

AI VIETNAM
All-in-One Course
(TA Session)

Data Manipulation and Crawling

Project



AI VIET NAM
[@aivietnam.edu.vn](http://aivietnam.edu.vn)

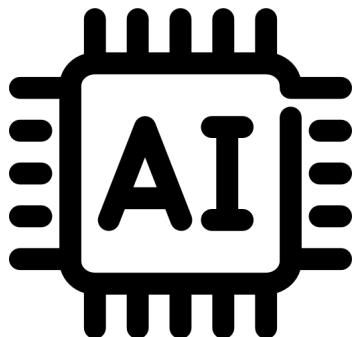
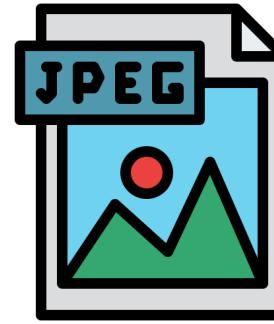
Dinh-Thang Duong – TA

Outline

- Introduction
- DataTable Handling
- Text Representation
- Data Crawling
- Question

Introduction

❖ Getting Started



Some data structure



When coding or learning AI

Data manipulation, crawling... is important

Outline

- Introduction
- DataTable Handling
- Text Representation
- Data Crawling
- Question

DataTable Handling

❖ Introduction

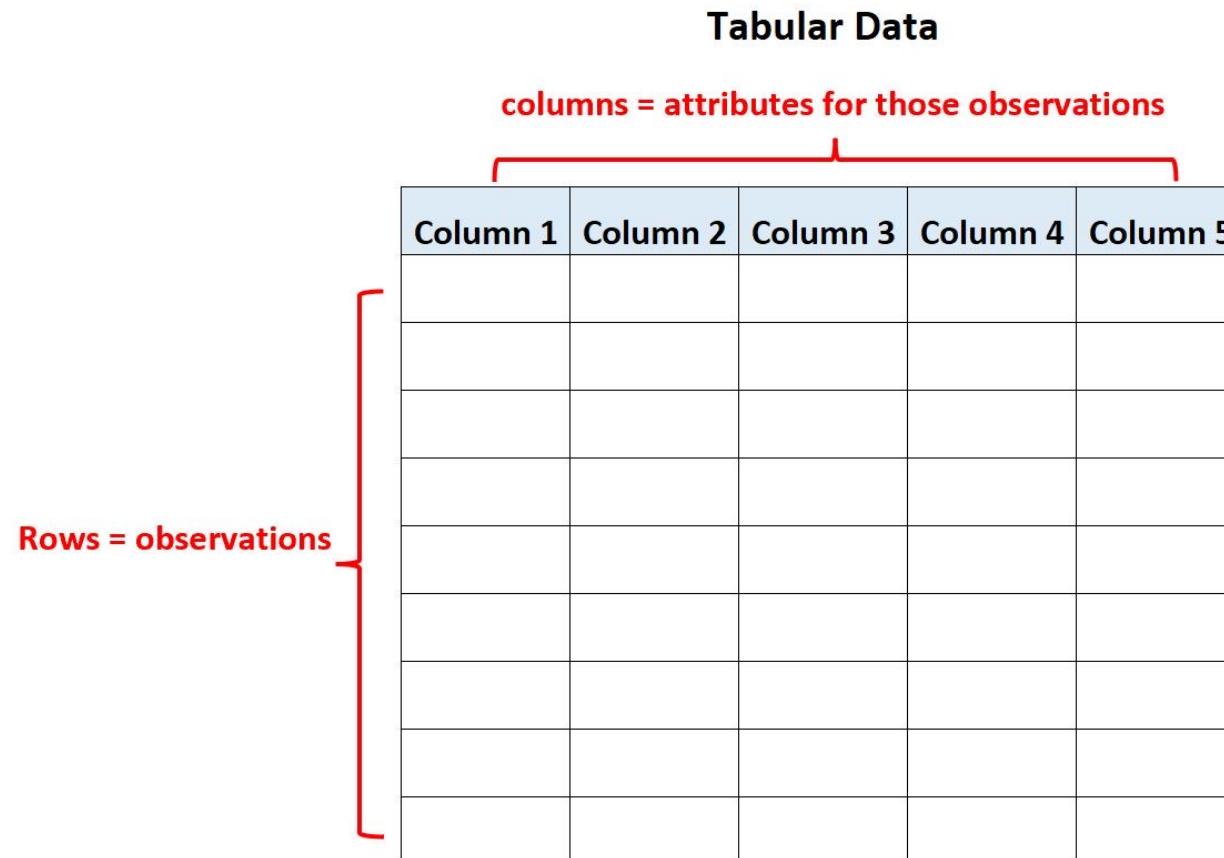
	B	C	D	E	F
1	trans_code	amount	ref_number	status	Account Type
2	1	10	123456789	Process Completed	Normal Account
3	1	10	123456789	Process Completed	Normal Accnt
4	1	10	123456789	Process Completed	Normal Account
5	1	10	123456789	Process Completed	Normal Accnt
6	1	10	123456789	Process Completed	Normal Accnt
7	1	10	123456789	Process Completed	Normal Account
8	1	10	123456789	Process Completed	Normal Account
9	1	10	123456789	Process Completed	Normal Accnt
10	1	10	123456789	Process Completed	Normal Account
11	1	10	123456789	Process Completed	DOSRI
12	1	10	123456789	Process Completed	Normal Account
13	1	10	123456789	Process Completed	Employee_Related
14	1	10	123456789	Process Completed	Normal Account
15	1	10	123456789	Process Completed	High-Risk
16	1	10	123456789	Process Completed	Normal Account
17	1	10	123456789	Process Completed	Normal Account
18	1	10	123456789	Process Completed	PEP
19	1	10	123456789	Process Completed	Normal Account
20	1	10	123456789	Process Completed	Normal Account
21	1	10	123456789	Process Completed	RPT
22	1	10	123456789	Process Completed	Normal Account
23					



DataTable (Tabular Data)

DataTable Handling

❖ Introduction



Tabular Data: Data that is organized in a table with rows and columns. We can think of it as a 2D array, list of lists...

- **Rows (records, samples):** Observation (cases).
- **Columns (fields, features):** Attributes of observations.

	A	B	C	D
1	First Name	Last Name	Age	Salary
2	Jon	Smith	36	26500
3	Helen	Mirren	22	21000
4	David	Cameron	29	39000
5	Brad	Pitt	52	45000
6	Anna	Starolsky	41	22500
7	Peter	Piper	20	31500
8	David	Duck	19	15700
9	Julie	Walters	33	19000

DataTable Handling

❖ Introduction



Tabular Data is very important



DataTable Handling

- ❖ How to interact tabular data in Python?



Pandas library: A fast, powerful, flexible and easy to use open source data analysis and manipulation tool

DataTable Handling

❖ Problem Statement

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

Given Advertising.csv file:

<https://raw.githubusercontent.com/justmarkham/scikit-learn-videos/master/data/Advertising.csv>

Use Python to read this file and do some basic operations. For each columns, calculate:

- Find sum.
- Find sum from row 2nd to row 10th.
- Find min and max.
- Find average.
- Find median.
- Find budget contribution percentage.

DataTable Handling

❖ Read .csv file

pandas.read_csv

```
pandas.read_csv(filepath_or_buffer, *, sep=_NoDefault.no_default,
delimiter=None, header='infer', names=_NoDefault.no_default,
index_col=None, usecols=None, dtype=None, engine=None, converters=None,
true_values=None, false_values=None, skipinitialspace=False,
skiprows=None, skipfooter=0, nrows=None, na_values=None,
keep_default_na=True, na_filter=True, verbose=False,
skip_blank_lines=True, parse_dates=None,
infer_datetime_format=_NoDefault.no_default, keep_date_col=False,
date_parser=_NoDefault.no_default, date_format=None, dayfirst=False,
cache_dates=True, iterator=False, chunksize=None, compression='infer',
thousands=None, decimal='.', lineterminator=None, quotechar='',
quoting=0, doublequote=True, escapechar=None, comment=None, encoding=None,
encoding_errors='strict', dialect=None, on_bad_lines='error',
delim_whitespace=False, low_memory=True, memory_map=False,
float_precision=None, storage_options=None,
dtype_backend=_NoDefault.no_default)
```

[source]

Read a comma-separated values (csv) file into DataFrame.

Also supports optionally iterating or breaking of the file into chunks.

Additional help can be found in the online docs for [IO Tools](#).

Parameters: **filepath_or_buffer** : str, path object or file-like object

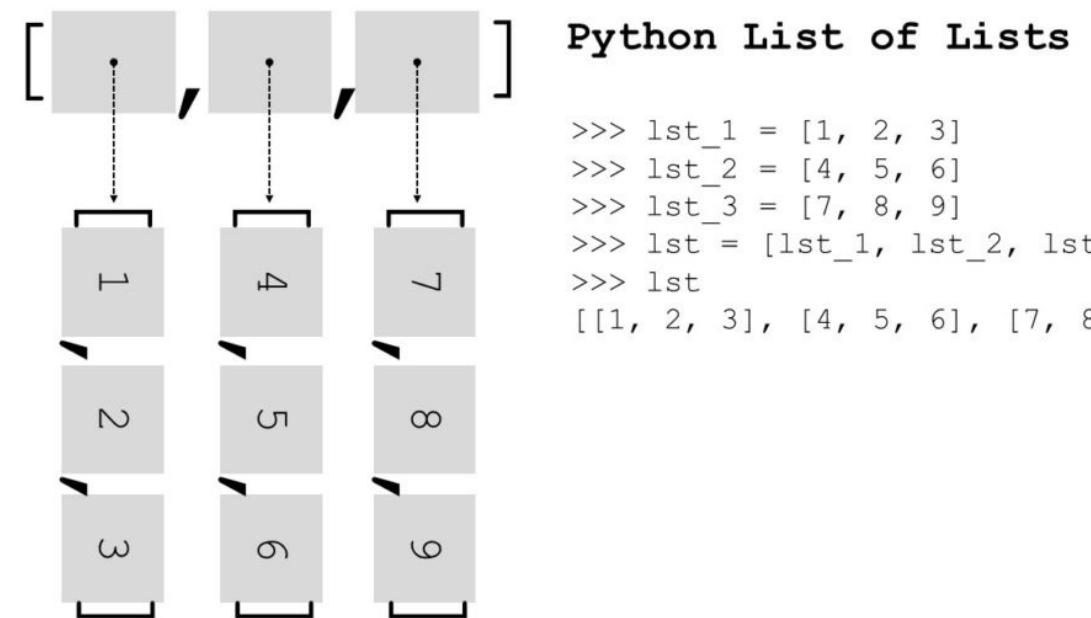
```
1 import pandas as pd
2
3 advertising_dataset_filepath = './Advertising.csv'
4 advertising_df = pd.read_csv(
5     advertising_dataset_filepath,
6     index_col=0
7 )
```

pd.read_csv() returns a DataFrame (a datatable object in Pandas)

DataTable Handling

❖ Convert DataFrame to Python List of Lists

Series		Series		DataFrame	
		apples	oranges	apples	oranges
0	3	0	0	0	0
1	2	1	3	1	3
2	0	2	7	2	7
3	1	3	2	3	2



Currently, we don't often use DataFrame

Convert DataFrame to List of Lists

DataTable Handling

❖ Convert DataFrame to Python List of Lists

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4



```
[[230.1, 37.8, 69.2, 22.1],  
 [44.5, 39.3, 45.1, 10.4],  
 [17.2, 45.9, 69.3, 9.3],  
 [151.5, 41.3, 58.5, 18.5],  
 [180.8, 10.8, 58.4, 12.9],  
 [8.7, 48.9, 75.0, 7.2],  
 [57.5, 32.8, 23.5, 11.8],  
 [120.2, 19.6, 11.6, 13.2],  
 [8.6, 2.1, 1.0, 4.8],  
 [199.8, 2.6, 21.2, 10.6],  
 [66.1, 5.8, 24.2, 8.6],  
 [214.7, 24.0, 4.0, 17.4],  
 [23.8, 35.1, 65.9, 9.2],  
 [97.5, 7.6, 7.2, 9.7],  
 [204.1, 32.9, 46.0, 19.0],  
 [195.4, 47.7, 52.9, 22.4],  
 [67.8, 36.6, 114.0, 12.5],
```

pandas.DataFrame.values

property `DataFrame.values`

[source]

Return a Numpy representation of the DataFrame.

⚠ Warning

We recommend using `DataFrame.to_numpy()` instead.

Only the values in the DataFrame will be returned, the axes labels will be removed.

Returns: `numpy.ndarray`

The values of the DataFrame.

```
1 advertising_lst = advertising_df.values.tolist()
```

DataTable Handling

❖ Find sum of columns

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

```
1 tv_sum = sum([lst[0] for lst in advertising_lst])
2 radio_sum = sum([lst[1] for lst in advertising_lst])
3 newspaper_sum = sum([lst[2] for lst in advertising_lst])
4 sales_sum = sum([lst[3] for lst in advertising_lst])
5 print('TV Sum: ', tv_sum)
6 print('Radio Sum: ', radio_sum)
7 print('Newspaper Sum: ', newspaper_sum)
8 print('Sales Sum: ', sales_sum)
```

TV Sum: 29408.499999999996
Radio Sum: 4652.800000000005
Newspaper Sum: 6110.799999999999
Sales Sum: 2804.500000000005

DataTable Handling

❖ Find sum of columns from row 2th to row 11th

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

L[start:stop:step]

Start position End position The increment

```
1 start_idx = 1
2 end_idx = 11
3 sub_lst = advertising_lst[start_idx:end_idx]
4 tv_sum = sum([lst[0] for lst in sub_lst])
5 radio_sum = sum([lst[1] for lst in sub_lst])
6 newspaper_sum = sum([lst[2] for lst in sub_lst])
7 sales_sum = sum([lst[3] for lst in sub_lst])
8 print('TV Sum: ', tv_sum)
9 print('Radio Sum: ', radio_sum)
10 print('Newspaper Sum: ', newspaper_sum)
11 print('Sales Sum: ', sales_sum)
```

TV Sum: 854.9
Radio Sum: 249.1
Newspaper Sum: 387.8
Sales Sum: 107.3

DataTable Handling

❖ Find min and max

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

```
1 tv_lst = [lst[0] for lst in advertising_lst]
2 radio_lst = [lst[1] for lst in advertising_lst]
3 newspaper_lst = [lst[2] for lst in advertising_lst]
4 sales_lst = [lst[3] for lst in advertising_lst]
5
6 tv_min, tv_max = min(tv_lst), max(tv_lst)
7 radio_min, radio_max = min(radio_lst), max(radio_lst)
8 newspaper_min, newspaper_max = min(newspaper_lst), max(newspaper_lst)
9 sales_min, sales_max = min(sales_lst), max(sales_lst)
10
11 print(f'TV min: {tv_min}, max: {tv_max}')
12 print(f'Radio min: {radio_min}, max: {radio_max}')
13 print(f'Newspaper min: {newspaper_min}, max: {newspaper_max}')
14 print(f'Sales min: {sales_min}, max: {sales_max}')
```

TV min: 0.7, max: 296.4
Radio min: 0.0, max: 49.6
Newspaper min: 0.3, max: 114.0
Sales min: 1.6, max: 27.0

DataTable Handling

❖ Find average

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

$$\text{average}(lst) = \frac{\text{sum}(lst)}{\text{length}(lst)}$$

```
1 tv_lst = [lst[0] for lst in advertising_lst]
2 radio_lst = [lst[1] for lst in advertising_lst]
3 newspaper_lst = [lst[2] for lst in advertising_lst]
4 sales_lst = [lst[3] for lst in advertising_lst]
5
6 tv_avg = sum(tv_lst) / len(tv_lst)
7 radio_avg = sum(radio_lst) / len(radio_lst)
8 newspaper_avg = sum(newspaper_lst) / len(newspaper_lst)
9 sales_avg = sum(sales_lst) / len(sales_lst)
10
11 print(f'TV Average: {tv_avg}')
12 print(f'Radio Average: {radio_avg}')
13 print(f'Newspaper Average: {newspaper_avg}')
14 print(f'Sales Average: {sales_avg}')
```

TV Average: 147.0425
Radio Average: 23.26400000000024
Newspaper Average: 30.553999999999995
Sales Average: 14.022500000000003

DataTable Handling

❖ Find median

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$
= **4.5**

```
1 def median(lst):
2     sorted_lst = sorted(lst)
3
4     n = len(lst)
5     mid = n // 2
6     if n % 2 == 0:
7         y = (sorted_lst[mid] + sorted_lst[mid - 1]) / 2
8     else:
9         y = sorted_lst[mid]
10
11    return y
```

- **If even:** $mid(lst) = lst\left[\frac{\text{length}(lst)}{2}\right]$
- **If odd:** $mid(lst) = \frac{lst\left[\frac{\text{length}(lst)}{2}\right] + lst\left[\frac{\text{length}(lst)}{2} - 1\right]}{2}$

DataTable Handling

❖ Find median

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

```
18 tv_median = median(tv_lst)
19 radio_median = median(radio_lst)
20 newspaper_median = median(newspaper_lst)
21 sales_median = median(sales_lst)
22
23 print(f'TV Median: {tv_median}')
24 print(f'Radio Median: {radio_median}')
25 print(f'Newspaper Median: {newspaper_median}')
26 print(f'Sales Median: {sales_median}')
```

```
TV Median: 149.75
Radio Median: 22.9
Newspaper Median: 25.75
Sales Median: 12.9
```

DataTable Handling

❖ Find contribution percentage

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

200 rows × 4 columns

```
1 tv_lst = [lst[0] for lst in advertising_lst]
2 radio_lst = [lst[1] for lst in advertising_lst]
3 newspaper_lst = [lst[2] for lst in advertising_lst]
4
5 total_budget = sum(tv_lst) + sum(radio_lst) + sum(newspaper_lst)
6
7 tv_percentage = (sum(tv_lst) / total_budget) * 100
8 radio_percentage = (sum(radio_lst) / total_budget) * 100
9 newspaper_percentage = (sum(newspaper_lst) / total_budget) * 100
10
11 print(f'TV Percentage: {tv_percentage:.3f}%')
12 print(f'Radio Percentage: {radio_percentage:.3f}%')
13 print(f'Newspaper Percentage: {newspaper_percentage:.3f}%')
```

TV Percentage: 73.206%
Radio Percentage: 11.582%
Newspaper Percentage: 15.212%

Outline

- Introduction
- DataTable Handling
- Text Representation
- Data Crawling
- Question

Text Representation

❖ Introduction

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt. Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.

