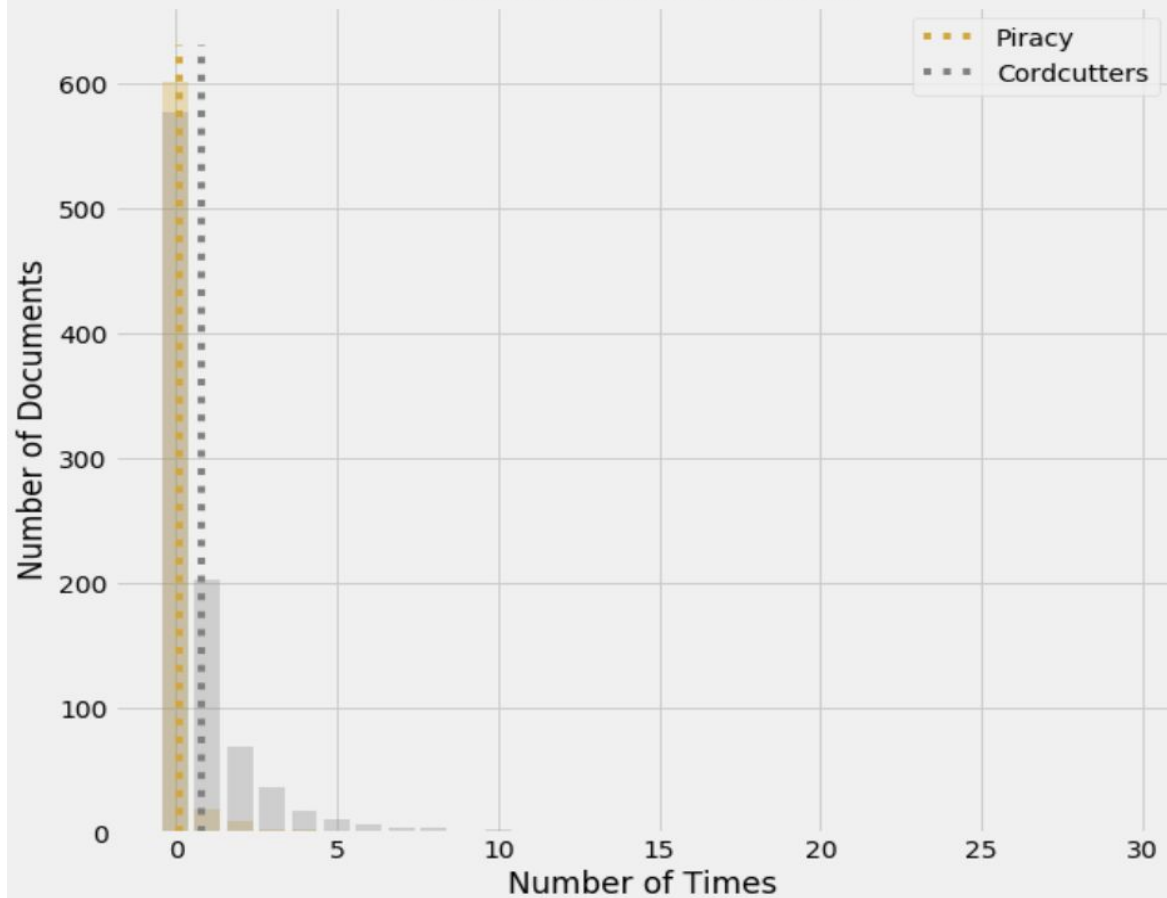- Ampxb
- Like
- Need
- Good
- Games
- Download
- Know

Hypothesis Test:

- *H0*: The subreddits for cordcutters and piracy have the same mean frequency for word x.
- *HA*: The subreddits for cordcutters and piracy have a different mean frequency for word x.

