PyCon MY 2024

USING PYTHON ON SPRM OFFENDERS DATA

Aissatou Diallo
Siti Nurliza Samsudin
Trinidad Carreno

Sinar Project
hacking democracy

INTRODUCTION

- **Corruption Offenders Database**
  - Official Portal Malaysian Anti-Corruption Commission (Suruhanjaya Pencegahan Rasuah Malaysia)
- Data includes the offenders' images and other details: personal information, summary of offense, penalty, and employer information.
- Used a Python crawler to convert data into a machine-readable format for public access.

Sinar Project
hacking democracy

SPRM

sprm.gov.my/index.php?id=21&page_id=96

Official Portal
MALAYSIAN ANTI-CORRUPTION COMMISSION
Independent • Transparent • Professional

Toll Free : 1-800-88-6000

REPORT CORRUPTION

HOME CORPORATE INFO CONTACT US

Search...

Corruption Offender Database

Corruption Offender Database

CORRUPTION OFFENDERS DATABASE

The Corruption Offenders Database primarily serves as a deterrent measure in sending a clear message in the fight against corruption. The database serves as a vital awareness tool to assist the public and organisations in assisting in the due diligence process concerning hiring, appointment and promotion of employees.

Disclaimer:"The information below for general reference only. The Malaysian Anti-Corruption Commission shall not be responsible for any error in or mission"

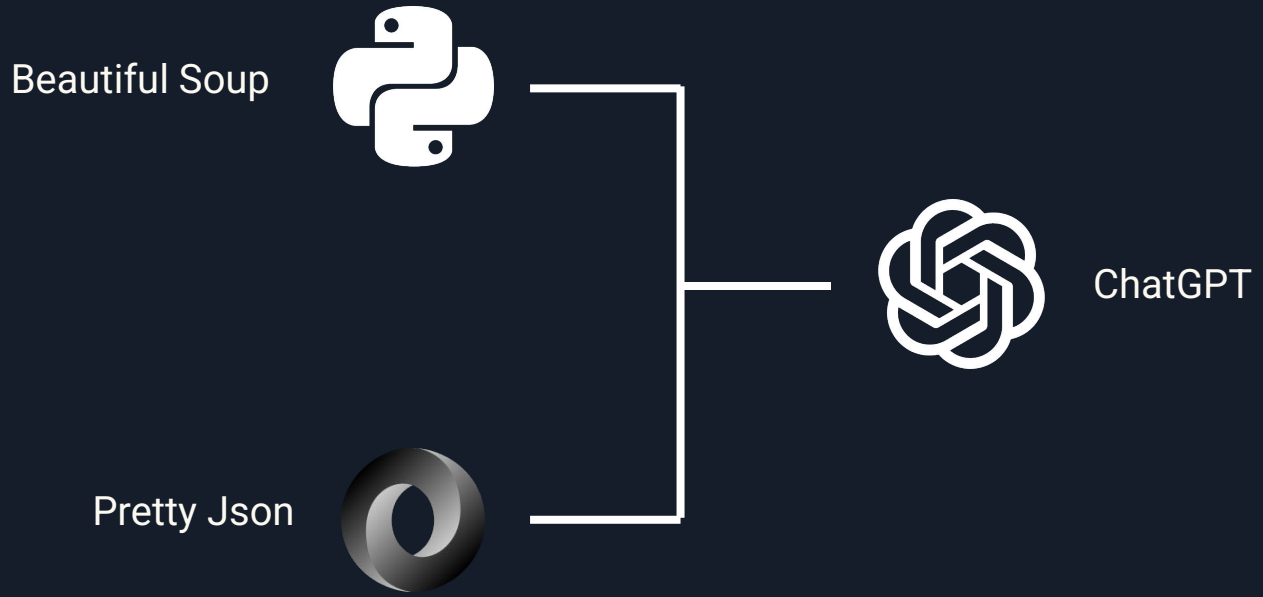
Year	All	Keyword	Gender
Type	Nationality	SEARCH	RESET STATISTICS

Showing **1-8** of **346** items.




TANYA SARA

DATA CRAWLING AND CLEANING



BEAUTIFUL SOUP

Utility

- Data extraction: used to scrape and extract HTML tables and images from the SPRM website.
- Pagination handling: automates the process of navigating through multiple pages of data.

Challenges & Lessons Learned

- Challenge: inconsistent HTML structures across pages, such as missing tables or images.
- Lesson: develop robust error handling in functions like get_person_data to ensure the script doesn't break when expected elements are missing.

SinarProject
hacking democracy

```python
def get_next_page_url(soup):
    """Return the URL of the next page or None if there is no next page."""
    next_page = soup.find('a', string='»')
    if next_page and 'href' in next_page.attrs:
        return "https://www.sprm.gov.my" + next_page['href']
    return None
```

```python
def get_person_data(soup):
    """Extract person data from the page content and return a DataFrame."""
    tables = soup.find_all('table')
    if not tables:
        return pd.DataFrame()
```

```python
    try:
        data_frames = pd.read_html(html_io)
    except Exception as e:
        print(f"Error reading HTML tables: {e}")
        return pd.DataFrame()
```

SinarProject
hacking democracy

PRETTY JSON

Utility

- Data cleaning: formats cleaned and reorganized data into JSON for easy export and sharing.
- Structure validation: ensures that the JSON output is well-structured and conforms to standard formats using the .to_json() method.

Challenges & Lessons Learned

- Challenge: managing complex nested data structures when converting DataFrames to JSON.
- Lesson: use Pretty JSON to validate and debug the JSON output, ensuring that nested structures and arrays are correctly formatted.

SinarProject
hacking democracy

```python
def reorganize_dataframe(df):
    """Reorganize the DataFrame to structure personal information and cases."""
    organized_data = []
    current_accused = {}

    for index, row in df.iterrows():
        if row[0] == 'Accused':
            if current_accused:
                organized_data.append(current_accused)
            current_accused = {'Cases': []}
        if pd.notna(row[0]):
            if row[0] == '#':
                current_accused['Cases'].append({
                    'No Kes': row[1],
                    'Ringkasan Pertuduhan': row[2],
                    'Kesalahan': row[3],
                    ..................

    final_df = pd.DataFrame(final_data)
    return final_df
```

```python
# Convert to JSON
json_filename = '1-2024-07-29_sprm_data.json'
final_df.to_json(json_filename, orient='records', lines=True)
```

CHAT GPT

Utility

- Code assistance: generated Python scripts and provided guidance on structuring the web scraping, data reorganization, and data export processes.
- Debugging support: helped identify and fix issues in the code, such as correctly matching names between images and data entries.

Challenges & Lessons Learned

- Challenge: adapting generic code suggestions to the specific needs of the SPRM data scraping task.
- Lesson: always review and customize generated code to ensure it fits the specific context and requirements of the project.

SinarProject
hacking democracy

KEY INSIGHTS

SinarProject
hacking democracy

> 300 offenders

since 2021

SinarProject
hacking democracy

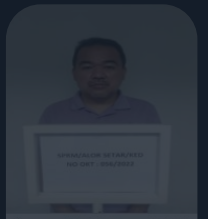
GENDER

87%

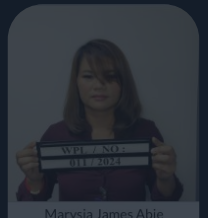
12%

Abd Rasid bin Mohamad
70040712XXXX
Malaysia

Roney Saimey bin Sakah
80081112XXXX
Malaysia



Ameyrudin bin Ahmad Zuki
81081914XXXX
Malaysia



Marysia James Abie
84111312XXXX
Malaysia

Mohd Syamsul bin Jikan
77100712XXXX
Malaysia

Abas bin Tuah
62082212XXXX
Malaysia

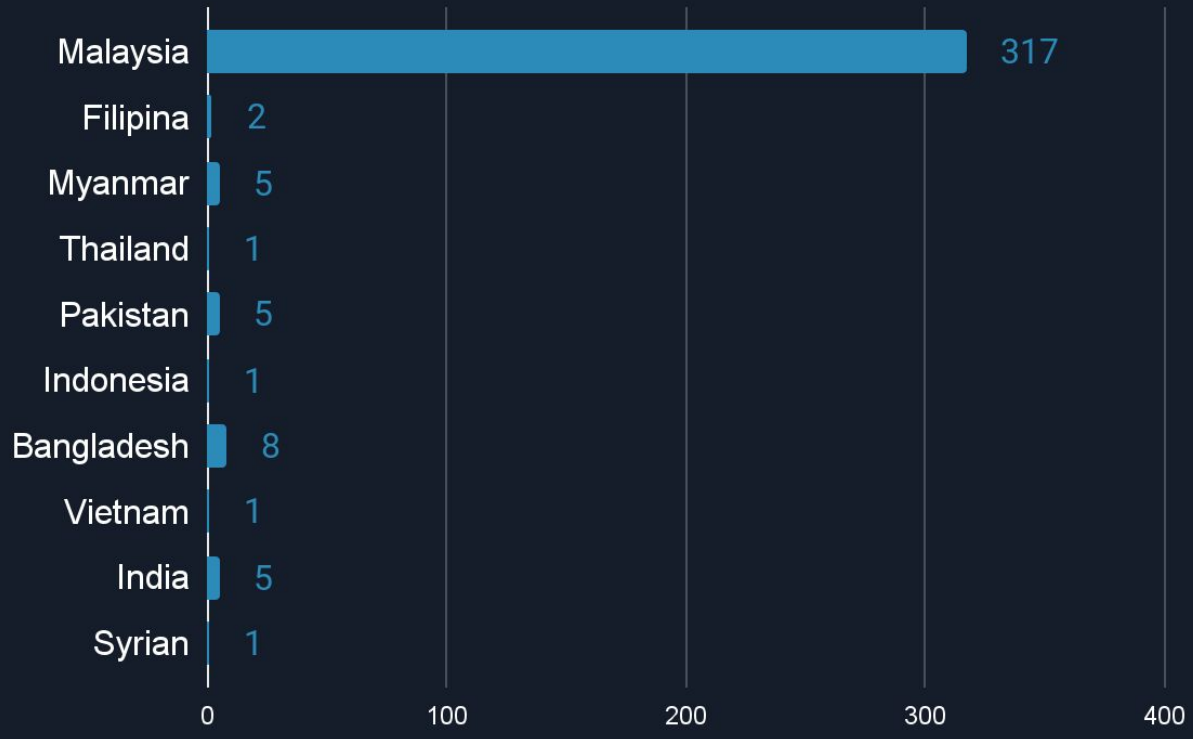
Shaharuddin bin Ahmad
79082508XXXX
Malaysia

Kamarulzaman bin Abdullah
59060708XXXX
Malaysia

Nor Azmi [illegible] Ali
Malaysia

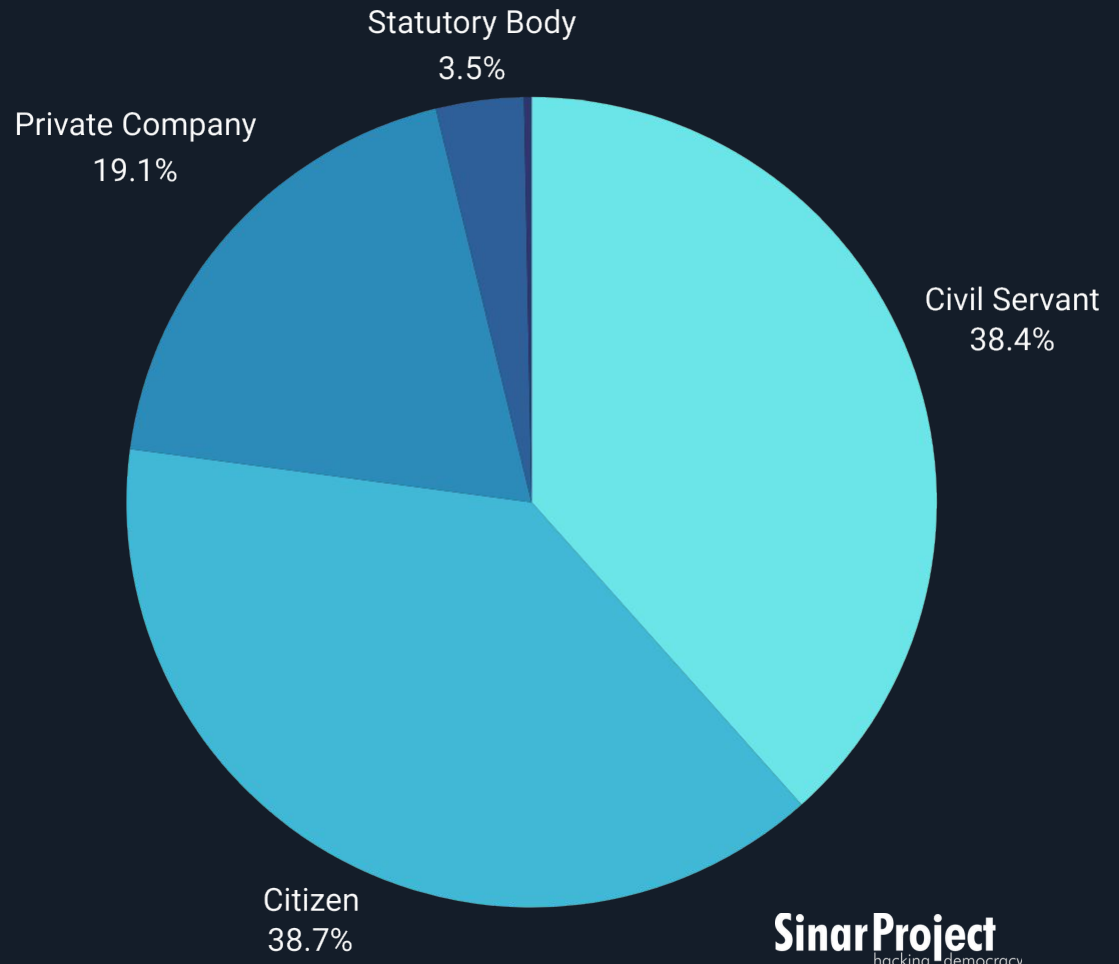
Sinar Project
hacking democracy

NATIONALITY



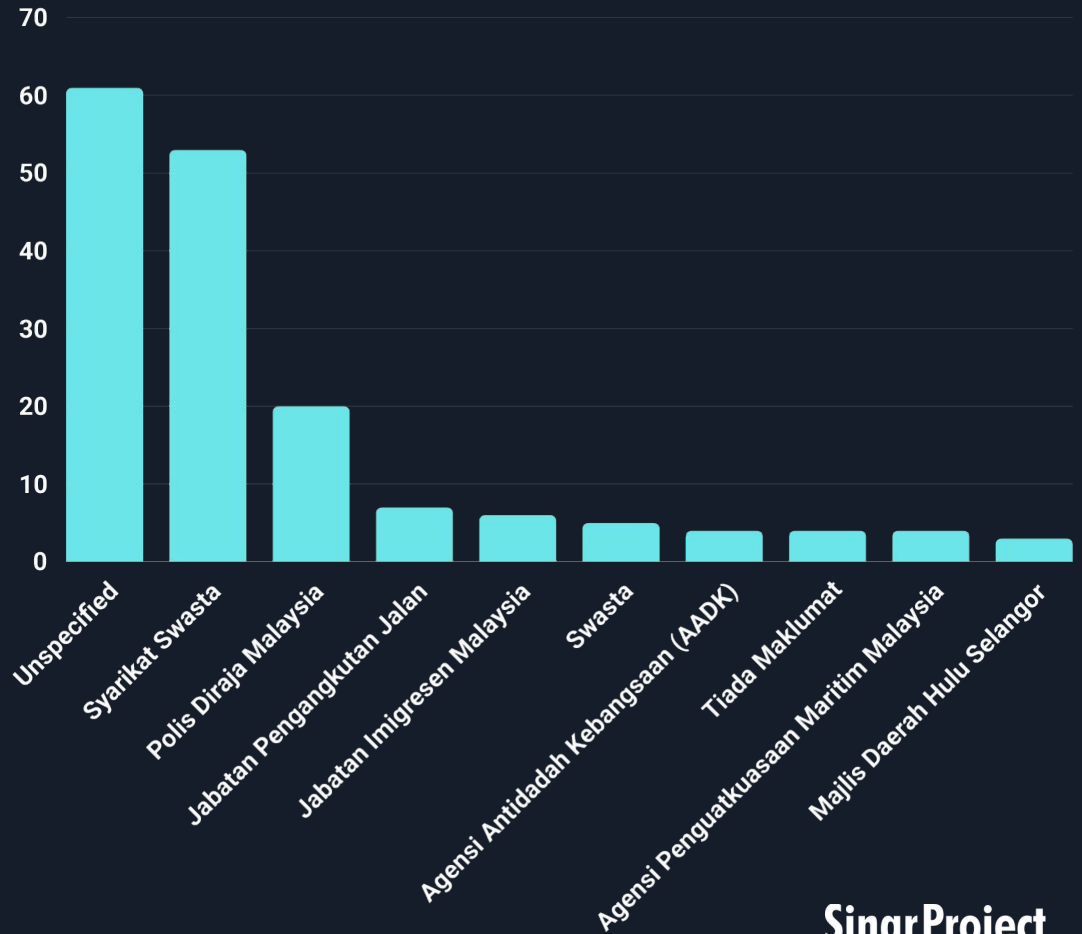
Sinar Project
hacking democracy

TYPES OF OFFENDERS



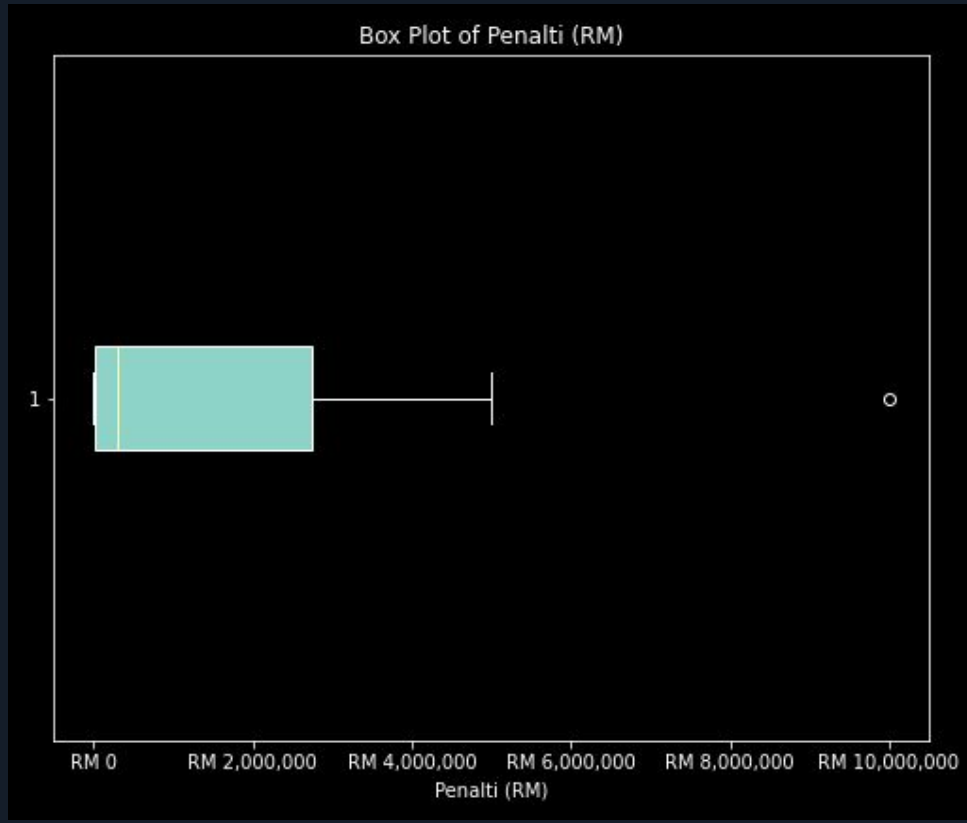
SinarProject
hacking democracy

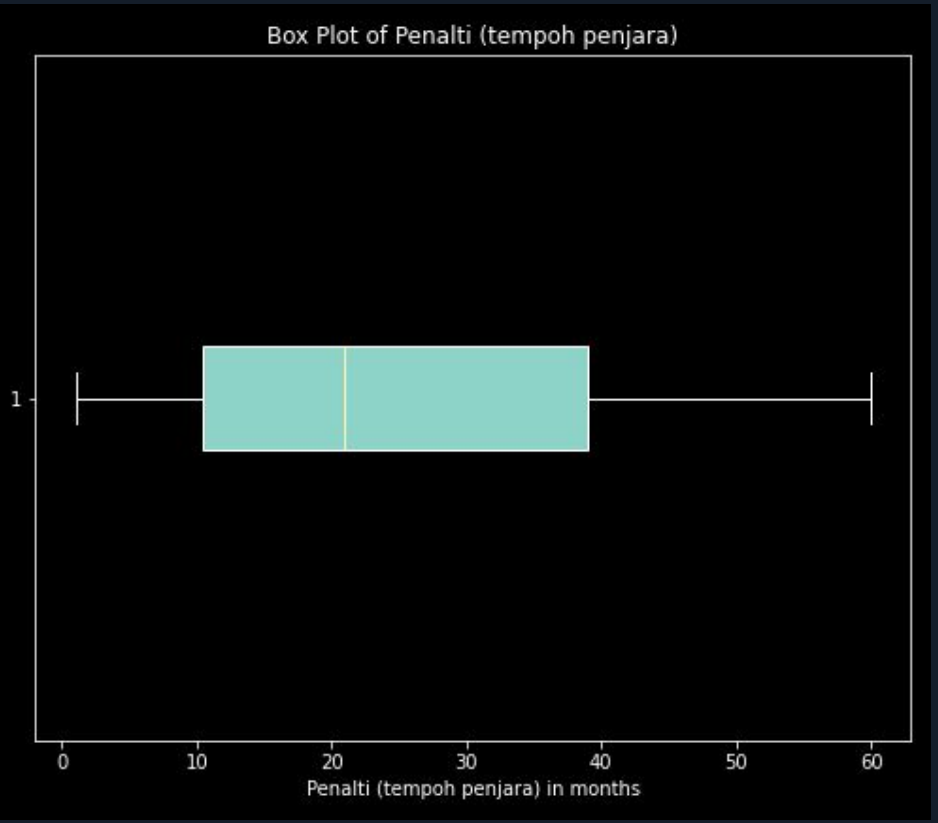
POPULAR EMPLOYERS



SinarProject
hacking democracy

PENALTY





SinarProject
hacking democracy

PENALTY

Metric		Penalti (RM)	Penalti (tempoh penjara)
0	Median	RM 300,000	21 months
1	Mean	RM 2,203,250	26 months
2	Max	RM 10,000,000	60 months
3	Min	RM 1,000	1 months

SinarProject
hacking democracy

CHARGES

Kanun Keseksaan
Penal Code

Akta Suruhanjaya Pencegahan Rasuah 2009
Malaysian Anti-Corruption Commission Act 2009

```
filtered_df['Kesalahan'].value_counts()
```

```
Kesalahan
Seksyen 471 Kanun Keseksaan
114
Seksyen 165 Kanun Keseksaan
69
Seksyen 214 Kanun Keseksaan
32
Seksyen 17(b) Akta SPRM 2009
29
Seksyen 415 Kanun Keseksaan
26
Seksyen 417 Kanun Keseksaan
26
Seksyen 17(a) Akta SPRM 2009
19
Seksyen 23(1) Akta SPRM 2009
16
Seksyen 25(1) Akta SPRM 2009
11
Seksyen 16(a)(A) Akta SPRM 2009
9
Seksyen 23(3) Akta SPRM 2009
8
Seksyen 16(a)(B) ASPRM 2009
7
Seksyen 18 Akta SPRM 2009
6
Seksyen 17(b) ASPRM 2009
5
Seksyen 16(a)(B) Akta SPRM 2009
5
```

CORRUPTION VS PENALTY

About 30 cases where the penalty was less than corruption amount?

Bahawa kamu pada 12 Mei 2019, di Pejabat Perikanan Negeri Kelantan, dalam Jajahan Kota Bharu sebagai pegawai badan awam iaitu Ketua Cawangan Akuakultur di Pejabat Perikanan Negeri Kelantan dan terikat di sisi undang-undang sebagai penjawat awam supaya tidak mengambil bahagian dalam perniagaan sebagaimana diperuntukkan di bawah peraturan 5(1)(a) Peraturan Pegawai Awam Kelakuan dan Tatatertib 1993, telah mengambil bahagian dalam perniagaan, iaitu melibatkan diri dalam perniagaan YY Indah Enterprise milik adik ipar kamu dengan meluluskan pesanan kerajaan No.C0190000000659705 bagi kerja-kerja membekal dan menghantar tangki akuakultur kepada Pejabat Perikanan Negeri Kelantan dengan nilai pembekalan sebanyak RM 71,694.00 dan dengan itu kamu telah melakukan kesalahan yang boleh dihukum di bawah Seksyen 168 Kanun Keseksaan.	Seksyen 168 Kanun Keseksaan	Denda RM2,000.00; id 6 bulan penjara – bagi setiap pertuduhan

SinarProject
hacking democracy

USE CASES OF THE DATA

- Convert the Penal Code into machine readable format and analyze exact charges of the offenders
- Contribute to databases e.g. Open Sanction database, which has persons and targets of interests
- Map to Beneficial Ownership schemes
- Match to datasets with individuals of interests e.g. CIDB (construction contractors) and ICIJ (offshore leaks)
- Convert images into relevant public domain formats to be used for facial recognition

CONCLUSION

It is a good exercise for Python beginners to learn about

- web crawling
- data structures
- data transformation

while contributing something meaningful for society & research 🌟

SinarProject
hacking democracy

THANK YOU

for your time and attention

Sinar Project
hacking democracy