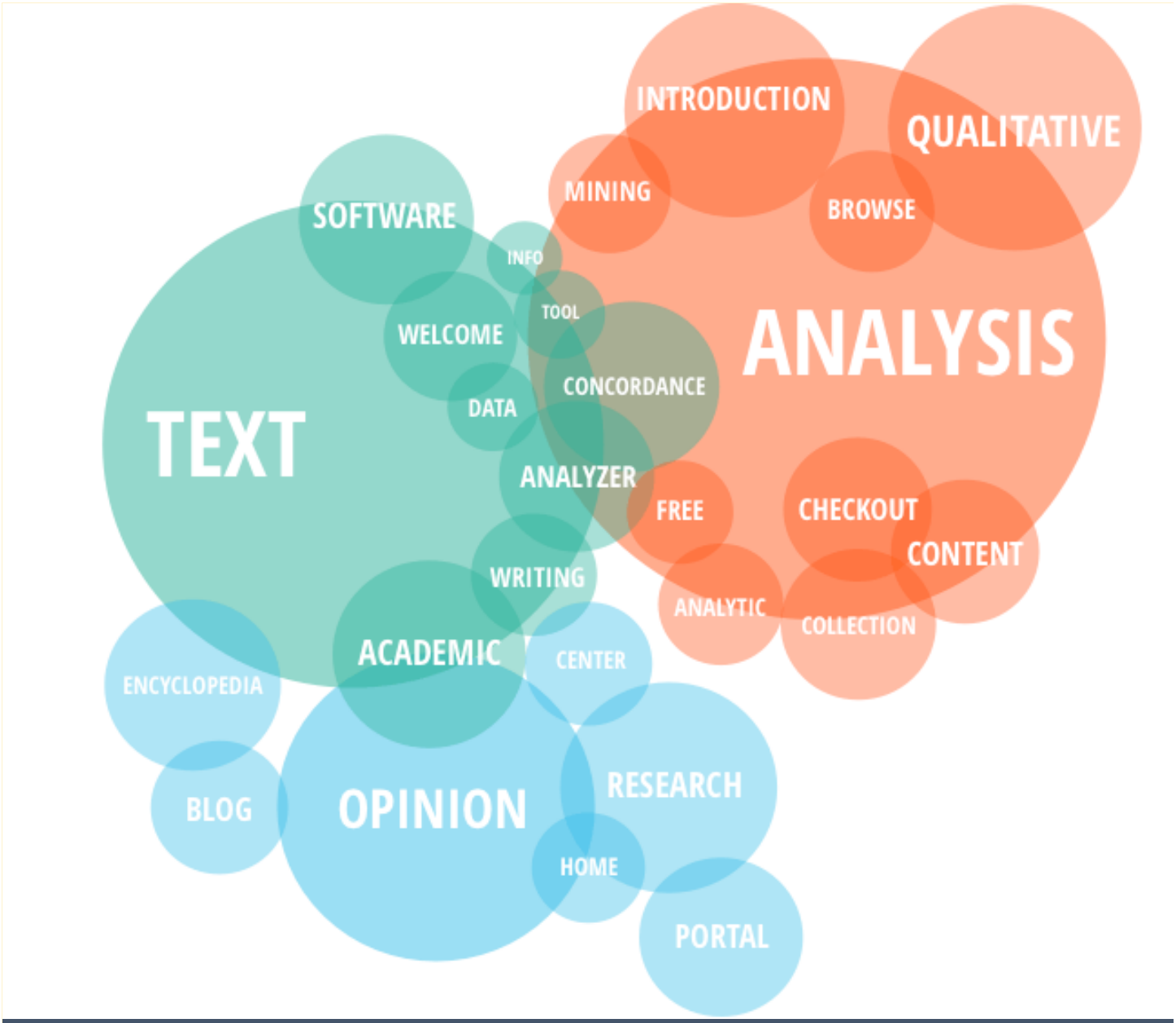


MDD Individual Portfolio

Text analysis

Thanh Dung Nguyen 1/1/23 MDDSDSC1A.8

Table of Contents

Business case

Topic Overview

In the Internet space, textual data appears widely (blogs, social networks, newspapers,...). Due to the nature of natural languages, taking advantage of these data sources is often not straightforward. With this problem, my idea is to build a model to help identify the topic of any text with the learning data source as the online newspaper. With this model, I hope to make it easier to take advantage of this huge data source by being able to turn from a long text into just their topic. Within the limitation of the short research and implementation time of this minor, this report only stops at the basic and not macro level of contributing to business or data science. However, I will continue to research and develop this project

To analyze a large amount of unstructured data in the form of text (emails, conversations on social networks, ...) is really a big problem. Manual analysis is often time-consuming, resource-intensive, and error-prone.

Text Analysis (TA) is a machine learning technique used to automatically extract valuable insights from unstructured text data. Many businesses use text analytics tools to quickly analyze online data and documents and convert them into useful insights. Text analytics can extract specific information, like keywords, names, or company information from thousands of emails, or categorize survey responses by intent, emotion, and topic.

Why is text analysis important?

Here are the outstanding advantages that text analysis AI tools can bring:

- Flexible scalability

Text analytics tools allow businesses to structure large amounts of information, like emails, chats, social networks, support requests, documents, and more, in seconds instead of days, so you can allocate resources to more important business tasks.

- Return results in real-time

Today, businesses have to deal with and deal with a flood of information and customer comments, appearing on many different channels and platforms. Text analytics is a promising game changer because it can detect issues urgently, wherever they appear, 24/7, and in real-time. By training text analytics models to detect issues, complaints, negative comments, and more, businesses can automatically flag tweets, reviews, videos, and more. and take early intervention action.

- AI text analytics delivers consistent data sets

By training text analysis models according to the unique needs and criteria of each business, algorithms can analyze, understand, and organize data much more accurately than humans.

Idea:

- Objective: serve to identify topics automatically.

My plan is to create a train model using articles from internet newspapers as a source. For instance, depending on the train set to identify based on the standard of nltimes.nl, the model recognizes text from a specific paragraph (random material, not from nltimes.nl) and determines the topic of it. The most practical use and related to business of this model is apply to analyzing customer reviews.

- Subjective: filter posts on social networks by topics that interested to avoid wasting time surfing online.

