APPLICATION OF SEAMLESS HYBRID GEOCODING SOLUTION FOR BUSINESS LOCATION USING KAWASANKU API

Ahmad Najmi Ariffin[1] and Mohamad Hamizan Abdullah[2]

ABSTRAK

Alamat boleh digeokodkan dengan mengubahnya kepada longitud dan latitud yang sepadan. Koordinat ini menyediakan cara bagi menentukan lokasi pada peta. Artikel ini menyediakan metodologi yang komprehensif untuk perkhidmatan geokod dalam talian yang digunakan secara meluas. Kajian ini bertujuan untuk mengesahkan taburan lokasi geokod yang diambil kira semasa menganalisis data alamat, dan menyediakan strategi memperkaya pangkalan data demografi dengan menggunakan repositori sumber data awam berpusat KAWASANKU API pada platform Github. Data Terbuka merupakan data yang boleh diakses, digunakan dan dikongsi oleh pengguna. Sektor swasta agak kurang yakin untuk menggunakan Data Terbuka. Jika perniagaan menggunakan Data Terbuka secara strategik, ia boleh menjadi faktor utama dalam menjana pelbagai peluang perniagaan, iaitu meningkatkan produk dan perkhidmatan baharu. Modul Python "geopy" membolehkan pemetaan koordinat global untuk alamat, bandar, negara dan mercu tanda. Penilaian risiko menggunakan Fuzzywuzzy (Python library) menghasilkan peratusan persamaan [1-100] antara dua jujukan rentetan alamat untuk tujuan padanan dengan ambang skor nisbah q yang ditetapkan iaitu melebihi 65. Alamat ini akan digeokodkan semula dan penilaian kesempurnaan. KAWASANKU API boleh menilai ciri sosio-demografi dan sempadan geospatial Malaysia hingga ke peringkat DUN, termasuk negara, negeri, daerah, parlimen dan dewan undangan negeri (Bahasa Melayu: Dewan Undangan Negeri, DUN). Kami mendapat 2,427 garisan mentah geojson untuk setiap ciri hartanah. Struktur ini dapat melancarkan gerak kerja dan kurang bergantung bagi mengurangkan risiko dan faedah pada pengayaan data. Menggunakan skrip yang disediakan, rangka kerja ini membolehkan SMD melakukan proses sendiri. Penyelidik mesti menyedari keistimewaan tertentu untuk menggunakan data secara berkesan, yang merupakan peluang bagi penyelidikan.

Kata kunci: Data terbuka, Geospatial, Python; Github API

ABSTRACT

Addresses can be geocoded by converting them into corresponding longitudes and latitudes. The coordinates offer a method for precisely locating a point on a map. This article presents a comprehensive methodology for widely-used online geocoding services. Our study aimed to validate the distribution of geocode locations to consider when analysing geocoded address data. Additionally, we sought to develop strategies

---

[1] Ahmad Najmi Ariffin is currently Assistant Director of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia and Postgraduate Student in the Centre for Restorative Dentistry, Faculty of Dentistry, The National University of Malaysia (UKM); and [2] Mohamad Hamizan Abdullah is currently Deputy Director of Agricultural and Environment Statistics Division, Department of Statistics Malaysia.

for enhancing demographic databases by leveraging the centralized public data sources repository, KAWASANKU API, on the Github platform. Open Data is information that is readily accessible, usable, and shareable by the public. Despite its potential benefits, the private sector has shown reluctance in embracing Open Data. If businesses strategically leverage Open Data, it can become a pivotal factor in creating diverse and potentially lucrative business opportunities, including the enhancement of 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 address strings for matching purpose with a q-ratio score threshold set at 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 obtain 2,427 geojson raw lines for each property feature. This framework facilitates a seamless, less-dependent workflow, reducing risks, and deriving benefits from data enrichment. By utilising the provided script, this framework empowers SMD to self-perform various processes. Researchers need to be aware of certain peculiarities to effectively leverage the data, presenting an opportunity for further research.

Keywords: Open data; Geospatial; Python; Github API

1. INTRODUCTION

Geocoding is the conversion of addresses, such as business addresses, into geographic coordinates, specifically longitudes and latitudes. These coordinates facilitate the mapping of locations and the placement of markers on a map, addressing the intricacies associated with geocoding addresses. In summary, this article contributes a comprehensive methodology for widely used online geocoding services. Our study aims to validate the distribution of geocode locations, providing insights for the analysis of geocoded address data. Additionally, we endeavor to develop methods for enhancing demographic databases, considering multiple administrative levels, including districts, parliamentary constituencies, 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 of the 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 (Veregin, 1999).

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.

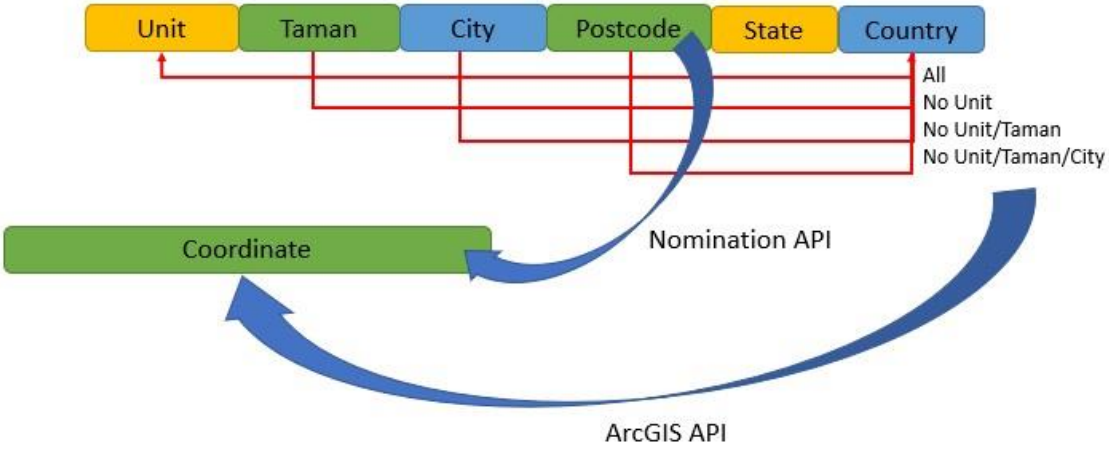


Figure 1: Ilustrates the Smart Search Mechanism implemented through the ArcGIS API.

The practice of mapping address records to physical locations is critical for understanding and leveraging geographic linkages that are fundamental to all statistics (MacEachren, 2017). 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 (Kirby, Delmelle, & Eberth, 2017).

1.2 The Advantages of Geocoding for Businesses

Facilitating the connection between businesses and their customers will be the main focus of the discussion. Employing geographical coordinates represents one approach. Consumers need a reliable method to determine their geographical coordinates. Historically, pinpointing an exact location posed considerable challenges (Cheng, Caverlee, & Lee, 2010). 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 Consulting (2018), address verification will be essential in the future. 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 Exploring the Availability of Open Data in 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 (Wendy et al., 2015). 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 (Khayyat & Bannister, 2015). 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 (Janssen, Charalabidis & Zuiderwijk, 2012). 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. According to Dawes, Vidiasova & Parkhimovich (2016), it provides a new channel for consumers to provide feedback for the purpose of enhancing the quality duplication on enhancing 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.

2. METHODOLOGY

A Geographic Information System (GIS) often includes a geocoding module that automates the geocoding process. This module is typically accessible on the Internet through a Web service interface. Data entry, such as a place name, street address, or zip code, is transmitted over the Internet using a communication protocol to interact with 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 (GeoPy, 2022). “Nominatim” is an OpenStreetMap data geocoder (Nominatim Documentation & OSM’s Nominatim Service, 2022). Installation of the “geopy” module is as 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__)
```

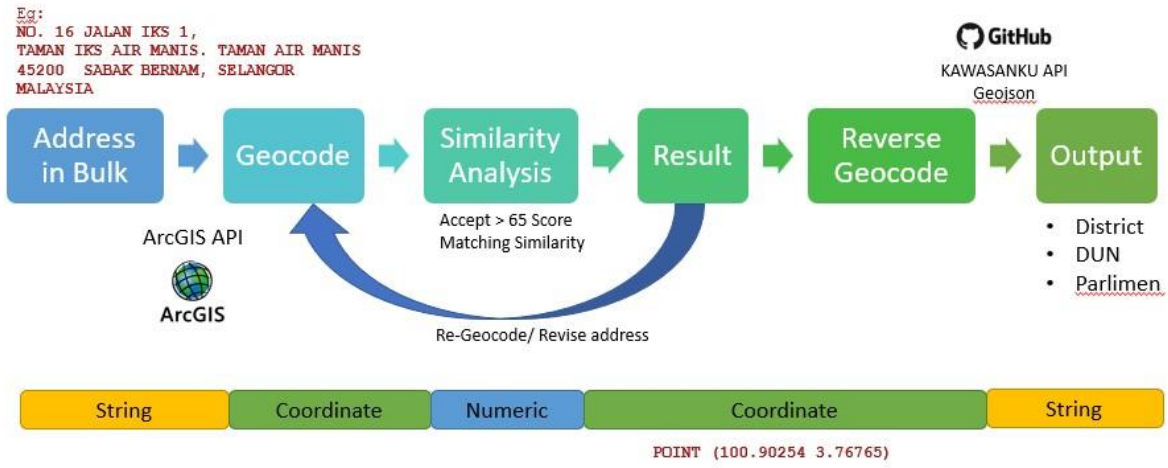
```
1.17.0
```

```
# 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),
```

Typically, each data entry takes just a few seconds to complete. The geocoding service transforms an input into coordinates and provides the user with the resulting information via the Internet. This information encompasses the coordinates, the geocoded address, and the corresponding accuracy level. Although this article employs the term "online geocoding," it's important to note that within the GIS community, other terms like real-time geocoding, address lookup, and address matching service are often used 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 (4) geocoding services offered by commercial geocoding companies in Krieger et al. (2001).

Figure 3: Workflow for Geocoding Address and Reverse Geocoding



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 (Roongpiboonsopit & Karimi, 2010). 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 geocoding do 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.

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 Application of Fuzzy String-Matching Technique for Address Matching

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 4: 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 (TheFuzz documentation, 2022) Python library, which is optimised for speed.

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

where 1(a*i*=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).

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

where |a| and |b| are the lengths of sequence a and sequence b, respectively (Krieger et al., 2001).

2.3 Mapping geocoded data using the 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, four (4) main components of data were fetched from the KAWASANKU GitHub API which were state, district, DUN and parliament data.

Figure 5: Integration setup for KAWASANKU Github API to Python Library

```
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()
```

```
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 expand the information at component levels (such as state, district, DUN, and parliament data).

3.1 Risk-assessment on Valid Geocode

The geocode addresses 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 6: 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 7: 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 Optimizing Data Enrichment by Extracting Data Using the KAWASANKU API from Open Data Sharing Sources

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 8: 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 aims to propose a more streamlined and independent workflow, minimizing risks and offering enhanced data enrichment. Before undergoing validation within the Subject Matter Division (SMD) at DOSM usually tasks the GIS team with performing geospatial analysis for address geocoding. The data retrieval process may span a week, delaying subsequent analysis within the SMD. However, this framework empowers the SMD to autonomously execute processes by leveraging a pre-designed script, thereby expediting the overall workflow.

4.1 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 the quality duplication on enhancing 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.

4.2 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 one (1) request per second is allowed.

In the future, we intend to utilize more advanced document embedding techniques, such as a Semantic Text Similarity Computing System based on Support Vector Machines (SVM). This approach aims to enhance our understanding of qualitative similarities between two datasets.

The output and application are only limited to the requirement of this study only and do not involve with outputs that require confidential geocode data/reversed geocode information (e.g. localities/ Mukim/ Villages/ *Taman*/ roads/ spatial features) for some details and do not involve analysis of measurements and spatial extent where the KAWASANKU application is limited in its use and displays information covering Malaysia, State, District, Parliament, DUN and postcode only. This is because the application methodology does not explain the extent to which the threshold has been set to meet the element of geospatial data quality involving the element of completeness (filling in a standard address) and the element of positional accuracy. The geocoding activity requires elements of quality control and conceptual geocoding is semi-automated processing (the verification process requires human intervention).

REFERENCES

Cheng, Z., Caverlee, J., & Lee, K. (2010). 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*, 759-768.

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.

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

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.

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

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

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.

Location Intelligence Drives Competitive Edge In The Digital Age. (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

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. *Springer*, Singapore, 139-155.

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/.

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

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

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

Wendy, C., Wae, S. C., Sander, F., & Eva, v. S. (Capgemini Consulting). (2015). Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources, https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf