Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Leveraging NLP and Social Network Analytic Techniques to Detect Censored Keywords: System Design and Experiments

Christopher S. Leberknight
Montclair State University
leberknightc@montclair.edu

Anna Feldman
Montclair State University
feldmana@montclair.edu

Abstract

Internet regulation in the form of online censorship and Internet shutdowns have been increasing over recent years. This paper presents a natural language processing (NLP) application for performing cross country probing that conceals the exact location of the originating request. A detailed discussion of the application aims to stimulate further investigation into new methods for measuring and quantifying Internet censorship practices around the world. In addition, results from two experiments involving search engine queries of banned keywords demonstrates censorship practices vary across different search engines. These results suggest opportunities for developing circumvention technologies that enable open and free access to information.

1. Introduction

Online censorship can be broadly segmented into two main categories: (1) polices and technologies to restrict access to information, and (2) policies and technologies used to restrict the publishing of content. This paper is solely focused on investigating the use of keywords as the criteria for deciding whether to allow or deny access to information. Keyword analysis has been extensively studied in the past, however research combining natural language processing and social network techniques is somewhat sparse. Several open challenges exist for understanding the consistency of censorship enforcement across different communication platforms such as blogs, search engines, and email. In addition there are several open questions for developing passive network monitors and for determining the accuracy of censorship systems with respect to different types of content such as text, images and video.

This research presents an application that applies natural language processing and social network analytic (SNA) techniques to investigate keyword censorship on several search engines in mainland China. The paper makes the following contributions:

1. provides a method to probe remote systems while obfuscating the location of originating requests
2. presents a classification algorithm for categorizing banned keywords and future research directions for enhancing the classification accuracy
3. demonstrates a hybrid approach combining NLP and social network analytic techniques for classifying and visualizing banned keywords

2. Related work

There are many examples that demonstrate the increasing threats to Internet Freedom for many citizens around the world [1, 2, 3]. Considering these threats and the power the Internet has to shape public opinion as is the case with Facebook and the US 2016 presidential elections [4] it is imperative to keep a watchful eye on policies and technologies that may transform the way in which we access information on the Internet. This paper aims to alert the research community to stay vigilant of government and authorities' attempt to enforce control over the Internet. These examples and many others highlight the need to study how censorship is enforced. In some societies that heavily enforce online censorship results from this work may shed light on the development of new circumvention technologies that enable free and open communication.

Information control is widely used to suppress public uprisings in response to disapproval with authoritative policies or actions. Censorship can come in the form of either restricting content from being published or restricting access to content. Restricting access is often enforced when the content or information resides on a machine outside of the geographical boundaries of the country practicing censorship. For example, access to Facebook is restricted in many countries outside of the United States. Social media

URI: https://hdl.handle.net/10125/59724
ISBN: 978-0-9981331-2-6

poses the greatest threat to maintaining the status quo and has sparked a great deal of research examining different factors of censorship on blogs such as the accuracy and time to detect certain keywords [5]. Restricting the publishing of content may involve both human intervention and automated methods to disrupt or block the transmission of specific phrases or keywords. Consequently, keyword extraction or analysis has become a burgeoning area of interest for Internet censorship research.

Several detection mechanisms have been proposed in the past to observe and categorize the type of content and keywords that are censored. However, recent examples claim there is no monolithic set of rules that govern how Internet censorship is being implemented in China [6, 7]. Advancements in cryptography that can conceal the transmission of banned keywords have also been examined, yet the use of existing cryptographic techniques [8, 9] to evade online censorship is not feasible. In many countries encryption technologies are heavily regulated and details regarding methods, code, and keys need to be disclosed to government authorities [10]. Automatic topic extraction is the process of identifying important terms in the text that are representative of the corpus as a whole. Prior research analyze topics of the deleted posts in Chinese social media in order to compare the deletion speeds for different keywords [7]. The authors report that the topics where mass removal happens the fastest are those that are hot topics in Weibo as a whole.

3. Architecture

To automate the banned words analysis process file IO, HTTP client technique, multi-threading and word embeddings are utilized.

