Cyber Security Analyzing
Phase report by Picasso group

Group Member: Shimeng Lin, Shaoying Tang, Bennong Liu, Chuanze Cai

Table of Contents

Introduction

Review on previous phase

In the previous phase, the other team aimed to analyze the National CyberSecurity Strategies of nearly 70 countries in the International Telecommunications Union who have made them publicly-available. The goal of the project is to make it possible for users to view for each strategy the pre-defined and not pre-defined key topics discussed, the keywords used, the main entities (e.g. people, places, organizations, etc.) mentioned, their scope and depth, and any other useful details derived from the text.

Importance

This project is important because it can help find commonalities, differences, and key characteristics among around 70 countries' CyberSecurity Strategies. Moreover, it can help achieve "Goal 8: Promote inclusive and sustainable economic growth, employment, and decent work for all.", by promoting development-oriented CyberSecurity Strategies that support productive activities, entrepreneurship, creativity and innovation, etc.

Text Mining

Each sentence in each file has been classified and corresponding topics have been assigned to it. Sentences that obtain main category or subcategory topics have been tagged. Many sentences can't be tagged at this time because they are not related to any of pre-defined CyberSecurity topics. As a result, sentence id, sentence content, and tags have been saved in the tag_to_sentence dataset.

Methodology

Data Processing

Different data cleaning techniques have been used for pre-defined topic matching and not pre-defined topic exploration. In preparation for pre-defined topic matching, each raw text file has been broken down into sentences. Each sentence has been given a unique id number. In the process of matching, the text files have been stemmed. For not pre-defined topic exploration, stop words, unprintable characters, punctuations, digits, and white spaces have been deleted.

Expending Pre-Defined Topics

In the previous phase, we can learn that they pre-defined six categories and around four to five sub-categories for each category.

Category	Subcategory
Legal Measures	Criminal Legislation, Regulation And Compliance
Technical Measures	CIRT, Standards, Certification
Organization Measures	Policy, Roadmap For Governance, Responsible Agency, National Benchmarking
Capacity Building	Standardisation development, manpower development, professional certification, agency certification
Cooperation	Intra-State Cooperation, Intra-Agency Cooperation, Public Sector Partnership, International Cooperation
Child Online Protection	National Legislation, Un Convention And Protocol, Institutional Support, Reporting Mechanism

Table 1:Pre-Defined Topics

However, within digging deeper into the research, we can figure out that the current classification criteria are not enough to provide the accurate result and we decided to learn more about the articles and managed to gain more insights from the strategies and expand the sub-categories and the result is shown as below.

Category	Corpus
Legal Measures	Criminal Legislation, Regulation and Compliance, government action, international action, democracy, the rule of law, rule, law, institution, human rights, freedom, privacy, secure, protection, minister, safeguard, national resilience, administration, government, data protection regulation
Technical Measures	CIRT, standards, certification, indigenous innovation, information infrastructures, Slovak digital, research, electronical mechanical
Capacity Building	behavioral change, infrastructure development, fundraising, training centers, exposure visit, office and documentation support, job training, learning center, consultant, enhance, education, Standardization development, manpower development, professional certification, agency certification.
Organization measures	policy, roadmap for governance, responsible agency, national benchmarking, social development, climate change, public administration, national information

Cooperation	intra-state cooperation, intra-agency cooperation, public sector partnership, international cooperation, business partnership, private sector partnership, government partnership, education partnership, GCHQ and business partnership, organization partnership, department partnership, national infrastructure partnership, cyber security hub, funding partners, national cybercrime capability, partnership between public and private sectors, partnership between government and private sector, government and industry partnership
Child online protection	national legislation, UN convention and protocol, Institutional support, reporting mechanism, UK council for child internet safety, National crime agency, shaping mainstream law, training mainstream law, government interface, police force, NCA, resilience to cyber-attacks, government support, education, technical capabilities, public service,

Table 2: Expended Pre-Defined Categories

Comparison

In the former presenting method, it is obvious to find out that the result of distribution is too general, which means the category was too vague and ambiguous to provide enough useful information that can be used in the following research. Hence after upgrading the database of sub-categories, we re-built the model and the code can be show in the image shown below.

```
# Author: Shaoying Tang

import os
os.chdir('E:\\UNProject\\tagged_excel2')
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
def tagged_dist(country_name):
    for i in os.listdir("E:\\UNProject\\tagged_excel"):
        if i.endswith('.csv') and i.startswith(country_name) and len(i) == len(country_name)+len('.csv'):
            df = pd.read_csv('{}.csv'.format(country_name))
            sns.factorplot(x='Subcategory',data=df, kind='count').add_legend()
            sns.plt.title(country_name)
            break

    else:
        print 'No file found !'
```

Figure 1: The realization with algorithm

Based on the result, it is obvious that the result is now more specific and detailed. In this case, we set the country name “Egypt” as an example and from the figure we can tell straight that the upgraded version is more detailed and comprehension friendly.

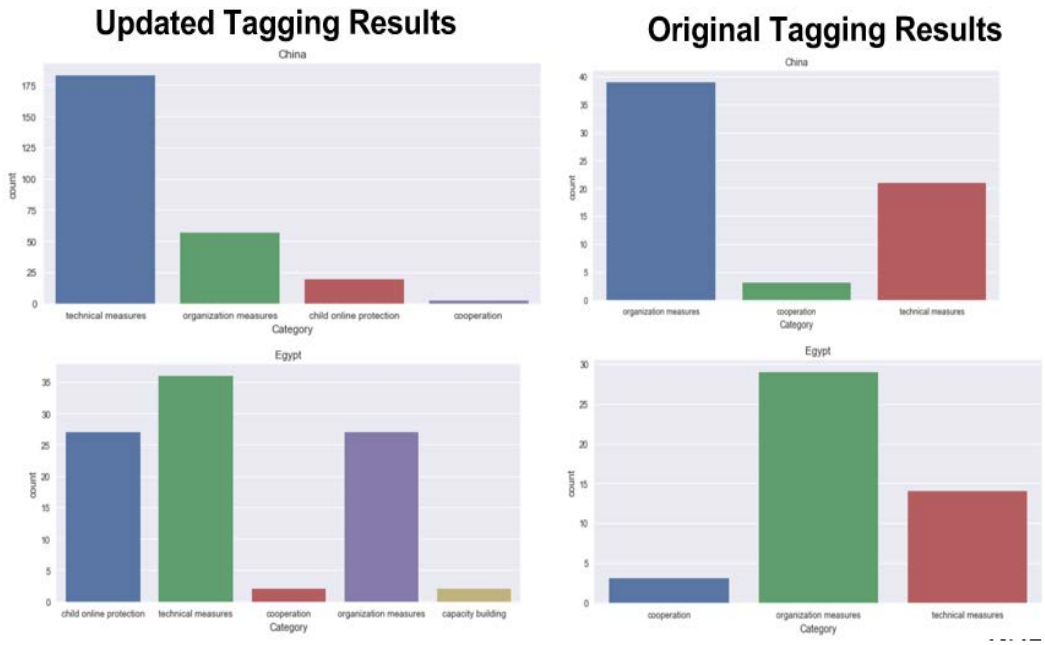

Figure 2: Comparison between two different criteria.

Accuracy

Around **83.3%** from the sample of 400 sentences

Problems

Sentences with keywords of mixed categories.

e.g. The government cooperates with a technical organization to implement a strategy…...

Organization Measure or Technical Measure or Cooperate ?

Future Study

From the above conclusions, we could observe that the results of distribution were only for informative purpose although we improved the results from Phase I.

According to the tagging results, the categories/keywords may not be accurate enough to provide solid evidence to support the next research stage.

Our future plan includes updating the model with more restrictive categorizing/clustering criteria to optimize the results. We will also calculate the tentative accuracy rates in terms of measurement. D3 templates will also be used to visualize the result.

It is also should be considered to change the method from classification to clustering and from supervised to unsupervised.

In the field of text mining, detailed information of each sentence need to be mined, including the main entities (e.g. people, places, organizations, etc.) mentioned, the scope and depth of the topic, and any other useful characteristics.

When it comes to unsupervised learning, LDA model or other models can be built to accurately extract new topics from each document. And the results can be used to improve the quality of classification.

Lastly, a user-friendly web page is needed to visualize the results of the whole project. Open source tools or platforms can be reached out.