For the time being, this project is just for my personal use of finding my favorite reading topics. However, if expanded, this model can be deployed, developed, and used for many different purposes. which will certainly serve in the business term

Data Science Techniques and Tools

I was able to apply different data science techniques and tools while working on this project

Tools	Techniques				
	Math & Stats	Data Visualization	Data Mining	Process Mining	Forecasting
R/RStudio	•	•	•		•
Python			•		
Github					
Web scraping					

Data mining

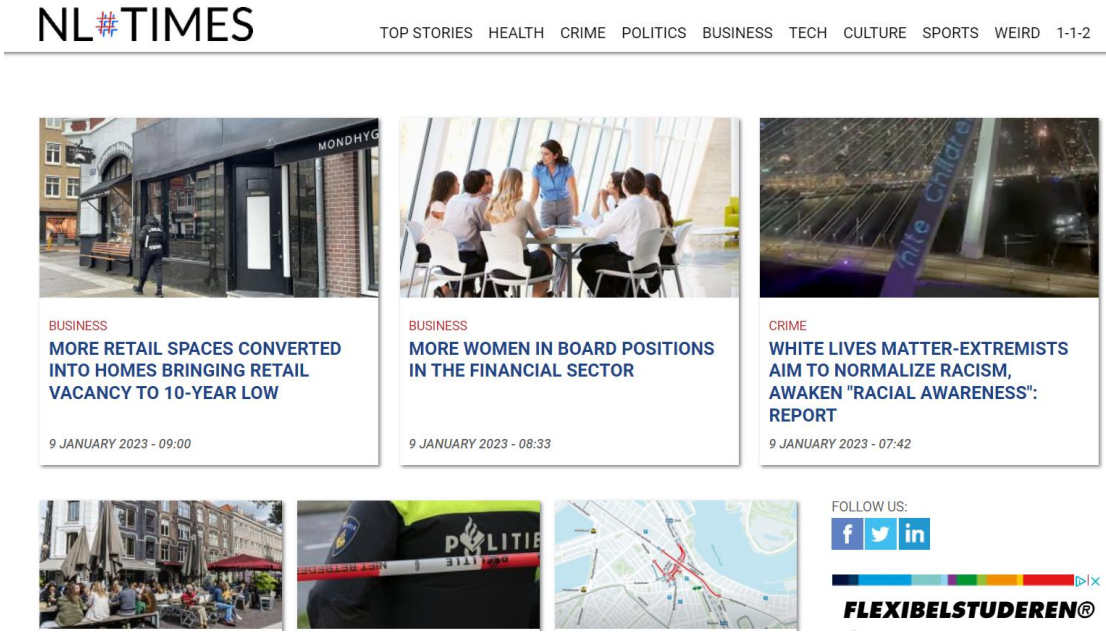
I create my own data by using web scraping to extract data and content from websites. The reality is that every business tries to protect its database, and every individual tries to protect privacy while the opportunity to work in large corporations with big data sources available is few. At that time, either we spend money to buy data from illegal sources, or we are forced to collect data from publicly available sources such as websites on the Internet. However, those data are often fragmentary, and difficult to mine manually by human power.

The fact is I didn't know how to start this project. It is quite difficult to find an existing dataset and build a new idea from it. Therefore, I chose to create it on my own. It was also a great experience for me to learn a new technique.

Why I choose an online newspaper? Firstly, with the feature of regularly updating information to readers, hundreds or thousands of articles are posted every hour, making online newspapers a huge data warehouse. Second, each published article has a certain category, so there is no need to spend too much effort to label the articles. And finally, as a newspaper, the articles will usually have quality and guarantee in terms of semantics as well as grammar.

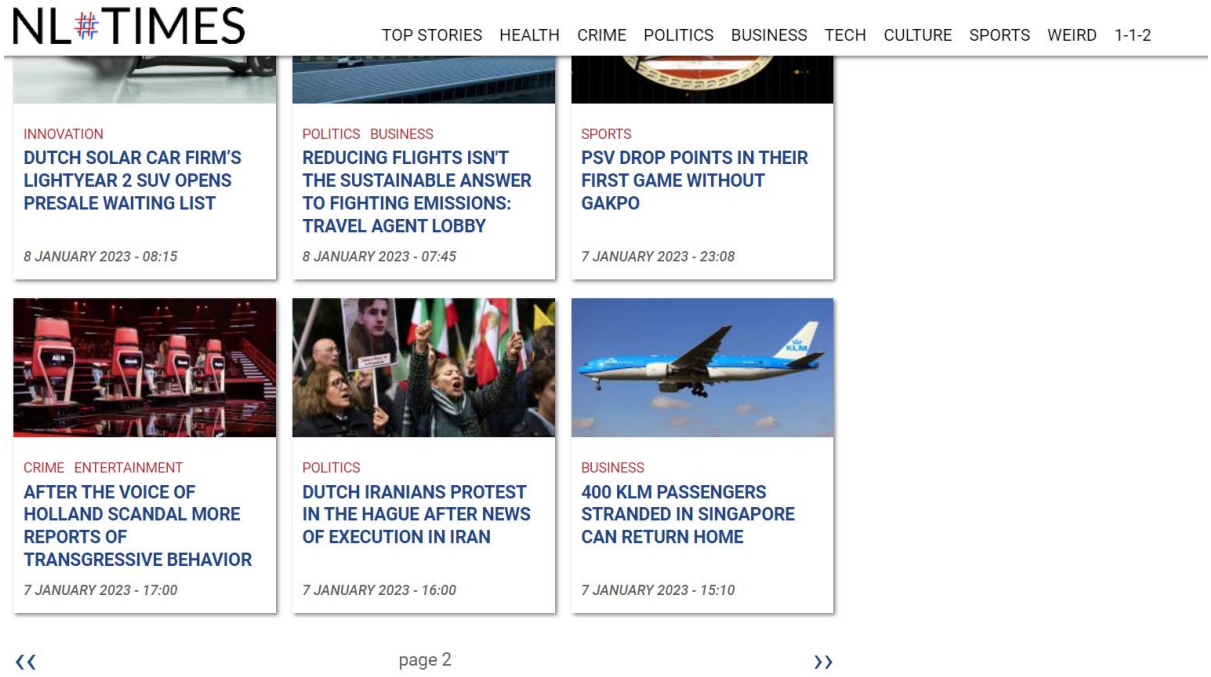
The data in this project were collected entirely from NLtimes (https:// nltimes.nl /).

When first entering the homepage, it can be seen that this newspaper has an uneven order and structure of information, which is difficult to collect.

