CSCI 1951A Spring 2019

The Sound of Censorship

dbayazit, npant, athoopth, chuang25

Midterm Report

I - Introduction

In our project "The Sound of Censorship" we've been investigating correlations between user features (country, age, gender… etc.) and a country's freedom of expression features (press freedom ranking, internet freedom status… etc.). We collected data on Internet freedom, press freedom and on Last FM users. We used the pandas package to drop/clean and investigate the data. The features we ended up choosing are the following for each of the datasets:

- Internet freedom:

country-id	free-status	access-block	content-limit	user-right-violation	total

- Press freedom:

country-id	2001	2002	2003	2004	2005	2006	2007
2008	2009	2012	2013	2014	2015	2016	

- Last FM Users:

user-id	country-id	age	gender	playcount
novelty_artist_avg_month	novelty_artist_avg_6months	novelty_artist_avg_year	mainstreaminess_avg_month	mainstreaminess_avg_6months
mainstreaminess_avg_year	mainstreaminess_global	cnt_listeningevents	cnt_distinct_tracks	cnt_distinct_artists
cnt_listeningevents_per_week				

When merging data from 3 sources, we realized how many users were removed due to the fact that they had no age or no country matching to the internet and press freedom datasets. There were essentially 3 types of joins we investigated with the user data:

- Internet only: due to the fact that this dataset dating from 2018 had 65 countries, when we left-joined it onto the Last FM users (where the Last FM dataset was the left side of the join), the amount of users decreased from 37510 to 21037.

- Press only: due to the fact that this dataset had more countries (178) it had a large amount of users (37286).
- Both: when combined with both datasets, the amount of users left were 20869, which is not a bad number of datapoints to train/do analysis on. We still get enough data to investigate our hypothesis.

However it wasn't really only the amount of users that was concerning for us but the variability in the type of freedom status per country that was left to study. After plotting a histogram of the free-status column given by the internet freedom dataset (which has 3 possible values, 0 meaning not free, 1 meaning partly free and 2 meaning free), we realized that this variability was lost. Here's a plot to make it clear, we can see that there aren't enough users from countries that are not free. The reasons behind this phenomenon can be that people from countries that are less free don't have the same opportunities to express themselves on the Internet because of Internet restrictions. That makes them less likely to be users on the Last FM streaming service.

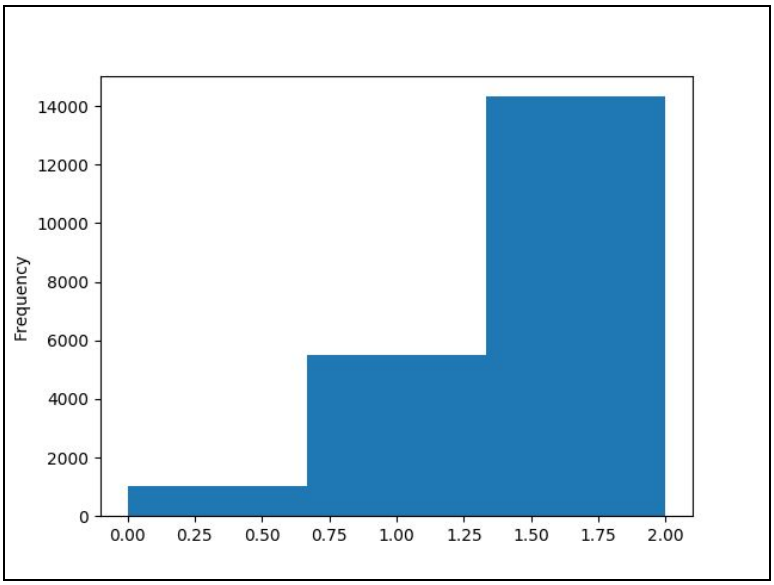

Fig. 1: Histogram of Freedom of Status
The first bar represents the NOT FREE countries, the second PARTLY FREE, and the third FREE. The classification is made by Freedom House, an independent watchdog organization.

II - Analysis

1. Country Internet Features vs. Novelty in Artist Listened:

We started our analysis with our second hypothesis mentioned in our blogpost that was the most related to our project mission. We wanted to investigate correlation between the variety of “new” musicians listened to by the user and the internet freedom features of the country they are from. We expected that a country with better freedom of Internet may have more medium to spread information about novel (to the user) music.

Shockingly we noticed that "novelty_artist_avg_month" and "novelty_artist_avg_year" didn’t even show up in the correlation matrix although we created a new dataframe to represent all of the columns shown in Figure 2. This means that they were so uncorrelated (probably less than 0.00…) that the pandas correlation matrix did not keep it. Nevertheless “novelty_artist_avg_6months” survived the correlation function. Sadly, the numbers do not seem that significant.

