# Preparation

## Import libraries needed.

In [1]:

```
import pandas as pd
import re
import xlrd # to read excel
import nltk # NLP toolkit
import matplotlib.pyplot as plt # for visualization
import string # for handling string
import seaborn as sns
import spacy
```

## Access the location of origin file by declaring the path

In [2]:

```
# give the location
data_path = '/Users/cathzzr2/Desktop/dataset/'
data_filename = 'dataset.xlsx'
# default path to the origin_file
summary = pd.read_excel(data_path + data_filename,header=0) # (default: 0)
```

## 1.1 Basic Properties - size, head, tail, etc.

In [3]:

```
print(summary.shape) # size of the excel: (row, colume)
print(summary.columns) # index of columns
print(summary.head(3)) # preview top 3 rows (default: 5)
print(summary.tail(3)) # preview top 3 rows (default: 5)
print(summary.describe) # only integer variables are shown
print(summary.nunique)
print(summary['Authors'].unique())
```

```
(1712, 8)
Index(['Title', 'Authors', 'Date', 'UID', 'Summary', 'PDF URL', 'Cyber_Risk',
       'Not_Cyber'],
      dtype='object')
                                              Title  \
0                  Cyber security and the Leviathan
1  Evaluation of Machine Learning Algorithms in N...
2  Getting Critical: Making Sense of the EU Cyber...

                                 Authors        Date           UID  \
0                   ['Joseph Da Silva'] 2022-03-10  2203.05256v1
1   ['Tuan-Hong Chua', 'Iftekhar Salam'] 2022-03-10  2203.05232v1
2  ['Ian Walden', 'Johan David Michels'] 2022-03-09  2203.04887v1

                                            Summary  \
0  Dedicated cyber-security functions are common ...
1  Cybersecurity has become one of the focuses of...
2  In this chapter, we review how the EU cybersec...

                                 PDF URL  Cyber_Risk  Not_Cyber
0  http://arxiv.org/pdf/2203.05256v1.pdf         1.0        0.0
1  http://arxiv.org/pdf/2203.05232v1.pdf         1.0        0.0
2  http://arxiv.org/pdf/2203.04887v1.pdf         1.0        0.0
```

```
                                                 Title  \
1709           Data Security Equals Graph Connectivity   
1710                Optimal Encryption of Quantum Bits   
1711  From quantum-codemaking to quantum code-breaking   

                                        Authors       Date        UID  \
1709                          ['Ming-Yang Kao'] 2001-01-27  0101034v1   
1710  ['P. Oscar Boykin', 'Vwani Roychowdhury'] 2000-03-16  0003059v2   
1711                            ['Artur Ekert'] 1997-03-19  9703035v1   

                                                 Summary  \
1709  To protect sensitive information in a cross ta...   
1710  We characterize the complete set of protocols ...   
1711  This is a semi-popular overview of quantum ent...   

                                          PDF URL  Cyber_Risk  Not_Cyber  
1709        http://arxiv.org/pdf/cs/0101034v1.pdf         NaN        NaN  
1710  http://arxiv.org/pdf/quant-ph/0003059v2.pdf         NaN        NaN  
1711  http://arxiv.org/pdf/quant-ph/9703035v1.pdf         NaN        NaN  
<bound method NDFrame.describe of 
Title  \
0                      Cyber security and the Leviathan   
1     Evaluation of Machine Learning Algorithms in N...   
2     Getting Critical: Making Sense of the EU Cyber...   
3     Adaptative Perturbation Patterns: Realistic Ad...   
4     Guidelines for cyber risk management in shipbo...   
...                                                 ...   
1707          Least Effort Strategies for Cybersecurity   
1708  On ASGS framework: general requirements and an...   
1709            Data Security Equals Graph Connectivity   
1710                 Optimal Encryption of Quantum Bits   
1711   From quantum-codemaking to quantum code-breaking   

                                                Authors       Date  \
0                                   ['Joseph Da Silva'] 2022-03-10   
1                  ['Tuan-Hong Chua', 'Iftekhar Salam'] 2022-03-10   
2                 ['Ian Walden', 'Johan David Michels'] 2022-03-09   
3     ['João Vitorino', 'Nuno Oliveira', 'Isabel Pra... 2022-03-08   
4     ['Priyanga Rajaram', 'Mark Goh', 'Jianying Zhou'] 2022-03-08   
...                                                 ...        ...   
1707  ['Sean P. Gorman', 'Rajendra G. Kulkarni', 'La... 2003-05-30   
1708             ['Kamil Kulesza', 'Zbigniew Kotulski'] 2002-11-18   
1709                                  ['Ming-Yang Kao'] 2001-01-27   
1710          ['P. Oscar Boykin', 'Vwani Roychowdhury'] 2000-03-16   
1711                                    ['Artur Ekert'] 1997-03-19   

               UID                                            Summary  \
0     2203.05256v1  Dedicated cyber-security functions are common ...   
1     2203.05232v1  Cybersecurity has become one of the focuses of...   
2     2203.04887v1  In this chapter, we review how the EU cybersec...   
3     2203.04234v1  Adversarial attacks pose a major threat to mac...   
4     2203.04072v2  Over the past few years, we have seen several ...   
...            ...                                                ...   
1707     0306002v3  Cybersecurity is an issue of increasing concer...   
1708     0211269v2  In the paper we propose general framework for ...   
1709     0101034v1  To protect sensitive information in a cross ta...   
1710     0003059v2  We characterize the complete set of protocols ...   
1711     9703035v1  This is a semi-popular overview of quantum ent...   

                                          PDF URL  Cyber_Risk  Not_Cyber  
```

```
0           http://arxiv.org/pdf/2203.05256v1.pdf         1.0        0.0
1           http://arxiv.org/pdf/2203.05232v1.pdf         1.0        0.0
2           http://arxiv.org/pdf/2203.04887v1.pdf         1.0        0.0
3           http://arxiv.org/pdf/2203.04234v1.pdf         NaN        NaN
4           http://arxiv.org/pdf/2203.04072v2.pdf         NaN        NaN
...                                           ...         ...        ...
1707  http://arxiv.org/pdf/cond-mat/0306002v3.pdf         NaN        NaN
1708      http://arxiv.org/pdf/math/0211269v2.pdf         NaN        NaN
1709        http://arxiv.org/pdf/cs/0101034v1.pdf         NaN        NaN
1710  http://arxiv.org/pdf/quant-ph/0003059v2.pdf         NaN        NaN
1711  http://arxiv.org/pdf/quant-ph/9703035v1.pdf         NaN        NaN

[1712 rows x 8 columns]>
<bound method DataFrame.nunique of 
Title  \
0                      Cyber security and the Leviathan   
1     Evaluation of Machine Learning Algorithms in N...   
2     Getting Critical: Making Sense of the EU Cyber...   
3     Adaptative Perturbation Patterns: Realistic Ad...   
4     Guidelines for cyber risk management in shipbo...   
...                                                 ...   
1707          Least Effort Strategies for Cybersecurity   
1708  On ASGS framework: general requirements and an...   
1709            Data Security Equals Graph Connectivity   
1710                 Optimal Encryption of Quantum Bits   
1711   From quantum-codemaking to quantum code-breaking   

                                                Authors       Date  \
0                                  ['Joseph Da Silva'] 2022-03-10   
1                  ['Tuan-Hong Chua', 'Iftekhar Salam'] 2022-03-10   
2                 ['Ian Walden', 'Johan David Michels'] 2022-03-09   
3     ['João Vitorino', 'Nuno Oliveira', 'Isabel Pra... 2022-03-08   
4     ['Priyanga Rajaram', 'Mark Goh', 'Jianying Zhou'] 2022-03-08   
...                                                 ...        ...   
1707  ['Sean P. Gorman', 'Rajendra G. Kulkarni', 'La... 2003-05-30   
1708             ['Kamil Kulesza', 'Zbigniew Kotulski'] 2002-11-18   
1709                                 ['Ming-Yang Kao'] 2001-01-27   
1710          ['P. Oscar Boykin', 'Vwani Roychowdhury'] 2000-03-16   
1711                                   ['Artur Ekert'] 1997-03-19   

            UID                                            Summary  \
0     2203.05256v1  Dedicated cyber-security functions are common ...   
1     2203.05232v1  Cybersecurity has become one of the focuses of...   
2     2203.04887v1  In this chapter, we review how the EU cybersec...   
3     2203.04234v1  Adversarial attacks pose a major threat to mac...   
4     2203.04072v2  Over the past few years, we have seen several ...   
...            ...                                                ...   
1707     0306002v3  Cybersecurity is an issue of increasing concer...   
1708     0211269v2  In the paper we propose general framework for ...   
1709     0101034v1  To protect sensitive information in a cross ta...   
1710     0003059v2  We characterize the complete set of protocols ...   
1711     9703035v1  This is a semi-popular overview of quantum ent...   

                                          PDF URL  Cyber_Risk  Not_Cyber  
0           http://arxiv.org/pdf/2203.05256v1.pdf         1.0        0.0  
1           http://arxiv.org/pdf/2203.05232v1.pdf         1.0        0.0  
2           http://arxiv.org/pdf/2203.04887v1.pdf         1.0        0.0  
3           http://arxiv.org/pdf/2203.04234v1.pdf         NaN        NaN  
4           http://arxiv.org/pdf/2203.04072v2.pdf         NaN        NaN  
...                                           ...         ...        ...  
```

```
1707  http://arxiv.org/pdf/cond-mat/0306002v3.pdf         NaN         NaN
1708      http://arxiv.org/pdf/math/0211269v2.pdf         NaN         NaN
1709        http://arxiv.org/pdf/cs/0101034v1.pdf         NaN         NaN
1710  http://arxiv.org/pdf/quant-ph/0003059v2.pdf         NaN         NaN
1711  http://arxiv.org/pdf/quant-ph/9703035v1.pdf         NaN         NaN

[1712 rows x 8 columns]>
["['Joseph Da Silva']" "['Tuan-Hong Chua', 'Iftekhar Salam']"
 "['Ian Walden', 'Johan David Michels']" ... "['Ming-Yang Kao']"
 "['P. Oscar Boykin', 'Vwani Roychowdhury']" "['Artur Ekert']"]
```