Spring Model-View-Controller (MVC) is used as the application framework since it facilitates parallel development without the need for an application server. This streamlined development from standard desktop to Web by allowing different front-end and back-end components to be developed and tested simultaneously. MVC is widely used by software developers to build flexible and loose coupled web applications.

Model is a Plain Old Java Object (POJO) to hold the application data. The View model generates the HTML based view with model data. The Controller handles user requests and builds the view for the client.

Figure 1 shows the functions provided by the Spring MVC framework. The dispatcher Servlet processes the data flow of an HTTP request. First the map controller l redirects the request to the associated controller. The controller then processes the data and passes the processed data to the view resolver. The view resolver then builds the view and passes it to the front end in HTTP response format.

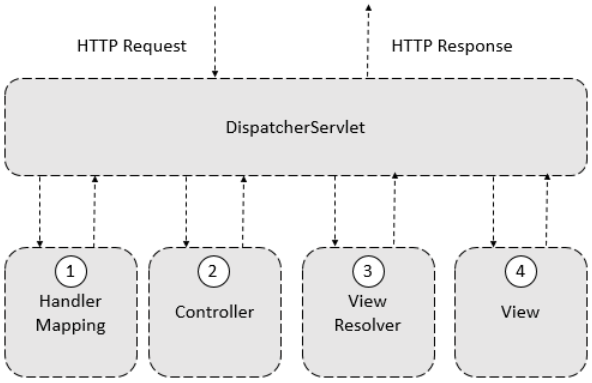

Figure 1 Spring MVC

On the front end, JSP, Bootstrap, JQuery, Ajax is used to build and render the view. The Mongodb NoSql database provides the backend data store and was chosen since it is easy to expand in the future and it can handle non-relational data very well. The application is hosted on a tomcat7 server. The architecture and technologies are used to develop the censor detection system that tests and records that status of banned words on several different popular search engines in China. The key modules that provide the core functionality of the system are described in the next section followed by a discussion of experiments and results. The experiments investigate the variability of censorship practices across the different search engines and study the categorical content most frequently censored.

4. Application

The censor detection system consists of two applications; one for data collection (ChCensor) and one for categorization. Figure 2 presents a high level overview of the ChCensor application. For efficiency the application is designed with three threads. One thread reads key words from text file while the second one checks if these words are blocked by Chinese search engine. The third thread writes data into database.

4.1. Data collector modules

ChCensor is a web application that used to collect data. With ChCensor, a user can search a single word and test if this word or its translation is blocked by Baidu, Sogou or Qihu. A user can also upload a word list file and the application will test every word in the file to determine if it is being blocked by each of the three search engines. The results are displayed in a data table and tag cloud format.

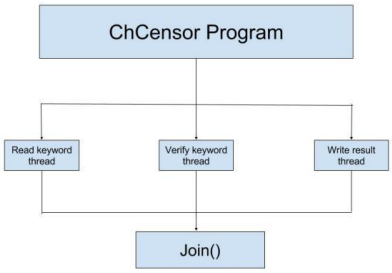

Figure 2 ChCensor Program

The data collector has three core modules: (1) IO module, (2) Search module and (3) Translation module. The collector also consists of three other Spring modules: (1) controller, (2) model and (3) repository.

The IO module reads word lists from a file and writes the searching result to a file. The search module checks if a word is banned by a specific search engine and the translation module translates from Chinese to English. Each of these modules are now discussed in greater detail.

The IO module can not only read and write a plain text file, but also can process Microsoft Excel files using the Apache poi package.

```
public IOExcel(String fileName) throws IOException {
    this.fileName = fileName;

    File file = new File(fileName);
    if (!file.exists()){
        file.createNewFile();
    }

    fs = new POIFSFileSystem(new FileInputStream(fileName))
    wb = new HSSFWorkbook(fs);
    sheet = wb.getSheetAt(0);
}
```

Figure 3 IOExcel Constructor

Figure 2 contains the code for the constructor of IOExcel class. The constructor accepts a file name that contains the path to the file. The constructor initializes the file input stream, POI file system, workbook and the first sheet.

