Gabelli School of Business

Fordham University

Machine-based Text Analytics of CyberSecurity Strategies-Phase 1 Report

Fordham Graduating Team

Writer: Liyi Li

Date: 12/12/2016

Phase1-Effort

1. Download and create a collection of text version of 73 cybersecurity strategy documents
2. Download a collection of Cyberwellness profiles, and confirm the categorical tags
3. Break down each cybersecurity strategy document into sentences
4. Classify each sentence and assign it a pre-defined tag.
5. Identify topics that are not pre-defined by Latent Dirichlet allocation (LDA) model

(Datasets and results are available in GitHub:

https://github.com/ICT4SD/CyberSecurity_Strategies/tree/Fordham_Designlab)

Report

1 Introduction

1.1 Objective

This project aims to analyze the National CyberSecurity Strategies of nearly 70 countries in the International Telecommunications Union who have made them publicly-available. This project will result on a web application available for the public enabling search and visualization of the CyberSecurity strategies. The web application will allow the user 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.

1.2 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.

2 Methodology

2.1 Data Collection

Data used in this project are sourced from the International Telecommunications Union. (http://www.itu.int/en/ITU-D/Cybersecurity/Pages/National-Strategies-repository.aspx) 65 countries' CyberSecurity Strategies are available. In the phase 1, 61 of them, which are English version, have been analyzed. The raw datasets are comprised of one text file per CyberSecurity Strategy file of each country.

2.2 Data Preprocessing

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. And a new dataset ('passage_to_sentence_json') has been generated to save all these files. 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.

Also, based on the results of human analysis files, pre-defined topics have been used to generate a topic dictionary with both category and subcategory.

Table 1. Pre-defined Topics

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

2.3 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.

2.4 Unsupervised Topic Extraction

Latent Dirichlet allocation (LDA) model has been used to extract unsupervised topics from CyberSecurity Strategy documents, with the help of Cyberwellness Profiles. LDA is a topic model that generates topics based on word frequency from a set of documents. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set. Cyberwellness profiles can be used to build a bigram model and a dictionary to make precise pattern recognition, thus supporting LDA model building.

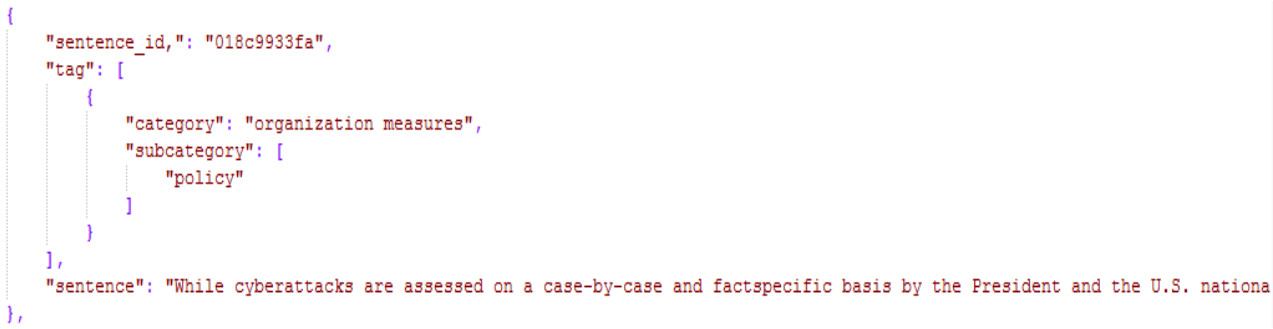
3 Results

In the phase 1 of Machine-based Text Analytics of CyberSecurity Strategies Project, document parsing has been done. All the available English CyberSecurity Strategy documents have been downloaded, converted into text file, broke down into sentences, and indexed in sentence level. Each sentence in each file has been tagged by pre-defined topics.

Figure 1. Example of Sentence Indexing Results


```
{
    "sentence": "In the past few years, threats in cyberspace have risen dramatically.",
    "sentence_id,": "5f2ba69aa3"
},
```

Figure 2. Example of Tag_to_Sentence Results



```
{
    "sentence_id,": "018c9933fa",
    "tag": [
        {
            "category": "organization measures",
            "subcategory": [
                "policy"
            ]
        }
    ],
    "sentence": "While cyberattacks are assessed on a case-by-case and factspecific basis by the President and the U.S. nationa
},
```


And 6 not pre-defined groups of topics with probabilities have been extracted by the LDA model. Group1 can be interpreted as network threat topic, which consists of network, policy, threat, business, access, digital, attack, etc. Group2 can be regarded as technology infrastructure, including technology, infrastructure, business, network, capability, and implementation. Similarly, group 3 to group 6 can be identified as public threat, digital threat, attack, and business sector internet threat, consisting of multiple subtopics. The results are shown in the followings:

Group1:

'0.008*"network" + 0.007*"policy" + 0.007*"threat" + 0.006*"business" + 0.006*"access" + 0.005*"organisation" + 0.005*"management" + 0.005*"digital" + 0.005*"level" + 0.005*"action" + 0.005*"attack" + 0.005*"infrastructure" + 0.005*"system" + 0.005*"cooperation" + 0.005*"ensure"'

Group2:

'0.011*"policy" + 0.008*"technology" + 0.006*"infrastructure" + 0.006*"business" + 0.006*"network" + 0.005*"sector" + 0.005*"internet" + 0.005*"ensure" + 0.005*"capability" + 0.005*"action" + 0.005*"implementation" + 0.005*"develop" + 0.005*"threat" + 0.005*"public" + 0.005*"objective"'

Topic3:

'0.009*"threat" + 0.008*"public" + 0.007*"sector" + 0.007*"technology" + 0.006*"network" + 0.006*"infrastructure" + 0.006*"system" + 0.006*"digital" + 0.005*"attack" + 0.005*"business" + 0.005*"policy" + 0.005*"state" + 0.005*"ensure" + 0.005*"level" + 0.005*"activity"'

Topic4:

'0.009*"technology" + 0.008*"digital" + 0.008*"threat" + 0.007*"policy" + 0.005*"capability" + 0.005*"sector" + 0.005*"state" + 0.005*"system" + 0.005*"network" + 0.005*"level" + 0.005*"business" + 0.005*"infrastructure" + 0.005*"develop" + 0.005*"support" + 0.004*"protection"'

Topic5:

'0.010*"threat" + 0.007*"attack" + 0.007*"policy" + 0.006*"technology" + 0.006*"infrastructure" + 0.006*"action" + 0.005*"level" + 0.005*"public" + 0.005*"business" + 0.005*"network" + 0.005*"internet" + 0.005*"state" + 0.005*"digital" + 0.005*"measure" + 0.005*"support"'

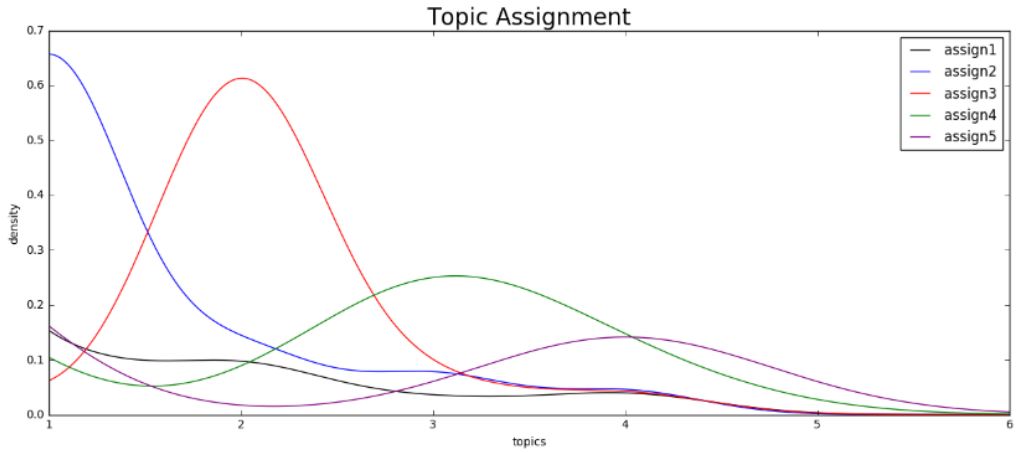
Topic6:

'0.009*"threat" + 0.009*"policy" + 0.006*"internet" + 0.006*"business" + 0.006*"sector" + 0.006*"technology" + 0.006*"digital" + 0.006*"infrastructure" + 0.005*"access" + 0.005*"attack" + 0.005*"network" + 0.005*"nacional" + 0.005*"capability" + 0.005*"system" + 0.005*"ensure"'

Additionally, the results of LDA model have been evaluated by topic assignment distribution. Assignment1 represents it is most accurate to assign documents to these 6 groups of topics. Assignment5 represent it is most inaccurate to assign documents to topics. Thus, these

two deserve more attention. Compared to the distributions of swing Assignment2, 3 and 4, Assignement1 and 5 are more likely to be uniform distribution, thus supporting the results of LDA model.

Figure 3. Distributions of Topic Assignment



4 Further Study

In the following phases, efforts from text mining, unsupervised learning, and visualization of results can be expected.

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.

Appendix

Team Members' Contribution

Section	Contributing Team Members
Download and create a collection of text version of 73 cybersecurity strategy documents	Jiachen Chen, Xingying Huang, Yuqi Jin, Nan Wang, Xianghua Su, Tianyang Zhang, Liyi Li
Download a collection of Cyberwellness profiles, and confirm the categorical tags	Jiachen Chen, Xingying Huang, Yuqi Jin, Nan Wang, Xianghua Su, Tianyang Zhang, Liyi Li
Text mining	Tianyang Zhang, Liyi Li
Unsupervised learning models	Nan Wang
Report writing	Liyi Li