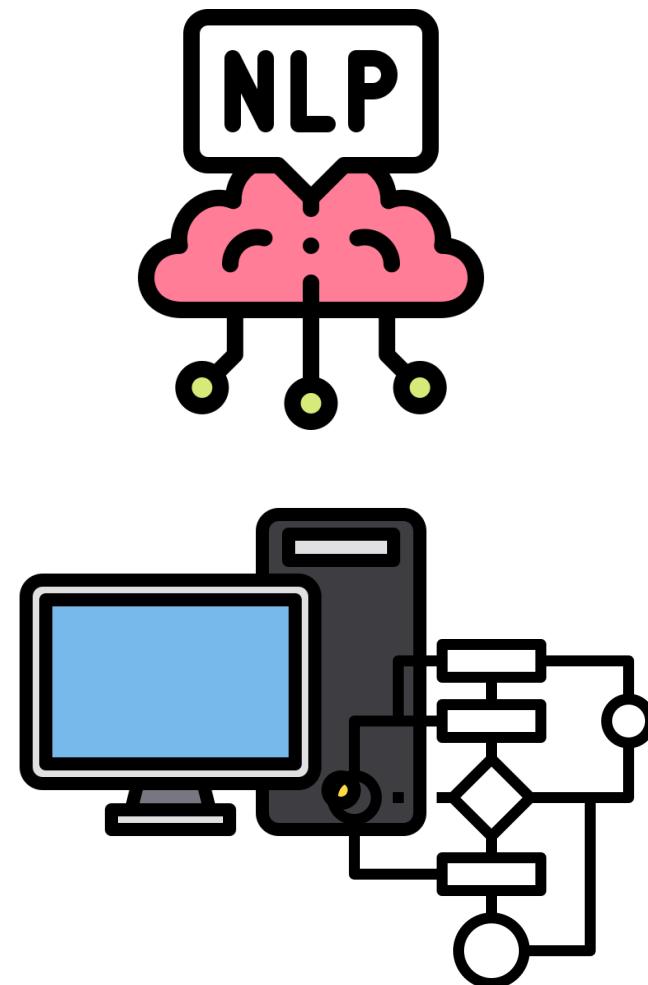
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh là một trung tâm đào tạo cán bộ kỹ thuật công nghệ và các nhà quản lý có trình độ ngang tầm với các nước tiên tiến trong khu vực Đông nam Á, đáp ứng nguồn nhân lực có chất lượng cao cho sự nghiệp công nghiệp hóa và hiện đại hóa đất nước cũng như khu vực phía Nam.

In general, we don't often use raw text when computing in some tasks.

Need a better representation for text.

Trời hôm nay đẹp quá

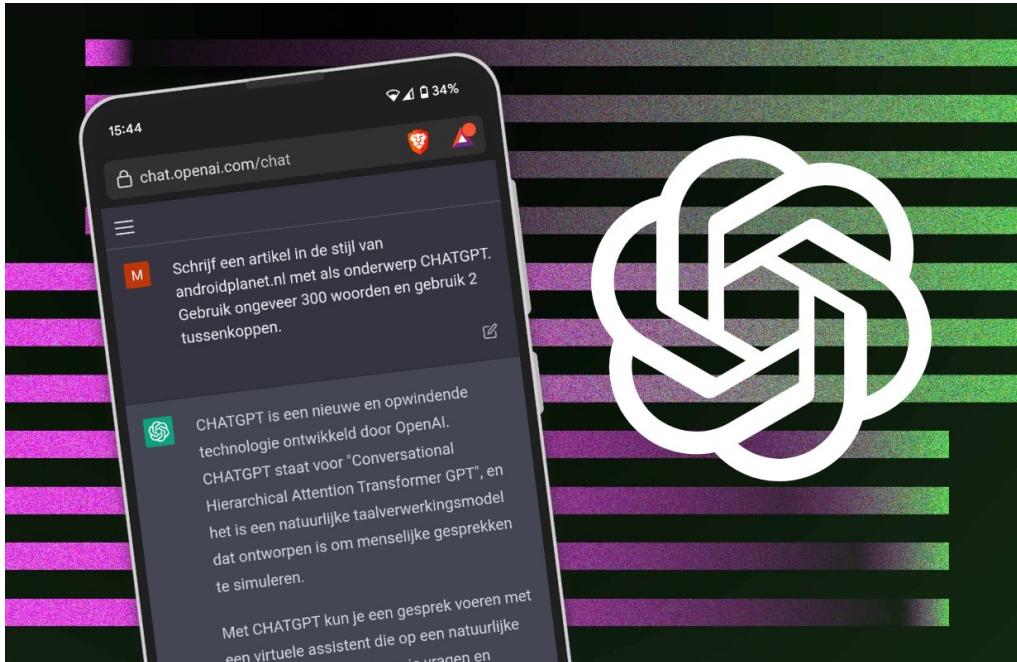
14	2	9	36	89
----	---	---	----	----



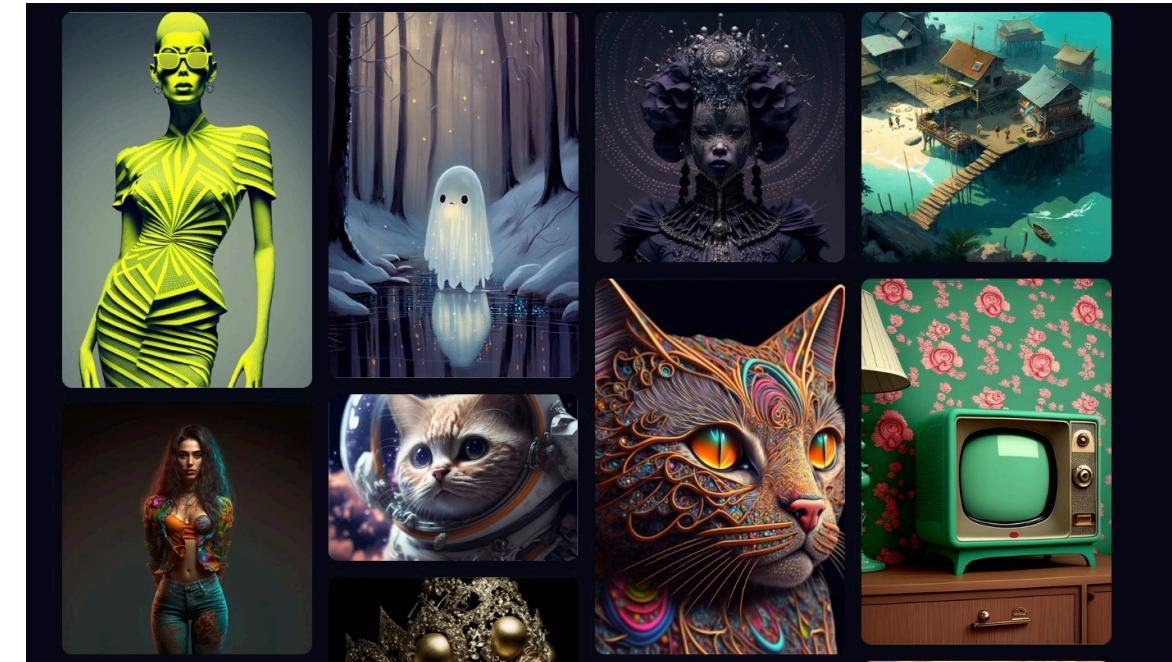
Computer/Algorithm might be more efficient

Text Representation

❖ Introduction



ChatGPT (Chatbot)

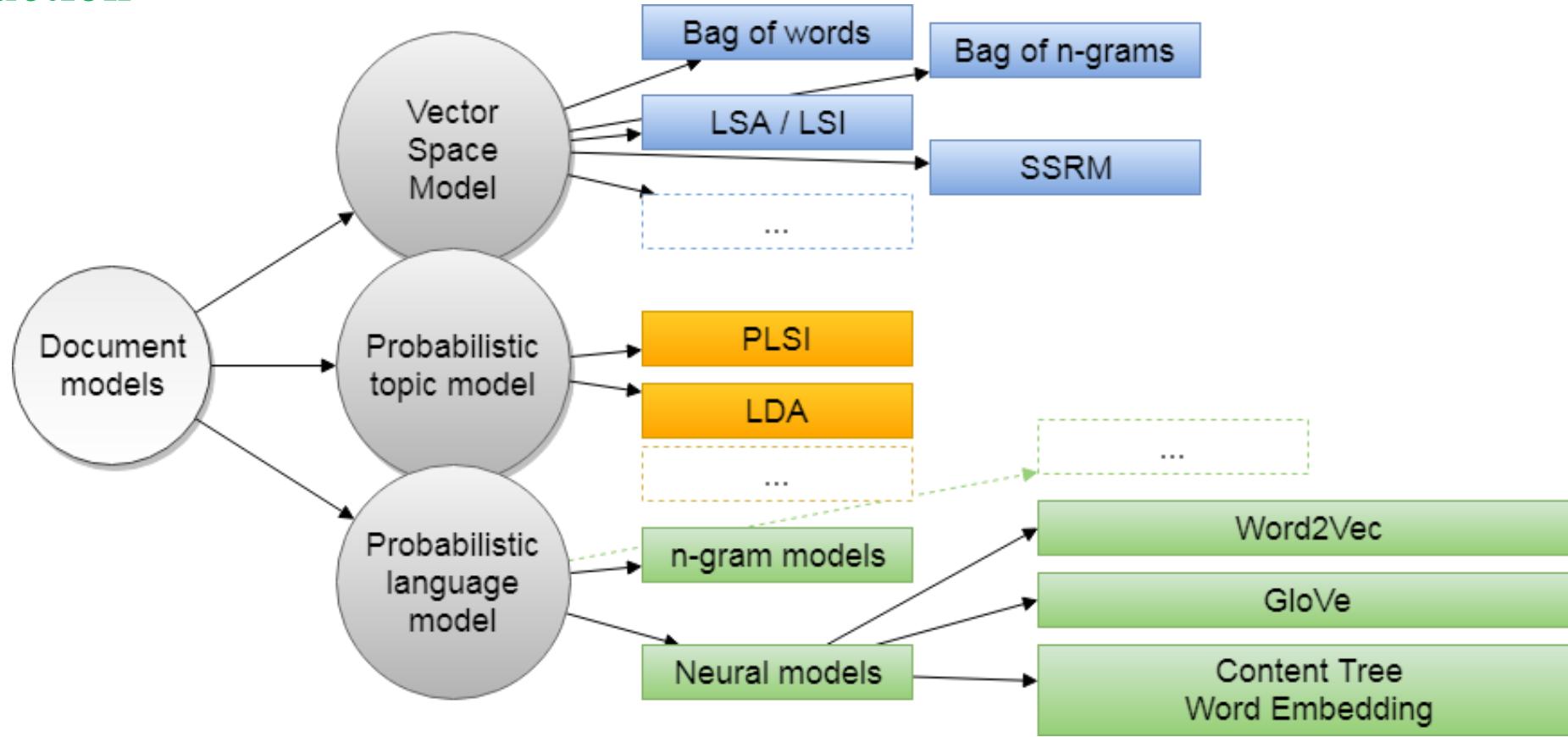


Midjourney (text2image application)

These successful text-related applications have their unique way to represent input text

Text Representation

❖ Introduction



Text Representation Category

Text Representation

❖ Introduction

Dictionary

Word	Index
xin	0
chào	1
tất	2
cả	3
...	...
word _n	n - 1



Trời hôm nay đẹp quá

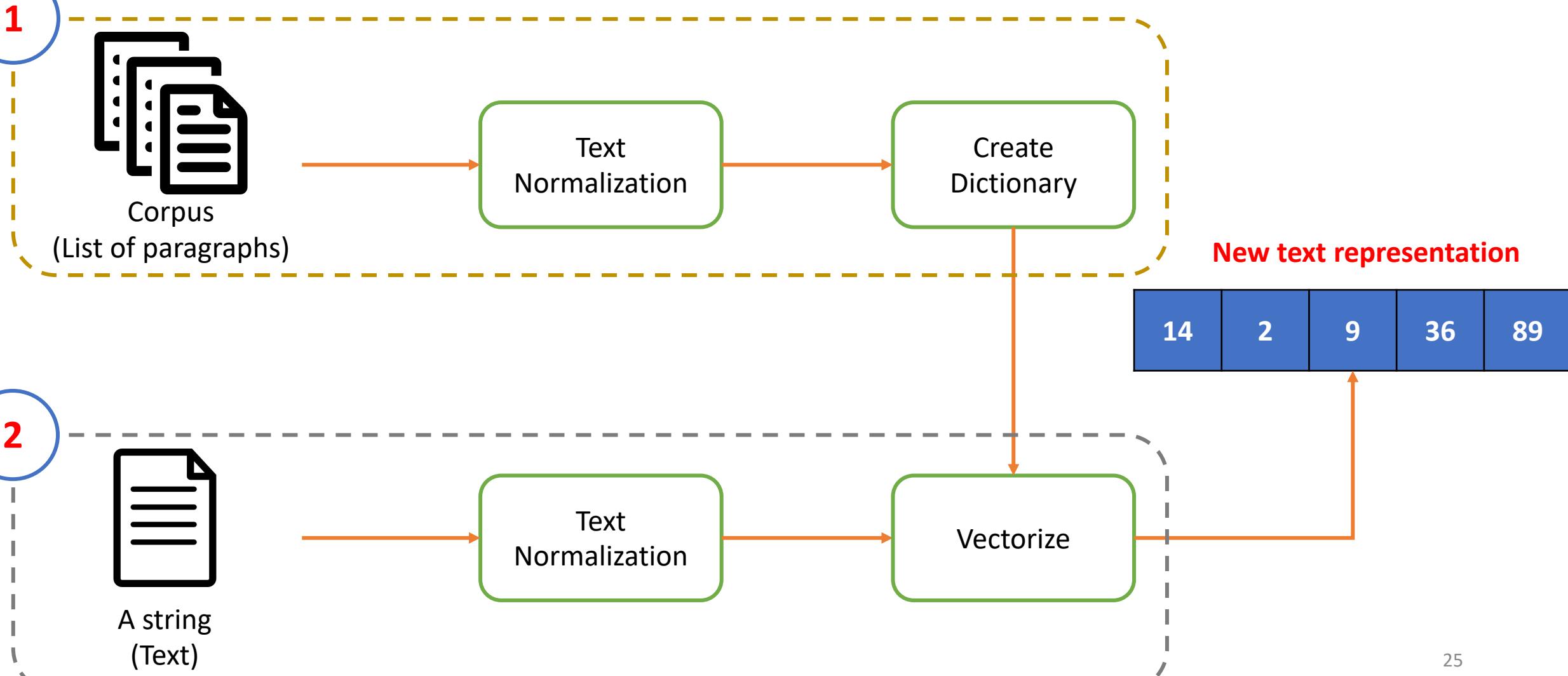
Trời hôm nay đẹp quá

14	2	9	36	89
----	---	---	----	----

Index-based Encoding

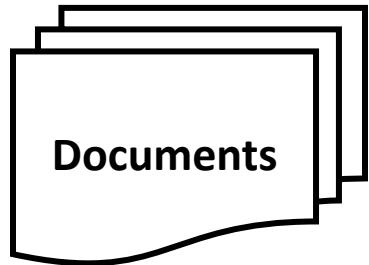
Text Representation

❖ Index-based Encoding Pipeline



Text Representation

❖ Text Normalization Introduction



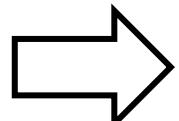
Document ID	Content
d1	Hello, we are learning artificial intelligence.
d2	tHIs iS a pRObLEM iN TexT rEPrEsEnTAtion
d3	#science?! <artificial intelligence> #deep learning!!!