Processing and checking each word requires good IO performance and parallel tasks. Reading words, searching words and saving results in sequence, sometimes caused the program to crash due to heavy IO and poor process synchronization. To avoid this kind of situation, multi-threading is used in the program. Multi-threading is implemented with the design of a word container called word bag as shown in Figure 3. The word bag holds a few words, so the word list reader can fill words in the bag and the searcher can retrieve word candidates from the bag. Basically, the program is built using a consumer-producer model. In this way, if a minor error occurs with one single word, the system will just throw an exception, record the exception and continue.

```
public synchronized void add(String line){
    if (bag.size() >= MAX_SIZE){
        try {
            wait();
        } catch (InterruptedException e) {
            e.printStackTrace();
        }
    }

    bag.add(line);
    System.out.println("Added: " + line);
    notify();
}

public synchronized String remove(){
    if (bag.isEmpty()){
        try {
            wait();
        } catch (InterruptedException e) {
            e.printStackTrace();
        }
    }

    System.out.println("Removing: " + bag.peek());
    notify();
    return bag.poll();
}
```

Figure 4 Word Bag

In the search module, the main method is the sendGet() method implemented with the Apache httpClient package. The function of this method is to send a get request with the keyword to the specific search engine and return the string of the response page from the search engine.

```
public boolean isBlocked(String queryStr){

    try {
        Matcher matcher;
        switch (engine){
            case QIHU:
                matcher = BlockedPattern_QIHU.matcher(sendGet(queryStr));
                break;
            case SOGOU:
                matcher = BlockedPattern_Sogou.matcher(sendGet(queryStr));
                break;
            case BAIDU:
                matcher = BlockedPattern_Baidu.matcher(sendGet(queryStr));
                break;
            default:
                matcher = BlockedPattern_Baidu.matcher(sendGet(queryStr));
                break;
        }
        String result = matcher.find()? matcher.group(1): "";
        return result.equals(BLOCKED_KEYWORD);

    } catch (IOException e) {
        e.printStackTrace();
```

Figure 5 Is Blocked Method

The isBlocked() method in Figure 4 takes the response page string and checks if the blocking keyword is inside the string. We used regular expressions to match the target keyword that show if a word is banned or not.

The translate() method in Figure 5 translates Chinese to English. The translation is accomplished

using an api provided by mymeory.com (https://mymemory.translated.net/doc/). We used the English word to perform the automatic categorization. The translation result is not ideal because a Chinese person's name and abstract concepts are very hard to translate.

```
public static String translate(String text){
    String ret = null;

    try {
        String query = URLEncoder.encode(text, "UTF-8");
        String langPair = URLEncoder.encode(srcLanguage.getLanguage()
        + "|" + dstLanguage.getLanguage(), "UTF-8");
        String url = "http://api.mymemory.translated.net/get?q="+query+"&langpair="
                +langPair;
        HttpClient hc = new DefaultHttpClient();
        HttpGet hg = new HttpGet(url);
        HttpResponse hr = hc.execute(hg);
        if (hr.getStatusLine().getStatusCode() == HttpStatus.SC_OK){
            JSONObject response = new JSONObject(EntityUtils.toString(hr.getEntity()));
            ret = response.getJSONObject("responseData").getString("translatedText");
        }
    } catch (Exception e) {
        e.printStackTrace();
    }
    return ret;
}
```

Figure 6 Translate Method

We model the word data using the Word class shown in Figure 6. This class holds all the information from the search query to each search engine.

```
@Document(collection = "word")
public class Word {
    private String word;
    private boolean baidu_ch;
    private boolean sogou_ch;
    private boolean qihu_ch;
    private String translation;
    private boolean baidu_en;
    private boolean sogou_en;
    private boolean qihu_en;
    @DateTimeFormat(iso = ISO.DATE_TIME)
    private Date createDate;

    public String getWord() { return word; }

    public void setWord(String word) { this.word = word; }

    public boolean isBaidu_ch() { return baidu_ch; }

    public void setBaidu_ch(boolean baidu_ch) { this.baidu_ch = baidu_ch;

    public boolean isSogou_ch() { return sogou_ch; }

    public void setSogou_ch(boolean sogou_ch) { this.sogou_ch = sogou_ch;

    public boolean isQihu_ch() { return qihu_ch; }

    public void setQihu_ch(boolean qihu_ch) { this.qihu_ch = qihu_ch; }

    public String getTranslation() { return translation; }
```