tv frequency counts
means: 0.088 vs 0.822

app frequency counts
means: 0.073 vs 0.137
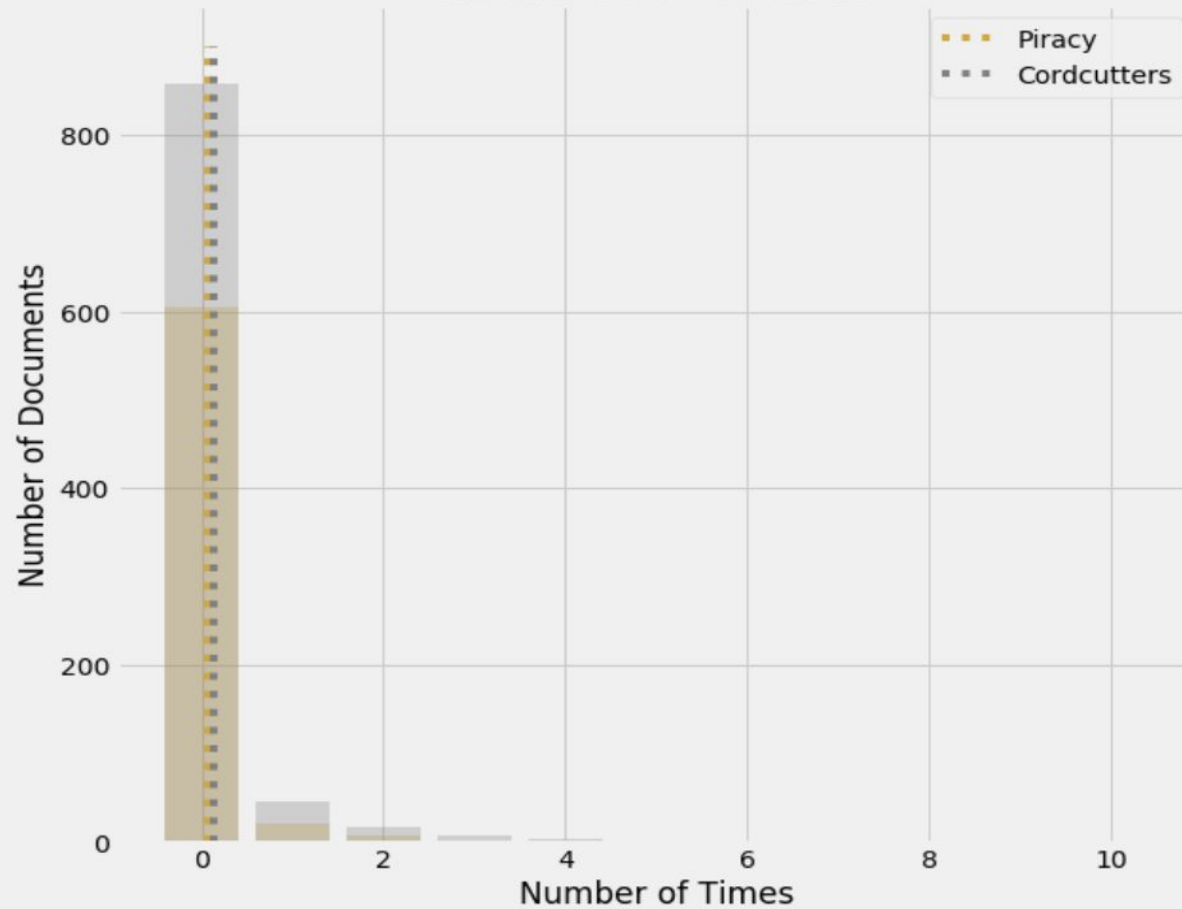
youtube frequency counts
means: 0.058 vs 0.109
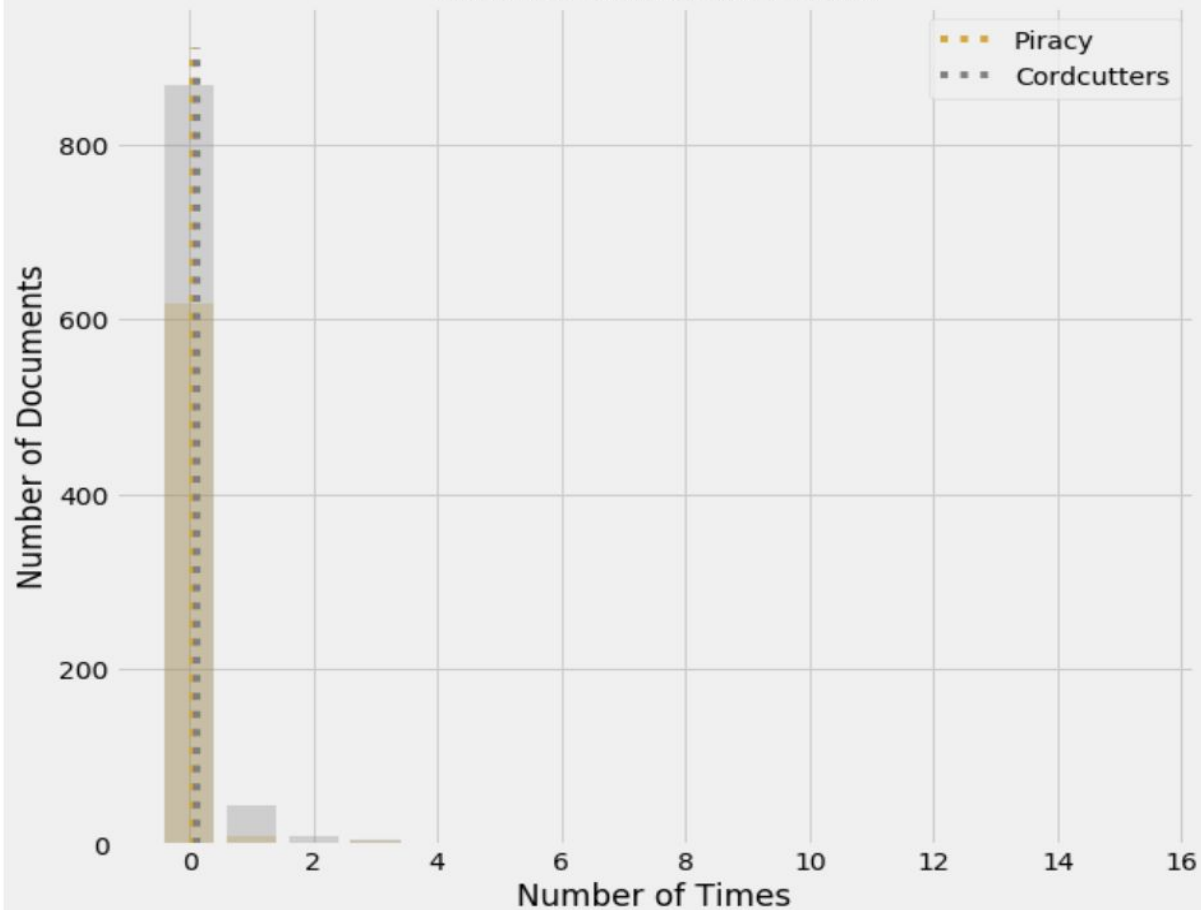
Model	Test Score	ROC AUC
Multinomial Naives Bayes	0.936	0.988
Logistic Regression	0.946	0.987
Random Forest	0.911	0.978
Decision Tree	0.842	0.880