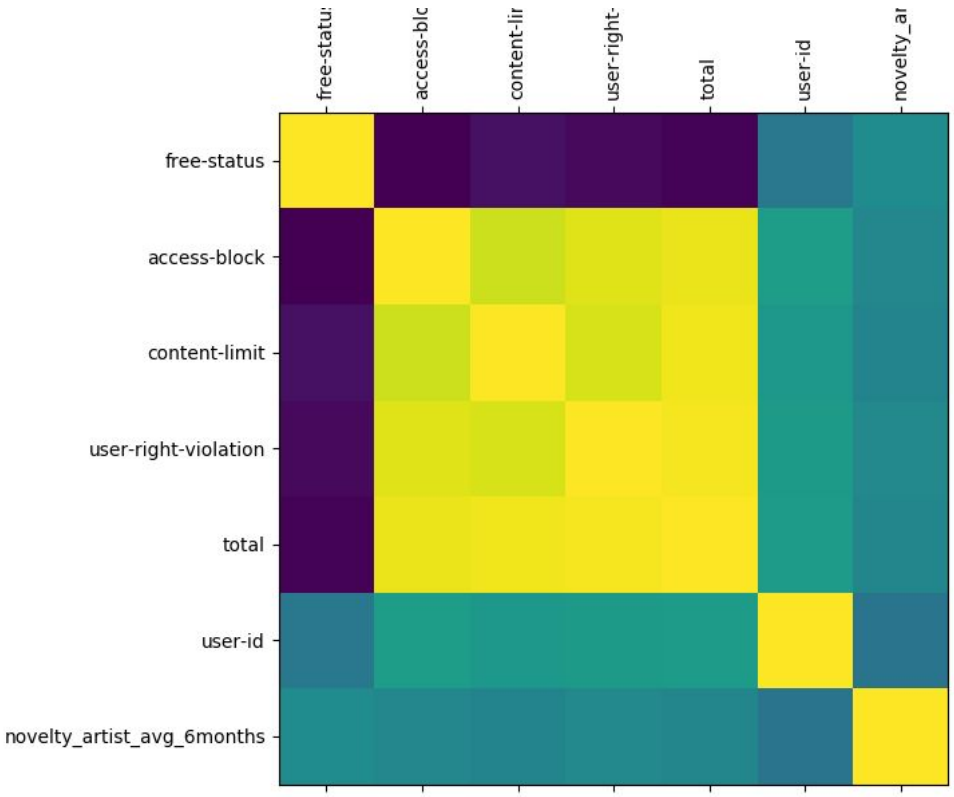

Fig. 2: Correlation Matrix Visualization

After brainstorming we had a few ideas as to why this might have happened. We had thought that Last FM's dataset was from 2018, but that was actually the date the paper got published. When we searched more we found out that the Last FM dataset is from 2011 while the Internet freedom dataset was from 2018. We did further investigation by making a correlation between press freedom in 2012 and still got no promising results.

After our dataset investigation we could probably relate this problem back to the lack countries that have less freedom of expression.

	free-status	access-block	content-limit	user-right-violation	total	user-id	novelty_artist_avg_6months
free-status	1	-0.91	-0.83	-0.86	-0.89	-0.15	0.01
access-block	-0.91	1	0.85	0.90	0.94	0.15	-0.02
content-limit	-0.83	0.85	1	0.88	0.96	0.10	-0.04
user-right-violation	-0.86	0.90	0.88	1	0.97	0.12	-0.01
total	-0.89	0.94	0.96	0.97	1	0.13	-0.03
user-id	-0.15	0.15	0.10	0.12	0.13	1	-0.18
novelty_artist_avg_6months	0.01	-0.02	-0.04	-0.01	-0.03	-0.18	1

Fig. 3: Correlation Matrix

2. Age vs. Internet Ranking vs. Number of Listening Events

Next we investigated our very first hypothesis from our blogpost by using a kMeans. We expected that younger generations (age 0 - 35) would find the obstacles put by the government less of a problem (ex: ease at using VPN) and have listening habits that are similar to the ones in their respective countries such as the number of listening events. This also means that we expected the opposite from older generations. For this we used scikit-learn's kMeans algorithm to achieve the kMeans analysis. 'Access-block' is used to represent the internet ranking while the other features already have their own corresponding columns. To find the optimal number of centroids, we plotted the Elbow Curve (showed in Fig. 4). The score is used to represent the accuracy of our model. And we see that the graph levels off rapidly after 3 clusters, implying that addition of more clusters do not improve our model but may increase the chance of overfitting.

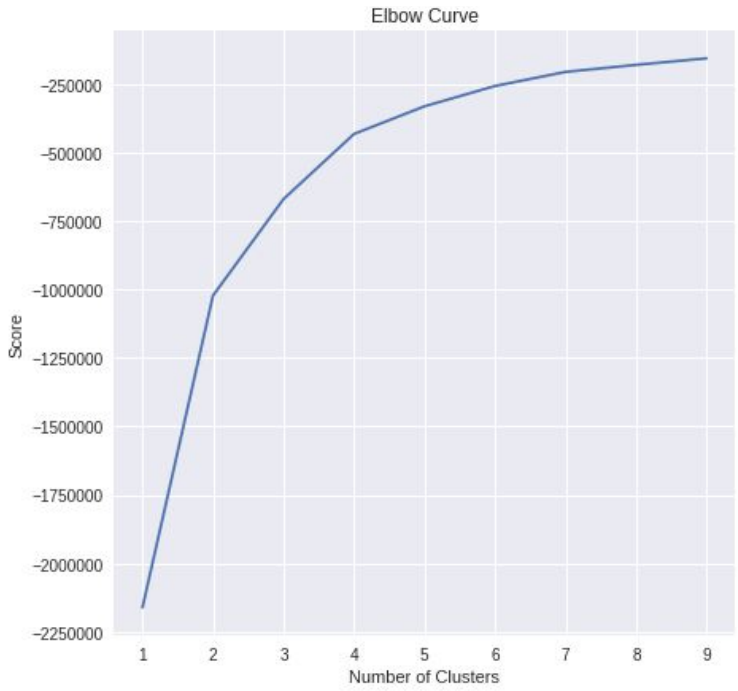

Fig 4: Elbow Curve of KMeans training

After we got the optimal number of centroids, we plot the result of kMeans using k=3 in Fig. 5. As we can see from the figure, the data points are simply categorized based upon the age of users regardless the number of 'access block'. This potentially implies that there is no significance relationship between a user's age, their internet accessibility and their listening habits. In other words, people are affected quite evenly by the government no matter how old they are.

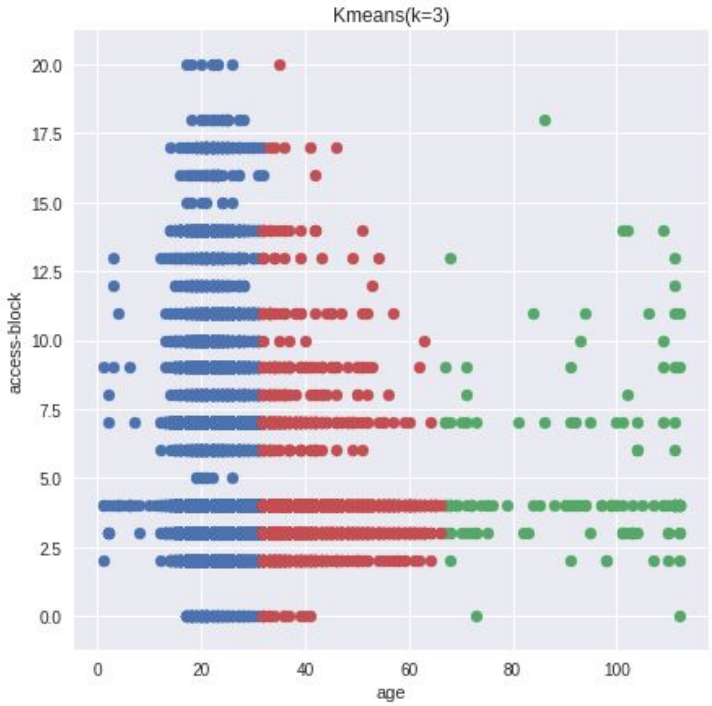

Fig. 5: Result of Kmeans (k=3)

III - Conclusion

a. There are 2 challenges that we have encountered so far. The first one was related to computation power and the largeness of our dataset. We finally got a team folder of 50 GB on the Computer Science department machines. Additionally we decided to use the additional user information dataset that has been extracted from the large dataset. The second challenge is discussed in (d).

b. Our initial insights and hypothesis were around finding correlations between freedom of expression features and user features.

c. We do have concrete results to show but they are not the results we would wish to have. We have noticed very low correlation between certain listening habits like the average amount of novel artists listened to in the past 6 months. This is largely due to the lack of a variety of freedom between the countries the users are from.

d. Now we face another challenge that is, the lack of users from countries that are not that free both in press and on the Internet. Going forward our biggest problem will probably to find data that accounts for that. We will probably have to find a new user dataset but a thorough investigation of the countries the users are from should be done. The lack of certain features like age or additional

user info could be related to this problem where certain users are being taken out of the data because of having null values in features that we might not care.

e. We are on track since we've been doing all of the checkpoints so far, but the lack of significance results from our analysis and the situation we have with our data will require us to do more work.
f. We still believe that there should be correlations between a user's listening events and the restrictions of expression put by their country. Changing the whole project is a complicated matter, although that might be the necessary step to make since it is true that if one's internet is restricted then they are less likely to be users on Internet services. We can only truly decide on that once we consult our mentor TA and run some other analysis on the rest of the hypothesis we had.