9th MALAYSIA STATISTICS CONFERENCE 2022
Dealing with Uncertainties: Unearthing Measures for Recovery
Institut Latihan Statistik Malaysia
5th October 2022

Application of seamless hybrid geocoding solution for business location using KAWASANKU API

***Ahmad Najmi ARIFFIN [1,2],** Mohamad Hamizan ABDULLAH [1]

[1] Core Team Big Data Analytics, Department of Statistics, Malaysia
[2] Faculty of Dentistry, Universiti Kebangsaan Malaysia (UKM);

Abstract:

Addresses can be geocoded by transposing them into the corresponding longitudes and latitudes. The coordinates provide a means of pinpointing a map's location. This article provides a comprehensive methodology for online geocoding services that are widely utilised. Our study aimed to validate the distributions of geocode locations to take into account when analysing geocoded address data, and to create strategies for enriching demographic databases by utilising the centralised public data sources repository KAWASANKU API on the Github platform. Open Data is data that is accessible, usable, and shareable by the public. The private sector has been very hesitant to adopt Open Data. If businesses utilise Open Data strategically, it can be a key factor in generating a variety of uncertain business opportunities, such as enhancing new products and services. The "geopy" Python module enables the mapping of global coordinates for addresses, cities, countries, and landmarks. Risk-assessment using Fuzzywuzzy (Python library) returns the similarity percentage [1-100] between two sequences of addresses strings to match. The q-ratio score threshold is over 65. These addresses will be re-geocoded and evaluated for completeness. KAWASANKU API can query socio-demographic features and Malaysia's geospatial boundaries down to the DUN level, including national, state, district, parliament, and state legislative assembly (Malay: Dewan Undangan Negeri, DUN). We get 2,427 geojson raw lines for each property feature. This framework for a seamless, less-dependent workflow to reduce risk and benefits on data enrichment. Using the provided script, this framework allows SMD to self-perform processes. Researchers must be aware of certain peculiarities to effectively use the data, which is a research opportunity.

Keywords:

Open data; Geospatial; Python; Github API

1. Introduction and background

Geocoding is the process of converting addresses, such as street addresses, into geographic coordinates, such as longitudes and latitudes. These coordinates can be used to locate a map or place markers on it. This clarifies the concerns involved with geocoding addresses.

In summary, the contributions of this article are a comprehensive methodology for widely used online geocoding services. The purpose of our study was to verify the distributions of geocode location to consider when analysing geocoded address data, as well as to develop methods for enriching demographic databases and representing multiple levels - district, parliament, and state legislative assembly (Malay: Dewan Undangan Negeri, DUN) - using centric public data sources repository KAWASANKU API from Github platform.

The rest of the article is organised as follows. The following section provides a brief overview of geocoding applications. The first section provides an overview of the advantages of geocoding in business and the availability of open data for business research as the baseline used in this study. Section 2 describes the method for performing analysis, and the evaluation results are presented in Section 3. Section 3 discusses results analysis, and Section 4 concludes with discussion and conclusions.

1.1 Benefits of Geocoding and Structuring Address Data

Address information is one of the most frequently collected types of data by businesses worldwide. Multiple businesses may share the same name, creating confusion regarding which addresses correspond to which locations. These are merely some of the data quality errors that may be present in address data[1]. This alternative is the Smart Search ArcGIS API, which provides an address data-cleansing structured geocoding call. The mechanism of this feature was simplified in Figure 1.

[1] Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.

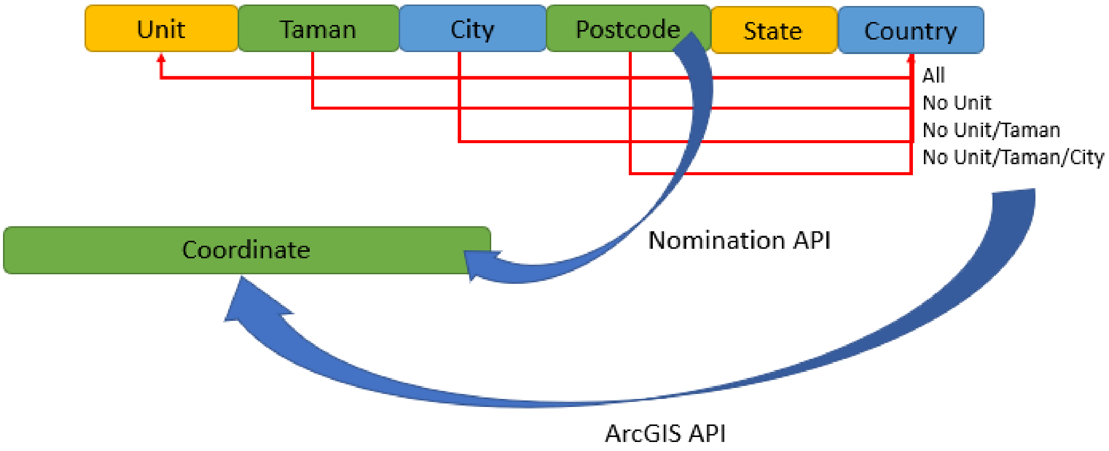


Figure 1 : Smart Search Mechanism using ArcGIS API

The practice of mapping address records to physical locations is critical for understanding and leveraging geographic linkages that are fundamental to all statistics.[2] These are available via the hybrid solution, the properties application, and the public web service. Modules are intended to support all reference information, access online to have complete control over reference information, including the application of geographic boundaries to resolution production. They accurately assign various geographic codes to each and every address. Using Smart Search or the ArcGIS Geocoding API in conjunction with the Geocoding API enables the development of applications that provide users with precise geocoding results and reduced latency.[3]

1.2 The advantages of geocoding for businesses

Connecting businesses with their customers will be the focus of the discussion. The use of geographical coordinates is one method. Consumers require a method for determining their geographical coordinates. Historically, locating an actual location was a difficult task.[4] An in-depth examination of the perceptions and applications of location intelligence across industries. Enhancing the user experience decreases friction and enhances the perception of brand awareness. According to the report Location Intelligence Drives Competitive Edge In The Digital Age by Forrester

[2] MacEachren, A. M. (2017). Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial data handling in big data era* (pp. 139-155). Springer, Singapore.
[3] Kirby, R. S., Delmelle, E., & Eberth, J. M. (2017). Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology*, *27*(1), 1-9.
[4] Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768).