Problem:

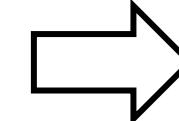
1. Documents contain unnecessary string (information).
2. Not well-represent natural language.

Input Text

'#ArTiFlciAL.
In?>TeLLi!g@ENce'



Text
Normalization



Output Text

'artificial intelligence'

Text Representation

❖ Text Normalization: Lowercasing

```
1 import string
2
3 remove_characters = '\t"' + string.punctuation
4 def text_normalize(text):
5     text = text.lower()
6     text = text.strip()
7     text = text.replace('\n', ' ')
8     for char in remove_characters:
9         text = text.replace(char, '')
10
11 return text
```

Lowercasing

“Hello we’re AIVN”

“hello we’re aivn”

Convert the given text to lowercase

Text Representation

❖ Text Normalization: Punctuation Removal

```
1 import string
2
3 remove_characters = '\t"' + string.punctuation
4 def text_normalize(text):
5     text = text.lower()
6     text = text.strip()
7     text = text.replace('\n', ' ')
8     for char in remove_characters:
9         text = text.replace(char, '')
10
11 return text
```

Punctuations Removal

Remove all punctuations in words

“Hello, welcome to AIVN.”

“Hello welcome to AIVN”

Text Representation

❖ Create Dictionary

1 dictionary

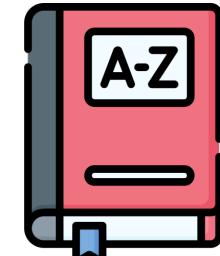
['Bế mạc Hội nghị Trung ương giữa nhiệm kỳ\n\nSáng 17/5, Hội nghị Trung ương 7 khóa XIII (Hội nghị Trung ương giữa nhiệm kỳ) họp
việc.\n\nHội nghị Trung ương 7 khóa XIII (Hội nghị Trung ương giữa nhiệm kỳ) chính thức khai mạc sáng 15/5 tại Thủ đô Hà Nội và c
rất quan trọng đối với việc hoàn thành thắng lợi Nghị quyết Đại hội XIII của Đảng; là dịp để nhìn lại và đánh giá một cách khách
đầu nhiệm kỳ đến nay, đồng thời chỉ ra những hạn chế, yếu kém còn tồn tại, nguyên nhân và bài học kinh nghiệm; dự báo bối cảnh tì
thách thức đan xen, để từ đó đề ra những chủ trương, quyết sách lớn cần phải tập trung lãnh đạo, chỉ đạo thực hiện trong nửa cuối

'Điều tra việc giả mạo văn bản UBND tỉnh Khánh Hòa cho thi công dự án\n\nChính quyền Khánh Hòa yêu cầu các đơn vị liên quan tror
thuận chủ trương thi công dự án, đang lan truyền trên mạng.\n\nTối 17/5, ông Nguyễn Thanh Hà, Chánh Văn phòng UBND tỉnh Khánh Hòa
án Ocean Hills đang lan truyền trên mạng xã hội là giả mạo. Văn bản giả ghi số 982 ngày 1/3/2023 của UBND tỉnh Khánh Hòa về việc
giả mạo chữ ký của Phó chủ tịch UBND tỉnh Khánh Hòa ông Đinh Văn Thiệu. Trong văn bản giả mạo có nội dung: "Qua xem xét báo cáo,
đoàn INCO về báo cáo tình trạng thực hiện dự án Ocean Hills sau khi chuyển giao về công ty Cổ phần tập đoàn INCO...UBND tỉnh có ch
ngành liên quan hỗ trợ tham mưu để dự án triển khai thi công". Ông Nguyễn Thanh Hà khẳng định, văn bản trên là giả mạo, không phâ
an tinh, Sở Thông tin và Truyền thông, cùng các sở ngành liên quan kiểm tra, xử lý. Đồng thời, UBND tỉnh chỉ đạo các sở, ban ngà
đăng tải, chia sẻ văn bản giả mạo trên các trang mạng xã hội; kịp thời thông báo nếu phát hiện cán bộ, công chức, viên chức, ngư
theo đúng quy định. \n\nXuân Ngọc',

'Gây tai nạn cho cụ ông 82 tuổi, tài xế ở Hải Dương 2 lần bỏ trốn\n\nSau khi va chạm với ông M., tài xế Bang lái xe bỏ trốn như
lại bỏ trốn tiếp. Sau đó, ông M, tử vong.\n\nCông an thị xã Kinh Môn, tỉnh Hải Dương sáng nay (19/5) thông tin, đơn vị đã điều tr
82 tuổi tử vong rồi bỏ trốn. Theo đó, khoảng 6h sáng ngày 17/5, tại ngã ba đường giao nhau giữa tỉnh lộ 389 với đường về xã Lạc I
ông Nguyễn Thanh M., 82 tuổi, trú tại thôn Phương Quất, xã Lạc Long, thị xã Kinh Môn. Thời điểm xảy ra tai nạn, ông Mai đi xe đạp
M. ngã ra đường. Tài xế xe tải liên quan đến vụ tai nạn không dừng lại đưa người bị nạn đi cấp cứu mà bỏ chạy về phía cầu Triều.



Given a list of paragraphs



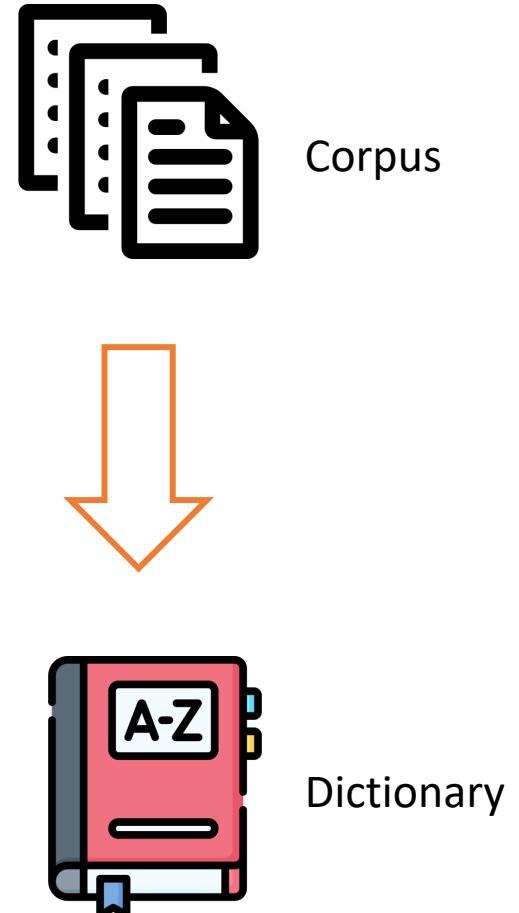
Get a list of unique words

['bê',
'mạc',
'hội',
'nghị',
'trung',
'ương',
'giữa',
'nhiệm',
'kỳ',
'sáng',
'175',
'7',
'khóa',
'xiii',
'hợp',
'phiên',
'sau',
'khi',
'hoàn',
'thành',
'chương',
'trình',
'làm',
'việc',
]

Text Representation

❖ Create Dictionary

```
13 def create_dictionary(corpus):  
14     dictionary = []  
15     for paragraph in corpus:  
16         paragraph = text_normalize(paragraph)  
17         tokens = paragraph.split()  
18         for token in tokens:  
19             if token not in dictionary:  
20                 dictionary.append(token)  
21  
22     return dictionary
```



Text Representation

❖ Create New Text Representation

```
24 def vectorize(  
25     text,  
26     dictionary,  
27     unknown_token_id  
28 ):  
29     text = text_normalize(text)  
30     tokens = text.split()  
31     vector = [  
32         dictionary.index(token) \  
33         if token in dictionary else unknown_token_id \  
34         for token in tokens  
35     ]  
36  
37     return vector
```

Trời hôm nay đẹp quá

Trời hôm nay đẹp quá

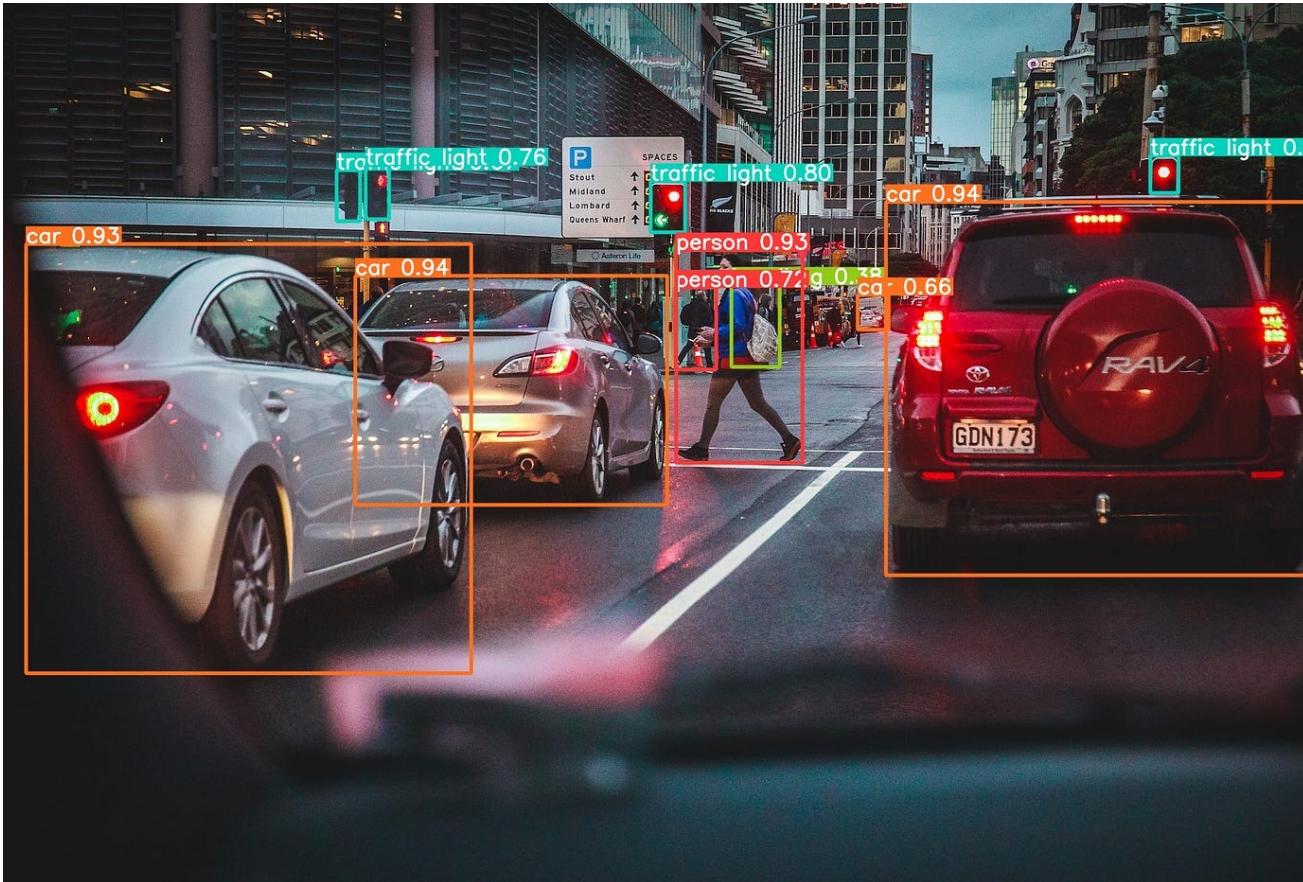
14	2	9	36	89
----	---	---	----	----

Outline

- Introduction
- DataTable Handling
- Text Representation
- Data Crawling
- Question

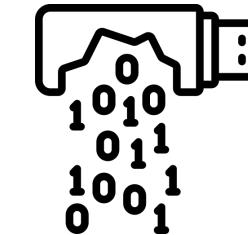
Data Crawling

❖ Introduction

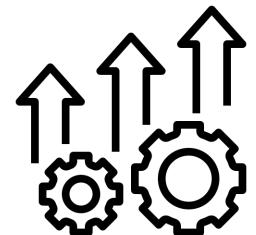
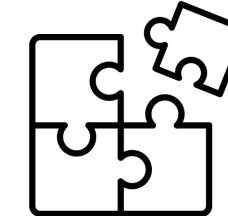


In Project 1, we need dataset to train YOLOv8

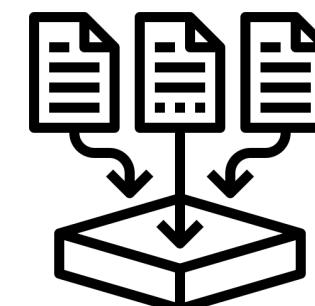
Problems:



Not enough data



Solution:



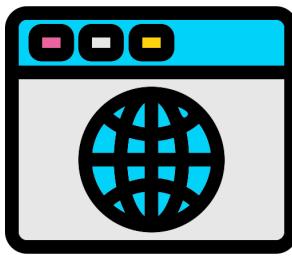
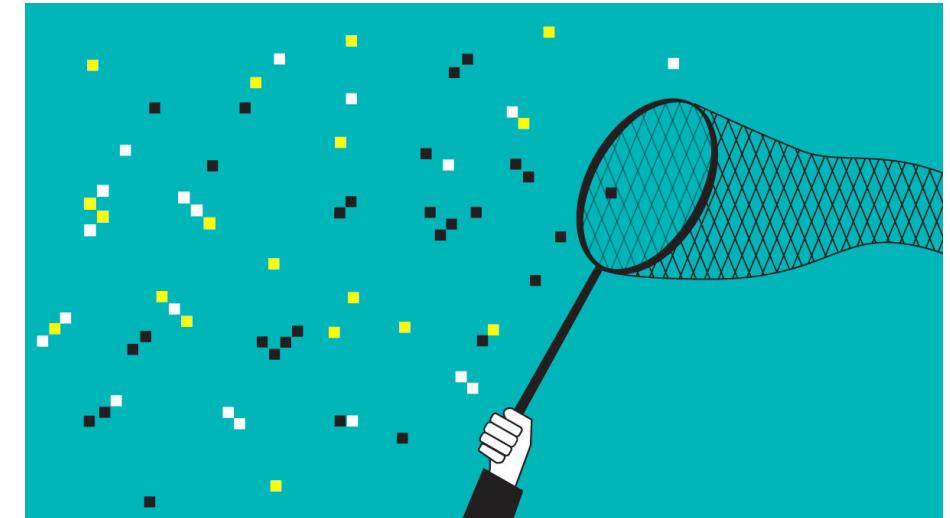
Data Crawling

❖ Introduction

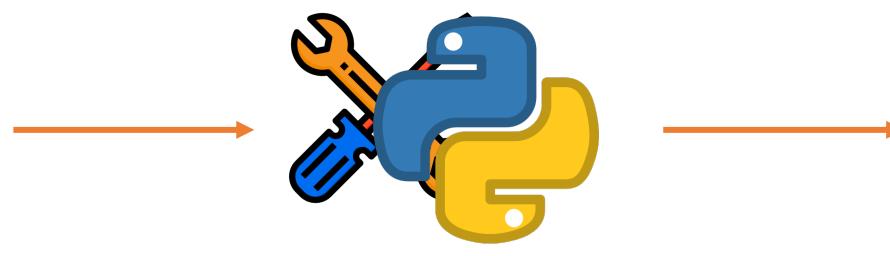


Data on Internet is huge!

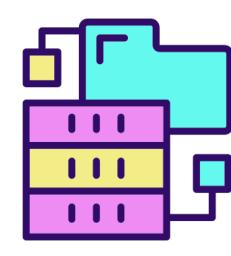
Is there a way to collect them all?



Webpages



Data Crawling



Structured Data

Data Crawling: Tools/Programs that collecting data from webpages

Data Crawling

❖ Motivation



The screenshot shows the homepage of Vietnamnet (vietnamnet.vn). At the top, there's a navigation bar with links like 'Podcast', 'Tin nóng', 'Tin tức 24h', 'PREMIUM*', 'Đăng nhập', and a search icon. Below the navigation is a banner for 'DÒNG CHẢY THÔNG TIN CHÍNH XÁC - TIN CẬY' (Flow of accurate information - reliable news). The main content area features a large image of a man speaking at a podium with a microphone, followed by several news thumbnails. One thumbnail on the right discusses 'TikTok Shop' and its relationship with Shopee and Lazada. Another thumbnail shows flags of the European Union in front of a building. A sidebar on the left contains a headline about digital signatures.

Chữ ký số cá nhân là một yếu tố đặc trưng của công dân số

Thứ trưởng Bộ TT&TT Nguyễn Huy Dũng cho rằng, một công dân số sẽ cần có 8 yếu tố đặc trưng, trong đó yếu tố đặc trưng quan trọng nhất là mỗi người dân có một chữ ký số.