Statistically Significant Words:

[internet, amazon, think, http, buy, content, money, play, hd, extra, provider, connection, figure, came, bundle, expensive, solution, needed, spend, purchased, entertainment, user, software, leave, accounts, windows, official, office, appear, fuck]

Y-Pred != Y-Test (False Positives):

	Raw Text	actual	predicted	probability
44	anyway to bypass it is there anyway to see on...	0	1	0.586170
93	it would possible to overcome netflixs protect...	0	1	0.544349
164	whenever i try to watch hour movies i always ...	0	1	0.534627
178	i am trying to watch the arrow but cant on net...	0	1	0.642231
202	do all rippers of amazon/netflix video use sc...	0	1	0.737383
216	looking for alternatives to nfl game mix i h...	0	1	0.631082
275	has anyone figured out how to do this yet been...	0	1	0.530763
355	i come from england and is it just me or at pe...	0	1	0.593745
359	is there any way to watch fite tv or njpw worl...	0	1	0.641161

1. do all rippers of **amazon**/netflix video use screen capture or is there something to intercept the **stream** direct
2. looking for alternatives to nfl game mix i have a splitter where i can display two inputs on my **tv** at once but im looking for something much easier mystreams has something similar with their subscription nfl game mix alternatives
3. whenever i try to watch hour movies i always have to pause and let it load a little but even then i still encounter buffers i have about /mbps at home and i still cant watch anything consistently stable i was wondering if leonflix would add an online feature where we can download movies and **tv** shows and watch it without the **internet** if theres another way of downloading movies and **tv** shows safely for now please link me to those sources thank you download/offline feature

Y-Pred != Y-Test (False Negatives):

	Raw Text	actual	predicted	probability
31	ive been using the crunchyroll free viewing fo...	1	0	0.448783
50	the film detective expands its library with n...	1	0	0.438082
88	nexus players are dying rapidly and no one kn...	1	0	0.420627
120	i am beyond ashamed for even asking almost mad...	1	0	0.404245
138	apologies if this is the wrong place for this ...	1	0	0.317555
192	cordcutting picks up steam despite cables bes...	1	0	0.487330
204	it seems that simpsons world is no longer a fe...	1	0	0.381714
248	so im with a relative in south carolina zip co...	1	0	0.487967
260	fubo rolling out more k beta broadcasts	1	0	0.450851
273	animal farm movie from andy serkis acquired b...	1	0	0.432202
277	parental controls added with new update versi...	1	0	0.413360
281	twentytwo states ask us appeals court to rein...	1	0	0.475047

1. ive been using the crunchyroll free viewing for some time now but the flash player they use is buggy slow and crashes a lot i was **thinking** of paying for the premium **account** soon but dont want to be disappointed with how it works does anyone have experience and can talk about it thanks for the help in advance crunchyroll flash player better in premium
2. i am beyond ashamed for even asking almost made a throwaway **account** to ask this there are dumb reasons i am asking is there a way to get the hallmark channel
3. apologies if this is the wrong place for this post i cannot find a website/app that does this and **know** it must exist you enter title in their search bar and it tells you if the movie or show is available to stream and where to find it does this exist website for finding where you can stream a movie/tv show

Conclusion:

- My analysis recognizes the intersecting words between the two subreddits and identifies the statistically significant words in both classes.
- This would benefit a SVoD company because using this data it can draw links between what cordcutters are watching and what is being pirated and adjust their decision-making accordingly.