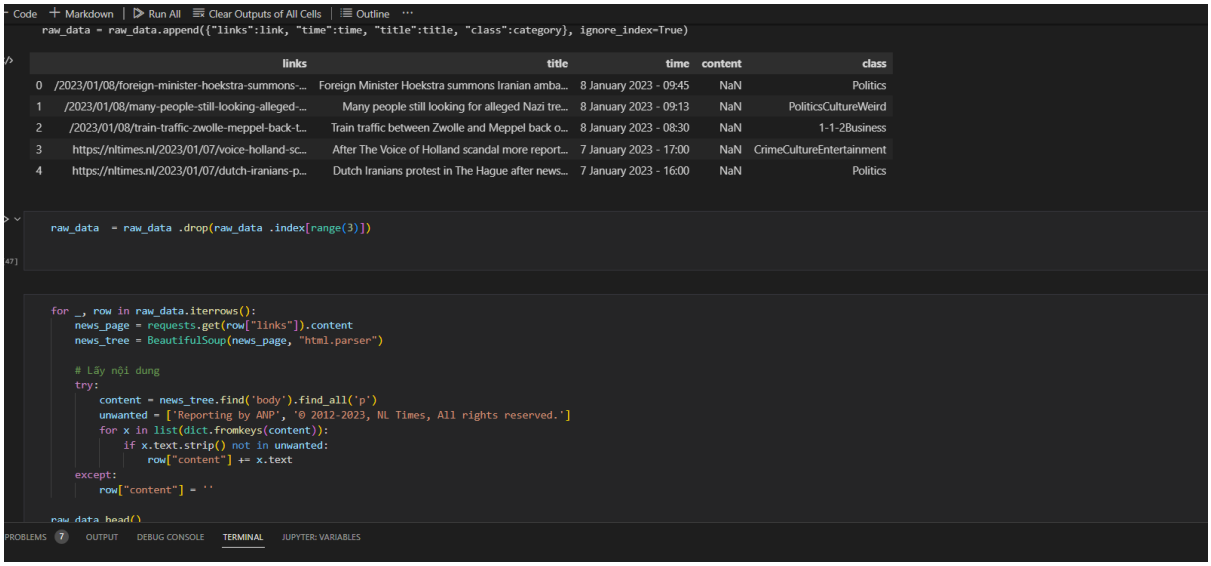


(https:// nltimes.nl /).

Fortunately, this site is structured as having the latest news articles containing all the news from the categories. The news surfing process will include: Scroll to the bottom of the page -> Click the arrow -> Scroll to the bottom of the page and repeat.



I use python to collect and aggregate data from the website. There are a number of web scraping tools out there to perform the task, and in a variety of languages, there are libraries that support web scraping. As far as I know, out of all these languages, Python is considered as one of the best for Web Scraping because of features like - a rich library, easy to use, dynamically typed, etc. Besides, I found several examples of this process in Python that I could learn from.



```
raw_data = raw_data.append({"links":link, "time":time, "title":title, "class":category}, ignore_index=True)
```

	links	title	time	content	class
0	/2023/01/08/foreign-minister-hoekstra-summons-...	Foreign Minister Hoekstra summons Iranian amba...	8 January 2023 - 09:45	NaN	Politics
1	/2023/01/08/many-people-still-looking-alleged-...	Many people still looking for alleged Nazi tre...	8 January 2023 - 09:13	NaN	PoliticsCultureWeird
2	/2023/01/08/train-traffic-zwolle-meppel-back-t...	Train traffic between Zwolle and Meppel back o...	8 January 2023 - 08:30	NaN	1-1-2Business
3	https://nltimes.nl/2023/01/07/voice-holland-sc...	After The Voice of Holland scandal more report...	7 January 2023 - 17:00	NaN	CrimeCultureEntertainment
4	https://nltimes.nl/2023/01/07/dutch-iranians-p...	Dutch Iranians protest in The Hague after news...	7 January 2023 - 16:00	NaN	Politics

```
raw_data  = raw_data .drop(raw_data .index[range(3)])
```

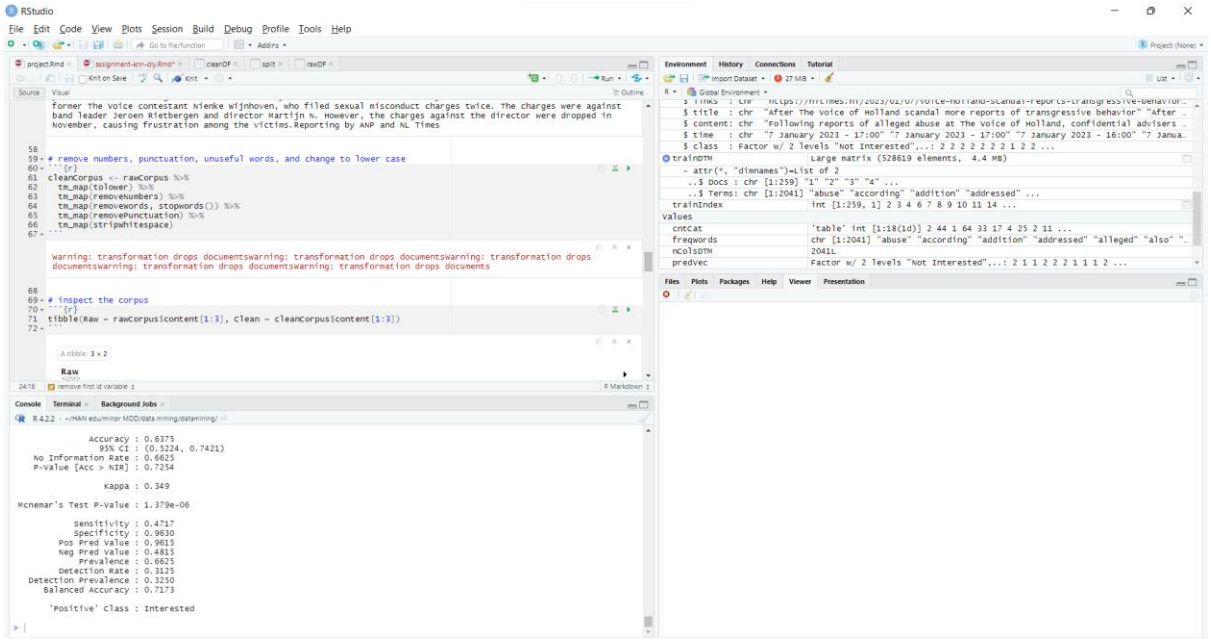
```
for _, row in raw_data.iterrows():
    news_page = requests.get(row["links"]).content
    news_tree = BeautifulSoup(news_page, "html.parser")

    # Lấy nội dung
    try:
        content = news_tree.find('body').find_all('p')
        unwanted = ['Reporting by ANP', '© 2012-2023, NL Times, All rights reserved.']
        for x in list(dict.fromkeys(content)):
            if x.text.strip() not in unwanted:
                row["content"] += x.text
    except:
        row["content"] = ''

raw_data.head()
```