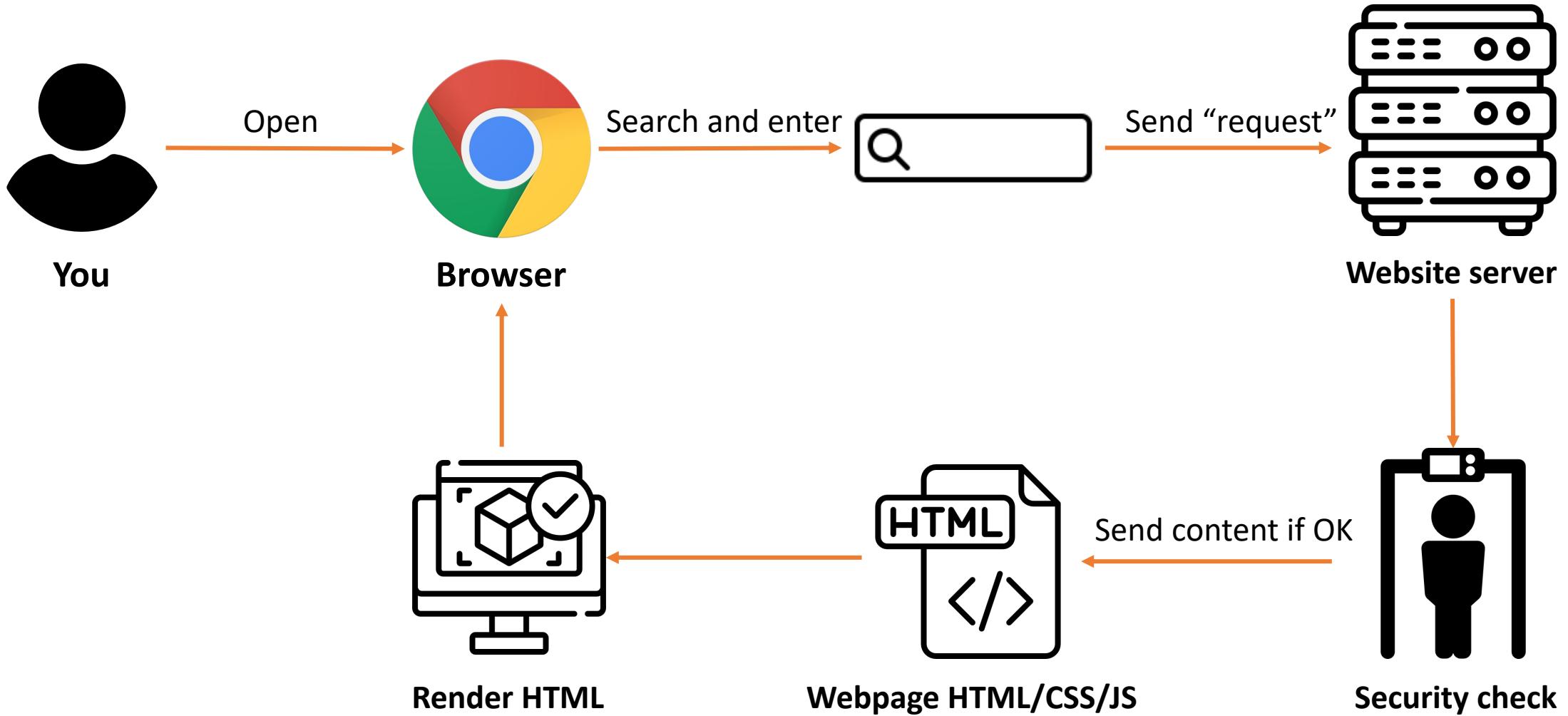
OpenAI sẽ không rút khỏi EU, treo thưởng 1 triệu USD cho sáng kiến quản lý AI

A webpage's content is represent in something called **HTML**

```
...<!DOCTYPE html> == $0
<html lang="vi" translate="no">
  <head>(...)</head>
  <body>
    <!-- BEGIN COMPONENT:: COMPONENT002199 -->
    <noscript>(...)</noscript>
    <!-- End Google Tag Manager (noscript) -->
    <!-- END COMPONENT:: COMPONENT002199 -->
    <!-- BEGIN COMPONENT:: COMPONENT002461 -->
    <!-- END COMPONENT:: COMPONENT002461 -->
    <div class="wrapper">(...)</div>
    <script vnn="vnnjs" type="text/javascript" src="https://res-files.vnncdn.net/files/public/2023/5/22/js-9b31d12...-desktop-41.js?v=1"></script>
    <script type="text/javascript" src="https://res-files.vnncdn.net/files/2023/5/22/vnnconfigate.js"></script>
    <script defer>(...)</script>
    <!-- BEGIN COMPONENT:: COMPONENT002475 -->
    <script type="text/javascript" src="https://res-files.vnncdn.net/files/2022/6/20/vnnvotemodulemobile.js"></script>
    <script defer>(...)</script>
    <!-- END COMPONENT:: COMPONENT002475 -->
    <!-- BEGIN COMPONENT:: COMPONENT002466 -->
    <script type="text/javascript" src="https://comment.vietnamnet.vn/js/vnnidmodule.js"></script>
    <!-- script tracking -->
    <script>(...)</script>
    <script>(...)</script>
    <iframe width="0" height="0" src="https://mic.gov.vn/Pages/PageEmbed/Vietnamnet/Default.aspx" marginwidth="0" allowtransparency="true" marginheight="0" hspace="0" vspace="0" frameborder="0" scrolling="no">(...)</iframe>
    <!-- END COMPONENT:: COMPONENT002466 -->
    <script type="text/javascript" src="https://vnn-res.vgcloud.vn/VietNamNet/Standard/js/vnn_trackerv3-0-3.js"></script>
    <script type="text/javascript" id>(...)</script>
    <noscript>(...)</noscript>
    <div>(...)</div>
    <div></div>
    ...
```

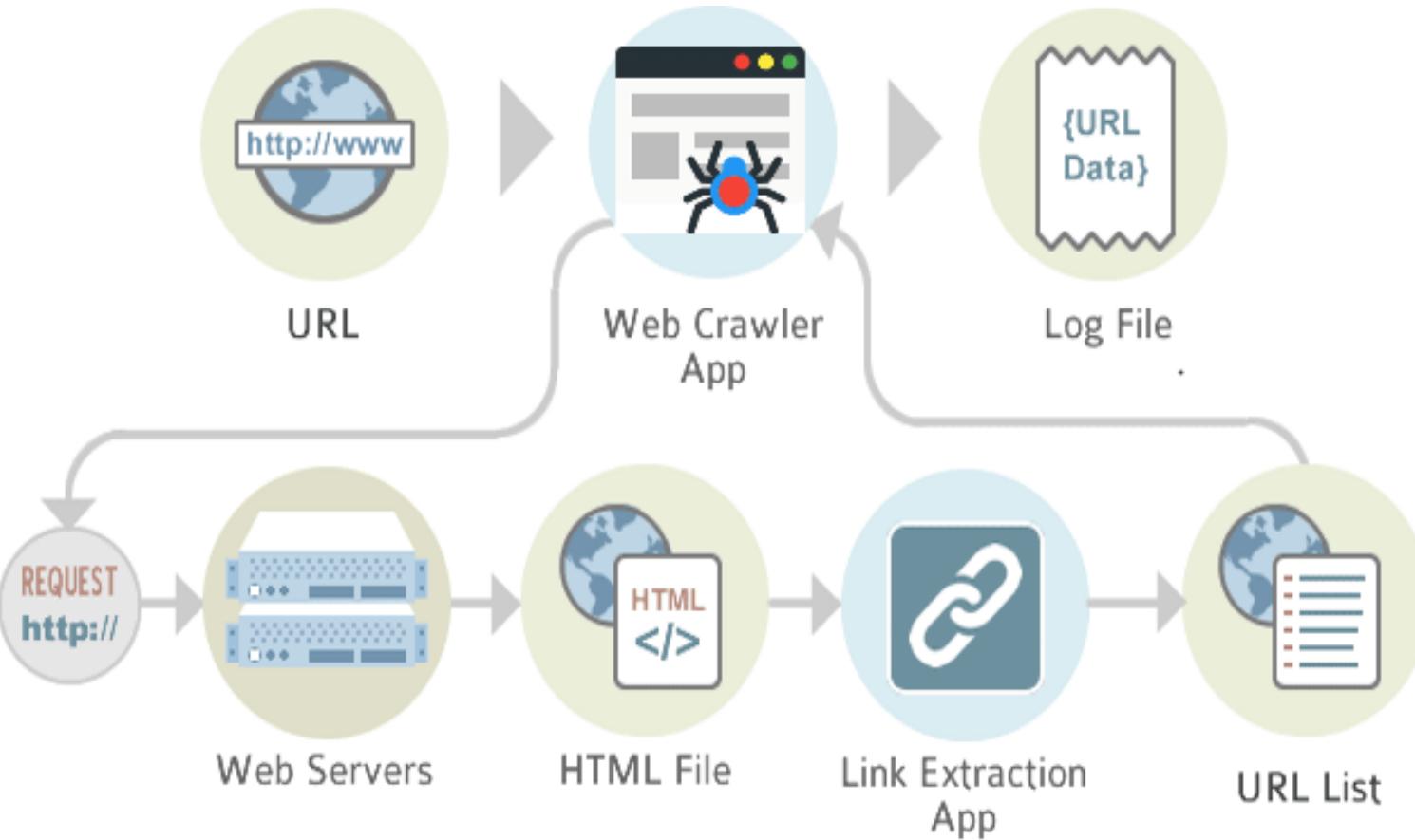
Data Crawling

❖ Motivation



Data Crawling

❖ Idea: General steps to crawling



A pipeline of crawling data

Data Crawling

❖ About HTML

HTML



HyperText Markup Language (HTML): The standard markup language for documents designed to be displayed in a web browser.

```
1  <!DOCTYPE html>
2  <html>
3    <head>
4      <meta charset="UTF-8">
5      <title>Title goes here</title>
6    </head>
7    <body>
8
9    </body>
10   </html>
```

Content of HTML



.html file

Data Crawling

❖ About HTML

HTML Page Structure

```
<!DOCTYPE html>      ← Tells version of HTML
<html>      ← HTML Root Element

<head>      ← Used to contain page HTML metadata
  <title>Page Title</title>  ← Title of HTML page
</head>

<body>      ← Hold content of HTML
  <h2>Heading Content</h2>  ← HTML heading tag
  <p>Paragraph Content</p>  ← HTML paragraph tag
</body>

</html>
```

Data Crawling

❖ About HTML: Example

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>Hello</h1>
<p>Xin chào mọi người.</p>

</body>
</html>
```

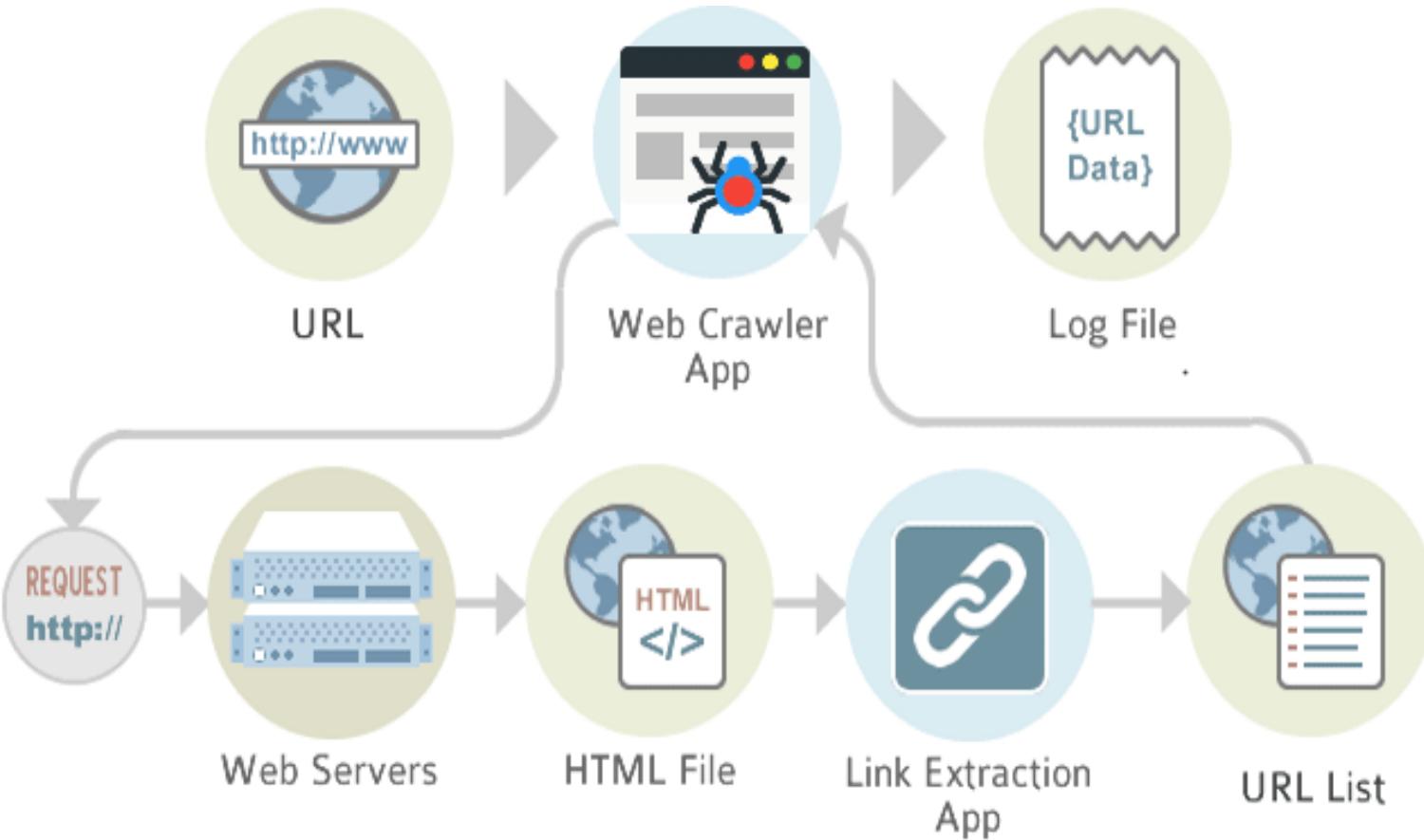
Hello

Xin chào mọi người.

Webpage Interface

Data Crawling

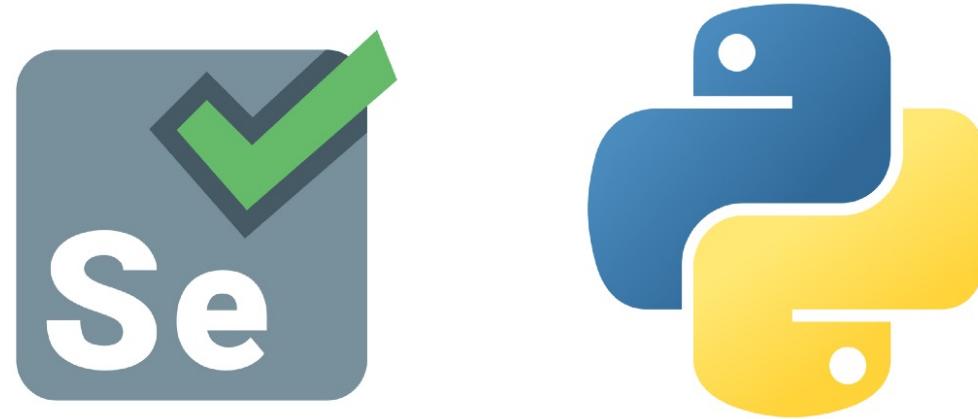
❖ How?



How to implement this in Python?

Data Crawling

❖ Selenium Package



Selenium Package: Used to automate web browser interaction from Python

Data Crawling

❖ Problem Statement

Vietnamnet
VĨ VIỆT NAM HÙNG CƯỜNG

Podcast | Tin nóng | Tin tức 24h | PREMIUM* | Vietnamnet TV | Đăng nhập |

Chính trị Thời sự Kinh doanh Thể thao Giải trí Thế giới Đời sống Giáo dục Sức khỏe Thông tin và Truyền thông Pháp luật Xe Bất động sản Tuần Việt Nam Du lịch Bạn đọc

MỞ TÀI KHOẢN TRỰC TUYẾN

- Nga khẳng định kiểm soát hoàn toàn Bakhmut, đánh chặn nhiều tên lửa Ukraine
- Học phí các trường thuộc ĐH Quốc gia TP.HCM, cao nhất lên tới cả trăm triệu
- Man City vô địch Ngoại hạng Anh: Chờ cú ăn 3 lịch sử
- Thanh Thúy, Đức Thịnh kỷ niệm ngày cưới, tiết lộ bí mật giấu kín 15 năm

* Tài ứng dụng Agribank E-Mobile Banking
* Đăng ký mở tài khoản
* Sử dụng ngay
Tải ngay ứng dụng tại

Thủ tướng thăm tàu vận chuyển hydro lỏng đầu tiên trên thế giới

Thủ tướng Phạm Minh Chính bất ngờ đến thăm cảng Itsukaichi tại Hiroshima, Nhật Bản và trực tiếp ngắm nhìn con tàu "Suiso Frontier" vận chuyển hydro lỏng đầu tiên trên thế giới.

THỜI SỰ

21/05/2023 05:38 (GMT+07:00)

Dự báo thời tiết 21/5: Bắc và Trung Bộ vẫn nắng đổ lửa, chiều tối mưa giông

Bảo Anh
Nhà báo

→ Xem các bài viết của tác giả



Theo dõi VietNamNet trên [Google News](#)

Dự báo thời tiết ngày 21/5, miền Bắc và Trung Bộ vẫn tiếp diễn nắng nóng đến đặc biệt gay gắt, nhiều nơi nhiệt độ vượt ngưỡng 40 độ. Tây Nguyên và Nam Bộ dịu mát dần về chiều tối.

Theo Trung tâm dự báo khí tượng thủy văn quốc gia, hôm nay (21/5), ở phía Đông Bắc Bộ tiếp diễn nắng nóng và nắng nóng gay gắt với nhiệt độ cao nhất 35-38 độ, có nơi trên 38 độ; độ ẩm tương đối thấp nhất 40-60%.

Phía Tây Bắc Bộ và khu vực từ Thanh Hóa đến Phú Yên có nắng nóng gay gắt, có nơi đặc biệt gay gắt với nhiệt độ cao nhất 37-40 độ, có nơi trên 40 độ; độ ẩm tương đối thấp nhất 30-55%.

Problem Statement: Collect articles from vietnamenet.vn

Data Crawling

❖ Problem Statement

THỜI SỰ

21/05/2023 05:38 (GMT+07:00)

Dự báo thời tiết 21/5: Bắc và Trung Bộ vẫn nắng đổ lửa, chiều tối mưa giông



Bảo Anh

Nhà báo

→ Xem các bài viết của tác giả

Article



f



Zalo



✉



🔗



🔖

Theo dõi VietNamNet trên Google News

Dự báo thời tiết ngày 21/5, miền Bắc và Trung Bộ vẫn tiếp diễn nắng nóng đặc biệt gay gắt, nhiều nơi nhiệt độ vượt ngưỡng 40 độ. Tây Nguyên và Nam Bộ dịu mát dần về chiều tối.

Author

Theo Trung tâm dự báo khí tượng thủy văn quốc gia, hôm nay (21/5), ở phía Đông Bắc Bộ tiếp diễn nắng nóng và nắng nóng gay gắt với nhiệt độ cao nhất 35-38 độ, có nơi trên 38 độ; độ ẩm tương đối thấp nhất 40-60%.

Abstract

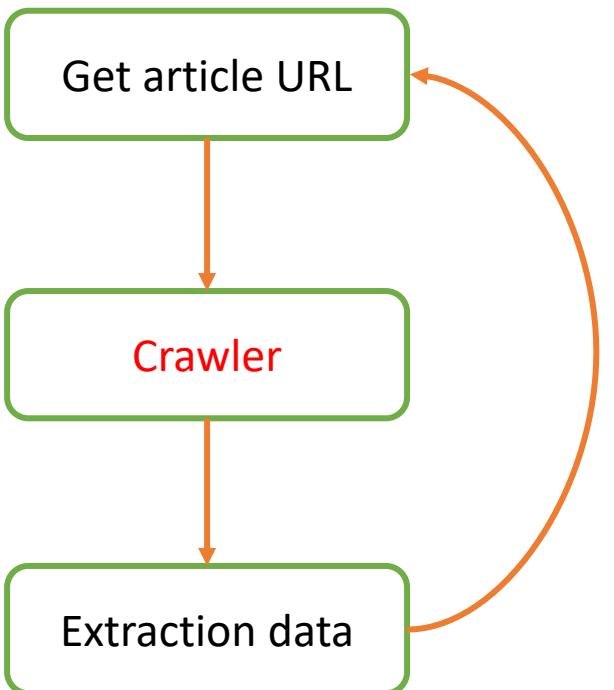
Phía Tây Bắc Bộ và khu vực từ Thanh Hóa đến Phú Yên có nắng nóng gay gắt, có nơi đặc biệt gay gắt với nhiệt độ cao nhất 37-40 độ, có nơi trên 40 độ; độ ẩm tương đối thấp nhất 30-55%.

Body

Things to extract in an article

Data Crawling

❖ Problem Statement



How to automatically access to many articles?

vietnamnet.vn/thoi-su-page1

Chính trị Thời sự Kinh doanh Thể thao Giải trí Thế giới Đời sống Giáo dục Sức khỏe Thông tin và Truyền thông Phá

Hà Nội: Cháy trong khu dân cư, khói đen bốc cao hàng chục mét
Lán tạm rộng khoảng 30m2 nằm sâu trong ngõ số 9 phố Lương Đình Của (Hà Nội) bắt ngờ bốc cháy, cột khói đen bốc cao hàng chục mét.

THỜI SỰ
Hà Nam: Công an xã là 'lá chắn thép' bảo vệ bình yên của Nhân dân
Lực lượng Công an xã chí lính chính trị, trình độ chuyên môn cao, là "lá chắn" vững chắc bảo đảm

THỜI SỰ
Doanh nghiệp làm kinh doanh ở Hà Nội yêu cầu 'nóng'
Ông Trần Sỹ Thành yêu cầu Sở Tư pháp phải xem xét việc doanh nghiệp, tổ chức, yêu cầu người lao động phải có phiếu lý lịch tư pháp mới sau 6 tháng có đúng luật hay không? Nếu không thì xử lý doanh nghiệp như thế nào?

THỜI SỰ
Dự báo thời tiết 19/5: Nắng nóng ở miền Bắc hạ xuống dưới 40 độ
Dự báo thời tiết ngày 19/5, miền Bắc và Trung Bộ tiếp diễn nắng nóng gay gắt diện rộng, nhưng nhiệt độ hạ nhẹ ở Bắc Bộ. Tây Nguyên và Nam Bộ ngày nắng, chiều tối mưa mát.

< 1 2 3 4 5 6 >

Xem tin theo ngày

Data Crawling

❖ Selenium Example

The screenshot shows the Python.org homepage. At the top, there's a navigation bar with links for Python, PSF, Docs, PyPI, Jobs, and Community. Below the header is the Python logo and a search bar. The main content area features a code snippet in a terminal window:

```
# Simple output (with Unicode)
>>> print("Hello, I'm Python!")
Hello, I'm Python!
# Input, assignment
>>> name = input('What is your name?\n')
What is your name?
Python
>>> print(f'Hi, {name}.')
Hi, Python.
```

Next to the code is a yellow button with a right-pointing arrow. To the right of the code, there's a section titled "Quick & Easy to Learn" with text about Python's learnability. Below the code and the "Quick & Easy to Learn" section is a green callout box containing the text: "Python is a programming language that lets you work quickly and integrate systems more effectively. [»» Learn More](#)". An orange arrow points from this callout down to another green callout at the bottom of the page.

Get Started
Whether you're new to programming or an experienced developer, it's easy to learn and use Python.
[Start with our Beginner's Guide](#)

Download
Python source code and installers are available for download for all versions!
Latest: [Python 3.11.3](#)

Docs
Documentation for Python's standard library, along with tutorials and guides, are available online.
[docs.python.org](#)

Jobs
Looking for work or have a Python related position that you're trying to hire for? Our [relaunched community-run job board](#) is the place to go.

Access to python.org and get this text

Data Crawling

❖ Selenium Example

Step 1: Initialize a browser and access to the website

```
chrome_options = webdriver.ChromeOptions()  
chrome_options.add_argument("start-maximized")  
chrome_options.add_argument("--disable-dev-shm-usage")  
  
driver = webdriver.Chrome(  
    service=Service(ChromeDriverManager().install()),  
    options=chrome_options  
)  
  
url = 'https://www.python.org/'  
driver.get(url)
```

Initialize a Google Chrome browser (webdriver)

driver.get(): Request (access) to a webpage given URL

Data Crawling

❖ Selenium Example

```
driver = webdriver.Chrome(  
    service=Service(ChromeDriverManager().install()),  
    options=chrome_options  
)  
  
url = 'https://www.python.org/'  
driver.get(url)  
print(driver.page_source)
```

```
<content>a>img,[data-ea-publisher]:not([data-ea-type]).loaded.horizontal .ea-content>a>img,,ea-type-image.horizontal .ea-content .ea-text,[data-ea-publisher]:not([data-ea-type]).loaded.horizontal .ea-content .ea-text,,ea-t  
uto}[data-ea-type="image"].loaded.horizontal .ea-callout,[data-ea-publisher]:not([data-ea-type]).loaded.horizontal t}[data-ea-type="text"].loaded,.ea-type-text{font-size:14px}[data-ea-type="text"].loaded .ea-content,.ea-type-text .ea-callout{margin:0.5em 1em 1em 1em;padding-left:1em;padding-right:1em;text-align:right;font-style:italic}[dat  
px;right:20px][data-ea-style="stickybox"].loaded .ea-type-image .ea-stickybox-hide{cursor:pointer;position:absolu  
-radius:50%;color:#088cdb;font-size:1em;text-align:center;height:1.5em;width:1.5em;line-height:1.5em}@media (max-  
tton:0;right:0;margin:auto;text-align:center}[data-ea-style="stickybox"].loaded .ea-type-image .ea-stickybox-hide  
pe-image .ea-content{background:#cdcdc}[data-ea-style="stickybox"].loaded.dark .ea-type-image .ea-content{backgr  
style="stickybox"].loaded.adaptive .ea-type-image .ea-content{background:#505050}  
</style></head>  
  
<body class="python home" id="homepage">  
  
    <div id="touchnav-wrapper">  
  
        <div id="nojs" class="do-not-print">  
            <p><strong>Notice:</strong> While JavaScript is not essential for this website, your interaction with  
p>  
            </div>  
  
        <!--[if lte IE 8]>  
        <div id="oldie-warning" class="do-not-print">  
            <p>  
                <strong>Notice:</strong> Your browser is <em>ancient</em>. Please  
                <a href="http://browsehappy.com/">upgrade to a different browser</a> to experience a better web.  
            </p>  
        </div>  
        <! [endif]-->  
  
        <!-- Sister Site Links -->  
        <div id="top" class="top-bar do-not-print">  
  
            <nav class="meta-navigation container" role="navigation">  
  
                <div class="skip-link screen-reader-text">  
                    <a href="#content" title="Skip to content">Skip to content</a>  
                </div>  
  
                <a id="close-python-network" class="jump-link" href="#python-network" aria-hidden="true">
```

Data Crawling

❖ Selenium Example

Python is a programming language that lets you work quickly and integrate systems more effectively. [»» Learn More](#)

```
<div class="introduction">
  <p>
    "Python is a programming language that lets you work quickly "
    <span class="breaker"></span>
    "and integrate systems more effectively. "
  <a class="readmore" href="/doc/">...</a>
</p>
</div>
```

```
url = 'https://www.python.org/'
driver.get(url)
text = driver.find_element(
    By.XPATH,
    '/html/body/div/header/div/div[3]/p'
).text
print(text)
```

```
(crawl_env) thangduong@Duongs-MacBook-Pro data_manipulation_crawling_project % python selenium_test.py
Python is a programming language that lets you work quickly
and integrate systems more effectively. Learn More
(crawl_env) thangduong@Duongs-MacBook-Pro data_manipulation_crawling_project %
```

driver.find_element(): Find element (tag) in HTML.

Data Crawling

❖ Selenium Example

```
url = 'https://www.python.org/'  
driver.get(url)  
text = driver.find_element(  
    By.XPATH,  
    '/html/body/div/header/div/div[3]/p'  
).text  
print(text)
```

Xpath: A path to navigate through elements and attributes in an HTML

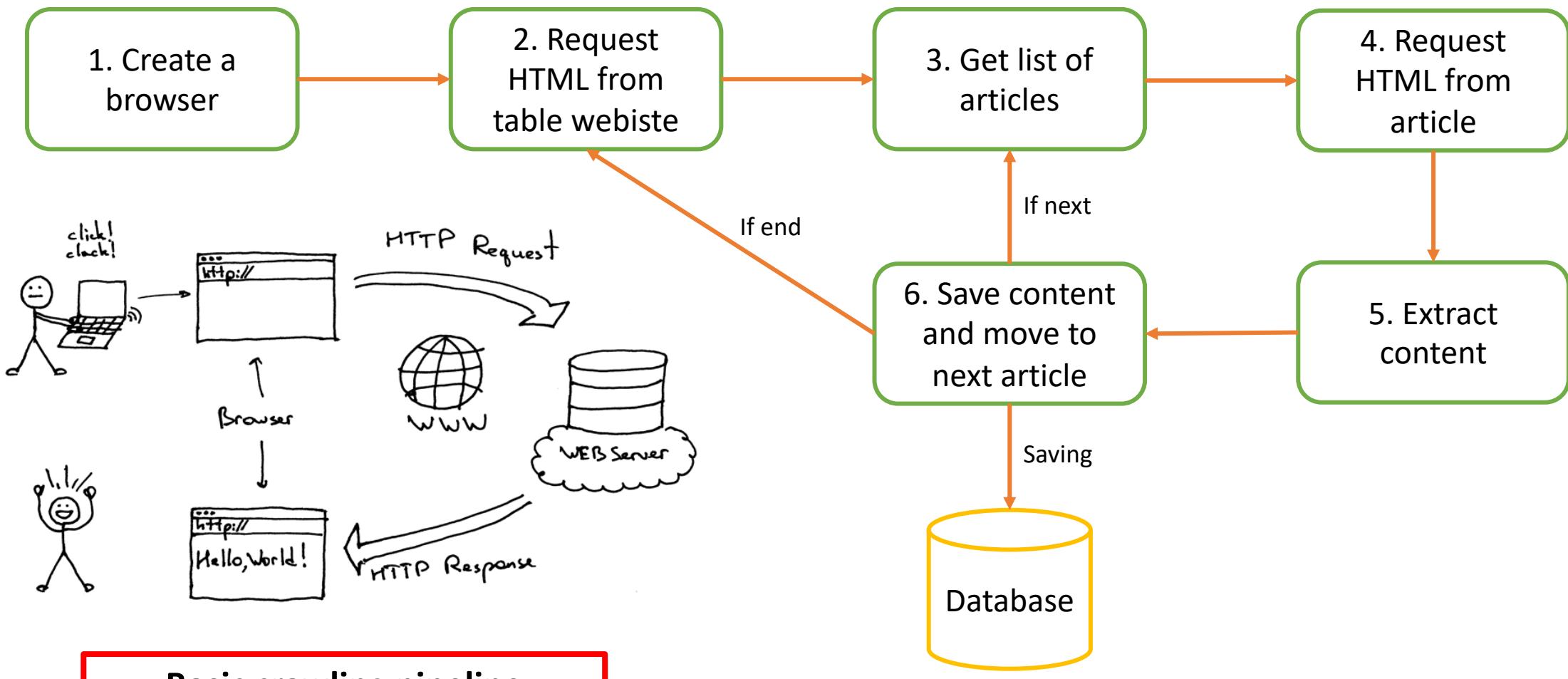
The diagram illustrates the XPATH navigation from the root `<html>` element to the `<h1>` element. The root `<html>` element is circled with a red circle labeled '1'. Below it is the `<head>` element, followed by the `<title>Page Title</title>` element. The `<body>` element is circled with a red circle labeled '2'. Inside the `<body>` element, the `<h1>Hello</h1>` element is circled with a red circle labeled '3'. Below the `<h1>` element is the `<p>Xin chào mọi người.</p>` element.