Consulting, address verification will be essential in the future [5]. It generates a variety of uncertain business opportunities, such as enhancing new products and services, increasing the organisation's productivity, and enabling entirely new business lines.

1.3 Availability of open data for business research

It is often said that "data is the new oil" because it is one of the most valuable resources available to businesses in the digital age. Open Data is data that is accessible, usable, and shareable by the public. According to the European Data Portal, the Open Data market size within the EU in 2016 was 55,3 billion Euros.[6] The private sector has been very hesitant to adopt Open Data. Historically, businesses have been more concerned with protecting the commercial value of their data, but they are missing out on numerous opportunities. Their argument is that since Open Data is freely accessible to the public, anything that is free has no value[7]. If businesses utilise Open Data strategically, it can be a key factor in the creation of a new product or service.

Open Data has substantial economic value, which includes opportunities for stimulating the development of new products and services, enhancing organisational efficiency, and generating consumer benefits – cost savings, convenience, and improved quality[8]. It aids in the development of the organisation's data impact initiatives by establishing a more transparent and adaptable platform, thereby facilitating creativity and experimentation. It provides a new channel for consumers to provide feedback for the purpose of enhancing and enhancing the quality of services and products. The benefits of Open Data can be increased if both private industry and public agencies advocate for the Open Data sharing platform and mindset, thereby fostering a thriving open data ecosystem[9].

[5] Location Intelligence Drives Competitive Edge In The Digital Age, July 2018, A Forrester Consulting Thought Leadership Paper Commissioned By Loqate, A GBG solution = https://info.loqate.com/hubfs/Loqate%202018/Reports/Location%20Intelligence%20Drives%20Competitive%20Edge%20In%20The%20Digital%20Age.pdf

[6] A study on the Impact of Re-use of Public Data Resources, November 2015, Wendy Carrara, Wae San Chan, Sander Fischer, Eva van Steenbergen (Capgemini Consulting), https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf

[7] Khayyat, M., & Bannister, F. (2015). Open data licensing: more than meets the eye. *Information Polity*, *20*(4), 231-252.

[8] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258-268.

[9] Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, *33*(1), 15-27.

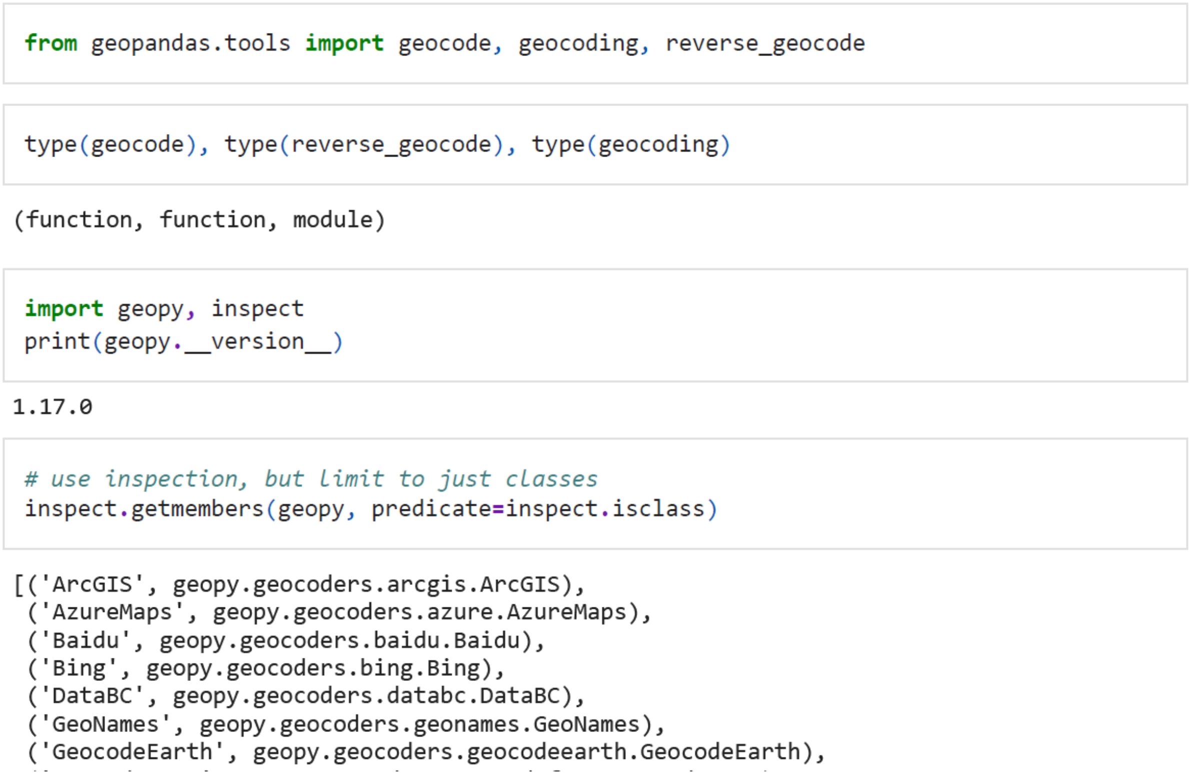
2. Methodology

Sometimes a module of a GIS, which automatically performs the geocode process. It is usually available on the Internet by utilising a Web service interface. The data entry, such as a place name, street address, or zip code, is passed over the Internet using a communication protocol to the geocoding service.

2.1 Geocoding Address and Reverse Geocoding Coordinate

Alternatively, for this study an online geocoding service like the module in Python using "geopy" is a network-accessible component which enables us to locate and identify the global coordinates of addresses, cities, countries, and landmarks. The "geopy" module utilises geo-coders and other data sources from third parties [10]. "Nominatim" is an OpenStreetMap data geocoder [11]. Installation of the "geopy" module was shown in Figure 2.

Figure 2 : Installation of the "geopy" module in Python IDE



```
from geopandas.tools import geocode, geocoding, reverse_geocode
```

```
type(geocode), type(reverse_geocode), type(geocoding)
```

```
(function, function, module)
```

```
import geopy, inspect
print(geopy.__version__)
```

```
1.17.0
```

```
# use inspection, but limit to just classes
inspect.getmembers(geopy, predicate=inspect.isclass)
```

```
[('ArcGIS', geopy.geocoders.arcgis.ArcGIS),
 ('AzureMaps', geopy.geocoders.azure.AzureMaps),
 ('Baidu', geopy.geocoders.baidu.Baidu),
 ('Bing', geopy.geocoders.bing.Bing),
 ('DataBC', geopy.geocoders.databc.DataBC),
 ('GeoNames', geopy.geocoders.geonames.GeoNames),
 ('GeocodeEarth', geopy.geocoders.geocodeearth.GeocodeEarth),
```

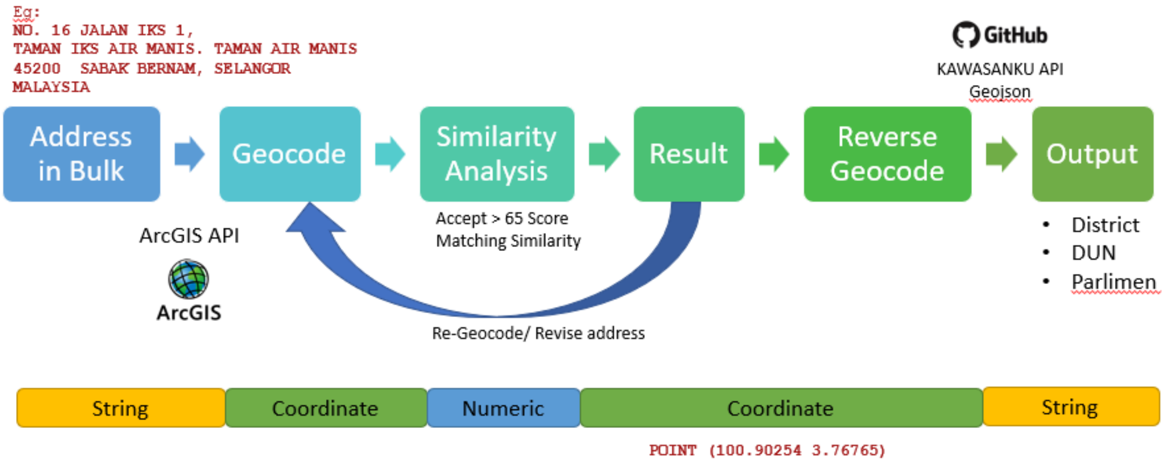
[10] GeoPy, Welcome to GeoPy documentation!, Retrieved on Sep. 2022, From https://geopy.readthedocs.io/en/stable/.

[11] Nominatim Documentation, Nominatim API, Retrieved on Sep. 2022, From https://nominatim.org/release-docs/develop/api/Overview/.

OSM’s Nominatim Service, Nominatim Usage Policy, Retrieved on Sep. 2022, From https://operations.osmfoundation.org/policies/nominatim/.

Typically, each data entry requires fewer than a few seconds to complete. The geocoding service converts an input into coordinates and returns the result, which includes the coordinates, the geocoded address, and the level of accuracy, to the user via the Internet. This article uses the term 'online geocoding,' but the GIS community has also used terms such as real-time geocoding, address lookup, and address matching service interchangeably. In our definition, we exclude services that require human intervention during the geocoding process or that do not provide users with immediate geocoded results, such as the four geocoding services offered by commercial geocoding companies in Krieger (2001)[12].

Figure 3: Workflow ot Geocoding Address and Reverse Geocode

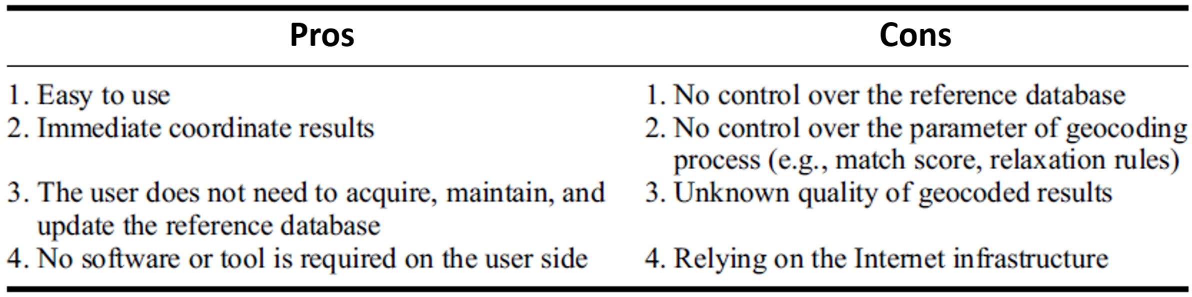


As depicted in Figure 3, the user of online geocoding and reverse geocode is not required to understand how geocoding works or how to acquire and maintain the target area's reference database.[13] The only thing the user must do is enter the required addresses and interpret the results. Second, any online geocoding service's reference database is stored, maintained, and updated by the service provider. Unlike conventional geocoding, which requires the user to provide reference databases, online geocoding and reverse geocode does not require the user to worry about reference databases. Using online geocoding differs in several ways from using conventional geocoding tools that come with GIS software packages, such as ArcView and Automatch. First, online geocoding services are user-friendly. Service providers predefine all geocoding process parameters and methods; consequently, they do not permit users to customise match scores and relaxation rules. In Web applications and location-based applications, latency in the geocoded result could delay subsequent analysis or processes. The advantages and disadvantages of utilising online geocoding services are summarised in Table 1.

[12] Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American journal of public health*, *91*(7), 1114.

[13] Roongpiboonsopit, D., & Karimi, H. A. (2010). Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science*, *24*(7), 1081-1100.

Table 1 : A summary of the pros and cons of utilising online geocoding services.

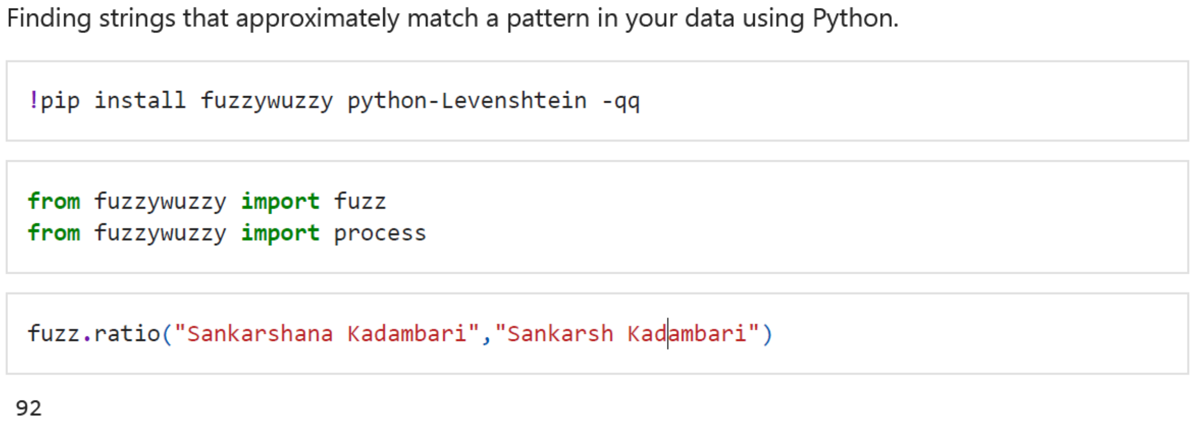


Pros	Cons
1. Easy to use	1. No control over the reference database
2. Immediate coordinate results	2. No control over the parameter of geocoding process (e.g., match score, relaxation rules)
3. The user does not need to acquire, maintain, and update the reference database	3. Unknown quality of geocoded results
4. No software or tool is required on the user side	4. Relying on the Internet infrastructure

2.2 Address Fuzzy string matching technique

Second, this study employs the standard metric of similarity string match ratio to determine the commonality of the reverse geocoded results. A measurement of the edit distance required to reconstruct a string from the original string is the simplest method for comparing two strings. Fuzzy string matching compares two strings containing spelling errors or incomplete words to find matches.

Figure 3 : Installation of the "fuzzywuzzy" module to Python Library



Finding strings that approximately match a pattern in your data using Python.

```
!pip install fuzzywuzzy python-Levenshtein -qq
```

```
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
```

```
fuzz.ratio("Sankarshana Kadambari","Sankarsh Kadambari")
```

```
92
```

It is called fuzzy because it uses an 'approximate' string matching technique based on Levenshtein Distance and the formula given in Equation. 2 to calculate the edit distance. For string matching, we use the Fuzzywuzzy[14] Python library, which is optimised for speed.

[14] TheFuzz documentation - Github repository, Retrieved on Sep. 2022, From https://github.com/seatgeek/fuzzywuzzy.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases} \tag{2}$$

where 1(*ai*=b*j*) denotes 0 when a = b and 1 otherwise. Finally, the Levenshtein similarity ratio is computed based on the Levenshtein distance, and is calculated using the formula in Equation. 3.

$$\frac{(|a|+|b|) - \mathrm{lev}_{a,b}(i,j)}{|a|+|b|} \tag{3}$$

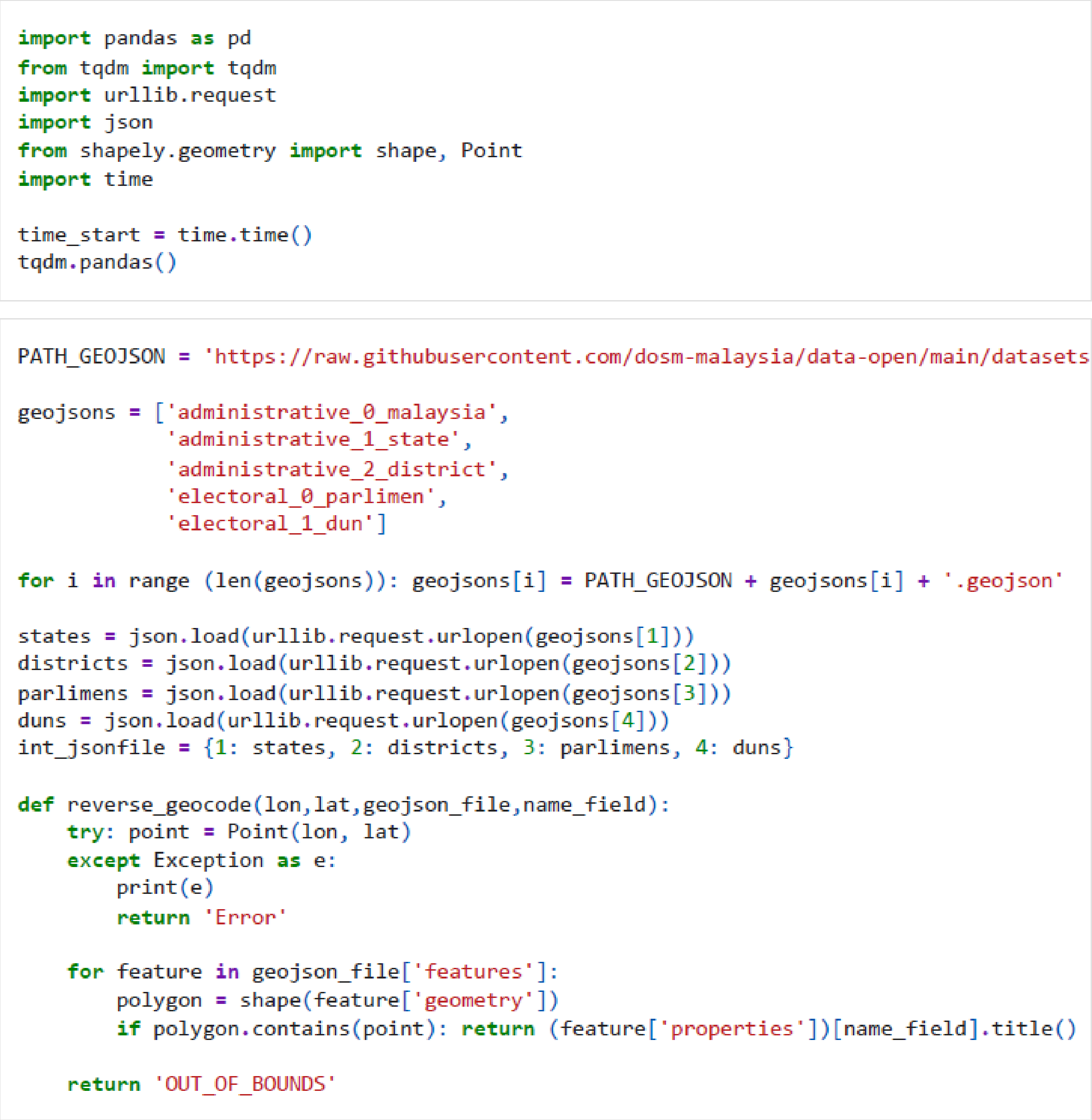
where |a| and |b| are the lengths of sequence a and sequence b, respectively[15].

2.3 Mapping geocoded data using KAWASANKU GitHub API

API stands for Application Programming Interface, a software interface that enables two applications to communicate with one another. GitHub offers the GitHub API to developers who wish to create GitHub-specific applications. We could retrieve a public repository and then conduct a search of the resulting documents. For instance, users could utilise the repositories endpoint, which fetches all public repositories, and then conduct our own search. For this study, 4 main components data were fetched from KAWASANKU GitHub API which were state, district, DUN, parliament data.

[15] Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American journal of public health*, *91*(7), 1114.

Figure 4 : Integration setup for KAWASANKU Github API to Python Library



```python
import pandas as pd
from tqdm import tqdm
import urllib.request
import json
from shapely.geometry import shape, Point
import time

time_start = time.time()
tqdm.pandas()
```

```python
PATH_GEOJSON = 'https://raw.githubusercontent.com/dosm-malaysia/data-open/main/datasets

geojsons = ['administrative_0_malaysia',
            'administrative_1_state',
            'administrative_2_district',
            'electoral_0_parlimen',
            'electoral_1_dun']

for i in range (len(geojsons)): geojsons[i] = PATH_GEOJSON + geojsons[i] + '.geojson'

states = json.load(urllib.request.urlopen(geojsons[1]))
districts = json.load(urllib.request.urlopen(geojsons[2]))
parlimens = json.load(urllib.request.urlopen(geojsons[3]))
duns = json.load(urllib.request.urlopen(geojsons[4]))
int_jsonfile = {1: states, 2: districts, 3: parlimens, 4: duns}

def reverse_geocode(lon,lat,geojson_file,name_field):
    try: point = Point(lon, lat)
    except Exception as e:
        print(e)
        return 'Error'

    for feature in geojson_file['features']:
        polygon = shape(feature['geometry'])
        if polygon.contains(point): return (feature['properties'])[name_field].title()

    return 'OUT_OF_BOUNDS'
```

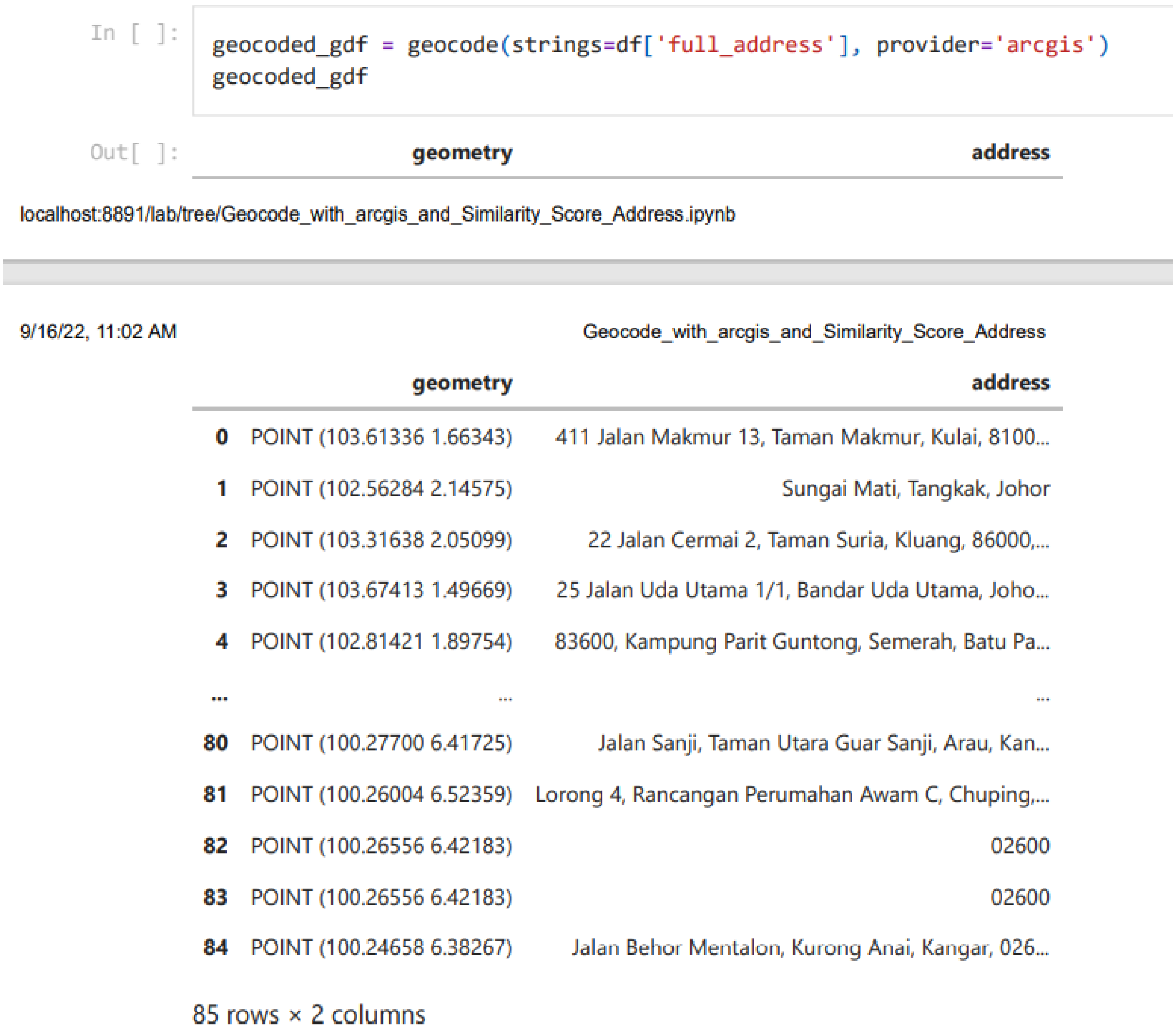
3. Result

This section will also discuss the results of a data science-based similarity analysis between the original and generated addresses, in addition to the geocoding results. The workflow is continued by mapping geocoded addresses using the KAWASANKU Github API to enrich geospatial data in order to extend level-level data (such as state, district, DUN, and parliament data).

3.1 Risk-assessment on valid geocode

The geocodeAddresses operation geocodes an entire list of addresses with a single request. The table can store addresses in a single field or multiple fields, one for each address component. The performance of batch geocoding is enhanced when the address components are stored in separate fields. These are advanced APIs that simplify the geocoding process in bulk. There is a maximum number of addresses that can be geocoded using the service in a single batch request. This parameter can be used to override the default city and street names returned in output fields for a geocoding transaction by specifying substitute city and street names. In this method, we used the arcGIS API as the valid provider to geocode the addresses so that they would coordinate with the data on latitude and longitude that was stored in a geospatial format.

Figure 5: Geocoded address(coordinate) and generated address from arcGIS API



```
In [ ]:  geocoded_gdf = geocode(strings=df['full_address'], provider='arcgis')
         geocoded_gdf
```

Out[ ]: geometry address

localhost:8891/lab/tree/Geocode_with_arcgis_and_Similarity_Score_Address.ipynb

9/16/22, 11:02 AM Geocode_with_arcgis_and_Similarity_Score_Address

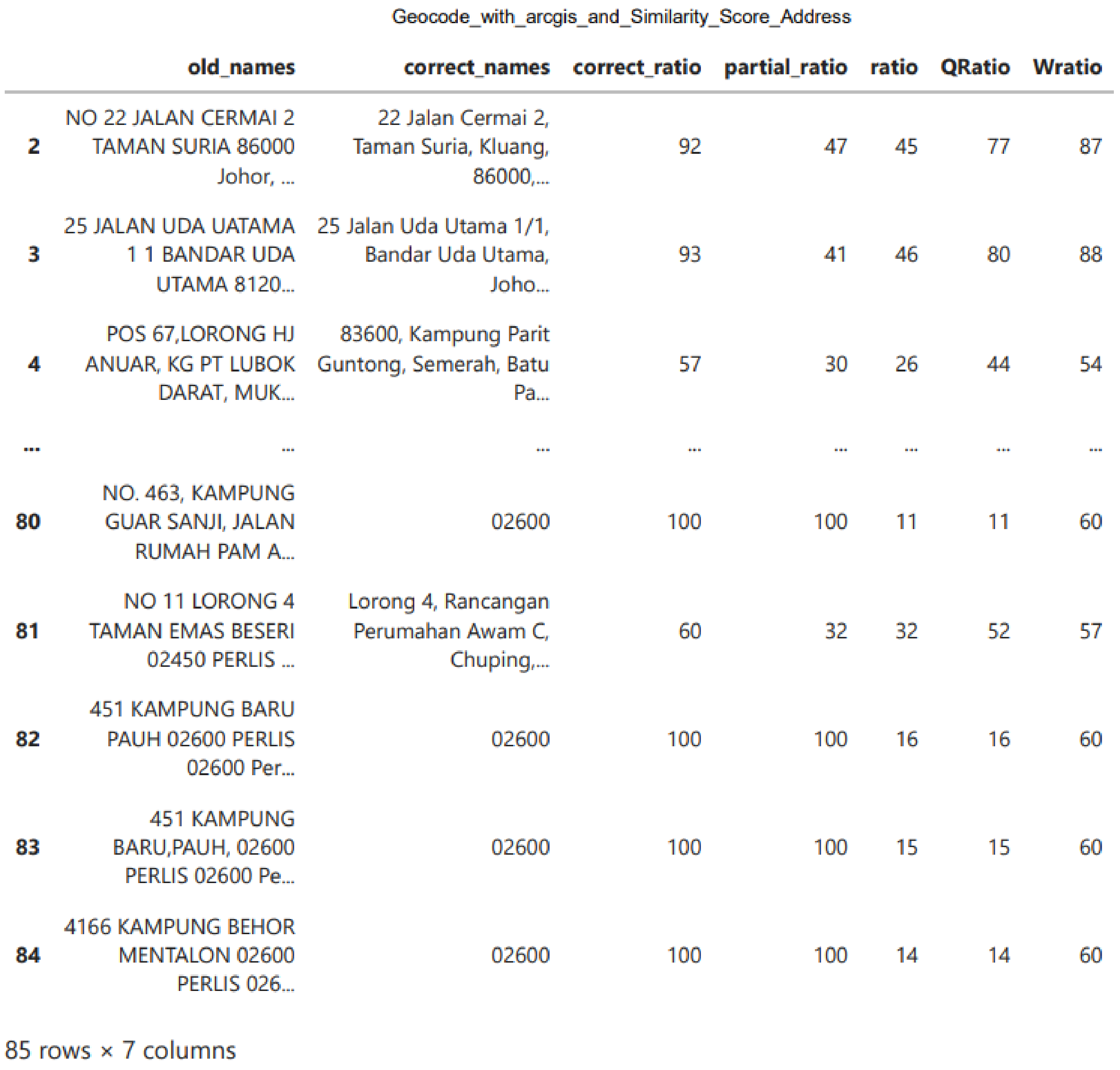
	geometry	address
0	POINT (103.61336 1.66343)	411 Jalan Makmur 13, Taman Makmur, Kulai, 8100...
1	POINT (102.56284 2.14575)	Sungai Mati, Tangkak, Johor
2	POINT (103.31638 2.05099)	22 Jalan Cermai 2, Taman Suria, Kluang, 86000,...
3	POINT (103.67413 1.49669)	25 Jalan Uda Utama 1/1, Bandar Uda Utama, Joho...
4	POINT (102.81421 1.89754)	83600, Kampung Parit Guntong, Semerah, Batu Pa...
...	...	...
80	POINT (100.27700 6.41725)	Jalan Sanji, Taman Utara Guar Sanji, Arau, Kan...
81	POINT (100.26004 6.52359)	Lorong 4, Rancangan Perumahan Awam C, Chuping,...
82	POINT (100.26556 6.42183)	02600
83	POINT (100.26556 6.42183)	02600
84	POINT (100.24658 6.38267)	Jalan Behor Mentalon, Kurong Anai, Kangar, 026...

85 rows × 2 columns

3.2 Risk-assessment on similarity analysis

The concatenation of original address and generated address text performs the best among the textual approaches for correct ratio, partial ratio, Q Ratio, and W Ratio. For Q Ratio, purely textual features produce better MSE than numerical scores. Fuzzywuzzy uses a similarity ratio between two sequences, rather than attempting to format the strings to match, and returns the similarity percentage [1-100]. The partial ratio() function allows substring matching to be performed. This is accomplished by matching the shortest string with all substrings of the same length. Included for completeness, the Qratio() function is merely a wrapper around fuzz.ratio with validation and short-circuiting. The Wratio() function attempts to weight (the name stands for 'Weighted Ratio') the results of various algorithms in order to determine the 'best' score. Using QRatio as a similarity indicator allows for a more accurate representation of similarity according to the dataset. The findings indicate that the acceptance threshold for valid geocoded addresses is greater than 65. The addresses that fall below this threshold will be re-geocoded and their completeness will be evaluated.

Figure 6 : Fuzzy matching analysis, Q-Ratio Score as Similarity Indicator



Geocode_with_arcgis_and_Similarity_Score_Address

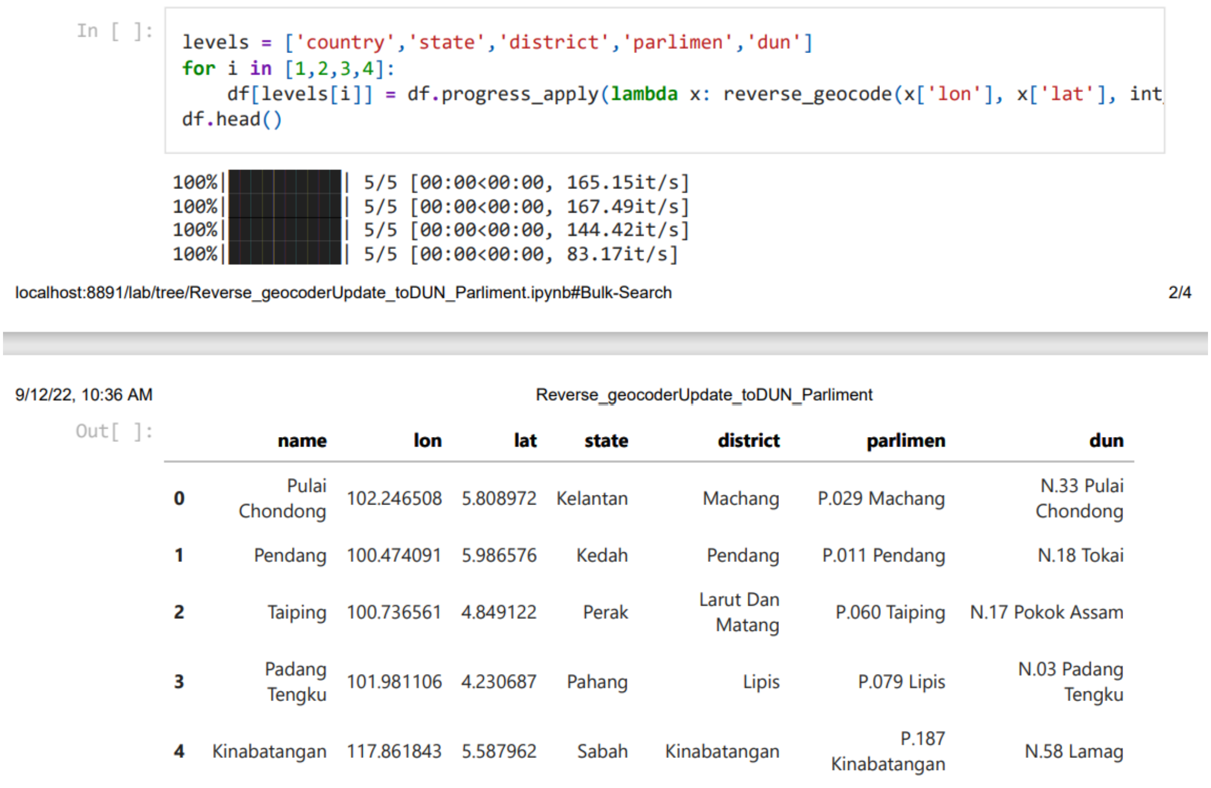
	old_names	correct_names	correct_ratio	partial_ratio	ratio	QRatio	Wratio
2	NO 22 JALAN CERMAI 2 TAMAN SURIA 86000 Johor, ...	22 Jalan Cermai 2, Taman Suria, Kluang, 86000,...	92	47	45	77	87
3	25 JALAN UDA UATAMA 1 1 BANDAR UDA UTAMA 8120...	25 Jalan Uda Utama 1/1, Bandar Uda Utama, Joho...	93	41	46	80	88
4	POS 67,LORONG HJ ANUAR, KG PT LUBOK DARAT, MUK...	83600, Kampung Parit Guntong, Semerah, Batu Pa...	57	30	26	44	54
...	...	...	...	...	...	...	...
80	NO. 463, KAMPUNG GUAR SANJI, JALAN RUMAH PAM A...	02600	100	100	11	11	60
81	NO 11 LORONG 4 TAMAN EMAS BESERI 02450 PERLIS ...	Lorong 4, Rancangan Perumahan Awam C, Chuping,...	60	32	32	52	57
82	451 KAMPUNG BARU PAUH 02600 PERLIS 02600 Per...	02600	100	100	16	16	60
83	451 KAMPUNG BARU,PAUH, 02600 PERLIS 02600 Pe...	02600	100	100	15	15	60
84	4166 KAMPUNG BEHOR MENTALON 02600 PERLIS 026...	02600	100	100	14	14	60

85 rows × 7 columns

3.3 Optimisation - Data Enrichment using open data sharing source, extract data using KAWASANKU API

We were able to retrieve a public repository data set from the Department of Statistics Malaysia's Github page (DOSM). KAWASANKU API Github offers its users an application programming interface (API) that can be used to query socio-demographic features and Malaysia's geospatial boundaries all the way down to the DUN level, which includes entities such as national, state, and parliament. We obtain approximately 2,427 geojson raw lines corresponding to the geospatial information for every property's features. Within the geometry boundary, we define coordinate points as those for which a matching repository is identified (which we call the mapping repository).

