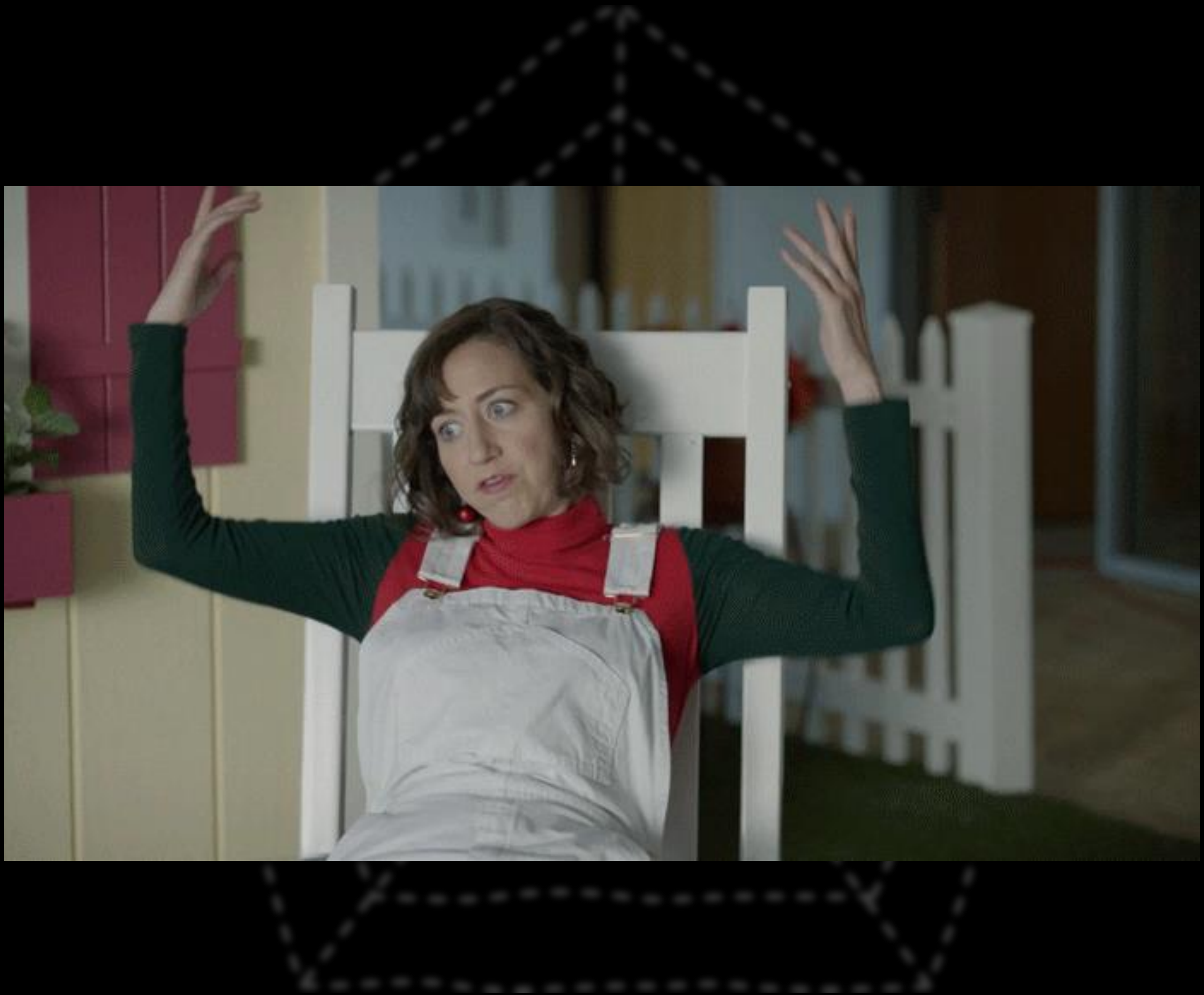
Alvin Alexander
Geo-Technician
JX Nippon Oil & Gas Exploration

PRACTICAL WEB SCRAPING

FOR E&P ACTIVITIES

POWERPOINT

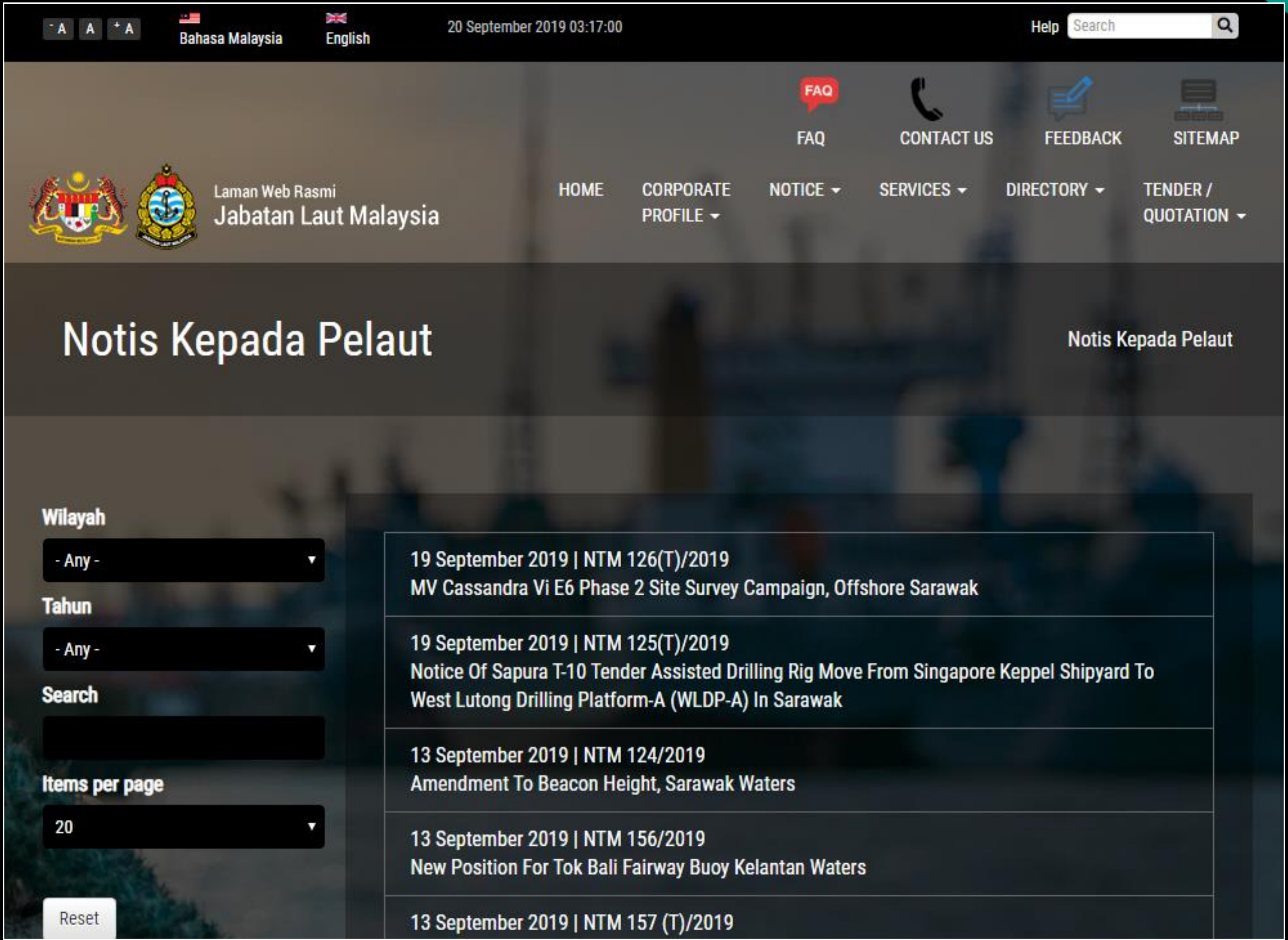


What we do if we don’t automate?

THE TRADITIONAL WAY

To do it manually

1. Visit the website
2. Filter using provided filter
3. Click each one of the link to download
4. Read the Notice to Mariners
5. Copy the coordinates of points
6. Transform to appropriate CRS
7. Convert to shapefiles
8. Rinse and repeat



Notis Kepada Pelaut

Notis Kepada Pelaut

Wilayah: - Any -

Tahun: - Any -

Search

Items per page: 20

19 September 2019 | NTM 126(T)/2019
MV Cassandra Vi E6 Phase 2 Site Survey Campaign, Offshore Sarawak

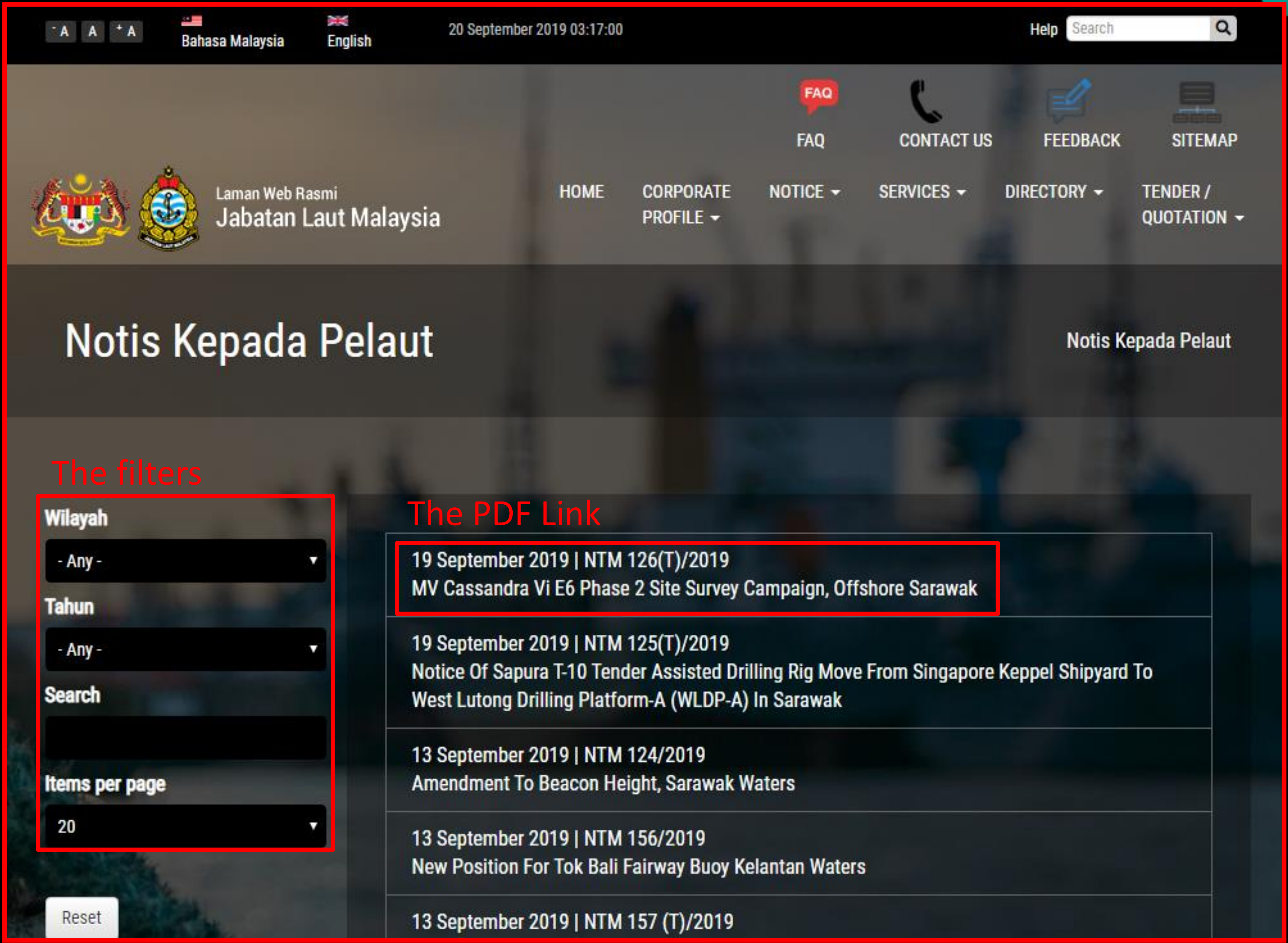
19 September 2019 | NTM 125(T)/2019
Notice Of Sapura T-10 Tender Assisted Drilling Rig Move From Singapore Keppel Shipyard To West Lutong Drilling Platform-A (WLDP-A) In Sarawak

13 September 2019 | NTM 124/2019
Amendment To Beacon Height, Sarawak Waters

13 September 2019 | NTM 156/2019
New Position For Tok Bali Fairway Buoy Kelantan Waters

13 September 2019 | NTM 157 (T)/2019

The Website



20 September 2019 03:17:00

Laman Web Rasmi
Jabatan Laut Malaysia

Notis Kepada Pelaut

Notis Kepada Pelaut

The filters

Wilayah: - Any -

Tahun: - Any -

Search:

Items per page: 20

The PDF Link

19 September 2019 | NTM 126(T)/2019
MV Cassandra Vi E6 Phase 2 Site Survey Campaign, Offshore Sarawak

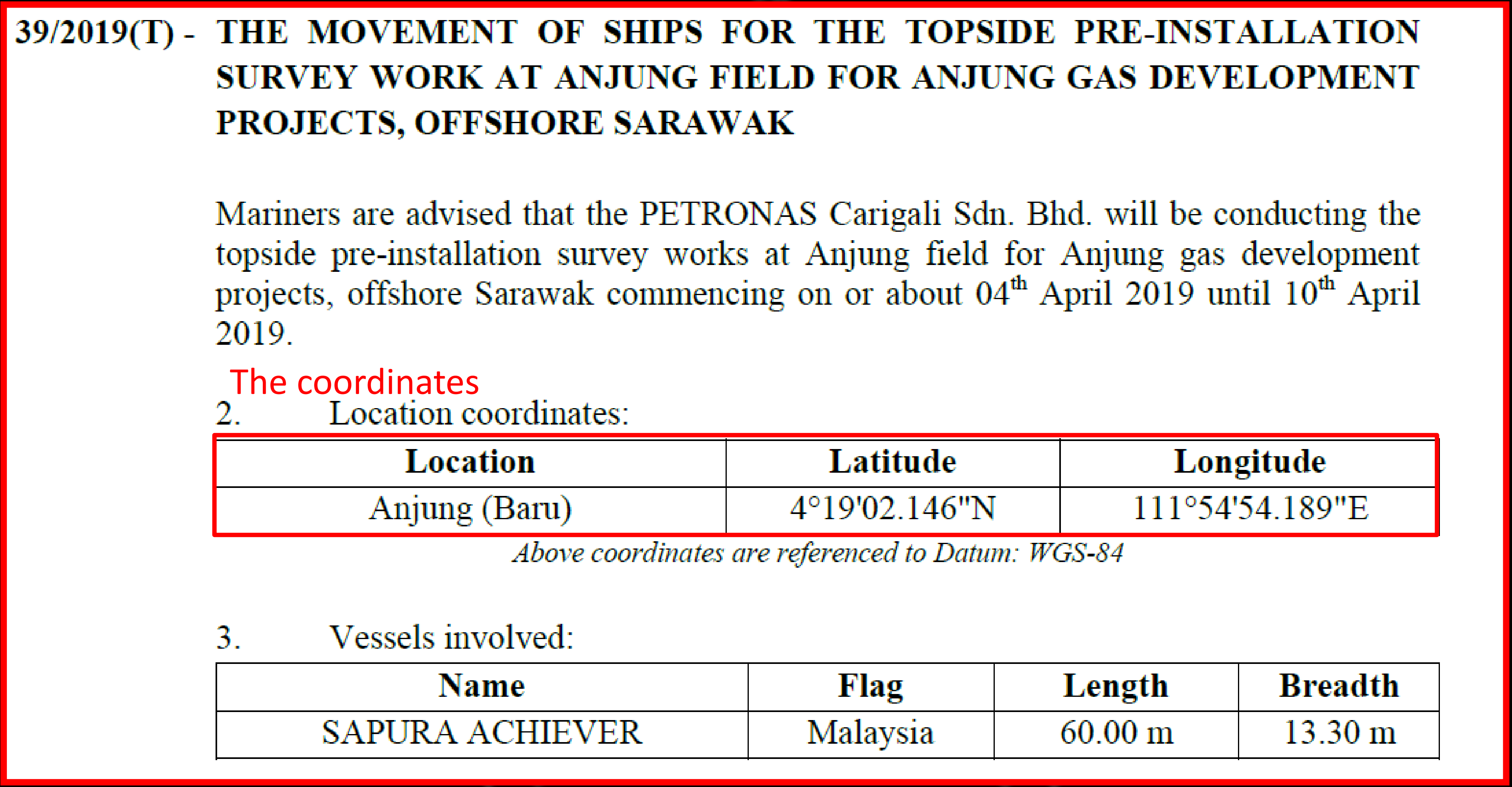
19 September 2019 | NTM 125(T)/2019
Notice Of Sapura T-10 Tender Assisted Drilling Rig Move From Singapore Keppel Shipyard To West Lutong Drilling Platform-A (WLDP-A) In Sarawak

13 September 2019 | NTM 124/2019
Amendment To Beacon Height, Sarawak Waters

13 September 2019 | NTM 156/2019
New Position For Tok Bali Fairway Buoy Kelantan Waters

13 September 2019 | NTM 157 (T)/2019

The PDF



39/2019(T) - THE MOVEMENT OF SHIPS FOR THE TOPSIDE PRE-INSTALLATION SURVEY WORK AT ANJUNG FIELD FOR ANJUNG GAS DEVELOPMENT PROJECTS, OFFSHORE SARAWAK

Mariners are advised that the PETRONAS Carigali Sdn. Bhd. will be conducting the topside pre-installation survey works at Anjung field for Anjung gas development projects, offshore Sarawak commencing on or about 04th April 2019 until 10th April 2019.

The coordinates

2. Location coordinates:

Location	Latitude	Longitude
Anjung (Baru)	4°19'02.146"N	111°54'54.189"E

Above coordinates are referenced to Datum: WGS-84

3. Vessels involved:

Name	Flag	Length	Breadth
SAPURA ACHIEVER	Malaysia	60.00 m	13.30 m

Thought, considerations, and observations while

DEVELOPING THE AUTOMATED WAY

Essential Steps

- Download PDF files
- Extract coordinates
- Process transform coordinates
- Output to shapefiles

***These steps involve usage of python language*

-Know your enemy, Know yourself-

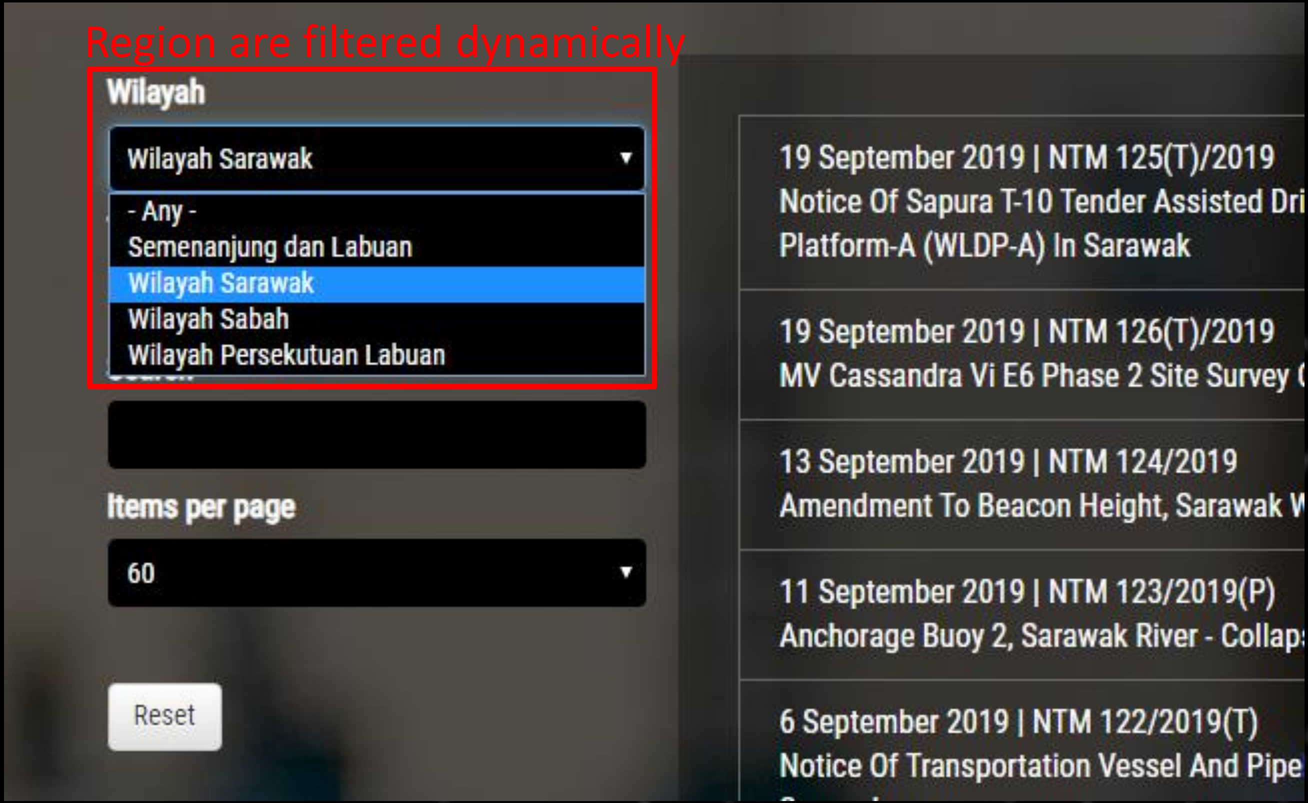
Understanding data and tools

IS IMPORTANT

Website

Considerations	Tools
• The website is **Dynamic**ally generated by JavaScript • The website is **Paginated** Dynamically	• **Chrome Browser** – to inspect HTML • **Requests** – Visiting website, downloading PDF • **BeautifulSoup** – Extracting HTML • **Selenium** – Handling dynamic website. (E.g. second page is loaded dynamically)

Region are filtered dynamically



Wilayah

Wilayah Sarawak ▾

- Any -
Semenanjung dan Labuan
Wilayah Sarawak
Wilayah Sabah
Wilayah Persekutuan Labuan

Items per page

60 ▾

Reset

19 September 2019 | NTM 125(T)/2019
Notice Of Sapura T-10 Tender Assisted Dri
Platform-A (WLDP-A) In Sarawak

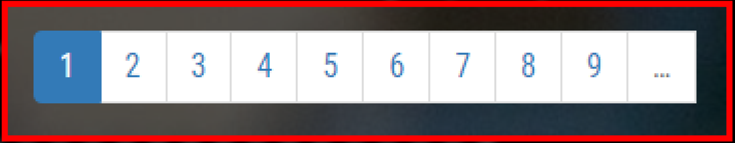
19 September 2019 | NTM 126(T)/2019
MV Cassandra Vi E6 Phase 2 Site Survey (

13 September 2019 | NTM 124/2019
Amendment To Beacon Height, Sarawak W

11 September 2019 | NTM 123/2019(P)
Anchorage Buoy 2, Sarawak River - Collaps

6 September 2019 | NTM 122/2019(T)
Notice Of Transportation Vessel And Pipe

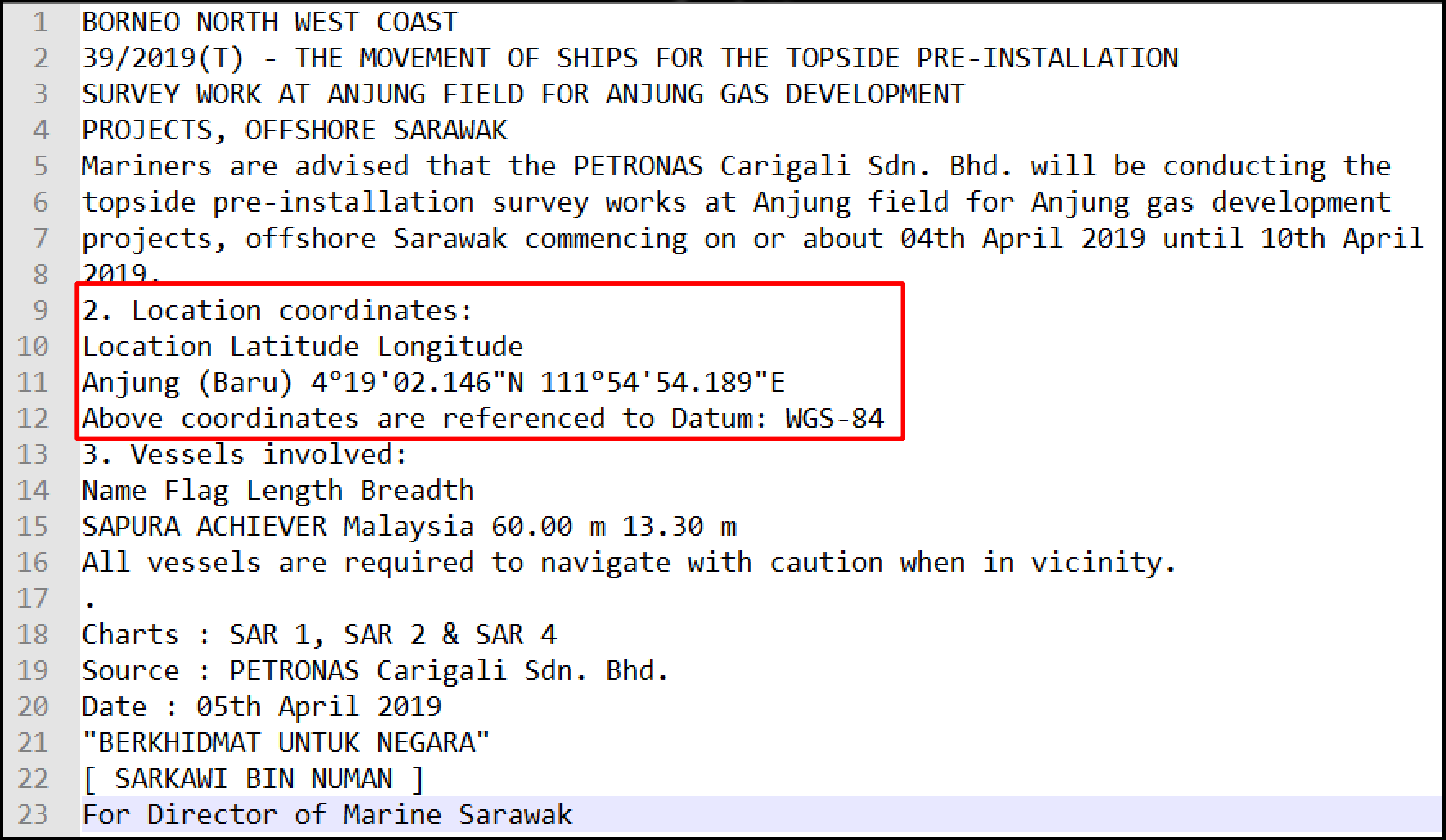
Results page are generated dynamically



PDF

Considerations	Tools
• PDF files are only visually sensible for human but the **visual structure breaks** when the texts are extracted.	• **Glob** – listing only PDF files from windows file system • **PyPDF2** – Reading text in PDF

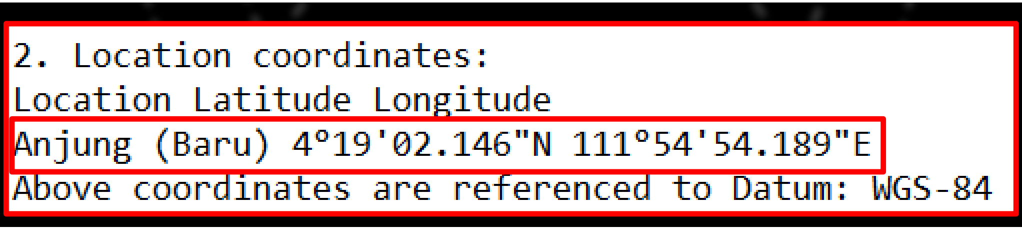
Location coordinates table becomes lines of texts



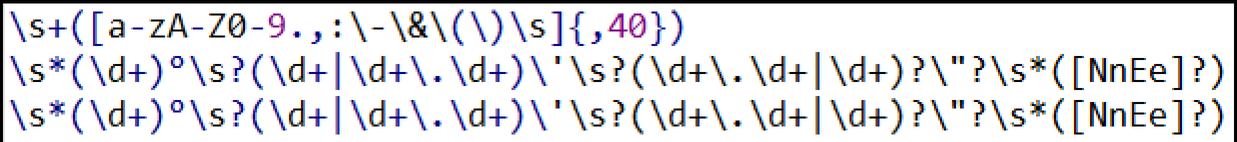
```
BORNEO NORTH WEST COAST
39/2019(T) - THE MOVEMENT OF SHIPS FOR THE TOPSIDE PRE-INSTALLATION
SURVEY WORK AT ANJUNG FIELD FOR ANJUNG GAS DEVELOPMENT
PROJECTS, OFFSHORE SARAWAK
Mariners are advised that the PETRONAS Carigali Sdn. Bhd. will be conducting the
topside pre-installation survey works at Anjung field for Anjung gas development
projects, offshore Sarawak commencing on or about 04th April 2019 until 10th April
2019.
2. Location coordinates:
Location Latitude Longitude
Anjung (Baru) 4°19'02.146"N 111°54'54.189"E
Above coordinates are referenced to Datum: WGS-84
3. Vessels involved:
Name Flag Length Breadth
SAPURA ACHIEVER Malaysia 60.00 m 13.30 m
All vessels are required to navigate with caution when in vicinity.
.
Charts : SAR 1, SAR 2 & SAR 4
Source : PETRONAS Carigali Sdn. Bhd.
Date : 05th April 2019
"BERKHIDMAT UNTUK NEGARA"
[ SARKAWI BIN NUMAN ]
For Director of Marine Sarawak
```

Coordinates

Considerations	Tools
• Designing sufficiently **specific** pattern to extract coordinates for the texts extracted from PDF. • Coordinate reference system **transformation**. • **Shapefiles** as output format	• **Regex** – Regular Expression for text pattern recognition • **Pandas** – Do MS Excel-like stuffs • **Pyproj** – Geographic projection & transformation • **Geopandas** & **Shapely** – Output to ESRI Shapefiles

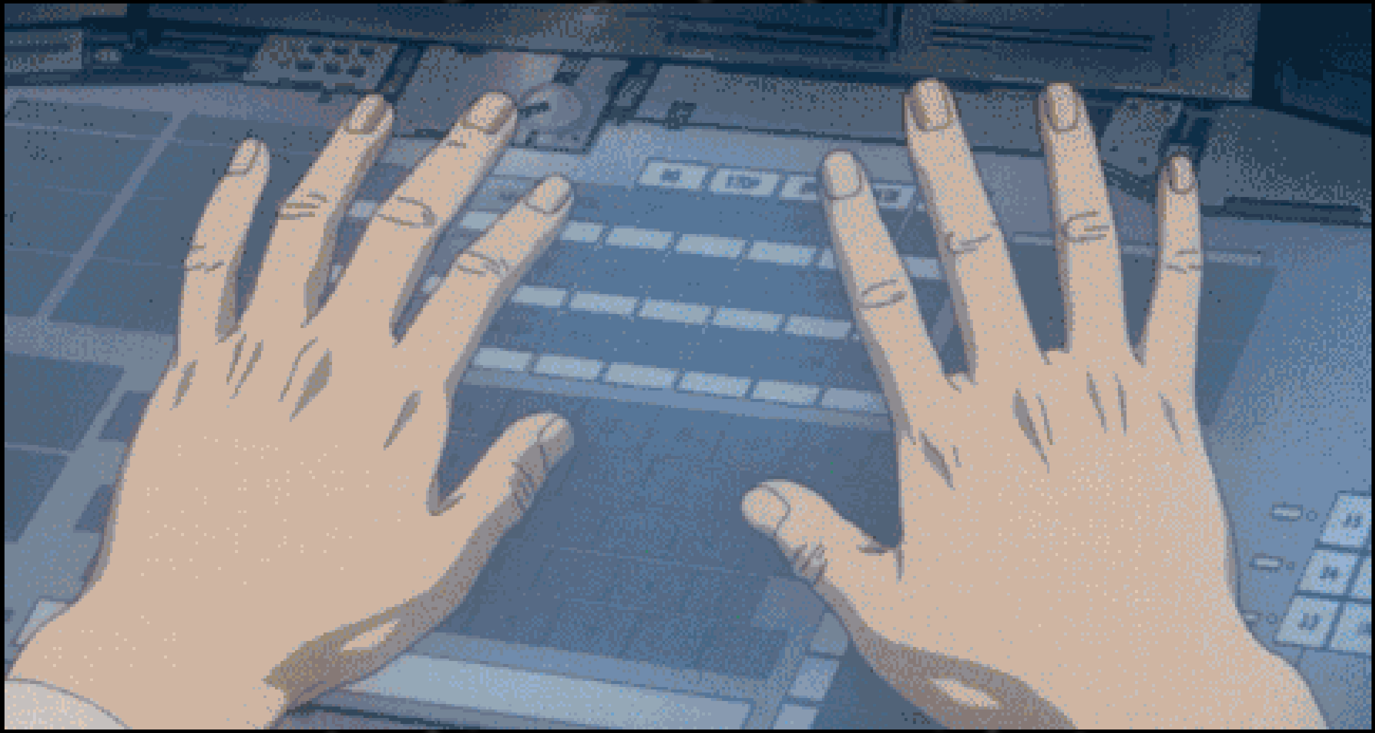


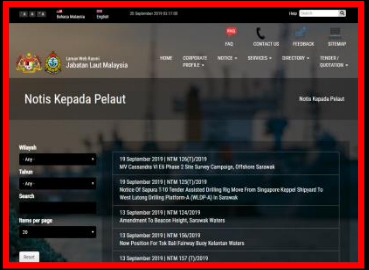
```
2. Location coordinates:
Location Latitude Longitude
Anjung (Baru) 4°19'02.146"N 111°54'54.189"E
Above coordinates are referenced to Datum: WGS-84
```



```
\s+([a-zA-Z0-9.,:\-\&\(\)\s]{,40})
\s*(\d+)°\s?(\d+|\d+\.\d+)\'\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)
\s*(\d+)°\s?(\d+|\d+\.\d+)\'\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)
```

A lot of typing later.





20 September 2019 03:17:08

Laman Web Rasmi
Jabatan Laut Malaysia

HOME | CORPORATE PROFILE ▾ | NOTICE ▾ | SERVICES ▾ | DIRECTORY ▾ | TENDER / QUOTATION ▾

FAQ | CONTACT US | FEEDBACK | SITEMAP

Notis Kepada Pelaut

Wilayah: - Any -
Tahun: - Any -
Search:
Items per page: 20
Reset

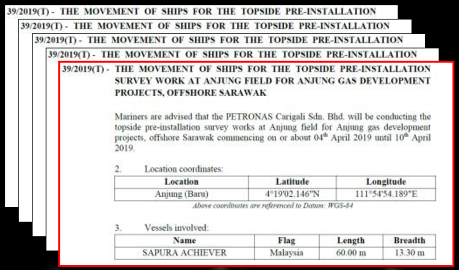
19 September 2019 | NTM 126(T)/2019
MV Cassandra VI E6 Phase 2 Site Survey Campaign, Offshore Sarawak

19 September 2019 | NTM 125(T)/2019
Notice Of Sapura T-10 Tender Assisted Drilling Rig Move From Singapore Keppel Shipyard To West Lutong Drilling Platform-A (WLDP-A) In Sarawak

13 September 2019 | NTM 124/2019
Amendment To Beacon Height, Sarawak Waters

13 September 2019 | NTM 156/2019
New Position For Tok Bali Fairway Buoy Kelantan Waters

13 September 2019 | NTM 157 (T)/2019



39/2019(T) - THE MOVEMENT OF SHIPS FOR THE TOPSIDE PRE-INSTALLATION SURVEY WORK AT ANJUNG FIELD FOR ANJUNG GAS DEVELOPMENT PROJECTS, OFFSHORE SARAWAK

Mariners are advised that the PETRONAS Carigali Sdn. Bhd. will be conducting the topside pre-installation survey works at Anjung field for Anjung gas development projects, offshore Sarawak commencing on or about 04th April 2019 until 10th April 2019.

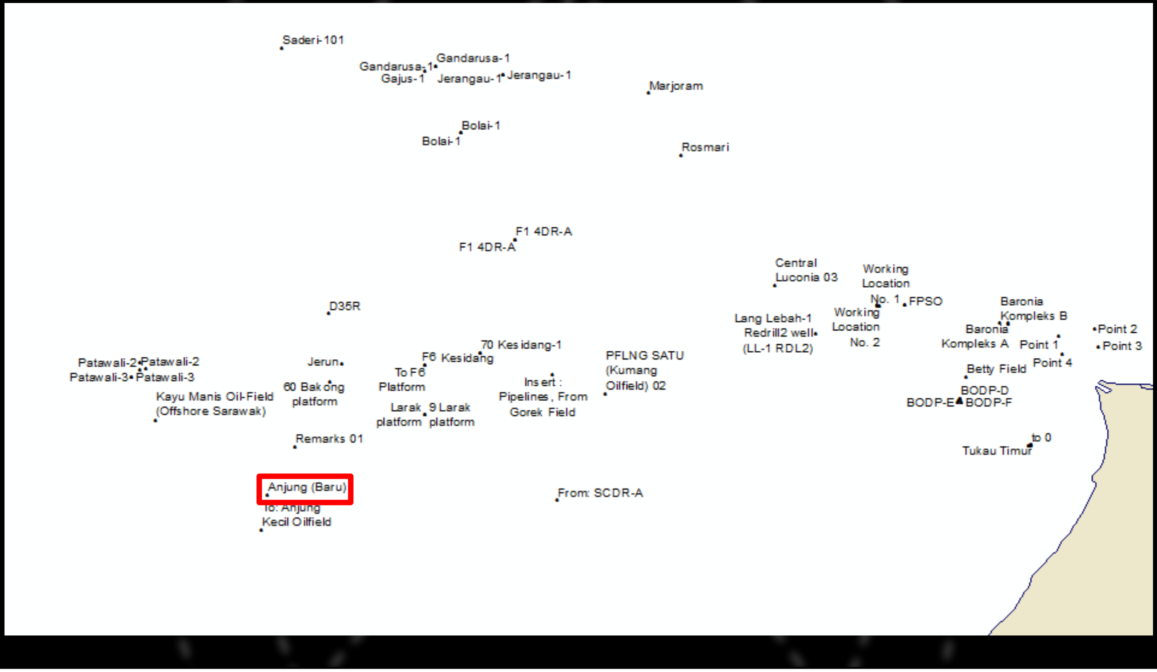
2. Location coordinates:

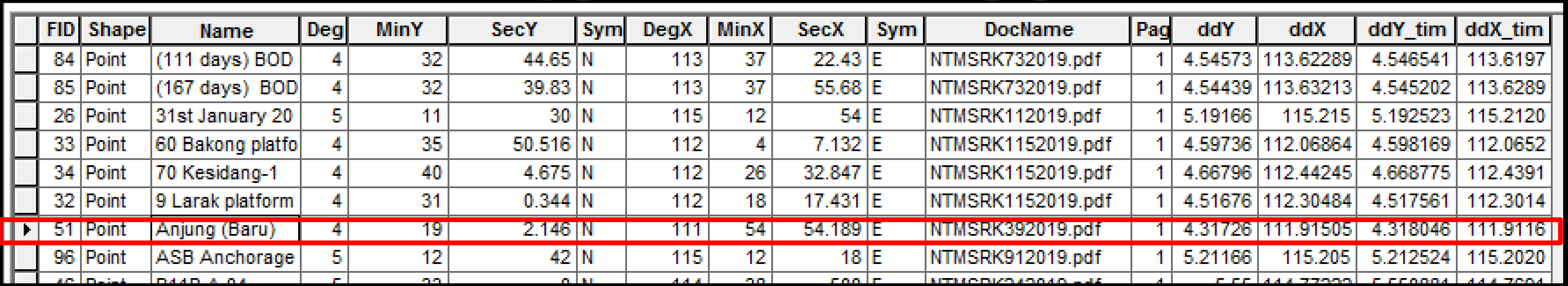
Location	Latitude	Longitude
Anjung (Baru)	4°19'02.146"N	111°54'54.189"E

Above coordinates are referenced to Datum: WGS-84

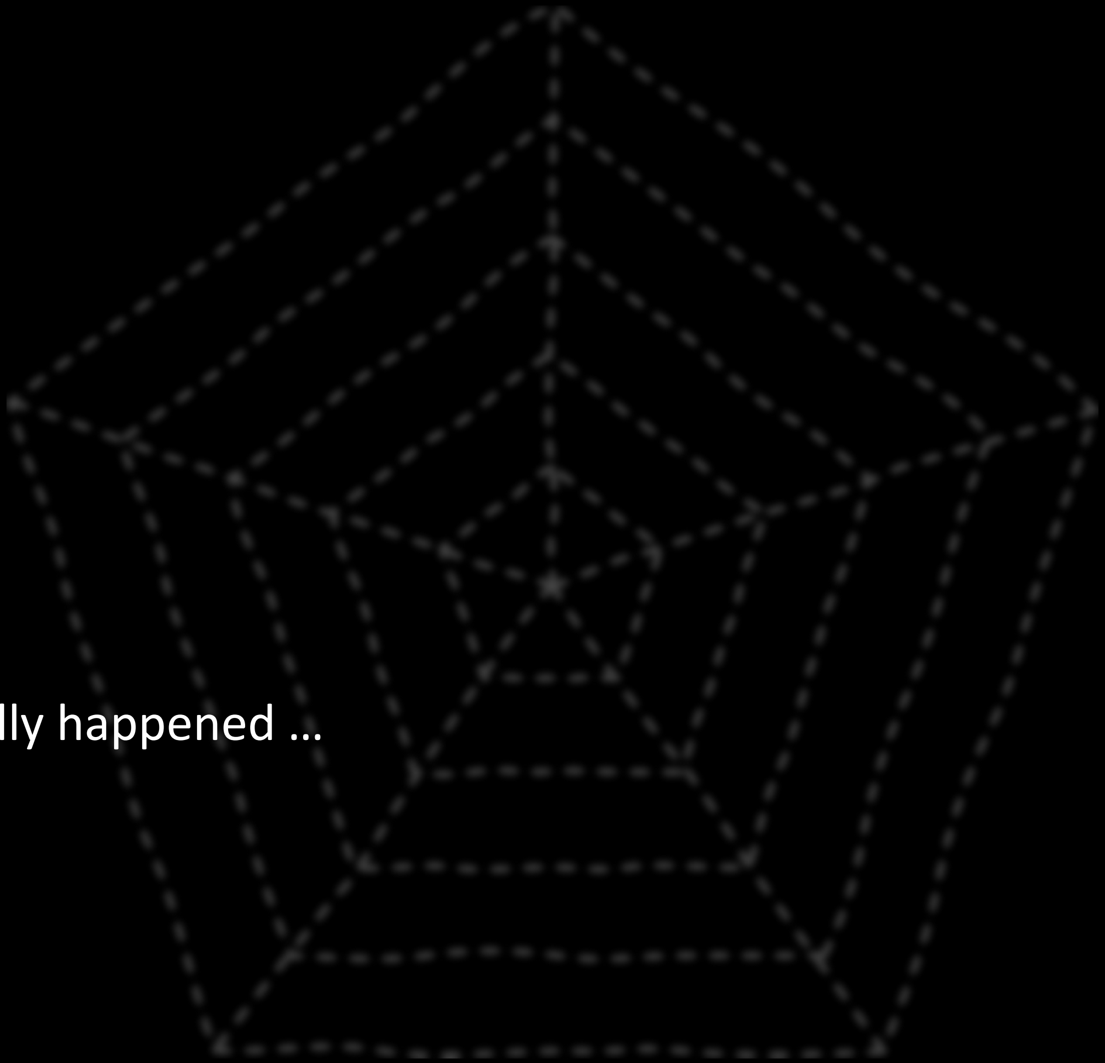
3. Vessels involved:

Name	Flag	Length	Breadth
SAPURA ACHIEVER	Malaysia	60.00 m	13.30 m





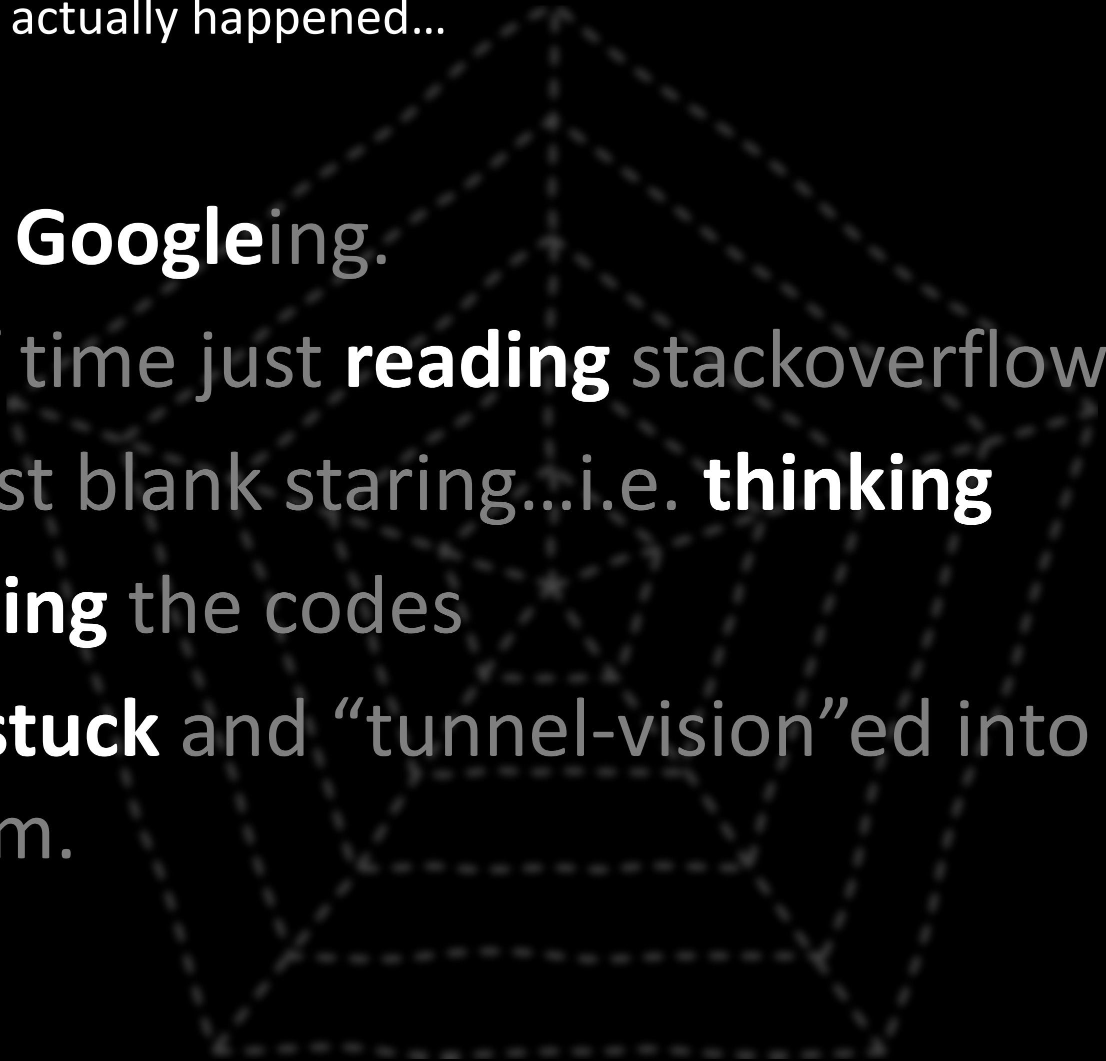
	FID	Shape	Name	Deg	MinY	SecY	Sym	DegX	MinX	SecX	Sym	DocName	Pag	ddY	ddX	ddY_tim	ddX_tim
	84	Point	(111 days) BOD	4	32	44.65	N	113	37	22.43	E	NTMSRK732019.pdf	1	4.54573	113.62289	4.546541	113.6197
	85	Point	(167 days) BOD	4	32	39.83	N	113	37	55.68	E	NTMSRK732019.pdf	1	4.54439	113.63213	4.545202	113.6289
	26	Point	31st January 20	5	11	30	N	115	12	54	E	NTMSRK112019.pdf	1	5.19166	115.215	5.192523	115.2120
	33	Point	60 Bakong platfo	4	35	50.516	N	112	4	7.132	E	NTMSRK1152019.pdf	1	4.59736	112.06864	4.598169	112.0652
	34	Point	70 Kesidang-1	4	40	4.675	N	112	26	32.847	E	NTMSRK1152019.pdf	1	4.66796	112.44245	4.668775	112.4391
	32	Point	9 Larak platform	4	31	0.344	N	112	18	17.431	E	NTMSRK1152019.pdf	1	4.51676	112.30484	4.517561	112.3014
▶	51	Point	Anjung (Baru)	4	19	2.146	N	111	54	54.189	E	NTMSRK392019.pdf	1	4.31726	111.91505	4.318046	111.9116
	96	Point	ASB Anchorage	5	12	42	N	115	12	18	E	NTMSRK912019.pdf	1	5.21166	115.205	5.212524	115.2020



What actually happened …

- **Find one** efficient way to do it manually.
- Write **pseudo-codes**.
- Automate **easiest ones first**. Small success build up confidence.
- **Repeat** repeat repeat. Iterate until all automation connected seamlessly.

What actually, actually happened…



- Lots of **Google**ing.
- 40% of time just **reading** stackoverflow posts
- 40% just blank staring…i.e. **thinking**
- 5% **typing** the codes
- Being **stuck** and “tunnel-vision”ed into a problem.

THE OPPORTUNITIES

1. *If you can see it , you can download it.*
2. *Websites with logins (can be downloaded with a tool named sessions)*
3. *Downloading published papers.*
4. *Extracting news announcement.*
5. *Browsing website which require many clickings. Automate the clicking actions.*

Take away

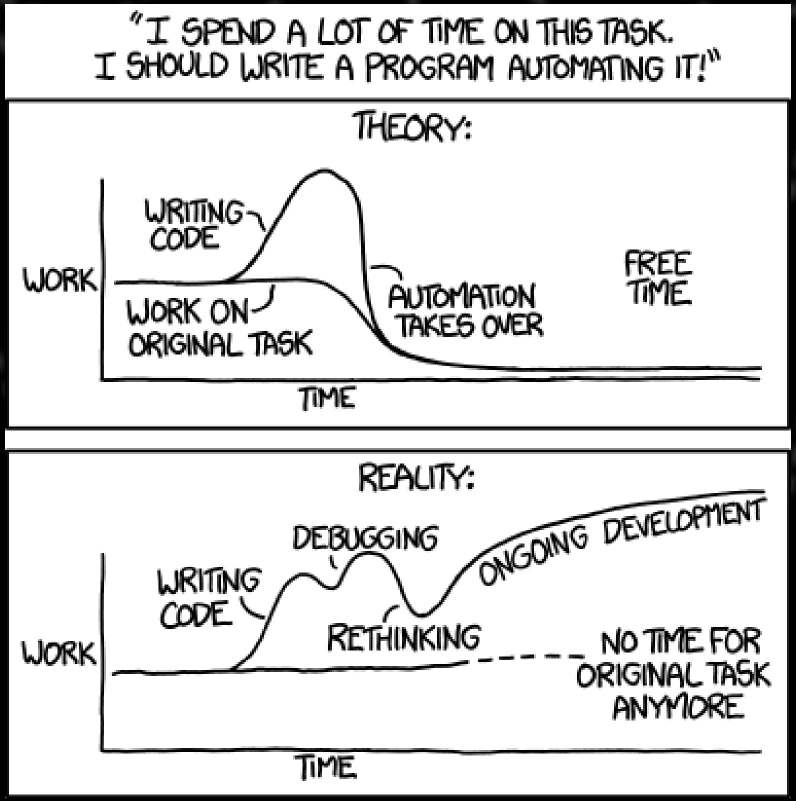
- Learn ALL these
  - For Web scraping, learn:
    - Requests
    - BeautifulSoup
    - Selenium
  - For GIS related, learn:
    - PyProj
    - Geopandas
    - Shapely
  - For General Data Manipulation, learn:
    - Pandas (by far the most useful and versatile tool)
- This simple little project teaches you many things
- Save your time and company's time
- More fun than doing it manually

Thank you!

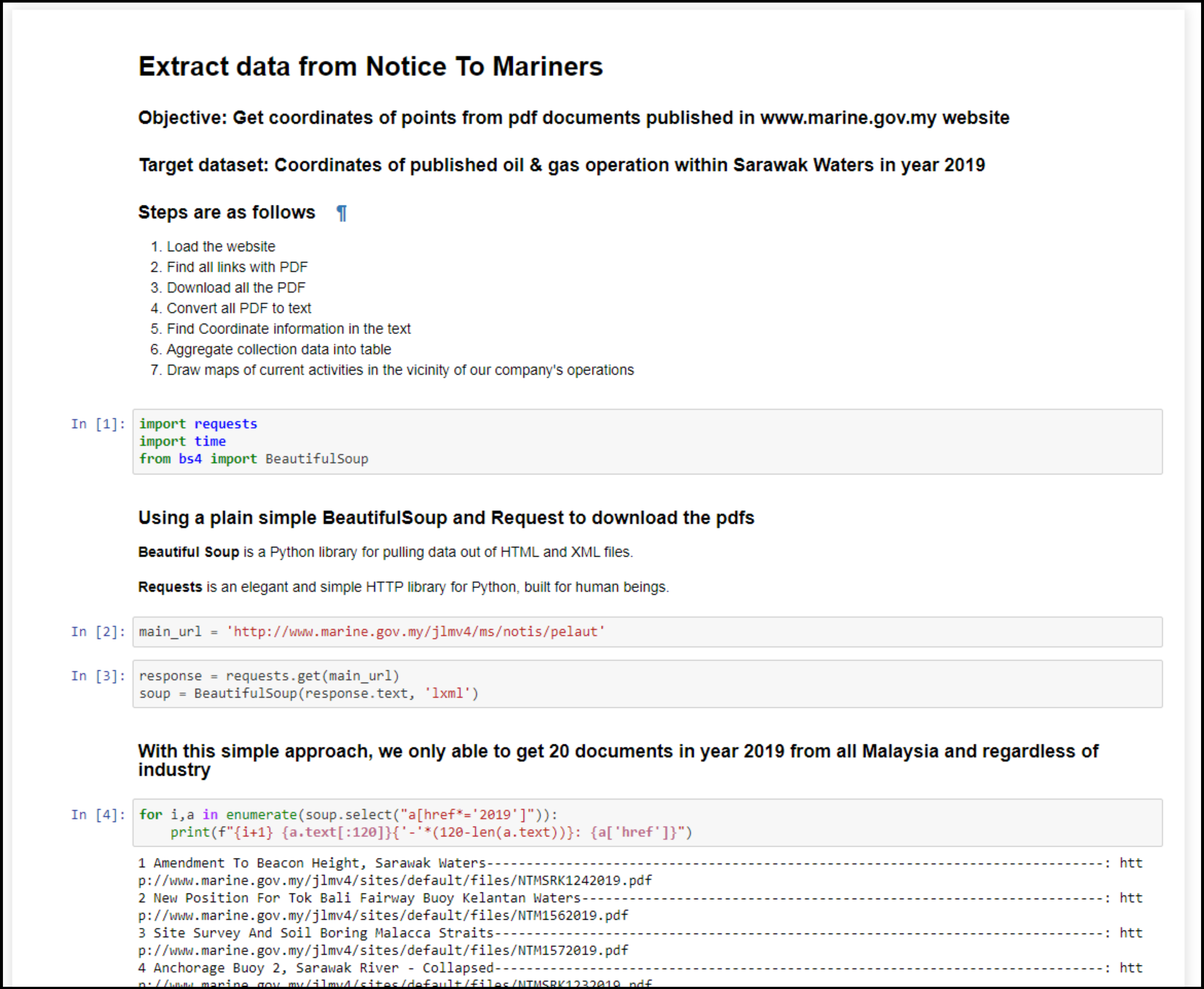
My Contact Info

Email: alvin@noex.com.my

GitHub: https://github.com/elvinado/Scraping-NTM-DEJ



xkcd.com/1319/



Extract data from Notice To Mariners

Objective: Get coordinates of points from pdf documents published in www.marine.gov.my website

Target dataset: Coordinates of published oil & gas operation within Sarawak Waters in year 2019

Steps are as follows ¶

1. Load the website
2. Find all links with PDF
3. Download all the PDF
4. Convert all PDF to text
5. Find Coordinate information in the text
6. Aggregate collection data into table
7. Draw maps of current activities in the vicinity of our company's operations

In [1]:
```
import requests
import time
from bs4 import BeautifulSoup
```

Using a plain simple BeautifulSoup and Request to download the pdfs

Beautiful Soup is a Python library for pulling data out of HTML and XML files.

Requests is an elegant and simple HTTP library for Python, built for human beings.

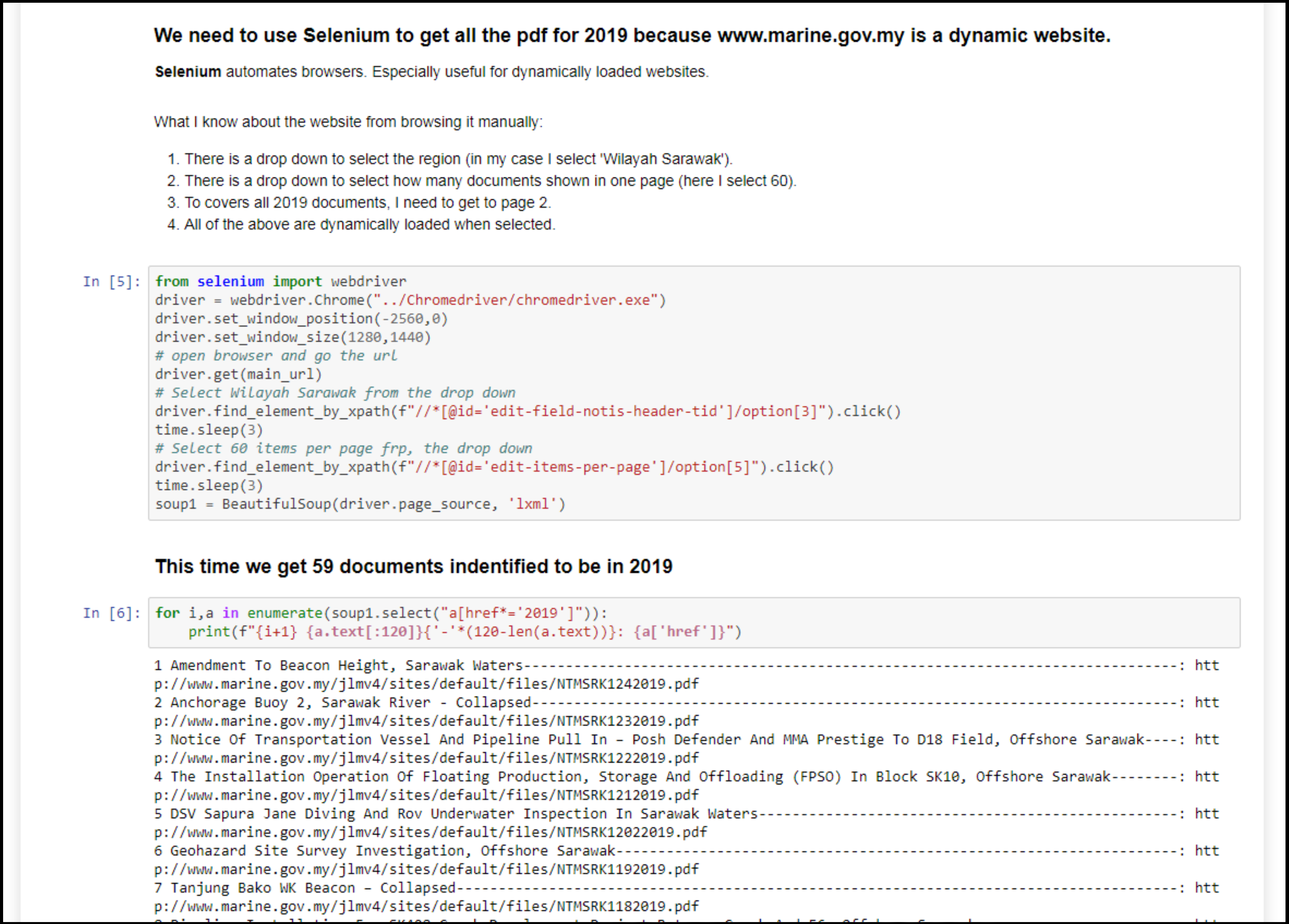
In [2]:
```
main_url = 'http://www.marine.gov.my/jlmv4/ms/notis/pelaut'
```

In [3]:
```
response = requests.get(main_url)
soup = BeautifulSoup(response.text, 'lxml')
```

With this simple approach, we only able to get 20 documents in year 2019 from all Malaysia and regardless of industry

In [4]:
```
for i,a in enumerate(soup.select("a[href*='2019']")):
    print(f"{i+1} {a.text[:120]}{'-'*(120-len(a.text))}: {a['href']}")
```

```
1 Amendment To Beacon Height, Sarawak Waters------------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1242019.pdf
2 New Position For Tok Bali Fairway Buoy Kelantan Waters------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTM1562019.pdf
3 Site Survey And Soil Boring Malacca Straits-----------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTM1572019.pdf
4 Anchorage Buoy 2, Sarawak River - Collapsed-----------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1232019.pdf
```



We need to use Selenium to get all the pdf for 2019 because www.marine.gov.my is a dynamic website.

Selenium automates browsers. Especially useful for dynamically loaded websites.

What I know about the website from browsing it manually:

1. There is a drop down to select the region (in my case I select 'Wilayah Sarawak').
2. There is a drop down to select how many documents shown in one page (here I select 60).
3. To covers all 2019 documents, I need to get to page 2.
4. All of the above are dynamically loaded when selected.