```
<!DOCTYPE html>  
<html>  
<head>  
<title>Page Title</title>  
</head>  
<body>  
<h1>Hello</h1>  
<p>Xin chào mọi người.</p>  
</body>  
</html>
```

Xpath of `<h1>`: /html/body/h1

Data Crawling

❖ Step-by-step: Program pipeline

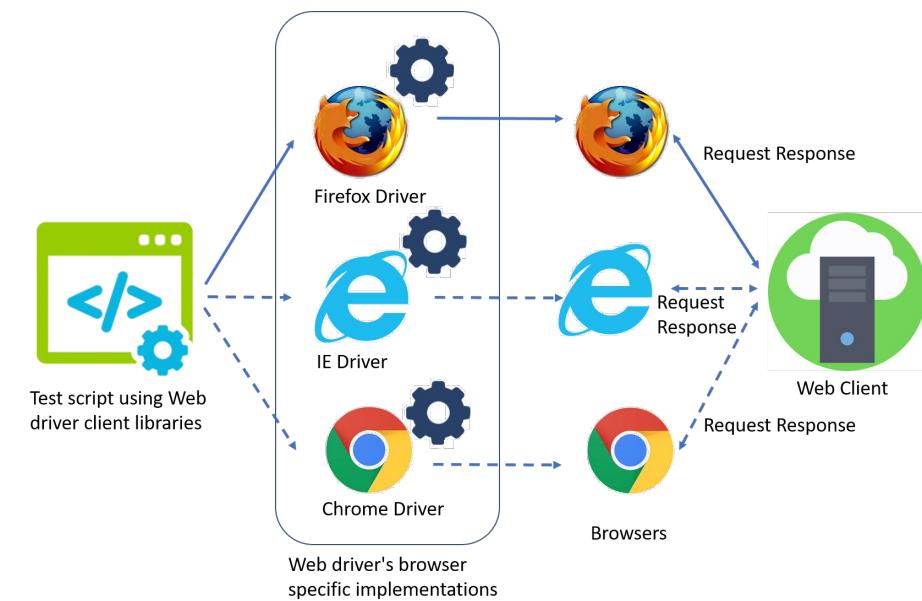


Data Crawling

❖ Step 1: Create browser

```
1 import os
2 import requests
3 import time
4 import random
5
6 from tqdm import tqdm
7 from selenium import webdriver
8 from selenium.webdriver.common.by import By
9 from selenium.webdriver.support.ui import WebDriverWait
10 from selenium.webdriver.support import expected_conditions as EC
```

```
1 # Initialize google chrome browser
2 chrome_options = webdriver.ChromeOptions()
3 chrome_options.add_argument('--headless=new')
4 chrome_options.add_argument('--no-sandbox')
5 driver = webdriver.Chrome(
6     'chromedriver',
7     options=chrome_options
8 )
```



Data Crawling

❖ Step 2: Create empty folder

```
os.makedirs(name, mode=0o777, exist_ok=False)
```

Recursive directory creation function. Like `mkdir()`, but makes all intermediate-level directories needed to contain the leaf directory.

Doc: <https://docs.python.org/3/library/os.html#os.makedirs>

```
10 # Create a folder for storing articles
11 root_dir = './vn_news_corpus'
12 os.makedirs(root_dir, exist_ok=True)
13 n_pages = 10 # Change if you want more articles
14 article_id = 0
```



Create a folder to store crawled articles (called a corpus)

Data Crawling

❖ Step 3: Get list of articles

Vietnamnet

VÌ VIỆT NAM HÙNG CƯỜNG

Thời sự | An toàn giao thông | Môi trường | BHXH - BHYT | Chống tham nhũng | Quốc phòng | Thời tiết

THỜI SỰ

Buổi tập luyện của lực lượng bảo vệ Lăng Chủ tịch Hồ Chí Minh

Để có thể thực hiện động tác đứng nghiêm, không nhúc nhích trong khi làm nhiệm vụ, những người lính tiêu binh Đoàn 275 - Bộ Tư lệnh Bảo vệ Lăng Chủ tịch Hồ Chí Minh đã phải trải qua quá trình rèn luyện thể lực trong nhiều giờ mỗi ngày.



THỜI SỰ

Cháy xưởng gỗ, hàng ngàn mét vuông bị thiêu rụi ở Đồng Nai

Xưởng gỗ mùn cưa rộng khoảng 2.000m² ở Đồng Nai cháy lớn khiến nhiều công nhân tháo chạy ra ngoài.



THỜI SỰ

Công cụ giám sát DAT trực trặc gây khó cho người học lái xe

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.



```
<!DOCTYPE html>
<html lang="vi" translate="no">
  <head> ... </head>
  <body>
    <!-- BEGIN COMPONENT:: COMPONENT002199 -->
    ><noscript>...</noscript>
    <!-- End Google Tag Manager (noscript) -->
    <!-- END COMPONENT:: COMPONENT002199 -->
    <!-- BEGIN COMPONENT:: COMPONENT002461 -->
    <!-- END COMPONENT:: COMPONENT002461 -->
    ><div class="wrapper">
      <!-- BEGIN COMPONENT:: COMPONENT002533 -->
      <!-- END COMPONENT:: COMPONENT002533 -->
      <!-- BEGIN COMPONENT:: COMPONENT002462 -->
      ><header class="header vnn-header">...</header>
      ><div class="wrapNav nav-warpper togglePinTop ">...</div>
      <!-- END COMPONENT:: COMPONENT002462 -->
      <!-- BEGIN COMPONENT:: COMPONENT000003 -->
      <!-- END COMPONENT:: COMPONENT000003 -->
      <!-- Css - js -->
      <!-- BEGIN COMPONENT:: COMPONENT002468 -->
      <!-- END COMPONENT:: COMPONENT002468 -->
      <!-- Content -->
      <!-- BEGIN COMPONENT:: COMPONENT002467 -->
      <!-- BEGIN COMPONENT:: COMPONENT002545 -->
      ><script>...</script>
      <input type="hidden" id="infoConfigGlobalId" data-websiteid="000003" category-id="000002" template-group-id="00001I" slugcatname="thoisu" data-utm-catname="thoisu">
    ><div class="main">
      <!-- BEGIN COMPONENT:: COMPONENT002471 -->
      ><div class="breadcrumbIsPin togglePinTop">...</div>
      <!-- END COMPONENT:: COMPONENT002471 -->
    ... ><div class="container sectionTopstory pb-0 align-start">...</div> == $0
    ><div class="boxRemarkable mb-20">...</div>
    ><div class="container newsStreams">...</div>

```

Ctrl/Command + Shift + C: Open page source of current webpage

Data Crawling

❖ Step 3: Get list of articles

```
▼<div class="main">
  <!-- BEGIN COMPONENT:: COMPONENT002471 -->
  ▶<div class="breadcrumbIsPin togglePinTop">...</div>
  <!-- END COMPONENT:: COMPONENT002471 -->
  ▼<div class="container sectionTopstory pb-0 align-start">
    ▼<div class="container_left"> flex
      <!-- begin:: ads adzone -->
      <!-- BEGIN COMPONENT:: COMPONENT001010 -->
      <!-- END COMPONENT:: COMPONENT001010 -->
      <!-- end:: ads adzone -->
      ▼<div class="topStory-15nd">
        <!-- BEGIN COMPONENT:: COMPONENT002469 -->
        ▶<div class="horizontalPost sm:lineSeparates mb-20">...</div>
        <!-- END COMPONENT:: COMPONENT002469 -->
        <!-- BEGIN COMPONENT:: COMPONENT002469 -->
        ▶<div class="horizontalPost sm:lineSeparates mb-20">...</div>
        <!-- END COMPONENT:: COMPONENT002469 -->
        <!-- BEGIN COMPONENT:: COMPONENT002469 -->
        ▶<div class="horizontalPost sm:lineSeparates mb-20">...</div>
        <!-- END COMPONENT:: COMPONENT002469 -->
        <!-- BEGIN COMPONENT:: COMPONENT002469 -->
        ▶<div class="horizontalPost sm:lineSeparates mb-20">...</div>
        <!-- END COMPONENT:: COMPONENT002469 -->
        <!-- BEGIN COMPONENT:: COMPONENT002469 -->
        ▶<div class="horizontalPost sm:lineSeparates mb-20">...</div>
        <!-- END COMPONENT:: COMPONENT002469 -->
```

div.topStory-15nd 895x2768.91

Môi trường BHXH - BHYT Chống tham nhũng Quốc phòng Thời tiết

THỜI SỰ

Công cụ giám sát DAT trực trặc gây khó cho người học lái xe 14

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

THỜI SỰ

Thanh niên cầm dao khống chế nữ chủ tiệm quần áo, đòi cướp tài sản

Nam thanh niên 23 tuổi bất ngờ khống chế, lôi chủ tiệm vào phòng vệ sinh đồng thời yêu cầu đưa tài sản. Khi nữ chủ tiệm chống cự, đối tượng dùng dao cắt vào cổ nạn nhân gây thương tích.

THỜI SỰ

Kiểm tra dấu hiệu vi phạm của Trưởng Công an phường đòi đánh người dân 12

Thường trực Thành ủy Hạ Long yêu cầu Đảng ủy phường Bãi Cháy tiến hành xác minh, kiểm tra dấu hiệu vi phạm và đề xuất xử lý theo quy định đối với Trưởng Công an phường Bãi Cháy.

THỜI SỰ

Khởi tố Giám đốc Trung tâm Đăng kiểm xe cơ giới 99-05D ở Bắc Ninh

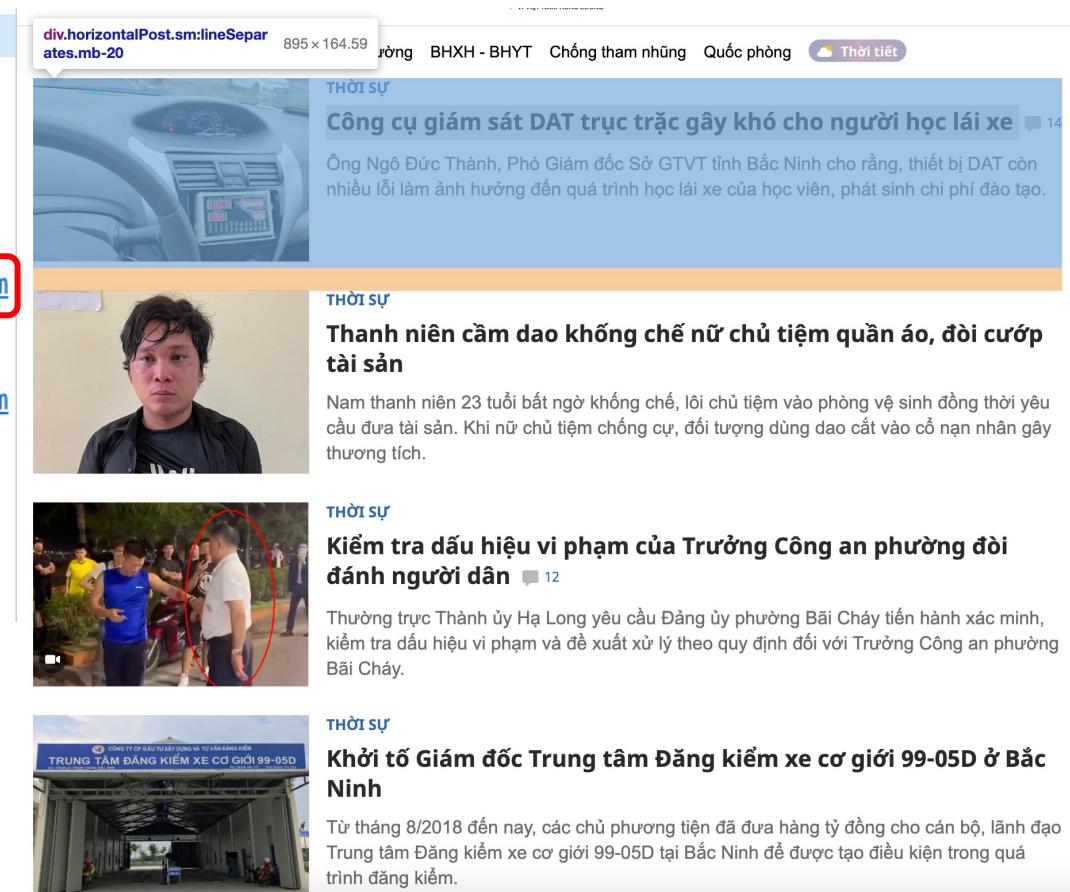
Từ tháng 8/2018 đến nay, các chủ phương tiện đã đưa hàng tỷ đồng cho cán bộ, lãnh đạo Trung tâm Đăng kiểm xe cơ giới 99-05D tại Bắc Ninh để được tạo điều kiện trong quá trình đăng kiểm.

Data Crawling

❖ Step 3: Get list of articles

```
▼<div class="horizontalPost sm:lineSeparates mb-20"> == $0
▶<div class="horizontalPost__avt avt-240"> ... </div>
▼<div class="horizontalPost__main">
  ▶<div class="horizontalPost__main-cate show"> ... </div>
  ▶<h3 data-id="2144796" class="horizontalPost__main-title vnn-title title-bold" ispr="False">
    ▶<a href="/cong-cu-giam-sat-dat-truc-trac-gay-kho-nguo-i-hoc-lai-xe-2144796.htm
      l" title="Công cụ giám sát DAT trực trắc gây khó cho người học lái xe" data-
      utm-source="#vnn_source=thoisu&vnn_medium=listtin1" data-limit> ... </a>
    ▶<a href="/cong-cu-giam-sat-dat-truc-trac-gay-kho-nguo-i-hoc-lai-xe-2144796.htm
      l#comment" class="icon-comment box-icon-text box-icon-comment"> ... </a>
  </h3>
  ▶<div class="horizontalPost__main-desc" data-limit> ... </div>
</div>
</div>
```

URL of the article is in <a> tag



div.horizontalPost.sm:lineSeparates.mb-20 895x164.59

BHNV - BHYT Chống tham nhũng Quốc phòng Thời tiết

THỜI SỰ

Công cụ giám sát DAT trực trắc gây khó cho người học lái xe 14

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

THỜI SỰ

Thanh niên cầm dao khống chế nữ chủ tiệm quần áo, đòi cướp tài sản

Nam thanh niên 23 tuổi bất ngờ khống chế, lôi chủ tiệm vào phòng vệ sinh đồng thời yêu cầu đưa tài sản. Khi nữ chủ tiệm chống cự, đối tượng dùng dao cắt vào cổ nạn nhân gây thương tích.

THỜI SỰ

Kiểm tra dấu hiệu vi phạm của Trưởng Công an phường đòn đánh người dân 12

Thường trực Thành ủy Hạ Long yêu cầu Đảng ủy phường Bãi Cháy tiến hành xác minh, kiểm tra dấu hiệu vi phạm và đề xuất xử lý theo quy định đối với Trưởng Công an phường Bãi Cháy.

THỜI SỰ

Khởi tố Giám đốc Trung tâm Đăng kiểm xe cơ giới 99-05D ở Bắc Ninh

Từ tháng 8/2018 đến nay, các chủ phương tiện đã đưa hàng tỷ đồng cho cán bộ, lãnh đạo Trung tâm Đăng kiểm xe cơ giới 99-05D tại Bắc Ninh để được tạo điều kiện trong quá trình đăng kiểm.

Data Crawling

❖ Step 3: Get list of articles

```
16 for page_idx in range(n_pages):
17     # Access to table page
18     main_url = f'https://vietnamnet.vn/thoi-su-page{page_idx}'
19     driver.get(main_url)
20
21     # Get list of articles (list of URLs)
22     news_lst_xpath = '//div[@class="topStory-15nd"]/div/div[1]/a'
23     news_tags = driver.find_elements(
24         By.XPATH,
25         news_lst_xpath
26     )
27     news_page_urls = [
28         news_tag.get_attribute('href') \
29         for news_tag in news_tags
30     ]
```

The screenshot shows a news article from vietnamnet.vn/thoi-su-page1. The URL in the address bar is https://vietnamnet.vn/thoi-su-page1. The page title is 'Chính trị Thời sự Kinh doanh Thể thao Giải trí Thể giới Đời sống Giáo dục Sức khỏe Thông tin và Truyền thông Phá...'. The main content area displays four news items:

- Hà Nội: Cháy trong khu dân cư, khói đen bốc cao hàng chục mét**
Lán tạm rộng khoảng 30m2 nằm sâu trong ngõ số 9 phố Lương Đình Của (Hà Nội) bất ngờ bốc cháy, cột khói đen bốc cao hàng chục mét.
- THỜI SỰ**
Hà Nam: Công an xã là 'lá chắn thép' bảo vệ bình yên của Nhân dân
Lực lượng Công an xã chính quy tại Hà Nam đã khẳng định được bản lĩnh chính trị, trình độ chuyên môn nghiệp vụ, thực sự là những "lá chắn" vững chắc bảo đảm an ninh, trật tự tại địa bàn cơ sở.
- THỜI SỰ**
Doanh nghiệp làm khó công nhân, Chủ tịch Hà Nội yêu cầu 'nóng'
Ông Trần Sỹ Thành yêu cầu Sở Tư pháp phải xem xét việc doanh nghiệp, tổ chức, yêu cầu người lao động phải có phiếu lý lịch tư pháp mới sau 6 tháng có đúng luật hay không? Nếu không thì xử lý doanh nghiệp như thế nào?
- THỜI SỰ**
Dự báo thời tiết 19/5: Nắng nóng ở miền Bắc hạ xuống dưới 40 độ
Dự báo thời tiết ngày 19/5, miền Bắc và Trung Bộ tiếp diễn nắng nóng gay gắt diện rộng, nhưng nhiệt độ hạ nhẹ ở Bắc Bộ. Tây Nguyên và Nam Bộ ngày nắng, chiều tối mưa mát.

A navigation bar at the bottom right includes buttons for < 1 2 3 4 5 6 > and a link 'Xem tin theo ngày'.

Data Crawling

❖ Step 4: Access to article contents

The screenshot shows a news article from VietnamNet. The title is "Công cụ giám sát DAT trực trắc gây khó cho người học lái xe". The author is N. Huyền, described as a reporter. Below the title, there is a link to "Xem các bài viết của tác giả". The article content discusses the impact of DAT monitoring on driving training. The browser's developer tools are open, specifically the Elements tab, which displays the raw HTML code of the page. The code includes standard HTML tags like <head> and <body>, as well as more complex components like noscript tags, Google Tag Manager, and scripts for tracking and comments.

THỜI SỰ

20/05/2023 06:30 (GMT+07:00)

Công cụ giám sát DAT trực trắc gây khó cho người học lái xe

N. Huyền
Nhà báo

Theo dõi VietNamNet trên [Google News](#)

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

Mới đây, ông Lương Duyên Thống - Trưởng phòng Quản lý vận tải phương tiện và người lái, Cục Đường bộ Việt Nam cho biết, kết quả kiểm tra tại các cơ sở đào tạo, sát hạch lái xe cho thấy việc theo dõi, kiểm tra, khai thác, sử dụng dữ liệu quản lý DAT còn nhiều hạn chế.

Theo đó, nhiều cơ sở đào tạo chưa theo dõi, giám sát, khai thác dữ liệu trên phần mềm hệ thống thông tin DAT để yêu cầu học viên học đủ số km lái xe ban đêm, học đủ thời gian trên xe số tự động.

Tương tự, vẫn còn có tình trạng không tập huấn cho giáo viên hướng dẫn học viên đăng nhập, đăng xuất các phiên học dẫn đến trùng học viên, trùng xe tập lái, trùng hành trình tại cùng một thời điểm.