Classification/Clustering: Naive Bayes

Although this was not the original goal of this project, due to limited skills and time the current results of the project according to this report stop at Probabilistic Learning with Naive Bayes Classification. And since this is the module that was taught in this minor in the R language, I have followed up and applied what I have learned and continued to implement it in this language.



```
former The Voice contestant Nienke Wijnhoven, who filed sexual misconduct charges twice. The charges were against
band leader Jeroen Kletbergen and director Martijn N. However, the charges against the director were dropped in
November, causing frustration among the victims.Reporting by ANP and NL Times

# remove numbers, punctuation, unuseful words, and change to lower case
```{r}
cleanCorpus <- rawCorpus %>%
  tm_map(tolower) %>%
  tm_map(removeNumbers) %>%
  tm_map(removeWords, stopwords()) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
```

Warning: transformation drops documentsWarning: transformation drops documentsWarning: transformation drops
documentsWarning: transformation drops documentsWarning: transformation drops documents

inspect the corpus
```{r}
tibble(Raw = rawCorpus$content[1:3], Clean = cleanCorpus$content[1:3])
```

A tibble: 3 x 2

Raw

Accuracy : 0.6375
95% CI : (0.5224, 0.7421)
No Information Rate : 0.6625
P-Value [Acc > NIR] : 0.7254

Kappa : 0.349

Mcnemar's Test P-Value : 1.379e-06

Sensitivity : 0.4717
Specificity : 0.9630
Pos Pred Value : 0.9615
Neg Pred Value : 0.4815
Prevalence : 0.6625
Detection Rate : 0.3125
Detection Prevalence : 0.3250
Balanced Accuracy : 0.7173

'Positive' class : Interested

trainIndex int [1:259, 1] 2 3 4 6 7 8 9 10 11 14 ...
values
cntCat 'table' int [1:18(1d)] 2 44 1 64 33 17 4 25 2 11 ...
freqwords chr [1:2041] "abuse" "according" "addition" "addressed" "alleged" "also" "...
nColsDTM 2041L
predVec Factor w/ 2 levels "Not Interested",..: 2 1 1 2 2 2 1 1 1 2 ...
```

Mathematics & Statistics, Data visualization, and Forecasting all appear in a step of the Naive Bayes technique and are done by the language of R due to the convenience of having practiced and gone through the learning process

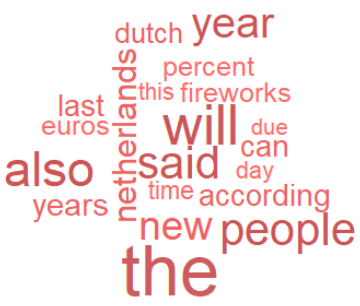
```r
predVec <- predict(nbayesModel, testDTM)
confusionMatrix(predVec, testDF$class, positive = "Interested", dnn = c("Prediction", "True"))
```

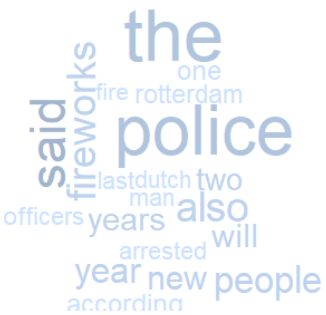
```
Confusion Matrix and Statistics

               True
Prediction      Not Interested Interested
  Not Interested             26         28
  Interested                  1         25

               Accuracy : 0.6375
                 95% CI : (0.5224, 0.7421)
    No Information Rate : 0.6625
    P-Value [Acc > NIR] : 0.7254

                  Kappa : 0.349

 Mcnemar's Test P-Value : 1.379e-06

            Sensitivity : 0.4717
            Specificity : 0.9630
         Pos Pred Value : 0.9615
         Neg Pred Value : 0.4815
             Prevalence : 0.6625
         Detection Rate : 0.3125
   Detection Prevalence : 0.3250
      Balanced Accuracy : 0.7173

       'Positive' Class : Interested
```

Process mining technique is not yet relevant for the moment of the project's execution