Figure 7 Word Model Class

4.2. Algorithm: Censor words categorization

Once the data is collected by ChCensor application the censor detection application automatically categorizes the banned words as (1) political, (2) sexual or (3) criminal. Categorization is accomplished using Word2vec which is an NLP vector space model that is used to train a text corpus into a model that contains a set of vectors. The index of a vector is a word and the elements of the word are numeric data that can represent the word. Once the words are vectorized we can compare the similarity between two vectors using cosine similarity. The cosine similarity measures the angle between the word and classes and places the word in the class with the smallest angle. Our categorization algorithm is shown in Figure 7. First, a word is retrieved from the dataset and searched on Wikipedia. The result from Wikipedia provides an article that is parsed in order to get a word list with corresponding frequencies. Second, the word retrieved from the word list is compared to each of the categorical names: political, sexual and criminal. The decision on the appropriate category is determined by calculating the cosine similarity between the retrieved word and each of the categorical names usingWord2vec. Various thresholds for the cosine similarity are tested. For example, if comparing a keyword from a list is determined to be blocked by one of the three search engines it gets categorized. If the cosine similarity is highest with the category name politics then we add the frequency of the current word to the counter of politics category, otherwise we skip the counter increment step. The same process is applied to the sexual and crime categories in parallel. If the current word is not the last word of the word list, we continue to retrieve next word from the word list and repeat the above steps, otherwise we compare the number of the three counters, and the category that has the maximum counter will be the result of the current word. After we get the result, we continue with the next word from dataset until we reach the end of the dataset.

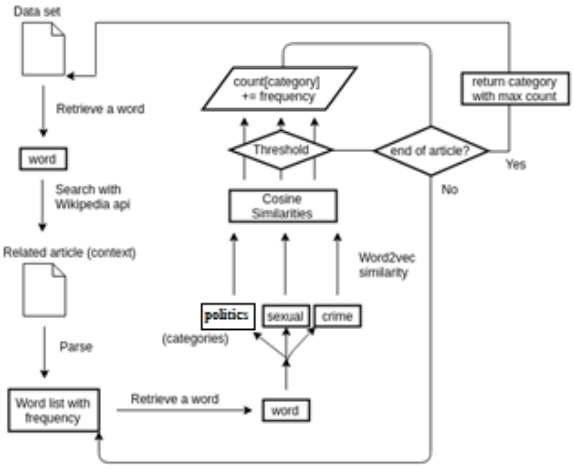

Figure 8 Algorithm

5. Experiments

Two experiments investigate the performance of the censor detection system and explore the variability of results from different search engines in mainland China. Another goal of the experiments is to examine type of content most frequently censored by classifying banned words into the most appropriate category by using word embeddings. Lastly the experiments use social analytic techniques to visualize the classification results.

6. Data

The dataset for experiment I consists of words banned by Tom-Skype and Sina UC in 2012 [12]. Tom-Skype and Sina UC are among top 10 instant messaging (IM) programs in China. The size of this dataset is 1863 words.

Experiment II applies social analytic techniques and word embeddings to categorize and visualize banned words consisting of a list of 31 words that were blacklisted in China on June 14, 2014 [11]. The categories used for classification are: sexual, political, economic and social.

7. Method

7.1. Experiment I

Experiment I uses the censor detection system to examine how different search engines block the same keyword and to determine if English translations of banned Chinese words are also blocked. Lastly, this experiment aims to classify the type of content or category associated with each blocked word. To detect banned keywords while concealing the location of the originating query, the censor detection system used a variety of IP proxy addresses in China. In this way, the queries appeared as they were being submitted within China. This provides for greater detection accuracy as queries outside of the country often return different results. Using the proxy IP address each banned word was submitted to the three search engines. It was discovered in preliminary tests, that when words are blocked the search engines respond with a text string that states the results cannot be displayed. The censor detection system used this text string to detect when a word was blocked. Figure 8 shows the banned word, "根据相关法律法规和政策，部分搜索结果未予显示。" submitted to Baidu and the following message is returned, **According to relevant laws, regulations and policies, some search results are not displayed."**

Figure 9 Baidu Blocked Keyword

7.2. Experiment II

A network analytic approach is applied to the data in experiment II to identify the top words in each category. Each node in the graph represents a word. For the experiment II dataset the political, social, economic and sexual topical categories were used [5, 6]. Next, all the words were vectorized using the open source word embedding application developed by Tomas Mikolov and colleagues [13]. The distance between each keyword and category was computed using the cosine distance similarity measure with a vocabulary of sixteen million words. Using the open source network visualization software Gephi [14] the matrix of nodes and corresponding edges was constructed to create the network.

8. Experiment I results

8.1. Platform and language

The TOM-Skype and Sina UC dataset, and their English translations are submitted as queries to three search engines (Baidu, Sogou and Qihu). According to the result in Figure 9, Baidu has highest overlap of banned words with IM clients (18%). This result suggests that censorship enforcement is not consistent across different platforms. Therefore, while information may be blocked on one system it may be available on others. This may be due to detection capabilities, enforcement policies, or the challenge of maintaining up to date banned lists. In any event this result suggests that if a circumvention tool submits queries to multiple search engines it may help to improve access to information. Another interesting result with respect to language is that on each search engine, the ratio of banned English translated words compared to the original Chinese characters is no more than 25%. The result indicates that blocking information for queries related to banned words in Chinese is greater compared to of the same word in English. That implies that people may be able to use different languages to access information. The accuracy of all results from this experiment needs to be more rigorously verified as they depend entirely on the automated detection of banned messages returned by the search engines. It may be the case that querying the systems with the English word did not trigger the blocked message and therefore the censor detection system did not flag the word as blocked. While this may be the case for the search engines tested in this experiment, the results do suggest that automated language translation may be a useful circumvention tool to enable open and free communication on different platforms such as email and SMS.

8.2. Automatic Categorization

The algorithm proposed in section 5.2 is used to determine the most likely category of content banned for the experiment I dataset. The categories are politics,

sexual, and crime categories. To assign each banned word to a category, the words were vectorized using Word2Vec and the cosine similarity was computed. The word was assigned to the word/category combination with the largest cosine similarity score. Figure 9, present results for a banned word using a threshold of 0.1, 0.3 and 0.5. The higher threshold indicates greater similarity between word and category. Regardless of the threshold value, results in Figure 9 are always greatest for the politics category. Hence this word is categorized as politics.

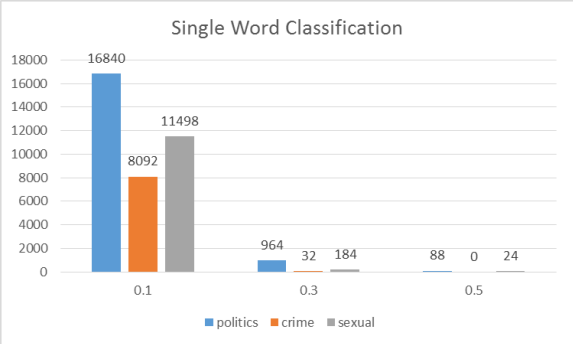

Figure 10 Single word automatic categorization

8.3. Manual Categorization

Automatic categorization did not always produce accurate results so in certain cases when the banned word was a name of a person, the word was manually classified.