```
<!DOCTYPE html>
<html lang="vi" translate="no">
  <head>...</head>
  ... <body> == $0
    <!-- BEGIN COMPONENT::: COMPONENT002199 -->
    <noscript>...</noscript>
    <!-- End Google Tag Manager (noscript) -->
    <!-- END COMPONENT::: COMPONENT002199 -->
    <!-- BEGIN COMPONENT::: COMPONENT002461 -->
    <!-- END COMPONENT::: COMPONENT002461 -->
    <div class="wrapper">...</div>
    <script vnn="vnnjs" type="text/javascript" src="https://res-files.vnncdn.net/files/publish/2023/5/11/js-ca8ee9a...desktop-153.js?v=1"></script>
    <script>...</script>
    <!-- BEGIN COMPONENT::: COMPONENT002475 -->
    <script type="text/javascript" src="https://res-files.vnncdn.net/files/2022/6/20/vnnvotemobilemobile.js"></script>
    <script defer>...</script>
    <!-- END COMPONENT::: COMPONENT002475 -->
    <!-- BEGIN COMPONENT::: COMPONENT002466 -->
    <script type="text/javascript" src="https://comment.vietnamnet.vn/js/vnnidmodule.js"></script>
    <!-- script tracking -->
    <script>...</script>
    <script>...</script>
    <iframe width="0" height="0" src="https://mic.gov.vn/Pages/PageEmbed/Vietnamnet/Default.aspx" marginwidth="0" allowtransparency="true" marginheight="0" hspace="0" vspace="0" frameborder="0" scrolling="no">...</iframe>
    <!-- END COMPONENT::: COMPONENT002466 -->
    
    <script type="text/javascript" id="...</script>
    <noscript>...</noscript>
    <div class="lg-container lg-vnn-custom" id="lg-container-1" tabindex="-1" aria-modal="true" role="dialog">...</div>
    <script type="text/javascript" src="https://vnn-res.vncloud.vn/VietNamNet/Standard/is/vnntr.html body
```

Data Crawling

❖ Step 4: Access to article contents

```
▼<div class="main-v1 bg-white">
  ▼<div class="container-v1 detail-page gap-40">
    ▼<div class="container_left not-pl">
      <!-- BEGIN COMPONENT:: COMPONENT002504 -->
      ▶<div class="bread-crumb-detail ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002504 -->
    ▼<div class="content-detail">
      <h1 class="content-detail-title">Công cụ giám sát DAT trực trặc gây khó cho người
        học lái xe</h1> == $0
      <!-- BEGIN COMPONENT:: COMPONENT002489 -->
      ▶<div class="article-detail-author-wrapper ">...</div>
      <!-- END COMPONENT:: COMPONENT002489 -->
      <!-- BEGIN COMPONENT:: COMPONENT002514 -->
      <!-- END COMPONENT:: COMPONENT002514 -->
      <!-- BEGIN COMPONENT:: COMPONENT002490 -->
      <!-- END COMPONENT:: COMPONENT002490 -->
      ▶<div class="vnn-share-social share-social ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002490 -->
      <!-- BEGIN COMPONENT:: COMPONENT002511 -->
      <!-- END COMPONENT:: COMPONENT002511 -->
      <!-- BEGIN COMPONENT:: COMPONENT002551 -->
      <!-- END COMPONENT:: COMPONENT002551 -->
      ▶<h2 class="content-detail-sapo">...</h2>
      <!-- BEGIN COMPONENT:: COMPONENT001017 -->
      <!-- END COMPONENT:: COMPONENT001017 -->
    ▼<div class="maincontent main-content" id="maincontent">...</div>
    ::after
  </div>
```



Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

Mới đây, ông Lương Duyên Thống - Trưởng phòng Quản lý vận tải phương tiện và người lái, Cục Đường bộ Việt Nam cho biết, kết quả kiểm tra tại các cơ sở đào tạo, sát hạch lái xe cho thấy việc theo dõi, kiểm tra, khai thác, sử dụng dữ liệu quản lý DAT còn nhiều hạn chế.

Theo đó, nhiều cơ sở đào tạo chưa theo dõi, giám sát, khai thác dữ liệu trên phần mềm hệ thống thông tin DAT để yêu cầu học viên học đủ số km lái xe ban đêm, học đủ thời gian trên xe số tự động.

Data Crawling

❖ Step 4: Access to article contents

```
<div class="main-v1 bg-white">
  <div class="container-v1 detail-page gap-40">
    <div class="container_left not-pl">
      <!-- BEGIN COMPONENT:: COMPONENT002504 -->
      ><div class="bread-crumb-detail ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002504 -->
    <div class="content-detail">
      <h1 class="content-detail-title">Công cụ giám sát DAT trực trặc gây khó cho người học lái xe</h1>
      <!-- BEGIN COMPONENT:: COMPONENT002489 -->
      ><div class="article-detail-author-wrapper ">...</div>
      <!-- END COMPONENT:: COMPONENT002489 -->
      <!-- BEGIN COMPONENT:: COMPONENT002514 -->
      <!-- END COMPONENT:: COMPONENT002514 -->
      <!-- BEGIN COMPONENT:: COMPONENT002490 -->
      ><div class="vnn-share-social share-social ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002490 -->
      <!-- BEGIN COMPONENT:: COMPONENT002511 -->
      <!-- END COMPONENT:: COMPONENT002511 -->
      <!-- BEGIN COMPONENT:: COMPONENT002551 -->
      <!-- END COMPONENT:: COMPONENT002551 -->
      ><h2 class="content-detail-sapo">...</h2> == $0
      <!-- BEGIN COMPONENT:: COMPONENT001017 -->
      <!-- END COMPONENT:: COMPONENT001017 -->
    <div class="maincontent main-content" id="maincontent">...</div>
    ::after
  </div>
```

The screenshot shows a news article from Vietnamnet. The header includes the logo, navigation menu, and date (20/05/2023 06:30 (GMT+07:00)). The main title is "Công cụ giám sát DAT trực trặc gây khó cho người học lái xe". Below the title is the author's profile (N. Huyền, Nhà báo) and a snippet of the article. The full text discusses the challenges faced by drivers learning to drive due to DAT monitoring equipment.

THỜI SỰ

20/05/2023 06:30 (GMT+07:00)

Công cụ giám sát DAT trực trặc gây khó cho người học lái xe

N. Huyền
Nhà báo

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

Mới đây, ông Lương Duyên Thống - Trưởng phòng Quản lý vận tải phương tiện và người lái, Cục Đường bộ Việt Nam cho biết, kết quả kiểm tra tại các cơ sở đào tạo, sát hạch lái xe cho thấy việc theo dõi, kiểm tra, khai thác, sử dụng dữ liệu quản lý DAT còn nhiều hạn chế.

Theo đó, nhiều cơ sở đào tạo chưa theo dõi, giám sát, khai thác dữ liệu trên phần mềm hệ thống thông tin DAT để yêu cầu học viên học đủ số km lái xe ban đêm, học đủ thời gian trên xe số tự động.

Tương tự, vẫn còn có tình trạng không tập huấn cho giáo viên hướng dẫn học viên đăng nhập, đăng xuất các phiên học dẫn đến trùng học viên, trùng xe tập lái, trùng hành trình tại cùng một thời điểm.

Data Crawling

❖ Step 4: Access to article contents

```
<div class="main-v1 bg-white">
  <div class="container-v1 detail-page gap-40">
    <div class="container_left not-pl">
      <!-- BEGIN COMPONENT:: COMPONENT002504 -->
      <div class="bread-crumb-detail ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002504 -->
    <div class="content-detail">
      <h1 class="content-detail-title">Công cụ giám sát DAT trực trặc gây khó cho người lái xe</h1>
      <!-- BEGIN COMPONENT:: COMPONENT002489 -->
      <div class="article-detail-author-wrapper ">...</div>
      <!-- END COMPONENT:: COMPONENT002489 -->
      <!-- BEGIN COMPONENT:: COMPONENT002514 -->
      <!-- END COMPONENT:: COMPONENT002514 -->
      <!-- BEGIN COMPONENT:: COMPONENT002490 -->
      <div class="vnn-share-social share-social ">...</div> flex
      <!-- END COMPONENT:: COMPONENT002490 -->
      <!-- BEGIN COMPONENT:: COMPONENT002511 -->
      <!-- END COMPONENT:: COMPONENT002511 -->
      <!-- BEGIN COMPONENT:: COMPONENT002551 -->
      <!-- END COMPONENT:: COMPONENT002551 -->
      <h2 class="content-detail-sapo">...</h2>
      <!-- BEGIN COMPONENT:: COMPONENT001017 -->
      <!-- END COMPONENT:: COMPONENT001017 -->
    <div class="maincontent main-content" id="maincontent">...</div> == $0
    ::after
  </div>
```

The screenshot shows a news article from VietnamNet (VIỆT NAM HÙNG CƯỜNG) dated 20/05/2023 at 06:30 (GMT+07:00). The article title is "Công cụ giám sát DAT trực trặc gây khó cho người lái xe". It is written by N. Huyền, a reporter. The article discusses the impact of DAT monitoring equipment on drivers. A callout box highlights a quote from Ông Ngô Đức Thành, Deputy Director of the Transportation Department of Bac Ninh province, stating that DAT monitoring equipment has caused many difficulties for drivers. Below the article, there are comments and a link to Google News.

THỜI SỰ

20/05/2023 06:30 (GMT+07:00)

Công cụ giám sát DAT trực trặc gây khó cho người lái xe

N. Huyền
Nhà báo

Theo dõi VietNamNet trên Google News

Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến

895 × 2006.25

sinh chi phí đào tạo.

Mới đây, ông Lương Duyên Thống - Trưởng phòng Quản lý vận tải phương tiện và người lái, Cục Đường bộ Việt Nam cho biết, kết quả kiểm tra tại các cơ sở đào tạo, sát hạch lái xe cho thấy việc theo dõi, kiểm tra, khai thác, sử dụng dữ liệu quản lý DAT còn nhiều hạn chế.

Theo đó, nhiều cơ sở đào tạo chưa theo dõi, giám sát, khai thác dữ liệu trên phần mềm hệ thống thông tin DAT để yêu cầu học viên học đủ số km lái xe ban đêm, học đủ thời gian trên xe số tự động.

Tương tự, vẫn còn có tình trạng không tập huấn cho giáo viên hướng dẫn học viên đăng nhập, đăng xuất các phiên học dẫn đến trùng học viên, trùng xe tập lái, trùng hành trình tại cùng một thời điểm.

Trao đổi với phóng viên VietNamNet về ý kiến trên, ông N.V.D. (giám đốc một trung tâm đào tạo sát hạch lái xe tại Bắc Giang) cho biết, khi áp dụng DAT tại cơ sở đã bộc lộ nhiều bất cập, làm khó cho các đơn vị thực hiện.

Theo đó, thiết bị DAT phát tín hiệu không ổn định, hay mất tín hiệu nên rất dễ truyền nhầm lẫn. Thiết bị cũng dễ hỏng, đơn vị phải sửa thường xuyên.

Data Crawling

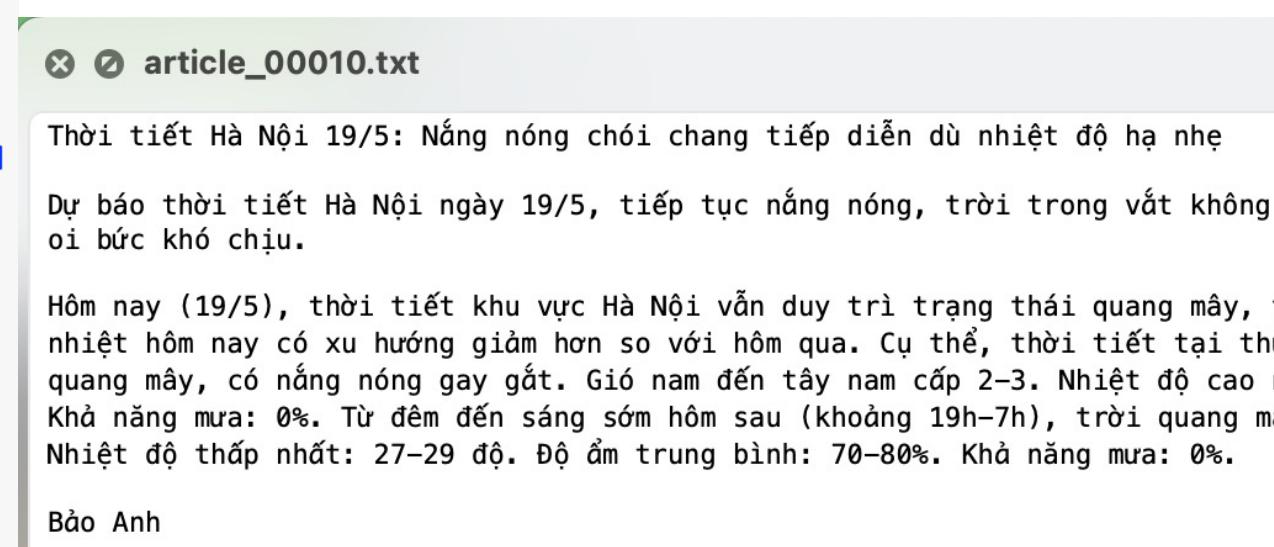
❖ Step 4: Access to article contents

```
32 for news_page_url in news_page_urls:
33     # Access to article page
34     driver.get(news_page_url)
35     time.sleep(1)
36
37     # Try to get main content tag
38     main_content_xpath = '//div[@class="content-detail"]'
39     try:
40         main_content_tag = driver.find_element(
41             By.XPATH,
42             main_content_xpath
43         )
44     except:
45         continue
46     # Ignore video article
47     video_content_xpath = '//div[@class="video-detail"]'
48     try:
49         video_content_tag = main_content_tag.find_element(
50             By.XPATH,
51             video_content_xpath
52         )
53         continue
54     except:
55         pass
56
57     # Get title (h1 tag)
58     title = main_content_tag.find_element(
59         By.TAG_NAME,
60         'h1'
61     ).text.strip()
62     # Get abstract (h2 tag)
63     abstract = main_content_tag.find_element(
64         By.TAG_NAME,
65         'h2'
66     ).text.strip()
67     # Get author name (span tag)
68     try:
69         author_xpath = '//span[@class="name"]'
70         author = main_content_tag.find_element(
71             By.XPATH,
72             author_xpath
73         ).text.strip()
74     except:
75         author = ''
76     # Get paragraphs (all p tags in div "maincontent main-content")
77     paragraphs_xpath = '//div[@class="maincontent main-content"]/p'
78     paragraphs_tags = main_content_tag.find_elements(
79         By.XPATH,
80         paragraphs_xpath
81     )
```

Data Crawling

❖ Step 5: Save content to a .txt file

```
81     paragraphs_lst = [
82         paragraphs_tag.text.strip() \
83             for paragraphs_tag in paragraphs_tags
84     ]
85     paragraphs = ' '.join(paragraphs_lst)
86     # Combine title, abstract, authoor and paragraphs
87     final_content_lst = [title, abstract, paragraphs, author]
88     final_content = '\n\n'.join(final_content_lst)
89
90     # Save artile to .txt file
91     article_filename = f'article_{article_id:05d}.txt'
92     article_savepath = os.path.join(
93         root_dir,
94         article_filename
95     )
96     article_id += 1
97     with open(article_savepath, 'w') as f:
98         f.write(final_content)
99
100    # Move back to previous page
101    driver.back()
```



Data Crawling

❖ Introduction

vn_news_corpus	--	Folder
article_00000.txt	1 KB	Plain Text
article_00001.txt	1 KB	Plain Text
article_00002.txt	5 KB	Plain Text
article_00003.txt	5 KB	Plain Text
article_00004.txt	2 KB	Plain Text
article_00005.txt	2 KB	Plain Text
article_00006.txt	2 KB	Plain Text
article_00007.txt	2 KB	Plain Text
article_00008.txt	3 KB	Plain Text
article_00009.txt	4 KB	Plain Text
article_00010.txt	978 bytes	Plain Text
article_00011.txt	2 KB	Plain Text
article_00012.txt	2 KB	Plain Text
article_00013.txt	2 KB	Plain Text
article_00014.txt	928 bytes	Plain Text
article_00015.txt	3 KB	Plain Text
article_00016.txt	6 KB	Plain Text
article_00017.txt	2 KB	Plain Text
article_00018.txt	1 KB	Plain Text
article_00019.txt	3 KB	Plain Text
article_00020.txt	1 KB	Plain Text

```
1 !zip -r vn_news_corpus.zip vn_news_corpus
adding: vn_news_corpus/ (stored 0%)
adding: vn_news_corpus/article_00070.txt (deflated 50%)
adding: vn_news_corpus/article_00030.txt (deflated 56%)
adding: vn_news_corpus/article_00006.txt (deflated 55%)
adding: vn_news_corpus/article_00091.txt (deflated 48%)
adding: vn_news_corpus/article_00060.txt (deflated 70%)
adding: vn_news_corpus/article_00037.txt (deflated 58%)
adding: vn_news_corpus/article_00073.txt (deflated 55%)
adding: vn_news_corpus/article_00114.txt (deflated 66%)
adding: vn_news_corpus/article_00110.txt (deflated 63%)
adding: vn_news_corpus/article_00111.txt (deflated 58%)
adding: vn_news_corpus/article_00034.txt (deflated 60%)
adding: vn_news_corpus/article_00090.txt (deflated 53%)
adding: vn_news_corpus/article_00011.txt (deflated 58%)
adding: vn_news_corpus/article_00069.txt (deflated 66%)
adding: vn_news_corpus/article_00099.txt (deflated 46%)
adding: vn_news_corpus/article_00039.txt (deflated 70%)
```

Save to a zip file for later use

Data Crawling

❖ Extend: How about images?

Vietnam.net
VÌ VIỆT NAM HÙNG CƯỜNG

Thời sự An toàn giao thông Môi trường BHXH - BHYT Chống tham nhũng Quốc phòng Thời tiết

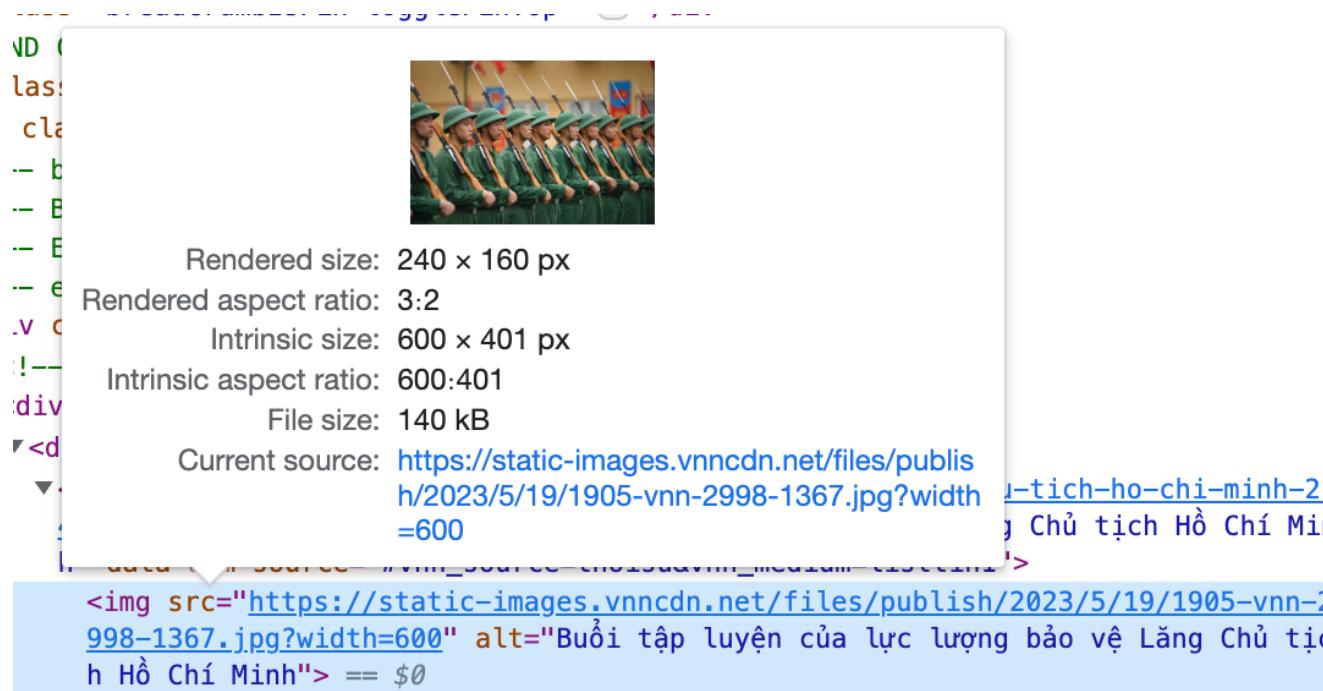
THỜI SỰ
Buổi tập luyện của lực lượng bảo vệ Lăng Chủ tịch Hồ Chí Minh
Để có thể thực hiện động tác đứng nghiêm, không nhúc nhích trong khi làm nhiệm vụ, những người lính tiêu binh Đoàn 275 - Bộ Tư lệnh Bảo vệ Lăng Chủ tịch Hồ Chí Minh đã phải trải qua quá trình rèn luyện thể lực trong nhiều giờ mỗi ngày.

THỜI SỰ
Cháy xưởng gỗ, hàng ngàn mét vuông bị thiêu rụi ở Đồng Nai
Xưởng gỗ mùn cưa rộng khoảng 2.000m² ở Đồng Nai cháy lớn khiến nhiều công nhân tháo chạy ra ngoài.

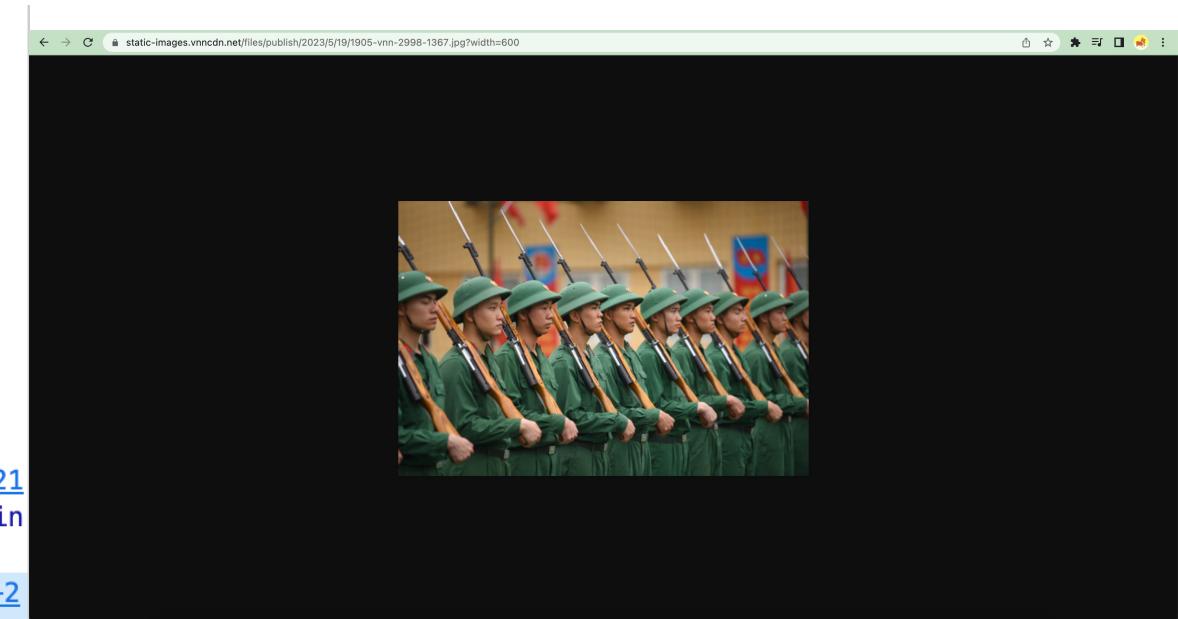
THỜI SỰ
Công cụ giám sát DAT trực trặc gây khó cho người học lái xe 14
Ông Ngô Đức Thành, Phó Giám đốc Sở GTVT tỉnh Bắc Ninh cho rằng, thiết bị DAT còn nhiều lỗi làm ảnh hưởng đến quá trình học lái xe của học viên, phát sinh chi phí đào tạo.

Data Crawling

❖ Extend: How about images?



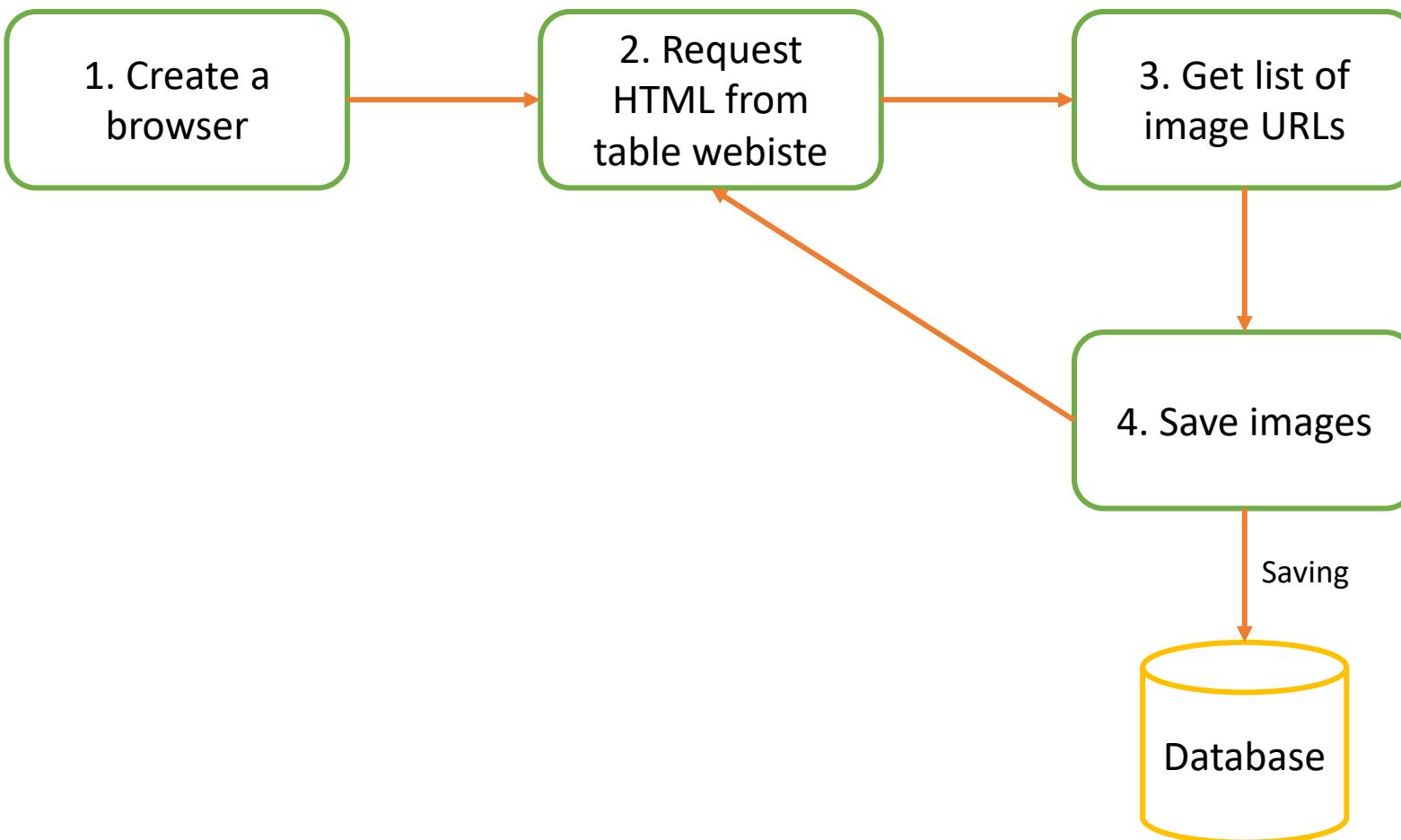
Images often present in tag



The URL leads to a page with only image

Data Crawling

❖ Crawl news thumbnail pipeline



Data Crawling

❖ Step 1, 2, 3: Create browser and get list of image URLs

```
1 # Initialize google chrome browser
2 chrome_options = webdriver.ChromeOptions()
3 chrome_options.add_argument('--headless=new')
4 chrome_options.add_argument('--no-sandbox')
5 driver = webdriver.Chrome(
6     'chromedriver',
7     options=chrome_options
8 )
9
10 # Create a folder for storing articles
11 root_dir = './vn_news_thumbnail'
12 os.makedirs(root_dir, exist_ok=True)
13 n_pages = 10 # Change if you want more articles
14 img_id = 0
```

Create browser and empty folder

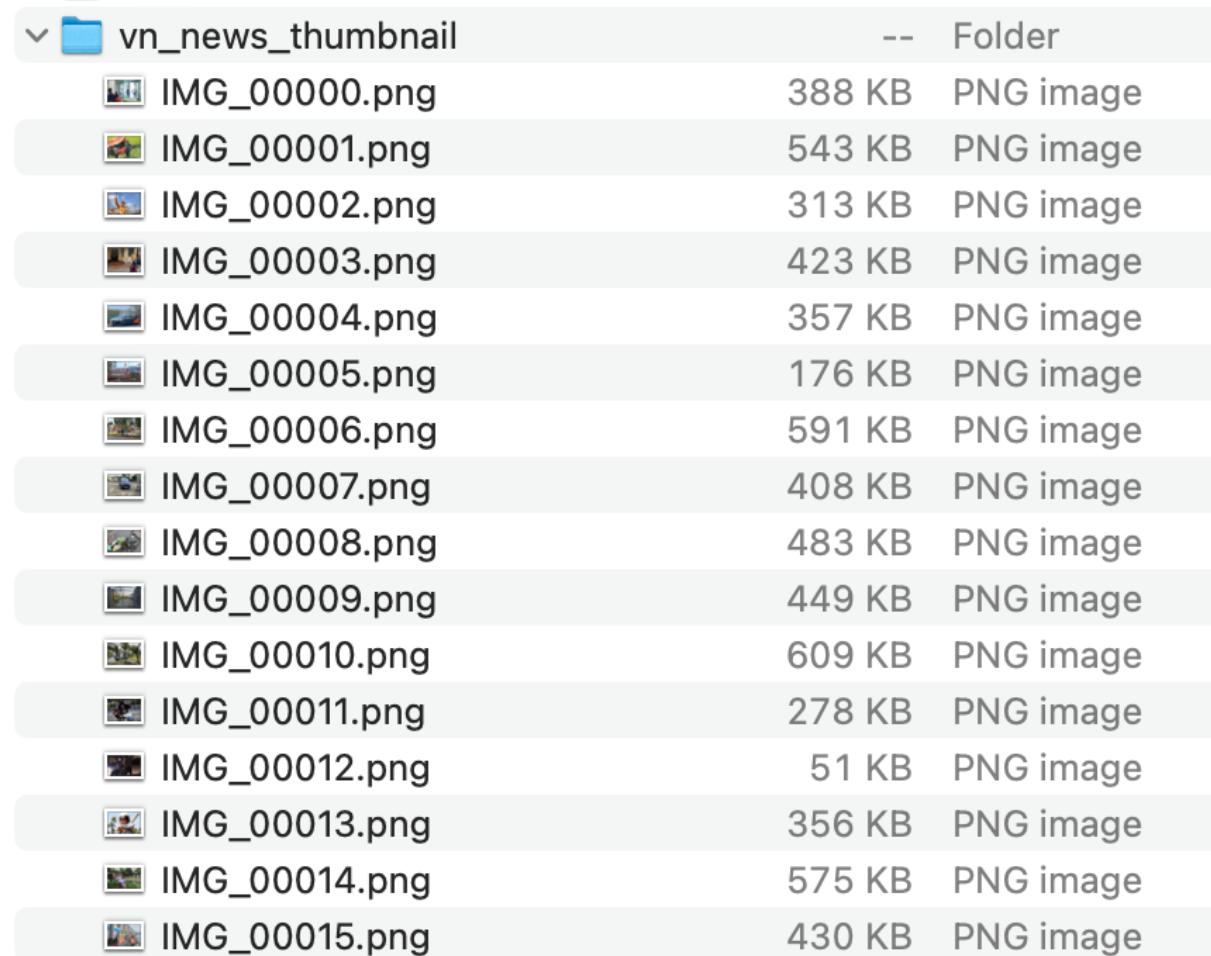
```
16 for page_idx in tqdm(range(n_pages)):
17     # Access to table page
18     main_url = f'https://vietnamnet.vn/thoi-su-page{page_idx}'
19     driver.get(main_url)
20
21     # Get list of articles (list of URLs)
22     imgs_lst_xpath = '//div[@class="topStory-15nd"]/div/div[1]/a/img'
23     imgs_tags = driver.find_elements(
24         By.XPATH,
25         imgs_lst_xpath
26     )
27     img_urls = [
28         imgs_tag.get_attribute('src') \
29         for imgs_tag in imgs_tags
30     ]
```

Get list of image URLs

Data Crawling

❖ Step 4: Download and save image

```
32     for img_url in img_urls:  
33         img_url_resp = requests.get(img_url)  
34         try:  
35             img = Image.open(  
36                 BytesIO(img_url_resp.content)  
37             )  
38         except:  
39             continue  
40  
41         if img.mode == 'P':  
42             img = img.convert('RGB')  
43  
44         img_name = f'IMG_{img_id:05}.png'  
45         img_save_path = os.path.join(  
46             root_dir,  
47             img_name  
48         )  
49         img.save(img_save_path)  
50         img_id += 1
```



Question