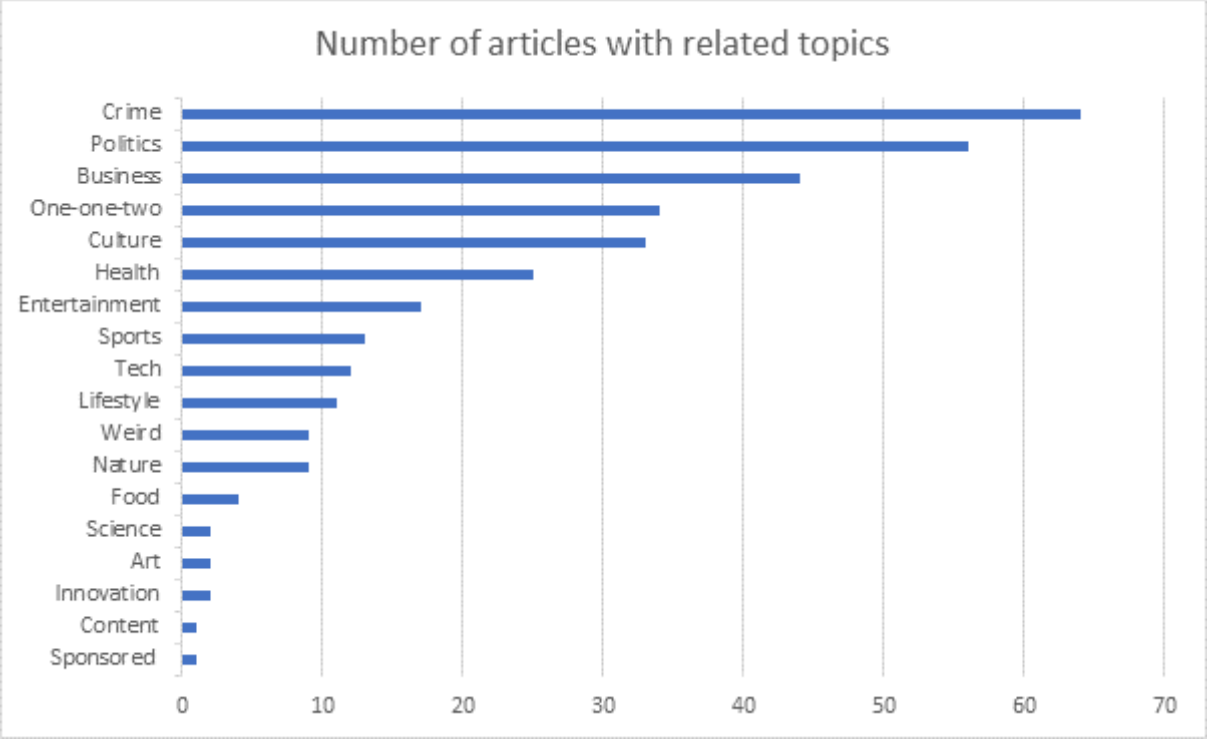
Visualisation

Visually inspect the data by creating wordclouds



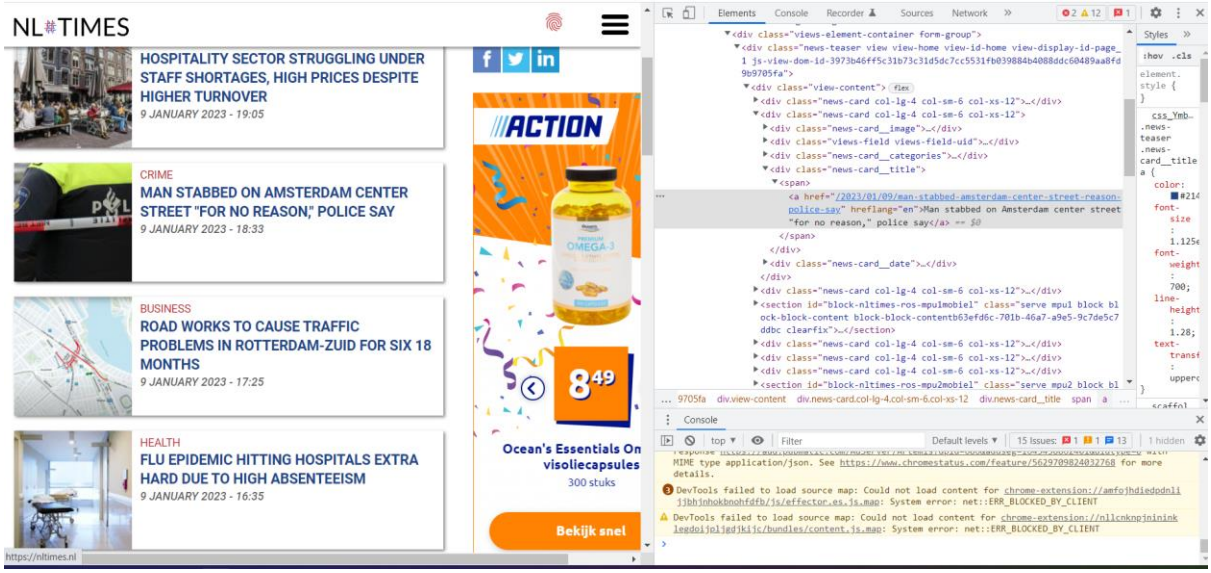


A keyword cloud is a graphical representation of the frequency of words that appear more frequently in the original text(s). The larger the word in the image, the more common and important the word is. The wordclouds above had been built before the dataset was processed. It is clear that the difference between the two types of information has not been pointed out. These are less meaningful words so the dataset needs to be clean to remove stop words.



Based on the bar chart, it can be seen that there are many classes with very few samples. This will result in the inaccuracy of the test. Some classes have a single record and these could be selected for the sample. Or else, with this, I could make a decision workaround that will remove the entire class with too few samples.

Justifying the choices throughout the process



Web scraping

When using the browser's network monitoring tool, I discovered that the website's new news request link when clicking the see more arrow for the first time was "https://nltimes.nl/?page=1". Chances are that when we replace page=1 with page=2, page=3,... we will get new results from the 2nd, 3rd,... clicks on the arrow button. " I have verified and this is true.

Taking this information and applying it to the code with this python language was my first experience. So it took me a lot of time to research I built the code that can retrieve data from a first test URL. The result is a complete dataset of a page with all the variables needed for analysis.

```python
for _, row in raw_data.iterrows():
    news_page = requests.get(row["links"]).content
    news_tree = BeautifulSoup(news_page, "html.parser")

    # Export content
    try:
        content = news_tree.find('body').find_all('p')
        unwanted = ['Reporting by ANP', '© 2012-2023, NL Times, All rights reserved.']
        for x in list(dict.fromkeys(content)):
            if x.text.strip() not in unwanted:
                row["content"] += x.text
    except:
        row["content"] = ''

raw_data.head()
```

[49]

	links	title	time	content	class
3	https://nltimes.nl/2023/01/07/voice-holland-sc...	After The Voice of Holland scandal more report...	7 January 2023 - 17:00	Following reports of alleged abuse at The Voic...	CrimeCultureEntertainment
4	https://nltimes.nl/2023/01/07/dutch-iranians-p...	Dutch Iranians protest in The Hague after news...	7 January 2023 - 16:00	The Association of Iranian Academics in the Ne...	Politics
5	https://nltimes.nl/2023/01/07/400-klm-passenge...	400 KLM passengers stranded in Singapore can r...	7 January 2023 - 15:10	The last of the 400 KLM passengers stranded in...	Business
6	https://nltimes.nl/2023/01/07/netherlands-co-o...	The Netherlands co-organizes summit in London ...	7 January 2023 - 14:22	Justice ministers from around the world will m...	Politics
7	https://nltimes.nl/2023/01/07/less-fireworks-d...	Less fireworks damage in Dutch municipalities ...	7 January 2023 - 13:30	This past New Year's Eve, significantly fewer ...	PoliticsCultureLifestyle