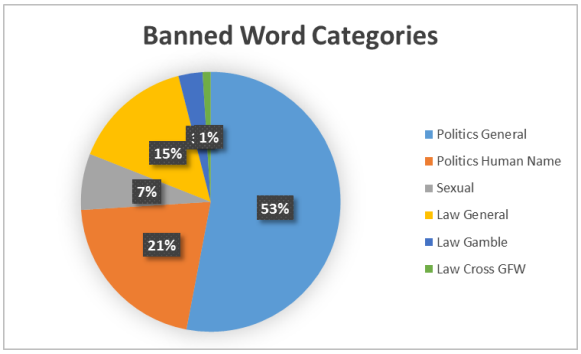

Figure 11 Banned Words Categories

Figure 10, shows the words banned by both platforms (IM clients and search engines). Results suggest that the majority of content blocked are related to politics (74%) while only 7% have a sexual connotation. This may be an artifact of the composition of the dataset and further longitudinal analysis is required.

8.4. Automatic Categorization Correctness

The accuracy of the automatic categorization and manual categorization was compared and it was determined that only 15% of automatic categorization correctly classified banned words when verified with the manual categorization result. That means our algorithm still has some problems.

Several key points may have caused poor classification performance. The context (Wikipedia article) may not produce the best dictionary for training the data, so a better search method for constructing the dictionary may yield better results. The category keywords may influence the result. For example, if "politic" instead of "politics", the result might be totally different. In this way, a group of words for each category as opposed to a single category may prove beneficial. The threshold might also influence the result. Low thresholds will not provide good discrimination quality and high thresholds may be too restrictive. More research is required to determine the optimal threshold.

9. Experiment II results

9.1. Network visualization

The overall topology for the banned words and categories is depicted in Figure 11.

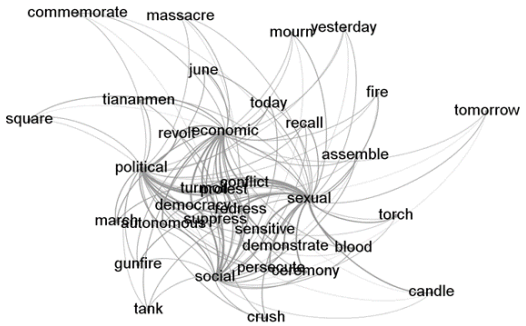

Figure 12 Keyword censorship network

Figure 12 presents a high level overview of the nodes as censored keywords and their relation, represented by curved edges, to the political, social, economic, and sexual categories. It can be observed that even with as few as 35 nodes the edges produce a great deal of noise and complexity that overshadow many details embedded in the network. However, the illustration provides a broad understanding of the diversity of words and their relative positions, but little can be discerned with respect to the importance of nodes in the network.

9.2. Node importance

To reveal node importance the weighted in-degree centrality based on the cosine similarity is computed

from vectorizing each word as part of the word embedding process. The weighted in-degree centrality is computed for each node or vertex, v, in the network as:

$$C_d = \deg(v) \quad \textbf{(1)}$$

The importance of each node as a function of the weighted in-degree is presented in Figure 12.

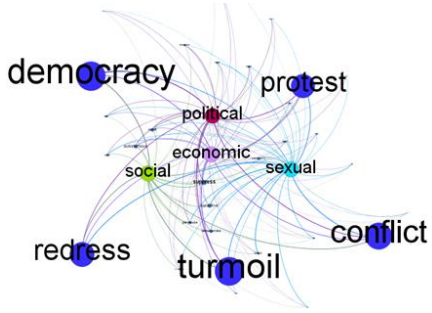

Figure 12 Most important keywords related to all categories

The blue nodes represent the importance of each word with respect to all categories. Consequently, one hypothesis is that greater emphasis and detection accuracy would be placed on these words compared to other words in the dataset.

The cosine similarity for the top 3 keywords for each category was computed to more closely examine the relation between the keywords in Figure 13 and the individual categories. Results demonstrate that the political category and the economic category share the exact same keywords with similar distance measures. However, the sexual category and political category both contained the keyword ``redress`` which was not present in the other categories. In addition, the sexual category contained the keyword ``blood`` which was not present in the top 3 most similar keywords in the other categories. The significant overlap in the social and economic categories may suggest that the ability to discriminate between these two categories may pose significant challenges compared to the sexual category. The lower discrimination quality for the social and economic categories may require further refinement into subcategories to more accurately classify keywords within these topical areas. The results for the top 3 keywords most closely related to each the social and economic category is provided in Tables 1-2.

Keyword	Cosine Similarity
democracy	1.0
turmoil	0.93
redress	0.78

Table 1. top 3 keywords closest to the social category

Keyword	Cosine Similarity
democracy	1.0
turmoil	0.99
redress	0.97

Table 2. top 3 keywords closest to the economic category

10. Conclusion

This paper presents an application for detecting censored keywords using three Chinese search engines. Results from two experiments suggest that restricting access to online information varies across different search platforms and languages. The significance of this result will help to inform the design of future circumvention tools that enable citizens to openly and freely communicate online. For example, features that provide for automatic language translation and probes that compare results from multiple search engines may help to provide increased access to information. Furthermore, while these results are limited to two small datasets they do provide insight into the nature of content most likely being blocked. In addition, it has been shown how network visualizations can be used to pinpoint relationships between different words. This technique could be used to automatically generate alternative words to construct different sentences that convey the same concept.

Overall, Chinese censorship is not same across different platforms and Chinese censorship focuses more on Chinese than English. Baidu's level of censorship is higher than Sogou and Qihu and from the datasets analyzed it appears online censorship in China may focus more on controlling access to political content.

11. Future work

Future work will focus on improving automatic text categorization by identifying more accurate search results. For example, the application could download the top 10 pages of the search results of the banned words from Google or Wikipedia. The content of all the downloaded pages could then serve as the data source of the Word2Vec training corpus. The application could use the trained model to categorize the banned words.

The percentage of the article category could be used to estimate the category of a word.

Future work will also investigate prediction techniques to determine if a word or concept will be censored. The censor detection system provides a means to probe remote systems with known banned words. This can be extended to search for words contained in news articles may be deemed controversial or have collective action potential [5]. Therefore, to discover new banned words or content, words from news articles that match certain criteria could be continually fed into the system.

Acknowledgements

The work is supported by the National Science Foundation under Grant No.: 1704113, Division of Computer and Networked Systems, Secure & Trustworthy Cyberspace (SaTC).

12. References

[1] Chen, T. M. (2011). Governments and the executive" internet kill switch"[Editor's Note]. Network, IEEE, 25(2), 2-3.

[2] Allagui, I. and Kuebler, J. (2011). The arab spring & the role of icts— introduction. International Journal of Communication, 5:8.

[3] Gurbuz, M. (2014). The Long Winter: Turkish Politics after the Corruption Scandal, volume 15. Rethink Institute

[4] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211-36.

[5] King, G., Pan, J., and Roberts, M. E. (2014). Reverse-engineering censorship in china: Randomized experimentation and participant observation. Science, 345(6199):1251722.

[6] Knockel J, Crete-Nishihata M, Ng J Q, et al. Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China[C]//5th USENIX Workshop on Free and Open Communications on the Internet (FOCI 15). 2015.

[7] Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., & Wallach, D. S. (2013). The velocity of censorship: High-fidelity detection of microblog post deletions.arXiv preprint arXiv:1303.0597.

[8] Clarke, I., Sandberg, O., Wiley, B., and Hong, T. W. (2001). Freenet: A distributed anonymous information storage and retrieval system. In Designing Privacy Enhancing Technologies, pages 46–66. Springer.

[9] Dingledine, R., Mathewson, N., and Syverson, P. (2004). Tor: The second-generation onion router. Technical report, DTIC Document.

[10] Moran, T. H. (2015). Cyber surveillance regulations: Is the United States asking china to accept a double standard?

[11] Engel, P., (2014 (accessed February 8, 2016)). 94 Search Terms That China Bans Because Of Tiananmen Square. http://www:businessinsider:com/wordschina- banned-from-search-enginesafter-tiananmen-square-2014-6.

[12] Knockel, J., Crandall, J. R., & Saia, J. (2011, August). Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance. In *FOCI*.

[13] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[14] Gephi, (2014 (accessed May 8, 2018)). Gephi Visualization Tool. http://gephi:org/.