Figure 7 : KAWASANKU API Matching - indicate geocode point within district, parliament and DUN



In [ ]:

```
levels = ['country','state','district','parlimen','dun']
for i in [1,2,3,4]:
    df[levels[i]] = df.progress_apply(lambda x: reverse_geocode(x['lon'], x['lat'], int
df.head()
```

```
100%|██████████| 5/5 [00:00<00:00, 165.15it/s]
100%|██████████| 5/5 [00:00<00:00, 167.49it/s]
100%|██████████| 5/5 [00:00<00:00, 144.42it/s]
100%|██████████| 5/5 [00:00<00:00, 83.17it/s]
```

localhost:8891/lab/tree/Reverse_geocoderUpdate_toDUN_Parliment.ipynb#Bulk-Search 2/4

9/12/22, 10:36 AM Reverse_geocoderUpdate_toDUN_Parliment

Out[ ]:

	name	lon	lat	state	district	parlimen	dun
0	Pulai Chondong	102.246508	5.808972	Kelantan	Machang	P.029 Machang	N.33 Pulai Chondong
1	Pendang	100.474091	5.986576	Kedah	Pendang	P.011 Pendang	N.18 Tokai
2	Taiping	100.736561	4.849122	Perak	Larut Dan Matang	P.060 Taiping	N.17 Pokok Assam
3	Padang Tengku	101.981106	4.230687	Pahang	Lipis	P.079 Lipis	N.03 Padang Tengku
4	Kinabatangan	117.861843	5.587962	Sabah	Kinabatangan	P.187 Kinabatangan	N.58 Lamag

4. Discussion and Conclusion

This framework to suggest a seamless and less-dependent workflow to reduce risk also gives advantages to the enrichment of data. Prior to validation by the methodology and research division, the Subject Matter Division (SMD) at DOSM typically requests the GIS team to conduct a geospatial analysis for address geocoding. It could take a week before the data are returned, and it will take that long to continue the analysis on the SMD side. This framework enables SMD to self-perform processes by utilising the provided ready-made script.

Open Data sharing platform fostering a thriving open data ecosystem.

It provides a new channel for consumers to provide feedback for the purpose of enhancing and enhancing the quality of services and products. The benefits of Open Data can be increased if both private industry and public agencies advocate for the Open Data sharing platform and mindset, thereby fostering a thriving open data ecosystem.

The work is motivated by the increasing popularity of GitHub as a collaborative platform for open source projects and ideas. In recent years, more academic and industrial researchers have shared the source code of their research on GitHub. In published papers, links to open source repositories are frequently included for research on machine learning and data mining, which is a particularly clear illustration of this trend in computer science.There are also a number of GitHub peculiarities that researchers must be aware of in order to utilise the data effectively, which represents an opportunity for research. The challenges and opportunities for the licences, community, development process, and product of the free/libre and open-source software communities hosted on GitHub are summarised.

In addition to the federal and state levels of government, micro level governments (district, DUN, and parliament) are also able to use a geospatial approach to plan for better strategies to enhance new uncertainty business entities. Planning a more advantageous location for an entrepreneur's business based on the distribution network using a map. Measuring the geographical distribution of economic activity is essential for scientific research and policy formation.

Limitation

Unstructured address data is a common obstacle. The Geocoding API request feature provides an easy-to-use solution for cleaning address data and building a database of geocoded locations, and makes it accessible via straightforward HTTP GET requests. The Geocoding API's address geocoding has significantly higher latency and produces less accurate results for incomplete or ambiguous queries; therefore, it is not recommended for real-time user input-responsive applications. If the automated system processes a high volume of ambiguous queries derived from user input, it may benefit from integrating the Places API into the app. According to the usage policy, heavy usage is not permitted, but 1 request per second is allowed.

In the future, we plan to employ more advanced document embedding techniques, such as a Semantic Text Similarity Computing System Based on SVM, in order to comprehend the qualitative similarities between two datasets more thoroughly.

5. References

1. Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.

2. MacEachren, A. M. (2017). Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial data handling in big data era* (pp. 139-155). Springer, Singapore.

3. Kirby, R. S., Delmelle, E., & Eberth, J. M. (2017). Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology*, *27*(1),1-9.

4. Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp.759-768).

5. Location Intelligence Drives Competitive Edge In The Digital Age, July 2018, A Forrester Consulting Thought Leadership Paper Commissioned By Loqate, A GBG solution, https://info.loqate.com/hubfs/Loqate%202018/Reports/Location%20Intelligence%20Drives%20Competitive%20Edge%20In%20The%20Digital%20Age.pdf

6. A study on the Impact of Re-use of Public Data Resources, November 2015, Wendy Carrara, Wae San Chan, Sander Fischer, Eva van Steenbergen (Capgemini Consulting), https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf

7. Khayyat, M., & Bannister, F. (2015). Open data licensing: more than meets the eye. *Information Polity*, *20*(4), 231-252.

8. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258-268.

9. Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, *33*(1), 15-27.

10. GeoPy, Welcome to GeoPy documentation!, Retrieved on Sep. 2022, From https://geopy.readthedocs.io/en/stable/.

11. Nominatim Documentation, Nominatim API, Retrieved on Sep. 2022, From https://nominatim.org/release-docs/develop/api/Overview/.

12. OSM's Nominatim Service, Nominatim Usage Policy, Retrieved on Sep. 2022, From https://operations.osmfoundation.org/policies/nominatim/.

13. Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American journal of public health*, *91*(7), 1114.

14. Roongpiboonsopit, D., & Karimi, H. A. (2010). Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science*, *24*(7),1081-1100.

15. TheFuzz documentation - Github repository, Retrieved on Sep. 2022, From https://github.com/seatgeek/fuzzywuzzy.