## 2.1 Cleaning the Data - remove null cells, convert to lowercases, remove punctuations

In [4]:

```
# number of null cells in each columns
print(summary.isnull().sum())
# drop unnecessary index and store it to a new file
new = summary.drop(['Date', 'UID', 'PDF URL', 'Cyber_Risk', 'Not_Cyber'], axis=1
# Lowercase the summary
new['summary_in_lowercase']=new['Summary'].apply(lambda x: x.lower())
new['summary_without_punkt']=new['summary_in_lowercase'].apply(lambda x: re.sub(
print(new.head(1))
```

```
Title             0
Authors           0
Date              0
UID               0
Summary           0
PDF URL           0
Cyber_Risk     1709
Not_Cyber      1709
dtype: int64
                              Title              Authors  \
0  Cyber security and the Leviathan  ['Joseph Da Silva']

                                             Summary  \
0  Dedicated cyber-security functions are common ...

                                summary_in_lowercase  \
0  dedicated cyber-security functions are common ...

                               summary_without_punkt
0  dedicated cybersecurity functions are common i...
```

## 2.2 Cleaning the Data - download stopwords

In [5]:

```
from nltk.corpus import stopwords
stoplist = set(stopwords.words("english"))
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /Users/cathzzr2/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[5]:

```
True
```

# Exploratory Data Analysis

## 3.1 Tokenization

In [6]:

```
from nltk.tokenize import word_tokenize

# tokenization
def preprocess(text):
  formatted_text = text.lower()
  tokens = []
  for token in nltk.word_tokenize(formatted_text):
    tokens.append(token)
  tokens = [word for word in tokens if word not in stoplist and word not in stri
  formatted_text1 = ' '.join(element for element in tokens)
  formatted_text2 = ''.join([i for i in formatted_text1 if not i.isdigit()])
  return formatted_text2

# remove stop words
new['summary_without_stopw']=new['summary_without_punkt'].apply(lambda x: prepro

# remove numbers

new['summary_without_num']=new['summary_without_stopw'].apply(lambda x: x.replac

# Loading model
nlp = spacy.load('en_core_web_sm',disable=['parser', 'ner'])

# Lemmatization with stopwords removal
new['lemmatized']=new['summary_without_num'].apply(lambda x: ' '.join([token.lem
```

## 4.1 Word Frequency - Document Term Matrix

In [7]:

```
# Creating Document Term Matrix
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(analyzer='word')
data=cv.fit_transform(new['lemmatized'])
new_dtm = pd.DataFrame(data.toarray(), columns=cv.get_feature_names())
new_dtm.index=new.index
new_dtm.head(10) # customize rows
```

Out[7]:

| | aa | aaai | aacn | aad | aadl | ab | abandon | abandonment | abatement | abc | ... | zkp | zksnark | zo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 2 | |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | |

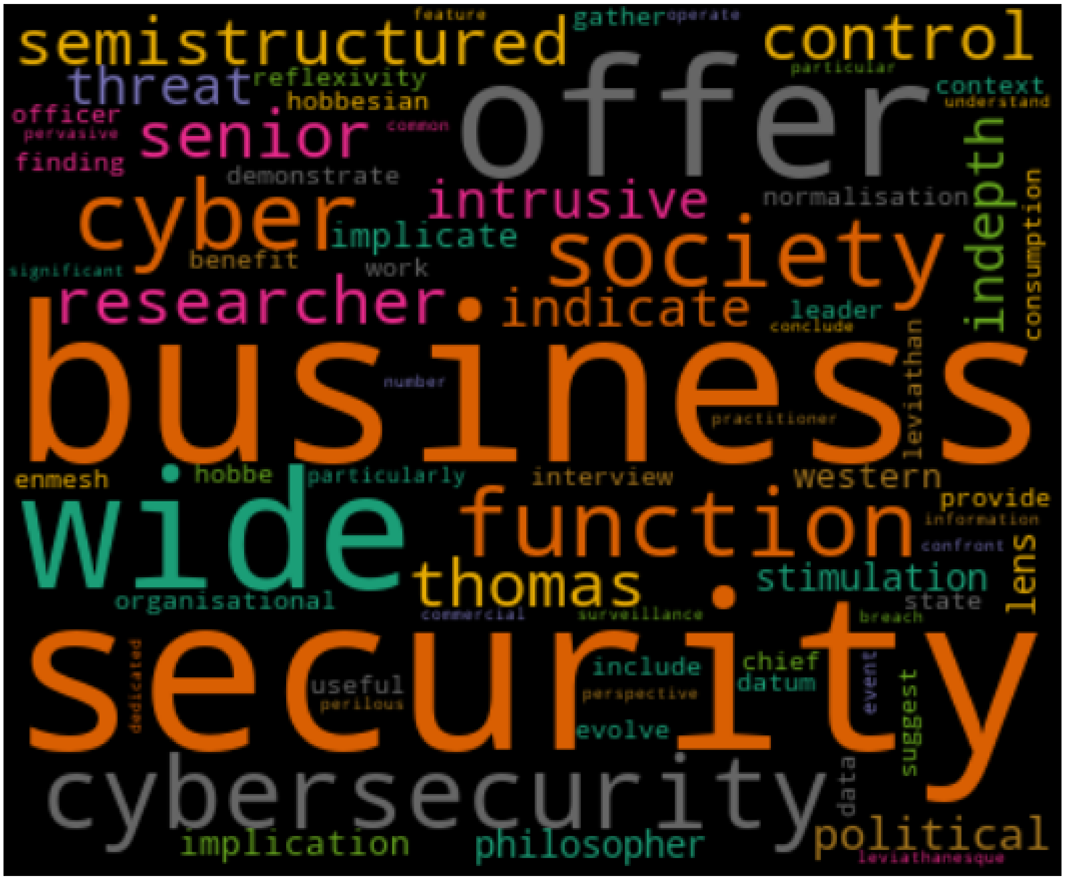
10 rows × 12283 columns

### 4.2.1 Word Frequency - Visualization - Word Cloud

In [16]:

```
from wordcloud import WordCloud
from textwrap import wrap
# Function for generating word clouds
def generate_wordcloud(data,title):
  wc = WordCloud(width=400, height=330, max_words=150,colormap="Dark2").generate
  plt.figure(figsize=(10,8))
  plt.imshow(wc, interpolation='bilinear')
  plt.axis("off")
  plt.title('\n'.join(wrap(title,60)),fontsize=13)
  plt.show()

# Transposing document term matrix
new_dtm=new_dtm.transpose()

# Plotting word cloud for each product
for index,product in enumerate(new_dtm.columns):
  generate_wordcloud(new_dtm[product].sort_values(ascending=False),product)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
/var/folders/7b/nnsn4tm15ns7j6qqstdz1src0000gn/T/ipykernel_82252/1865818958.py i
n <module>
     15 # Plotting word cloud for each product
     16 for index,product in enumerate(new_dtm.columns):
---> 17   generate_wordcloud(new_dtm[product].sort_values(ascending=False),produ
ct)

/var/folders/7b/nnsn4tm15ns7j6qqstdz1src0000gn/T/ipykernel_82252/1865818958.py i
n generate_wordcloud(data, title)
      7   plt.imshow(wc, interpolation='bilinear')
      8   plt.axis("off")
----> 9   plt.title('\n'.join(wrap(title,60)),fontsize=13)
     10   plt.show()
     11

~/opt/anaconda3/lib/python3.9/textwrap.py in wrap(text, width, **kwargs)
    377     """
    378     w = TextWrapper(width=width, **kwargs)
--> 379     return w.wrap(text)
    380
    381 def fill(text, width=70, **kwargs):

~/opt/anaconda3/lib/python3.9/textwrap.py in wrap(self, text)
    349         converted to space.
    350         """
--> 351         chunks = self._split_chunks(text)
    352         if self.fix_sentence_endings:
    353             self._fix_sentence_endings(chunks)

~/opt/anaconda3/lib/python3.9/textwrap.py in _split_chunks(self, text)
    335
    336     def _split_chunks(self, text):
--> 337         text = self._munge_whitespace(text)
    338         return self._split(text)
    339 

~/opt/anaconda3/lib/python3.9/textwrap.py in _munge_whitespace(self, text)
    152         """
    153         if self.expand_tabs:
--> 154             text = text.expandtabs(self.tabsize)
    155         if self.replace_whitespace:
    156             text = text.translate(self.unicode_whitespace_trans)

AttributeError: 'int' object has no attribute 'expandtabs'
```



### 4.2.2 Word Frequency - Visualization - Plot with MATLAB

In [ ]: