Gabelli School of

Business Fordham

University

Machine-based Text Analytics of CyberSecurity Strategies-Progress Report

Group: HANDEL

Date:2/3/2017

Group Member:

Youwei Xiao

Chuanjian Deng

Nianting Ouyang

Haixuan Zhu

Xiaoyu Wang

Report

Phasel-Effort

1. Set up the training dataset by doing human analysis to tag the sentences from the USA cybersecurity strategies.
2. Coding the core algorithm of machine-based test analytics to tag the testing dataset by using Python.
3. Convert all dataset into JSON version and wrote basic code to visualize the result.
4. Identify topics are not pre-defined through unsupervised machine learning models (LDA).

Methodology

1. Set up the training dataset by human analysis


```
[
  {
    "sentence": "INTERNATIONAL COOPERATION The Department of State will lead federal efforts to enhance international cyberspace security cooperation.",
    "tag": [
      {
        "category": "cooperation",
        "subcategory": [
          "international cooperation"
        ]
      }
    ],
    "keywords": [
      {
        "international cooperation": [
          "international cyberspace security cooperation"
        ]
      }
    ],
    "country": "united states",
    "sentence_id,": "1f61a8c912"
  },
  {
    "sentence": "Work through International Organizations and with Industry to Facilitate and to Promote a Global Culture of Security Americas interest in promoting global c
    "tag": [
      {
        "category": "cooperation",
        "subcategory": [
          "intra-state cooperation"
        ]
      }
    ],
    "keywords": [
      {
        "INTERNATIONAL COOPERATION": [
          "International Organizations","Industry"
        ]
      }
    ],
    "country": "united states",
    "sentence_id,": "e8b0070918"
  },
  {
    "sentence": "Our 50 T H E N A T I O N A L S T R A T E G Y T O S E C U R E C Y B E R S P A C E P R I O R I T Y information infrastructure is directly linked with Canada,
    "tag": [
      {
        "category": "cooperation",
        "subcategory": [
          "intra-agency cooperation","international cooperation"
        ]
      }
    ]
```

2. Collect all keywords we used to set up a package which helps to identify the testing dataset

```
[{
    "category": "legal measures",
    "subcategory": ["criminal legislation", "regulation and compliance"],
    "keywords": [["cyber options", "cyber plan", "cybersecurity", "confidentiality", "integrity of data", "security", "steal data", "destroying data", "disrupting businesses

    ["cyber options", "cyber plan", "cyber options and plan", "security", "cybersecurity", "confidentiality", "integrity of data", "security", "send data", "steal data", "de
},
{
    "category": "technical measures",
    "subcategory": ["cirt", "standards", "certification"],
    "keywords": [["ICS", "control system"], ["Computer code", "industrial control systems", "control system", "ICS", "technical", "security standards", "standards", "OECDs",

    ["an open, secure", "interoperable", "and reliable Internet", "the Internet", "Internet-related", "Internet", "Internet", "the Internets", "Internet", "the Internet", "c
},
{
    "category": "organization measures",
    "subcategory": ["policy", "roadmap for governance", "responsible agency", "national benchmarking"],
    "keywords": [["cyberspace policy", "cyberspace national", "cyberspace security", "cyber options", "u s", "cybersecurity policy", "cybersecurity policy official", "policy
cybersecurity-related policies", "of policies", "foreseeable future national strategy", "system national", "actions and recommendations a r", "program national", "program na

    ["cyber systems", "responsibility", "Federal government", "policy objectives", "missions and activities", "trustworthy systems", "roadmap", "organizational construct", "

    ["the department of", "the department of defense", "the white house", "of homeland security", "department of", "department of defense", "department of homeland security"

    ["an open, secure", "interoperable", "and reliable Internet", "the Internet", "social and economic value", "billions of people", "Internet-related", "Within the U.S", "L
},
{
    "category": "capacity building",
    "subcategory": ["standardisation development", "manpower development", "professional certification", "agency certification"],
    "keywords": [["Research, cybersecurity vulnerabilities", "information technology", "comprehensive framework", "this framework", "benefits of technology", "interdependent
sses", "formulation", "develop tools", "test theories", "clarity", "accountability", "defense", "critical infrastructures", "defense", "critical infrastructures", "extent pr

    ["power", "investments", "data systems", "defense investments", "investments", "private sector", "application developers", "developers", "Providers", "Internet Services
els", "Competition", "cybersecurity expertise", "training and rotational assignments", "cybersecurity awareness", "information exchange"], ["certification programs", "certif
},
{
    "category": "cooperation",
    "subcategory": ["intra-state cooperation", "intra-agency cooperation", "public sector partnership", "international cooperation"],
    "keywords": [["crosses borders", "North American", "Canada", "Mexico", "society", "federal government", "coordinated", "focused effort", "society", "federal government",
information", "across agencies", "civil liberties", "threat and risks", "public and private sectors", "reliable infrastructure", "sensitive or proprietary business informati

    ["private sector", "Transatlantic Business Dialogue", "private hands", "public and private sectors", "public-private partnerships", "unique partnership", "DHS", "public
zation", "public and private institutions", "individual enterprises", "infrastructure sectors alone", "voluntary efforts", "active partnership", "unprecedented partnership",
"service providers", "public-private partnership", "organizational redundancies", "policy gaps", "industry partners", "collaboration", "industry groups", "venues", "communit
st of the world", "international effort", "international venues", "international activities", "international allies", "international organizations", "ITU", "ISO", "internati
},
{
    "category": "child online protection",
    "subcategory": ["national legislation", "un convention and protocol", "institutional support", "reporting mechanism"],
    "keywords": [["cyberspace"], [], [], ["defense response", "response"]]
```

3. **Through the core algorithm, we can tag any sentence in the rest dataset.**

```
# -*- coding:utf-8 -*-
_author_ = "Youwei Xiao"

import nltk
import json
import sys
import glob

#load the data and set up the Porter
porter = nltk.PorterStemmer()
keywords = json.load(open("keywords-new-1.json"))
print keywords

# Transfer the keywords into stem
keywords_stem = list()
for i in keywords:
    temp = i.copy()
    for k in range(len(temp["keywords"])):
        temp['keywords'][k] = [' '.join([porter.stem(w) for w
```

```
in s.split()]) for s in temp['keywords'][k]]
    keywords_stem.append(temp)
print keywords_stem
# for t in keywords_stem:
#     print t
keywords = json.load(open("keywords-new-1.json"))
print keywords

# set up the function of tag
def tag(sentence):
    label = list()
    # sig_label = {"category": [], "subcategory": [],
"keywords": []}
    sig_label = {"category": [], "subcategory": [],
"keywords": []}
    for i in range(len(keywords_stem)):
        for j in range(len(keywords_stem[i]["keywords"])):
            for z in
range(len(keywords_stem[i]["keywords"][j])):
                if keywords_stem[i]['keywords'][j][z] in
sentence:
                    k_1 = keywords[i]["subcategory"][j]
                    k_2 = keywords[i]["keywords"][j][z]
                    k = {k_1 : k_2}
                    sig_label["keywords"].append(k)
                    if not keywords[i]["category"] in
sig_label["category"]:

sig_label["category"].append(keywords[i]["category"])
                    if not keywords[i]["subcategory"][j] in
sig_label["subcategory"]:

sig_label["subcategory"].append(keywords[i]["subcategory"][j])
    if sig_label["keywords"]:
        label.append(sig_label)
    else:
        sig_label2 = {"category": [], "subcategory": [],
"keywords": []}
        label.append(sig_label2)
    return label
#
# json_files = glob.glob("C:\Users\yxiao\Desktop\Cyber
Security\USA\*.txt")
# for name in json_files:
test_passage = json.load(open("US-1.json"))
#passage = open(name,"r").read()
result = list()
```

```
for line in test_passage:
    sentence_stem = " ".join(porter.stem(w)for w in
line["sentence"].lower().split())
    # print sentence_stem
    label = tag(sentence_stem)
    if label:
        line['tag'] = label
        result.append(line)


print json.dumps(result, indent=4)
#data = json.dumps(result, indent = 4)
output_name = "US-1-tag.json"
with open(output_name, "w") as f:
    json.dump(result, f, indent= 4)
```

4. Example of result


```
{
    "sentence_id,": "5f2ba69aa3",
    "tag": [
        {
            "category": [
                "cooperation",
                "child online protection"
            ],
            "keywords": [
                {
                    "international cooperation": "cyberspace"
                },
                {
                    "national legislation": "cyberspace"
                }
            ],
            "subcategory": [
                "international cooperation",
                "national legislation"
            ]
        }
    ],
    "sentence": "In the past few years, threats in cyberspace have risen dramatically."
},
```

Further Study

1. To reduce the bias of samples, we decided to select sentences from the whole dataset randomly (Previously, we only used the sentences from the USA.)
2. Set up an improved training dataset by using the sentences selected randomly.
3. Trying to use more unsupervised models to identify the un-defined topics and compare the results from different models.
4. Finished the code of result visualization by using HTML, Javascript and jQuery.

Team Members' Contribution

Section	Contributing Team Members
Human analysis to set up the training dataset	Youwei Xiao, Chuanjian Deng, Nianting Ouyang, Haixuan Zhu, Xiaoyu Wang
Core algorithm	Youwei Xiao
Convert Data into JSON file.	Youwei Xiao, Chuanjian Deng, Nianting Ouyang, Haixuan Zhu, Xiaoyu Wang
Doing test mining to testing dataset	Youwei Xiao, Chuanjian Deng, Nianting Ouyang, Haixuan Zhu, Xiaoyu Wang
Data visualization and search engine	Chuanjian Deng, Xiaoyu Wang, Youwei Xiao
Report writing	Youwei Xiao