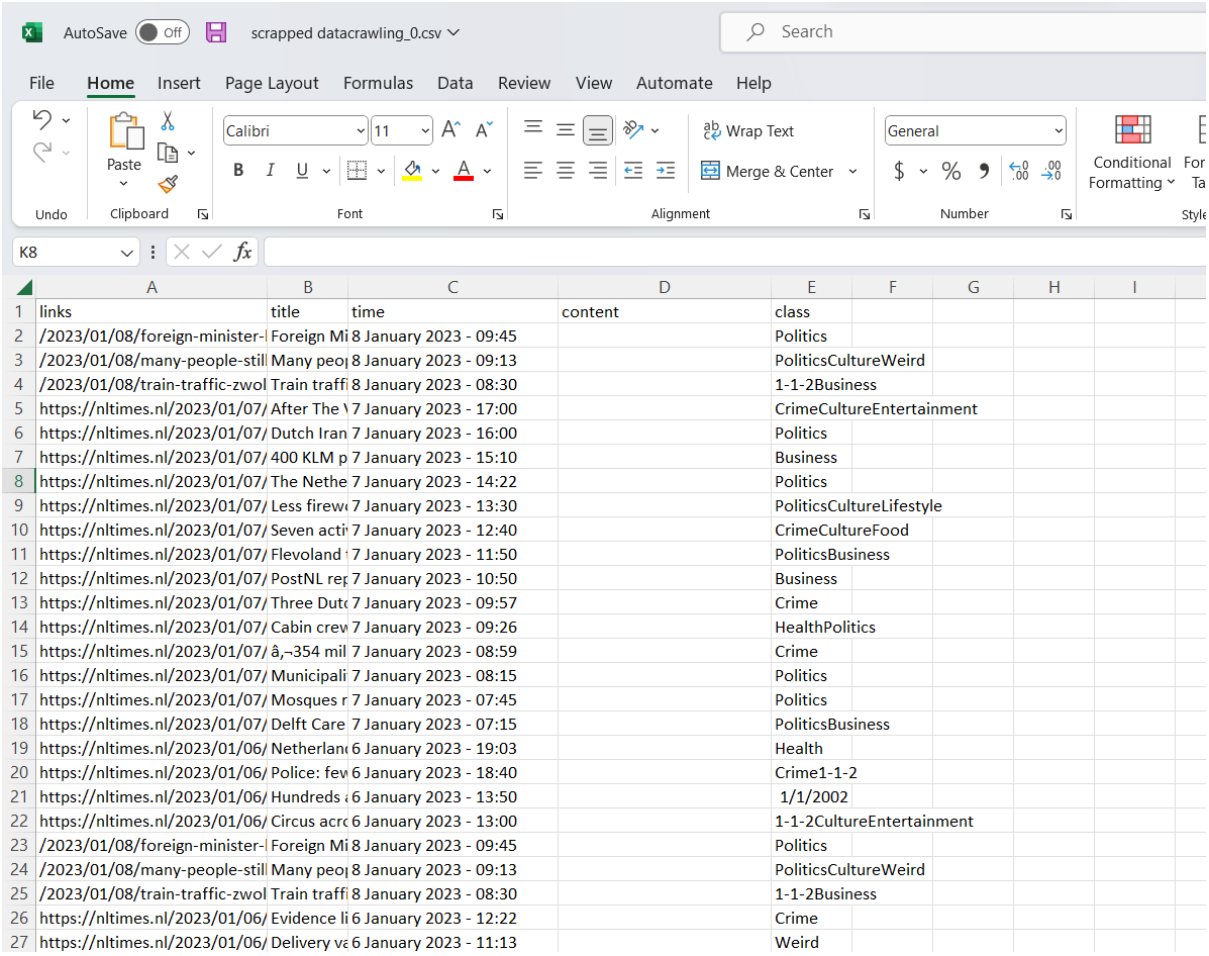
The next step is to automatically export from not just one page but multiple pages in real-time from just the first link. I also succeeded in extracting subsequent links based on a single first link automatically.

```python
#Batch scraping
def batch_scraping(num = 150, output_dir = ""):
    '''
    Collect all data on nltimes.nl

    Param:
        num: number of request pages for 1 batch
    '''
    iter_num = 0  # Start batch number
    continue_flag = True # The loop termination flag when an error occurs

    while (continue_flag):

        '''
        Initialize empty dataframe.
        Then get enough pages for 1 batch.
        Then export the csv . file.
        '''
        batch_df = pd.DataFrame(columns=["links","title","time","content","class"])
        for index in range(iter_num*num+1,(iter_num+1)*num+1):
            data = single_request_scraping(index)
            if (data is None):
                continue_flag = False
                break
            print(f"Page {index} complete!")
            batch_df = batch_df.append(data)
        batch_df.to_csv(output_dir + f'crawling_{iter_num}.csv',index=False,encoding="utf-8")
        iter_num+=1
```

Python

However, an error occurred when outputting the article content of the links when the construction structure of the source site was inconsistent. Every page returns the same 3 latest articles without a proper link. The output file does not have the articles' content. Loading this data took a lot of time, more than an hour for 200 pages, but there was an error that made me really lose my temper.

```
Request error link /2023/01/08/foreign-minister-hoekstra-summons-iranian-ambassador-executions
Request error link /2023/01/08/many-people-still-looking-alleged-nazi-treasure-near-ommeren
Request error link /2023/01/08/train-traffic-zwolle-meppel-back-track-repairs
Page 210 complete!
```



AutoSave Off scrapped datacrawling_0.csv Search

File Home Insert Page Layout Formulas Data Review View Automate Help

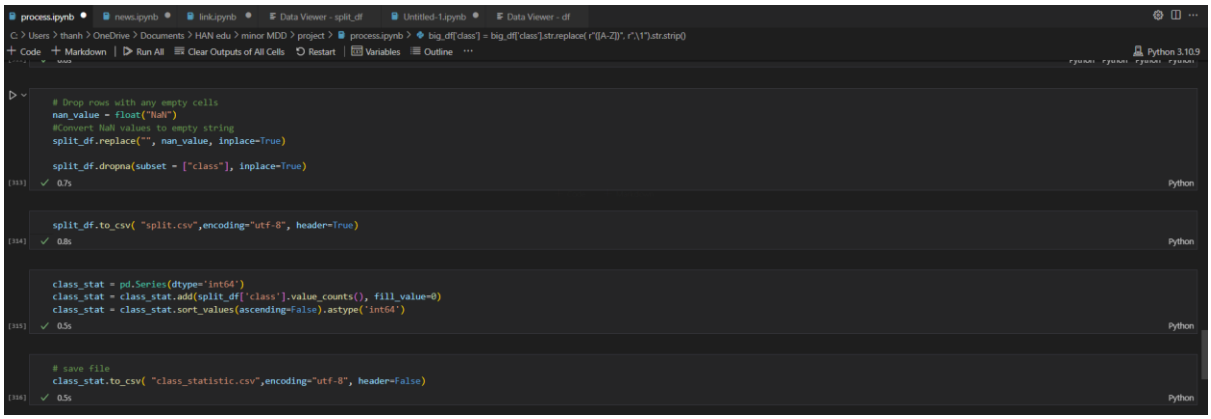
Paste Calibri 11 Wrap Text General Conditional Formatting

Undo Clipboard Font Alignment Number Style

K8

	A	B	C	D	E
1	links	title	time	content	class
2	/2023/01/08/foreign-minister-	Foreign Mi	8 January 2023 - 09:45		Politics
3	/2023/01/08/many-people-still	Many peo	8 January 2023 - 09:13		PoliticsCultureWeird
4	/2023/01/08/train-traffic-zwol	Train traffi	8 January 2023 - 08:30		1-1-2Business
5	https://nltimes.nl/2023/01/07/	After The \	7 January 2023 - 17:00		CrimeCultureEntertainment
6	https://nltimes.nl/2023/01/07/	Dutch Iran	7 January 2023 - 16:00		Politics
7	https://nltimes.nl/2023/01/07/	400 KLM p	7 January 2023 - 15:10		Business
8	https://nltimes.nl/2023/01/07/	The Nethe	7 January 2023 - 14:22		Politics
9	https://nltimes.nl/2023/01/07/	Less firew	7 January 2023 - 13:30		PoliticsCultureLifestyle
10	https://nltimes.nl/2023/01/07/	Seven acti	7 January 2023 - 12:40		CrimeCultureFood
11	https://nltimes.nl/2023/01/07/	Flevoland	7 January 2023 - 11:50		PoliticsBusiness
12	https://nltimes.nl/2023/01/07/	PostNL rep	7 January 2023 - 10:50		Business
13	https://nltimes.nl/2023/01/07/	Three Dut	7 January 2023 - 09:57		Crime
14	https://nltimes.nl/2023/01/07/	Cabin crew	7 January 2023 - 09:26		HealthPolitics
15	https://nltimes.nl/2023/01/07/	â,¬354 mil	7 January 2023 - 08:59		Crime
16	https://nltimes.nl/2023/01/07/	Municipali	7 January 2023 - 08:15		Politics
17	https://nltimes.nl/2023/01/07/	Mosques r	7 January 2023 - 07:45		Politics
18	https://nltimes.nl/2023/01/07/	Delft Care	7 January 2023 - 07:15		PoliticsBusiness
19	https://nltimes.nl/2023/01/06/	Netherlan	6 January 2023 - 19:03		Health
20	https://nltimes.nl/2023/01/06/	Police: few	6 January 2023 - 18:40		Crime1-1-2
21	https://nltimes.nl/2023/01/06/	Hundreds	6 January 2023 - 13:50		1/1/2002
22	https://nltimes.nl/2023/01/06/	Circus acro	6 January 2023 - 13:00		1-1-2CultureEntertainment
23	/2023/01/08/foreign-minister-	Foreign Mi	8 January 2023 - 09:45		Politics
24	/2023/01/08/many-people-still	Many peo	8 January 2023 - 09:13		PoliticsCultureWeird
25	/2023/01/08/train-traffic-zwol	Train traffi	8 January 2023 - 08:30		1-1-2Business
26	https://nltimes.nl/2023/01/06/	Evidence li	6 January 2023 - 12:22		Crime
27	https://nltimes.nl/2023/01/06/	Delivery va	6 January 2023 - 11:13		Weird

For now, I couldn't find a solution for exporting automatically real-time data but I have to move on with the project. Therefore, I decide to retrieve data from each page manually like the one I successfully made with one first test URL. I repeated the code 12 times to export 12 data files and then merge them into a single dataset.

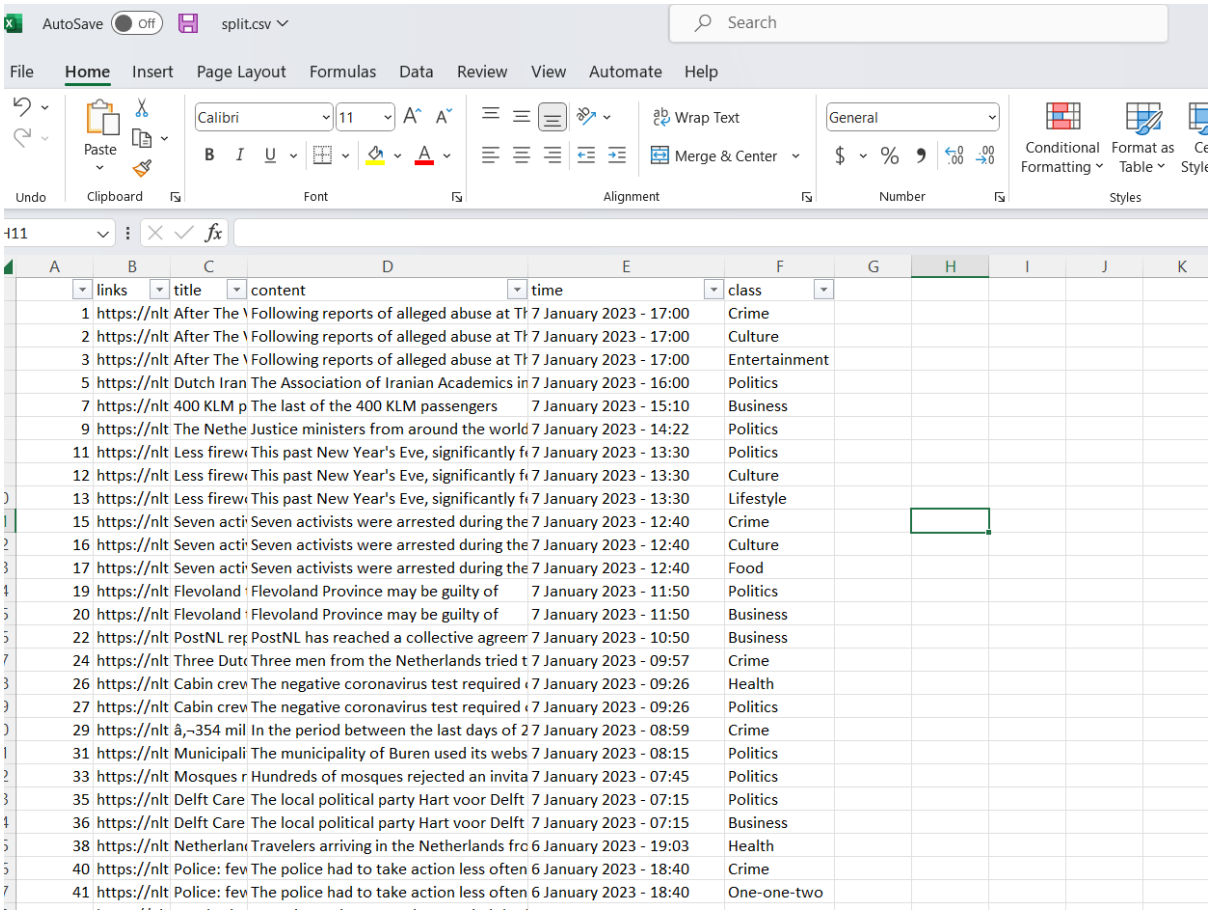


```
# Drop rows with any empty cells
nan_value = float("NaN")
#Convert NaN values to empty string
split_df.replace("", nan_value, inplace=True)

split_df.dropna(subset = ["class"], inplace=True)
```

```
split_df.to_csv( "split.csv",encoding="utf-8", header=True)
```

```
class_stat = pd.Series(dtype='int64')
class_stat = class_stat.add(split_df['class'].value_counts(), fill_value=0)
class_stat = class_stat.sort_values(ascending=False).astype('int64')
```

```
# save file
class_stat.to_csv( "class_statistic.csv",encoding="utf-8", header=False)
```



