Censorship Machine Learning and Phylogeny Proposal

Introduction

The research aims at understanding private actors' motivation and behaviour in censorship. Based on existing Citizen Lab research, Chinese online censorship is decentralized without much official announcement or centralized wordlist[1]. We will analyze censored wordlists posted by various game and app developers[2] to answer these questions: **what is the mechanism of word lists dissemination? What factors affect private entities' decisions on what to censor?** Insights into these questions have both theoretical and practical value. Existing surveys and theories on censorship focuses on state-actors and centralized decisions, instead of private actors. Findings on motivations and practice of private entities enrich the theoretical field and provide a more complete and nuanced picture of information control. Practically, better understanding the working of censors will inform and inspire activists to design more effective circumvention.

Research Question

What is the mechanism of word lists dissemination? What factors affect private entities' decisions on what to censor?

We aim to answer these questions by measuring the similarity between blacklists to identify clusters of similar censorship practices between applications, and identify how these vary with the project scale and business model, if applicable. In particular, we aim to study the proportion of political keywords in each blacklist, and see which contextual factors influence this.

Hypothesis

We will take two steps to process the dependent variable — similarity between lists. First, political keywords are separated from the non-political ones. If there is a similarity in ratio of political/non-political between lists, we will take notice. Second, political keywords will be the focus of analysis. We will discover the similarities between lists, by using certain strings (blocks of words in the same order) that clusters of lists have in common.

Independent variables include the characteristics of the developers, namely their scale, profit-orientation, and affiliation. We hypothesize that, if a developer team is 1) large scale, 2) profit driven, it is likely to have a higher political to non-political keyword ratio, and be the parent of other lists. If a developer team is 1) small scale, 2) non-profit (e.g. personal/student/experimental projects), and/or 3) a subsidiary of a larger company, it is likely to have a lower political to non-political ratio, and be the progeny of other lists. The larger, thus more resourceful, a developer team is, the more likely its lists or part of its list, being copied. The larger, thus more influential, a developer team is, the more likely it being strict about socio-political content to avoid major fallouts.

---

[1] "Measuring Decentralization of Chinese Keyword ... - Usenix." Accessed October 29, 2020. https://www.usenix.org/system/files/conference/foci17/foci17-paper-knockel.pdf.

[2] "The effect of information controls on developers ... - Citizen Lab." Accessed October 29, 2020. https://citizenlab.ca/wp-content/uploads/2018/08/nlp4if2018github-1.pdf.

Methodology

For the first step of the process, we will be working to come up with a precise definition of what entails a political vs a non-political term. This definition can be used to come up with parameters that would differentiate either term. Using these parameters, we can classify lists of terms from the databases provided. Here, there is an opportunity to build a machine learning algorithm and train it to differentiate between the two types of terms. Alongside this, there is also potential to classify these lists into more than just political and non-political terms, but also potentially tag them sentiments, or perhaps action vs non-action statements, etc.

Within the second step, we aim to find similarity amongst these lists. This will likely come in the form of how much overlap is found between any two given lists. We will work with our team to develop precise metrics, such as the longest string (block of words in the same order) in common between blacklists. Out of these similarity classifications, we will be able to create a cluster map (with potential use of graph theory concepts), where the more similar any lists are, the closer they are to each other.

Ideally, we actually find lists with high levels of similarity, which would warrant the next sub-step of gathering further data about which companies those lists have originated from. Through the use of the mobile games and Github. We are able to gather some information about the companies that created those lists based on further research from the datasets. For example, within the mobile games database, we can see in many instances which company created said mobile games, and there are many repeats (e.g. games with Baidu in the blacklist filename) (e.g. games with Baidu in the blacklist filename). If we find there are large clusters, we can look into the metrics of said companies and find further research into their demographics.

Alternatively, if there is not that much similarity, we can change gears and continue further with our classification step and see if there are some categories of words more prevalent than others. For the similarity idea, the separation of political and non-political terms is mainly practical, and not necessarily for analysis. However, if similarity yields no results, that classification can be used for a variety of other metrics for many different insights.

QUESTIONS:

- How much information do you believe we would be able to get about the companies we are analysing? Which database would be best to use for this?
- If we do switch gears from similarity, and aim to find potentially different categorizations, what are the top three metrics you would be most interested in being explored and why?

Notes from the meeting:

1. Metrics for similarity: same words and same word order, indicates copying
    - Eliminate false positives: eliminate bigrams
2. Simplify independent variables: profit vs. non-profit; big companies have a larger list, but is not definitive of social vs. political ratio?
3. Company metric: company and independently-working teams, could be hard to find in the internet.
4. Do independent teams share lists more than large companies?

Classification based on radicals/roots:
Find out the wordroots of currently political censored words
To track words with the same roots to see if they are censored