In [5]:

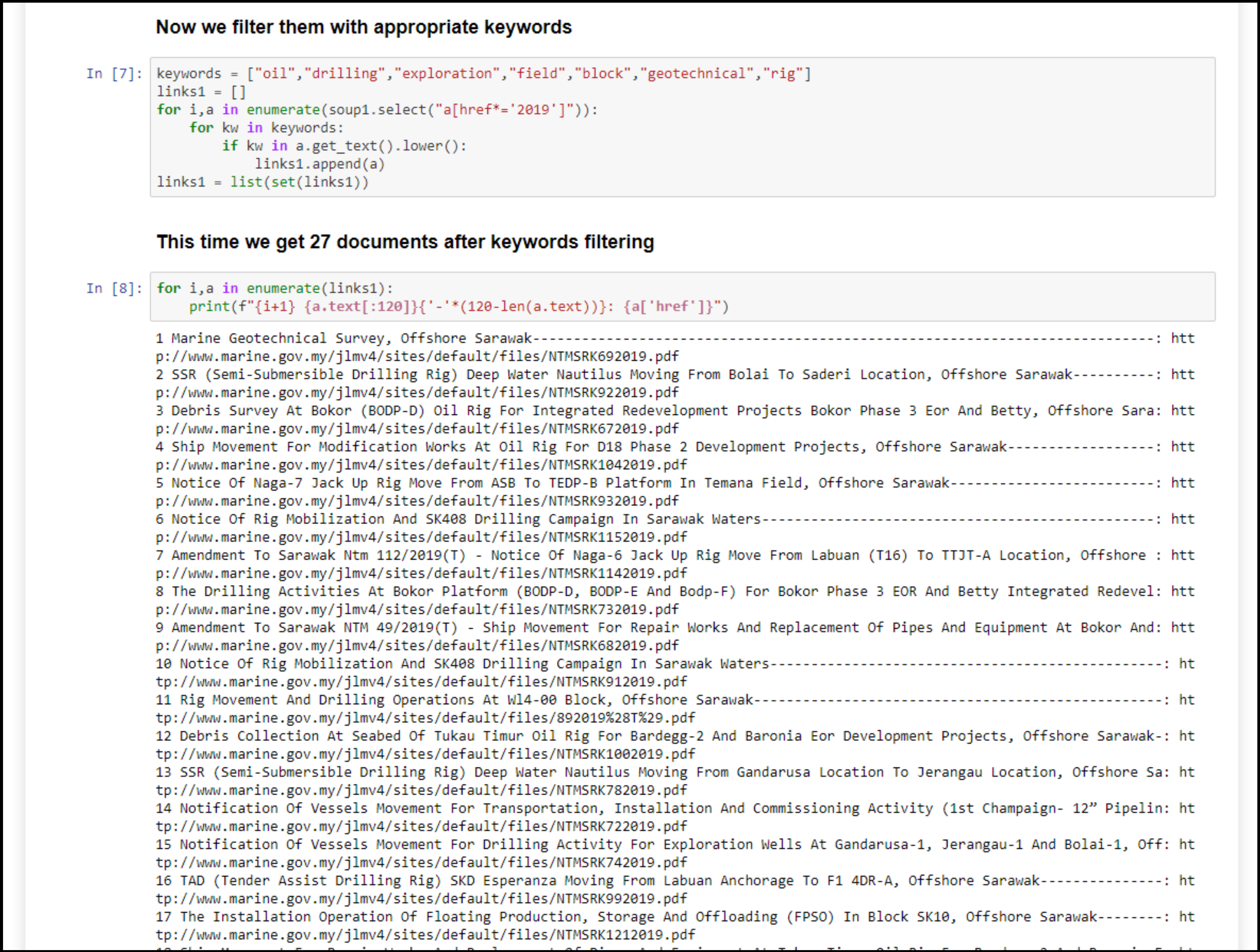
```python
from selenium import webdriver
driver = webdriver.Chrome("../Chromedriver/chromedriver.exe")
driver.set_window_position(-2560,0)
driver.set_window_size(1280,1440)
# open browser and go the url
driver.get(main_url)
# Select Wilayah Sarawak from the drop down
driver.find_element_by_xpath(f"//*[@id='edit-field-notis-header-tid']/option[3]").click()
time.sleep(3)
# Select 60 items per page frp, the drop down
driver.find_element_by_xpath(f"//*[@id='edit-items-per-page']/option[5]").click()
time.sleep(3)
soup1 = BeautifulSoup(driver.page_source, 'lxml')
```

This time we get 59 documents indentified to be in 2019

In [6]:

```python
for i,a in enumerate(soup1.select("a[href*='2019']")):
    print(f"{i+1} {a.text[:120]}{'-'*(120-len(a.text))}: {a['href']}")
```

```
1 Amendment To Beacon Height, Sarawak Waters------------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1242019.pdf
2 Anchorage Buoy 2, Sarawak River - Collapsed-----------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1232019.pdf
3 Notice Of Transportation Vessel And Pipeline Pull In – Posh Defender And MMA Prestige To D18 Field, Offshore Sarawak----: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1222019.pdf
4 The Installation Operation Of Floating Production, Storage And Offloading (FPSO) In Block SK10, Offshore Sarawak--------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1212019.pdf
5 DSV Sapura Jane Diving And Rov Underwater Inspection In Sarawak Waters--------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK12022019.pdf
6 Geohazard Site Survey Investigation, Offshore Sarawak-------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1192019.pdf
7 Tanjung Bako WK Beacon – Collapsed--------------------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1182019.pdf
```



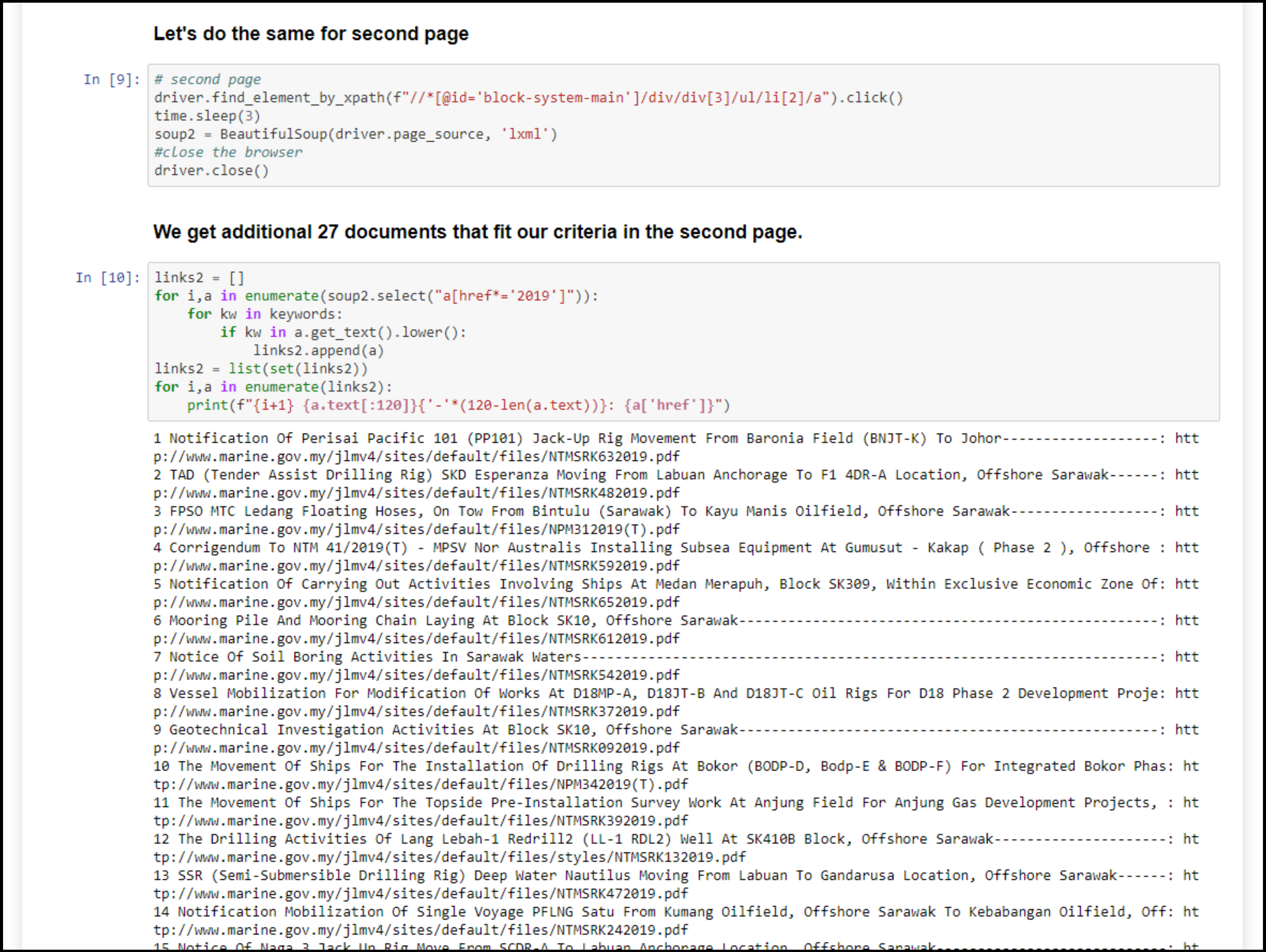
Now we filter them with appropriate keywords

```
In [7]: keywords = ["oil","drilling","exploration","field","block","geotechnical","rig"]
        links1 = []
        for i,a in enumerate(soup1.select("a[href*='2019']")):
            for kw in keywords:
                if kw in a.get_text().lower():
                    links1.append(a)
        links1 = list(set(links1))
```

This time we get 27 documents after keywords filtering

```
In [8]: for i,a in enumerate(links1):
            print(f"{i+1} {a.text[:120]}{'-'*(120-len(a.text))}: {a['href']}")
```

```
1 Marine Geotechnical Survey, Offshore Sarawak----------------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK692019.pdf
2 SSR (Semi-Submersible Drilling Rig) Deep Water Nautilus Moving From Bolai To Saderi Location, Offshore Sarawak----------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK922019.pdf
3 Debris Survey At Bokor (BODP-D) Oil Rig For Integrated Redevelopment Projects Bokor Phase 3 Eor And Betty, Offshore Sara: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK672019.pdf
4 Ship Movement For Modification Works At Oil Rig For D18 Phase 2 Development Projects, Offshore Sarawak------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1042019.pdf
5 Notice Of Naga-7 Jack Up Rig Move From ASB To TEDP-B Platform In Temana Field, Offshore Sarawak-------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK932019.pdf
6 Notice Of Rig Mobilization And SK408 Drilling Campaign In Sarawak Waters------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1152019.pdf
7 Amendment To Sarawak Ntm 112/2019(T) - Notice Of Naga-6 Jack Up Rig Move From Labuan (T16) To TTJT-A Location, Offshore : htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1142019.pdf
8 The Drilling Activities At Bokor Platform (BODP-D, BODP-E And Bodp-F) For Bokor Phase 3 EOR And Betty Integrated Redevel: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK732019.pdf
9 Amendment To Sarawak NTM 49/2019(T) - Ship Movement For Repair Works And Replacement Of Pipes And Equipment At Bokor And: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK682019.pdf
10 Notice Of Rig Mobilization And SK408 Drilling Campaign In Sarawak Waters------------------------------------------------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK912019.pdf
11 Rig Movement And Drilling Operations At Wl4-00 Block, Offshore Sarawak--------------------------------------------------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/892019%28T%29.pdf
12 Debris Collection At Seabed Of Tukau Timur Oil Rig For Bardegg-2 And Baronia Eor Development Projects, Offshore Sarawak-: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1002019.pdf
13 SSR (Semi-Submersible Drilling Rig) Deep Water Nautilus Moving From Gandarusa Location To Jerangau Location, Offshore Sa: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK782019.pdf
14 Notification Of Vessels Movement For Transportation, Installation And Commissioning Activity (1st Champaign- 12” Pipelin: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK722019.pdf
15 Notification Of Vessels Movement For Drilling Activity For Exploration Wells At Gandarusa-1, Jerangau-1 And Bolai-1, Off: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK742019.pdf
16 TAD (Tender Assist Drilling Rig) SKD Esperanza Moving From Labuan Anchorage To F1 4DR-A, Offshore Sarawak---------------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK992019.pdf
17 The Installation Operation Of Floating Production, Storage And Offloading (FPSO) In Block SK10, Offshore Sarawak--------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1212019.pdf
```



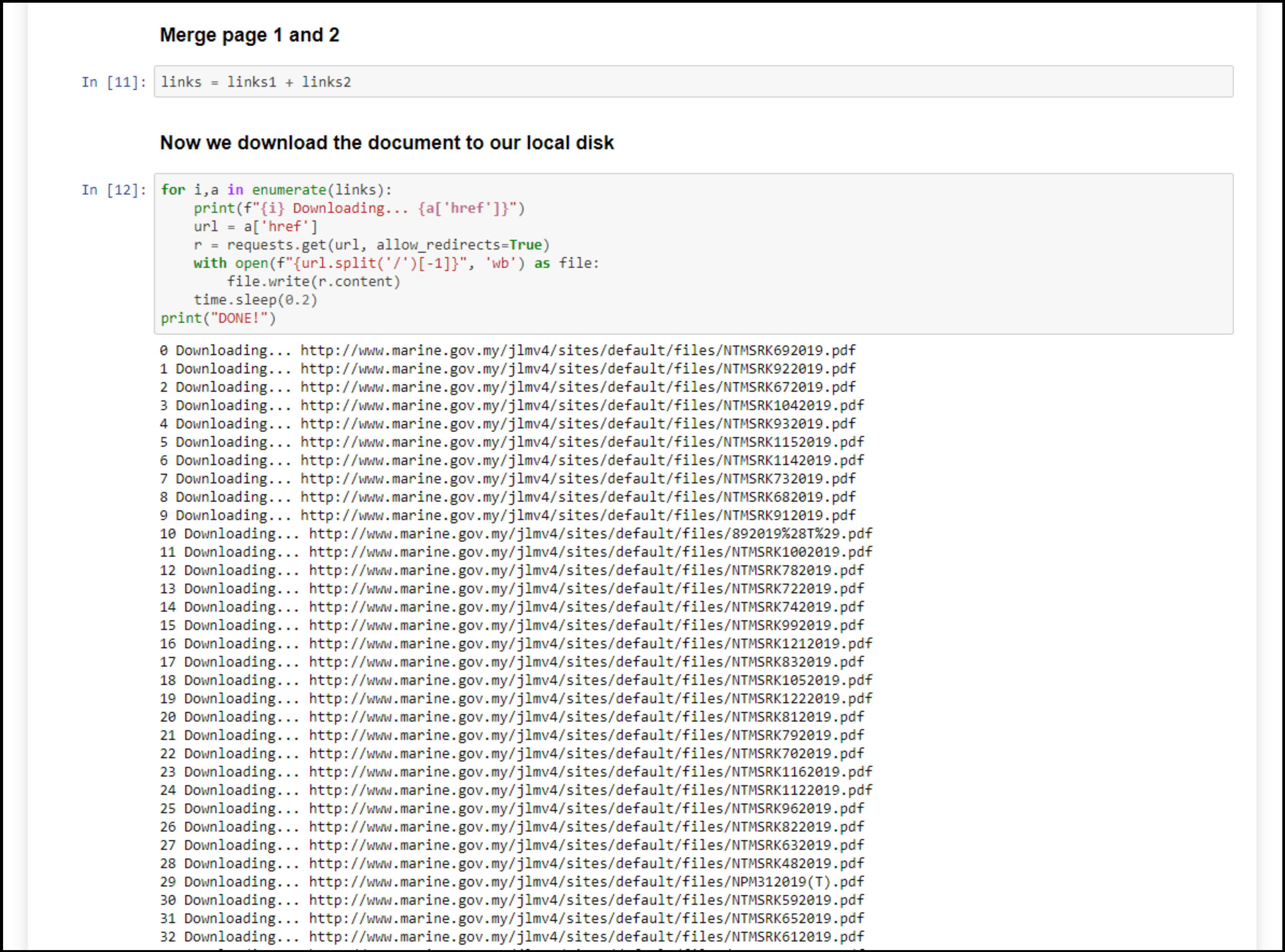
Let's do the same for second page

In [9]:
```
# second page
driver.find_element_by_xpath(f"//*[@id='block-system-main']/div/div[3]/ul/li[2]/a").click()
time.sleep(3)
soup2 = BeautifulSoup(driver.page_source, 'lxml')
#close the browser
driver.close()
```

We get additional 27 documents that fit our criteria in the second page.

In [10]:
```
links2 = []
for i,a in enumerate(soup2.select("a[href*='2019']")):
    for kw in keywords:
        if kw in a.get_text().lower():
            links2.append(a)
links2 = list(set(links2))
for i,a in enumerate(links2):
    print(f"{i+1} {a.text[:120]}{'-'*(120-len(a.text))}: {a['href']}")
```

```
1 Notification Of Perisai Pacific 101 (PP101) Jack-Up Rig Movement From Baronia Field (BNJT-K) To Johor-------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK632019.pdf
2 TAD (Tender Assist Drilling Rig) SKD Esperanza Moving From Labuan Anchorage To F1 4DR-A Location, Offshore Sarawak------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK482019.pdf
3 FPSO MTC Ledang Floating Hoses, On Tow From Bintulu (Sarawak) To Kayu Manis Oilfield, Offshore Sarawak------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NPM312019(T).pdf
4 Corrigendum To NTM 41/2019(T) - MPSV Nor Australis Installing Subsea Equipment At Gumusut - Kakap ( Phase 2 ), Offshore : htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK592019.pdf
5 Notification Of Carrying Out Activities Involving Ships At Medan Merapuh, Block SK309, Within Exclusive Economic Zone Of: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK652019.pdf
6 Mooring Pile And Mooring Chain Laying At Block SK10, Offshore Sarawak---------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK612019.pdf
7 Notice Of Soil Boring Activities In Sarawak Waters----------------------------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK542019.pdf
8 Vessel Mobilization For Modification Of Works At D18MP-A, D18JT-B And D18JT-C Oil Rigs For D18 Phase 2 Development Proje: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK372019.pdf
9 Geotechnical Investigation Activities At Block SK10, Offshore Sarawak---------------------------------------------------: htt
p://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK092019.pdf
10 The Movement Of Ships For The Installation Of Drilling Rigs At Bokor (BODP-D, Bodp-E & BODP-F) For Integrated Bokor Phas: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NPM342019(T).pdf
11 The Movement Of Ships For The Topside Pre-Installation Survey Work At Anjung Field For Anjung Gas Development Projects, : ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK392019.pdf
12 The Drilling Activities Of Lang Lebah-1 Redrill2 (LL-1 RDL2) Well At SK410B Block, Offshore Sarawak---------------------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/styles/NTMSRK132019.pdf
13 SSR (Semi-Submersible Drilling Rig) Deep Water Nautilus Moving From Labuan To Gandarusa Location, Offshore Sarawak------: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK472019.pdf
14 Notification Mobilization Of Single Voyage PFLNG Satu From Kumang Oilfield, Offshore Sarawak To Kebabangan Oilfield, Off: ht
tp://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK242019.pdf
15 Notice Of Naga 3 Jack Up Rig Move From SCDP-A To Labuan Anchorage Location, Offshore Sarawak----------------------------: ht
```



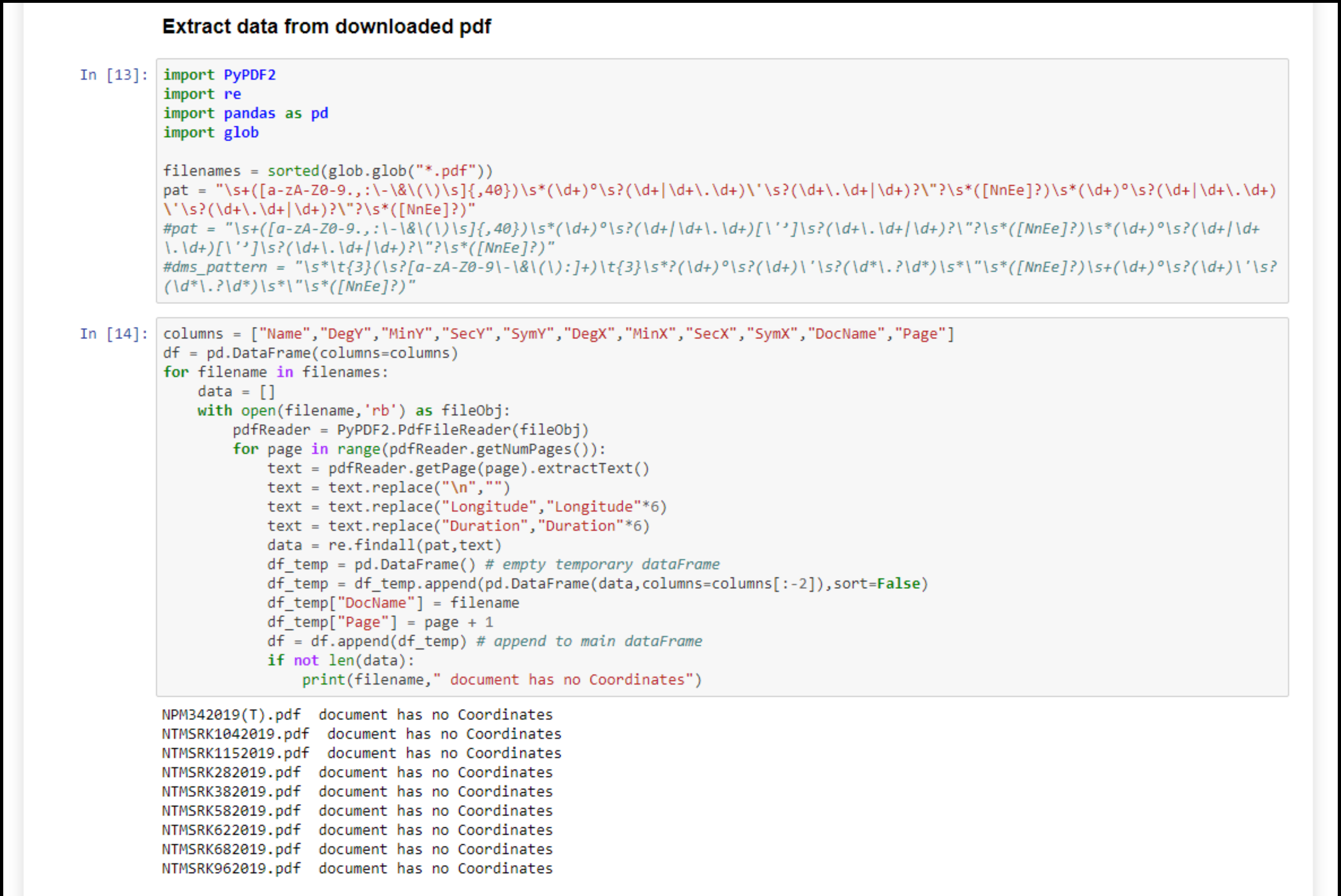
Merge page 1 and 2

```
In [11]: links = links1 + links2
```

Now we download the document to our local disk

```
In [12]: for i,a in enumerate(links):
             print(f"{i} Downloading... {a['href']}")
             url = a['href']
             r = requests.get(url, allow_redirects=True)
             with open(f"{url.split('/')[-1]}", 'wb') as file:
                 file.write(r.content)
             time.sleep(0.2)
         print("DONE!")
```

```
0 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK692019.pdf
1 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK922019.pdf
2 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK672019.pdf
3 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1042019.pdf
4 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK932019.pdf
5 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1152019.pdf
6 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1142019.pdf
7 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK732019.pdf
8 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK682019.pdf
9 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK912019.pdf
10 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/892019%28T%29.pdf
11 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1002019.pdf
12 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK782019.pdf
13 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK722019.pdf
14 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK742019.pdf
15 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK992019.pdf
16 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1212019.pdf
17 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK832019.pdf
18 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1052019.pdf
19 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1222019.pdf
20 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK812019.pdf
21 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK792019.pdf
22 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK702019.pdf
23 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1162019.pdf
24 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK1122019.pdf
25 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK962019.pdf
26 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK822019.pdf
27 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK632019.pdf
28 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK482019.pdf
29 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NPM312019(T).pdf
30 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK592019.pdf
31 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK652019.pdf
32 Downloading... http://www.marine.gov.my/jlmv4/sites/default/files/NTMSRK612019.pdf
```



Extract data from downloaded pdf

In [13]:

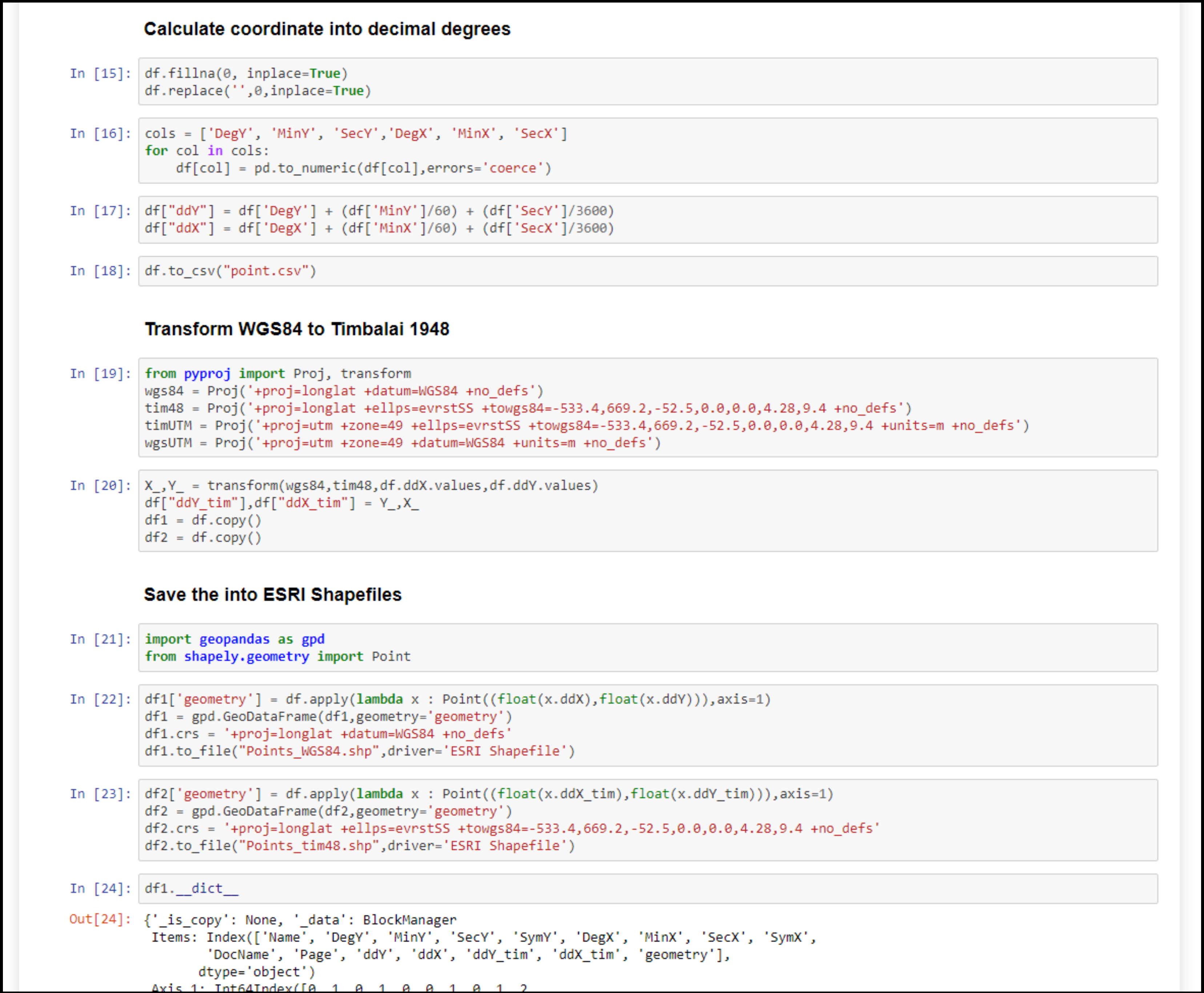
```
import PyPDF2
import re
import pandas as pd
import glob

filenames = sorted(glob.glob("*.pdf"))
pat = "\s+([a-zA-Z0-9.,:\-\&\(\)\s]{,40})\s*(\d+)°\s?(\d+|\d+\.\d+)\'\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)\s*(\d+)°\s?(\d+|\d+\.\d+)
\'\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)"
#pat = "\s+([a-zA-Z0-9.,:\-\&\(\)\s]{,40})\s*(\d+)°\s?(\d+|\d+\.\d+)[\'’]\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)\s*(\d+)°\s?(\d+|\d+
\.\d+)[\'’]\s?(\d+\.\d+|\d+)?\"?\s*([NnEe]?)"
#dms_pattern = "\s*\t{3}(\s?[a-zA-Z0-9\-\&\(\):]+)\t{3}\s*?(\d+)°\s?(\d+)\'\s?(\d*\.?\d*)\s*\"\s*([NnEe]?)\s+(\d+)°\s?(\d+)\'\s?
(\d*\.?\d*)\s*\"\s*([NnEe]?)"
```

In [14]:

```
columns = ["Name","DegY","MinY","SecY","SymY","DegX","MinX","SecX","SymX","DocName","Page"]
df = pd.DataFrame(columns=columns)
for filename in filenames:
    data = []
    with open(filename,'rb') as fileObj:
        pdfReader = PyPDF2.PdfFileReader(fileObj)
        for page in range(pdfReader.getNumPages()):
            text = pdfReader.getPage(page).extractText()
            text = text.replace("\n","")
            text = text.replace("Longitude","Longitude"*6)
            text = text.replace("Duration","Duration"*6)
            data = re.findall(pat,text)
            df_temp = pd.DataFrame() # empty temporary dataFrame
            df_temp = df_temp.append(pd.DataFrame(data,columns=columns[:-2]),sort=False)
            df_temp["DocName"] = filename
            df_temp["Page"] = page + 1
            df = df.append(df_temp) # append to main dataFrame
            if not len(data):
                print(filename," document has no Coordinates")
```

```
NPM342019(T).pdf  document has no Coordinates
NTMSRK1042019.pdf  document has no Coordinates
NTMSRK1152019.pdf  document has no Coordinates
NTMSRK282019.pdf  document has no Coordinates
NTMSRK382019.pdf  document has no Coordinates
NTMSRK582019.pdf  document has no Coordinates
NTMSRK622019.pdf  document has no Coordinates
NTMSRK682019.pdf  document has no Coordinates
NTMSRK962019.pdf  document has no Coordinates
```



Calculate coordinate into decimal degrees

```
In [15]: df.fillna(0, inplace=True)
         df.replace('',0,inplace=True)
```

```
In [16]: cols = ['DegY', 'MinY', 'SecY','DegX', 'MinX', 'SecX']
         for col in cols:
             df[col] = pd.to_numeric(df[col],errors='coerce')
```

```
In [17]: df["ddY"] = df['DegY'] + (df['MinY']/60) + (df['SecY']/3600)
         df["ddX"] = df['DegX'] + (df['MinX']/60) + (df['SecX']/3600)
```

```
In [18]: df.to_csv("point.csv")
```

Transform WGS84 to Timbalai 1948

```
In [19]: from pyproj import Proj, transform
         wgs84 = Proj('+proj=longlat +datum=WGS84 +no_defs')
         tim48 = Proj('+proj=longlat +ellps=evrstSS +towgs84=-533.4,669.2,-52.5,0.0,0.0,4.28,9.4 +no_defs')
         timUTM = Proj('+proj=utm +zone=49 +ellps=evrstSS +towgs84=-533.4,669.2,-52.5,0.0,0.0,4.28,9.4 +units=m +no_defs')
         wgsUTM = Proj('+proj=utm +zone=49 +datum=WGS84 +units=m +no_defs')
```

```
In [20]: X_,Y_ = transform(wgs84,tim48,df.ddX.values,df.ddY.values)
         df["ddY_tim"],df["ddX_tim"] = Y_,X_
         df1 = df.copy()
         df2 = df.copy()
```

Save the into ESRI Shapefiles

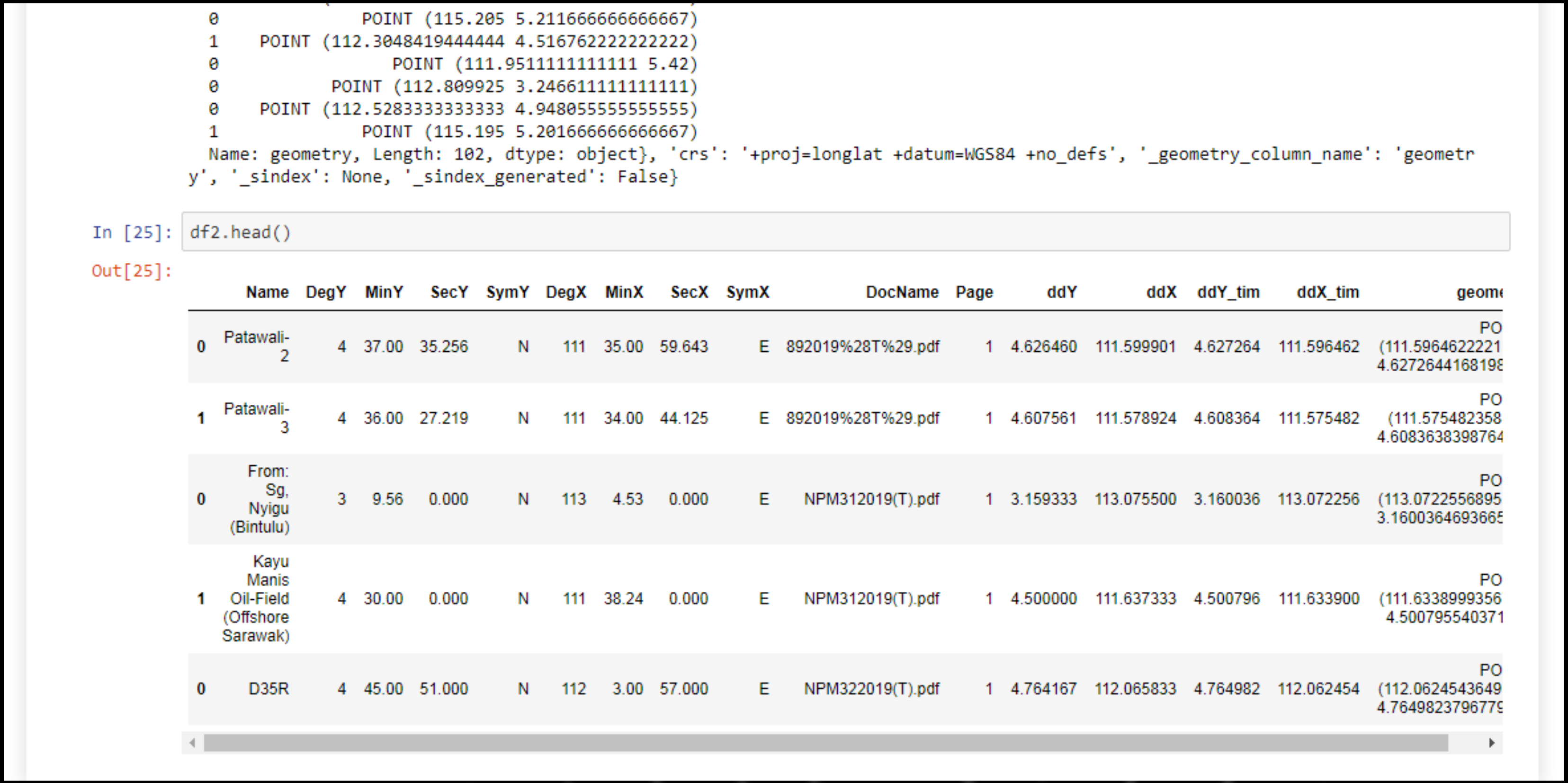
```
In [21]: import geopandas as gpd
         from shapely.geometry import Point
```

```
In [22]: df1['geometry'] = df.apply(lambda x : Point((float(x.ddX),float(x.ddY))),axis=1)
         df1 = gpd.GeoDataFrame(df1,geometry='geometry')
         df1.crs = '+proj=longlat +datum=WGS84 +no_defs'
         df1.to_file("Points_WGS84.shp",driver='ESRI Shapefile')
```

```
In [23]: df2['geometry'] = df.apply(lambda x : Point((float(x.ddX_tim),float(x.ddY_tim))),axis=1)
         df2 = gpd.GeoDataFrame(df2,geometry='geometry')
         df2.crs = '+proj=longlat +ellps=evrstSS +towgs84=-533.4,669.2,-52.5,0.0,0.0,4.28,9.4 +no_defs'
         df2.to_file("Points_tim48.shp",driver='ESRI Shapefile')
```

```
In [24]: df1.__dict__
```

```
Out[24]: {'_is_copy': None, '_data': BlockManager
          Items: Index(['Name', 'DegY', 'MinY', 'SecY', 'SymY', 'DegX', 'MinX', 'SecX', 'SymX',
                 'DocName', 'Page', 'ddY', 'ddX', 'ddY_tim', 'ddX_tim', 'geometry'],
                dtype='object')
          Axis 1: Int64Index([0, 1, 0, 1, 0, 0, 1, 0, 1, 2,
```



```
0            POINT (115.205 5.211666666666667)
1    POINT (112.3048419444444 4.516762222222222)
0               POINT (111.9511111111111 5.42)
0         POINT (112.809925 3.246611111111111)
0    POINT (112.5283333333333 4.948055555555555)
1            POINT (115.195 5.201666666666667)
Name: geometry, Length: 102, dtype: object}, 'crs': '+proj=longlat +datum=WGS84 +no_defs', '_geometry_column_name': 'geometr
y', '_sindex': None, '_sindex_generated': False}
```

In [25]:

```
df2.head()
```

Out[25]:

	Name	DegY	MinY	SecY	SymY	DegX	MinX	SecX	SymX	DocName	Page	ddY	ddX	ddY_tim	ddX_tim	geome
0	Patawali-2	4	37.00	35.256	N	111	35.00	59.643	E	892019%28T%29.pdf	1	4.626460	111.599901	4.627264	111.596462	PO (111.5964622221 4.6272644168198
1	Patawali-3	4	36.00	27.219	N	111	34.00	44.125	E	892019%28T%29.pdf	1	4.607561	111.578924	4.608364	111.575482	PO (111.575482358 4.6083638398764
0	From: Sg, Nyigu (Bintulu)	3	9.56	0.000	N	113	4.53	0.000	E	NPM312019(T).pdf	1	3.159333	113.075500	3.160036	113.072256	PO (113.0722556895 3.1600364693665
1	Kayu Manis Oil-Field (Offshore Sarawak)	4	30.00	0.000	N	111	38.24	0.000	E	NPM312019(T).pdf	1	4.500000	111.637333	4.500796	111.633900	PO (111.6338999356 4.500795540371
0	D35R	4	45.00	51.000	N	112	3.00	57.000	E	NPM322019(T).pdf	1	4.764167	112.065833	4.764982	112.062454	PO (112.0624543649 4.7649823796779