	links	title	content	time	class
1	https://nlt	After The \	Following reports of alleged abuse at Tl	7 January 2023 - 17:00	Crime
2	https://nlt	After The \	Following reports of alleged abuse at Tl	7 January 2023 - 17:00	Culture
3	https://nlt	After The \	Following reports of alleged abuse at Tl	7 January 2023 - 17:00	Entertainment
5	https://nlt	Dutch Iran	The Association of Iranian Academics in	7 January 2023 - 16:00	Politics
7	https://nlt	400 KLM p	The last of the 400 KLM passengers	7 January 2023 - 15:10	Business
9	https://nlt	The Nethe	Justice ministers from around the world	7 January 2023 - 14:22	Politics
11	https://nlt	Less firew	This past New Year's Eve, significantly f	7 January 2023 - 13:30	Politics
12	https://nlt	Less firew	This past New Year's Eve, significantly f	7 January 2023 - 13:30	Culture
13	https://nlt	Less firew	This past New Year's Eve, significantly f	7 January 2023 - 13:30	Lifestyle
15	https://nlt	Seven acti	Seven activists were arrested during the	7 January 2023 - 12:40	Crime
16	https://nlt	Seven acti	Seven activists were arrested during the	7 January 2023 - 12:40	Culture
17	https://nlt	Seven acti	Seven activists were arrested during the	7 January 2023 - 12:40	Food
19	https://nlt	Flevoland	Flevoland Province may be guilty of	7 January 2023 - 11:50	Politics
20	https://nlt	Flevoland	Flevoland Province may be guilty of	7 January 2023 - 11:50	Business
22	https://nlt	PostNL rep	PostNL has reached a collective agreem	7 January 2023 - 10:50	Business
24	https://nlt	Three Dut	Three men from the Netherlands tried t	7 January 2023 - 09:57	Crime
26	https://nlt	Cabin crev	The negative coronavirus test required	7 January 2023 - 09:26	Health
27	https://nlt	Cabin crev	The negative coronavirus test required	7 January 2023 - 09:26	Politics
29	https://nlt	â,¬354 mil	In the period between the last days of 2	7 January 2023 - 08:59	Crime
31	https://nlt	Municipali	The municipality of Buren used its webs	7 January 2023 - 08:15	Politics
33	https://nlt	Mosques r	Hundreds of mosques rejected an invita	7 January 2023 - 07:45	Politics
35	https://nlt	Delft Care	The local political party Hart voor Delft	7 January 2023 - 07:15	Politics
36	https://nlt	Delft Care	The local political party Hart voor Delft	7 January 2023 - 07:15	Business
38	https://nlt	Netherlan	Travelers arriving in the Netherlands fro	6 January 2023 - 19:03	Health
40	https://nlt	Police: few	The police had to take action less often	6 January 2023 - 18:40	Crime
41	https://nlt	Police: few	The police had to take action less often	6 January 2023 - 18:40	One-one-two

I must say that being able to process a complete data file is already an achievement for me who have no background or experience in coding. However, I know that this is just the very first step of the data analysis process

My idea is to use content from online newspapers as a source to build a train model. For example, the model detects text from a particular paragraph (random text, not from nltimes.nl) and identifies the topic of it as “politics”, “lifestyle”, or “health”, etc. based on what it has been trained based on the standard of nltimes.nl topics. Nonetheless, the data available is not enough to build an accurate model as I got some bugs during the process. More than that, lacking skill makes me unable to achieve the goal in a short period of time. I do make some research and found some useful demos, yet they are in python, and I don’t really know about this language.

I also found that the model that can be used for this case are the Naïve Bayes model, Logistic Regression model, or Bagging MLP. And since we have learned about the Naïve Bayes model in R, I used the data file exported by python to run the simple Naïve Bayes model although the outcome doesn’t fit with the goal.

For now, the only substantial outcome of this project so far is having the data from the web scraping technique. But of course, I will not stop at that. My project is still being developed. For more updates, please visit my Github.

Sources

sparkbyexamples https://sparkbyexamples.com/pandas/pandas-read-multiple-csv-files/

Witek ten Hove https://businessdatasolutions.github.io/courses/data%20mining/gitbook/book-output/index.html#purpose-of-this-course

several questions on geeksforgeeks https://www.geeksforgeeks.org/

several questions on stackoverflow https://stackoverflow.com/questions/

several topics pythondaddy https://www.pythondaddy.com/

Shanelynn https://www.shanelynn.ie/pandas-drop-delete-dataframe-rows-columns/

Tutorialspoint
https://www.tutorialspoint.com/python_data_science/python_word_tokenization.htm

Proxyscrape https://proxyscrape.com/blog/web-scraping-for-news-articles-using-python