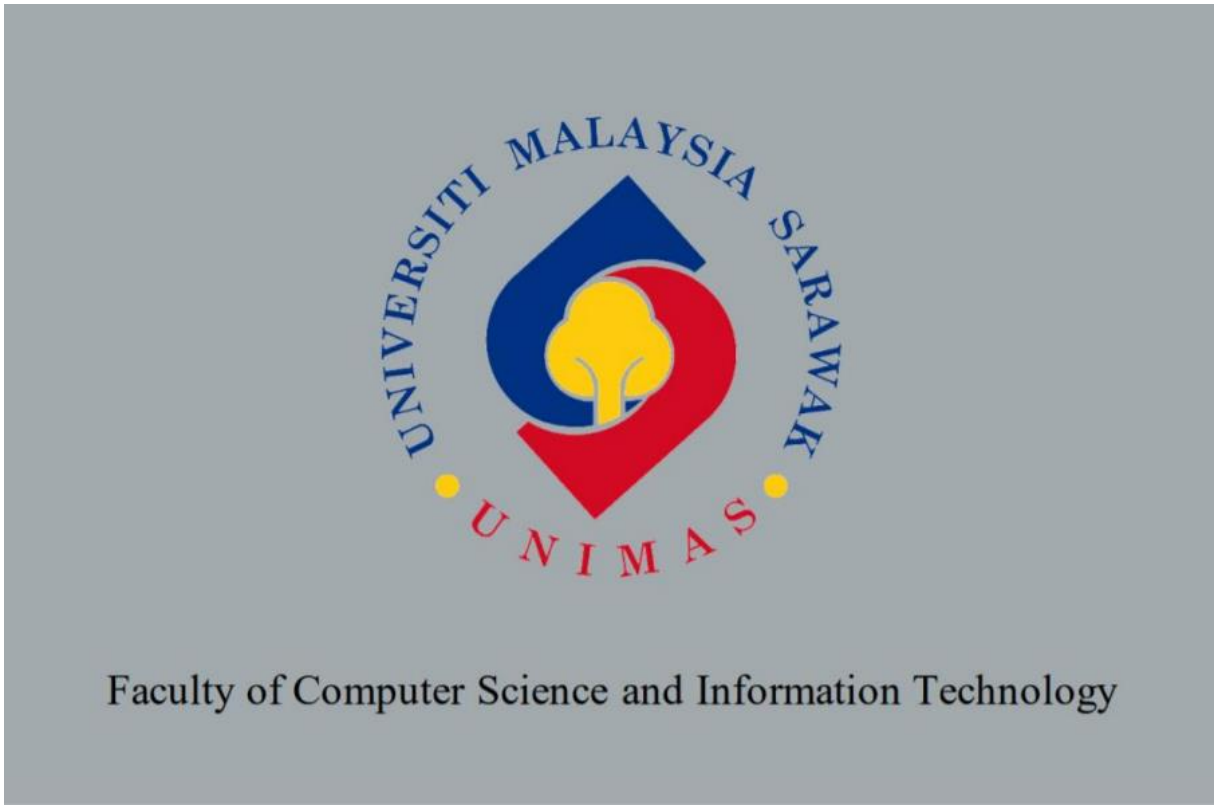

GEOSPATIAL TEXT ANALYSIS ON HISTORICAL DOCUMENTS: SARAWAK GAZETTE

Winnona Jane Binti Jesnik

Bachelor of Computer Science with Honours (Software Engineering)

2023

GEOSPATIAL TEXT ANALYSIS ON HISTORICAL DOCUMENTS:

SARAWAK GAZETTE

WINNONA JANE BINTI JESNIK

This project is submitted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science with Honours (Software Engineering)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2023

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to Dr. Suhaila Binti Saee, my advisor, for her guidance, support, and encouragement throughout my project in conducting geospatial text analysis on historical documents: Sarawak Gazette. Her expertise and valuable insights have been instrumental in the successful completion of this work. I am also grateful to Prof. Wang Yin Chai for providing clear instructions and access to all the necessary materials for us to complete our final year project. Without his assistance, this work would not have been possible. I would also like to thank my friends and family for their support and encouragement during the course of this work. Finally, I would like to express my appreciation to all those who have contributed directly or indirectly to this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This study conducted geospatial text analysis on historical documents, specifically the Sarawak Gazette, a colonial-era publication in Sarawak, Malaysia. A deep-learning model successfully extracted location named entities from structured historical documents in XML format. Geospatial visualization techniques mapped the distribution of these location names across the region on the Sarawak map by utilizing advanced NLP techniques, including Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM).

The analysis sheds light on the geographical context of historical places during Sarawak's colonial period, contributing to a better understanding of the historical significance and presence of specific locations. However, limitations include potential biases within the historical document collection and challenges in interpreting historical texts. Future research can enhance named entity recognition accuracy, refine location entity extraction, and explore innovative geospatial analysis methods to gain deeper insights into historical geospatial contexts. Overall, this study establishes a foundation for further investigations into geospatial analysis of historical documents, opening avenues for understanding the historical geography of Sarawak and beyond.

Keywords: Geospatial Text Analysis, Natural Language Processing, Deep-Learning Model, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Map Visualization

ABSTRAK

Kajian ini melakukan analisis teks geospasial pada dokumen sejarah, khususnya Sarawak Gazette, sebuah penerbitan era kolonial di Sarawak, Malaysia. Model pembelajaran mendalam berjaya mengekstrak lokasi bernama entiti dari dokumen sejarah berstruktur dalam format XML. Teknik visualisasi geospasial memetakan sebaran nama lokasi ini di seluruh wilayah di peta Sarawak dengan menggunakan teknik NLP canggih, termasuk Rangkaian Neural Berulang (RNN) dan Memori Jangka Pendek Panjang (LSTM).

Analisis ini menjelaskan konteks geografi tempat bersejarah semasa zaman penjajahan Sarawak, menyumbang kepada pemahaman yang lebih baik mengenai kepentingan sejarah dan kehadiran lokasi tertentu. Walau bagaimanapun, batasan merangkumi kemungkinan bias dalam pengumpulan dokumen sejarah dan cabaran dalam menafsirkan teks sejarah. Penyelidikan masa depan dapat menumpukan pada peningkatan ketepatan pengiktirafan entiti bernama, menyempurnakan pengekstrakan entiti lokasi, dan meneroka kaedah analisis geospasial yang inovatif untuk mendapatkan pandangan yang lebih mendalam mengenai konteks geospasial sejarah. Secara keseluruhan, kajian ini menetapkan asas untuk penyelidikan lebih lanjut mengenai analisis geospasial dokumen sejarah, membuka jalan untuk memahami geografi sejarah Sarawak dan seterusnya.

Kata kunci: Analisis Teks Geospatial, Pemprosesan Bahasa Asli, Model Pembelajaran Dalam, Rangkaian Neural Berulang (RNN), Memori Jangka Pendek Panjang (LSTM), Visualisasi Peta

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND OF STUDY

Sarawak Gazette is one of the Government Printing Office's first publications and was intended to promulgate the Rajah's orders and outstation officers (Pustaka Negeri Sarawak, n.d.). It is the largest single historical source of published information on the nation during the reigns of the second and third Rajahs (1868-1941) (Cotter, 1966).

The Electronic Sarawak Gazette is the first project created by White Hornbill's Innovative and Creative Circle with the purpose of specifically improving Sarawak Gazette consumption. The creation of The Electronic Sarawak Gazette will take place in phases. This website's Electronic Sarawak Gazette contains a digital version of the printed Sarawak Gazette's content. It is accessible from everywhere in the world with an internet connection, 24 hours a day, 7 days a week (Pustaka Negeri Sarawak, n.d.). Additionally, several of the page graphics are unclear and do not yet have part-of-speech tags, which aid in classifying words as nouns, verbs, adverbs, etc.

In order to improve information search speed, word, and context recognition, this project will map the historical locations and events that were extracted from the historical text documents. The approach getting applied will also process the data gathered from Sarawak Gazette and use it to train a model to recognize the context of the texts published within.

1.2 PROBLEM STATEMENT

The first problem faced by Sarawak Gazette is the location name ambiguity. It is challenging for the current system to identify the toponyms and to provide the proper coordinates for the Sarawak map display as a result of place name identification errors. Problems such as ambiguous, inconsistent, homonymies, and different spelling variations of the same location name in the historical documents of the Sarawak Gazette can be seen in the figure shown below which was found in Sarawak Gazette Vol. 77, No. 1125, December 28th, 1951, page 235:

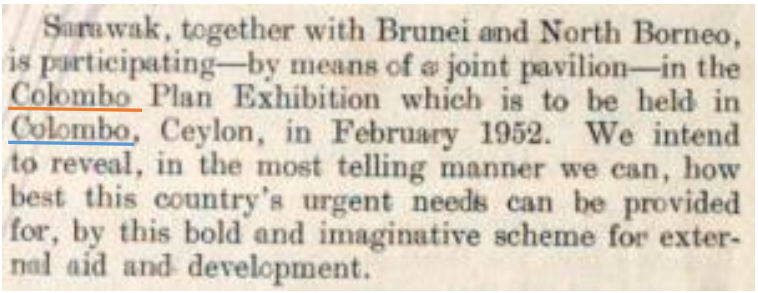
Sarawak, together with Brunei and North Borneo, is participating—by means of a joint pavilion—in the Colombo Plan Exhibition which is to be held in Colombo, Ceylon, in February 1952. We intend to reveal, in the most telling manner we can, how best this country's urgent needs can be provided for, by this bold and imaginative scheme for external aid and development.

Figure 1.1 Screenshot of a paragraph from Sarawak Gazette Vol. 77, No.1125, December 28th, 1951, page 235

The phrase "Colombo" underlined in blue relates to a site in Ceylon, whereas the word "Colombo" underlined in red is part of the name of an event. However, as the machine might not accurately identify the entities based on their context, this necessitates that the people in charge of event and location entity extraction have knowledge of historical events and location names. Additionally, for the system to accurately recognize each entity's context, both entities must be correctly tagged.

As the years passed, certain locations are outdated or have since changed their names, making it difficult to keep track of currently available place names in the official geographical database such as OpenStreetMap (OSM). According to Rayson et al. (2017), locational errors occur when a place name might not correspond to the actual site. This may be due to disambiguation errors (Lancaster England, Lancaster Pennsylvania, or Lancaster California) or the challenges associated with accurately depicting features like rivers, lakes, or regions using point data.

ID	Location	Address	Lat	Lon
88	Great Britain	(Great Britain, United Kingdom, (54.31536155, ...	54.315362	-1.918023
189	Kuala Saribas	None	NaN	NaN
56	12th Mile, Simanggang Road	None	NaN	NaN
52	Sarawak	(Sarawak, Malaysia, (2.5023855, 112.9547283))	2.502385	112.954728
63	Tarat	(中国, (35.000074, 104.999927))	35.000074	104.999927

Figure 1.2 Screenshot of automatically retrieved address and coordinate samples using GeoPy and Nominatim

As shown in Figure 1.2, the address and coordinates for Kuala Saribas and 12th Mile, Simanggang Road are not found and hence, listed as 'None' and 'NaN' respectively. This may be due to the fact that Kuala Saribas is a stream of mouth located on a body of water while according to an article by Lian Cheng (2019), Sri Aman town was originally called Simanggang in the early days, and the change in the town's name occurred in 1974. Multiple sources of historical and current locations may need to be compared in order to obtain the accurate location names, address and coordinates.

1.3 PROJECT OBJECTIVES

This project aims to visualize historical text documents of the Sarawak Gazette on the Sarawak map. The objectives of this project are as follows:

- To create a geospatial corpus from unstructured historical text documents of Sarawak Gazette,
- To extract name entities from the annotated historical text documents using deep learning model,
- To visualize the name entities on a Sarawak map.

1.4 METHODOLOGY

1.4.1 Geographical Text Analysis

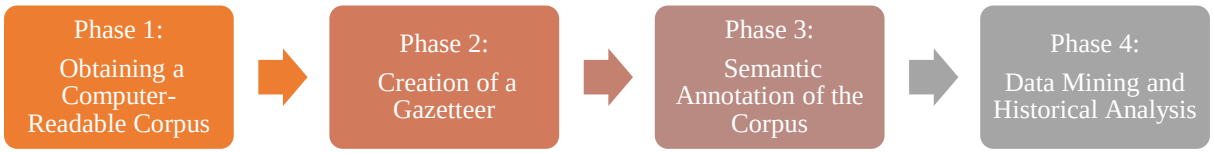

Figure 1.3 Geographical Text Analysis (GTA) methodology phases

Geographical Text Analysis (GTA) is a computer-assisted methodology originally produced to reveal new, unsuspected patterns of information by focusing on the geographic components of the historical narrative (Cooper and Gregory 2011) (Murietta-Flores et al. 2015) (Murrieta-Flores et al. 2017) (Gregory et al. 2018). There are 4 phases involved: obtaining a computer-readable corpus, the creation of a gazetteer, the semantic annotation and geoparsing of the corpus, and data mining and historical analysis.

Phase One: Obtaining a computer-readable corpus

Identify and correct significant errors caused by Optical Character Recognition (OCR) software, which is a common problem with old documents.

Phase Two: Creation of a Gazetteer

1. **Compilation of all the place names that exist in the corpus**

   Extracting toponym indices from published editions of historical texts and comprehensive studies on historical geography.

2. **Linguistic Disambiguation**

   Resolving spelling inconsistencies and name variations of the same toponym, the application of Castilian Transliterations to the names of ancient settlements, or dealing with homonyms.

3. **Geographic Disambiguation**

   Detecting and merging duplicate entries by comparing the place name list with records found online in geographical databases. The entries will be used for the gazetteer and the structure will be available in JSON (JavaScript Object Notation) format, an accepted standard in many textual annotation and analysis platforms as well as in a standalone downloadable GIS (Geographic Information System) format.

Phase Three: Semantic Annotation of the Corpus

1. **Defining the Annotation Schema (Ontology)**

   Selection of the ontology categories.

2. **Manual Annotation of Text Samples for Training the NLP Model**

   Manual annotation of historical text documents.

3. **Development of the NLP Deep-Learning Model**

   a. **Neural Networks**

   It is a computational model applied in the NLP (Natural Language Processing) and CL (Corpus Linguistics) fields to perform word classification, or the assigning of a label to each word in an input text. Indicating whether it belongs to a certain semantic category or a "part-of-speech" syntactic category.

   b. **Word-Embedding**

   The word-embedding algorithm makes the neural network easily distinguish all of the different semantic contexts in which the word might be used.

Phase Four: Data Mining and Historical Analysis

Several data mining and analysis procedures can be carried out during this stage.

1.5 PROJECT SCOPE

This project's scope involves selecting dataset samples from historical documents of Sarawak Gazette to identify the name entities found and to extract toponyms to be visualized on the Sarawak map.

1.6 SIGNIFICANCE OF PROJECT

Ambiguity, inconsistency, homonymies, and spelling variations of the same toponyms in the historical documents of Sarawak Gazette lead to difficulties for the system in identifying the toponyms and assigning correct coordinates. The outcome of this project will be used in visualizing the Sarawak map using trained Sarawak Gazette datasets of extracted locations and events from the unstructured historical documents of Sarawak Gazette which are associated with each other using NLP model training. Furthermore, the extracted toponyms will be compiled in a gazetteer for future references and works related to Sarawak Gazette.

1.7 PROJECT SCHEDULE

Figure 1.4 shows the Gantt chart of the project for Semester 1 and Semester 2 of 2022/2023 session in APPENDIX.

1.8 PROJECT OUTCOME

The expected outcome for this project is a visualized Sarawak map using trained Sarawak Gazette datasets of extracted locations and events from the unstructured historical text documents of Sarawak Gazette which are associated with each other using NLP model training and the extracted toponyms will be compiled in a gazetteer for future references and works.

1.9 PROJECT OUTLINE

***Chapter 1**: Introduction*

Chapter 1 introduces and explains the project in brief.

***Chapter 2**: Literature Review*

Chapter 2 conducts a thorough study of related research, articles, documentation, or applications of Geospatial Text Analysis (GTA) and its background study will be explained.

***Chapter 3**: Methodology*

Chapter 3 discusses the usage, advantages, and how the applied methods, and algorithms will assist in reaching this project's objectives.

***Chapter 4**: Results and Discussions*

Chapter 4 collects the results and carries out a discussion to analyze the accuracy and precision once the methods and algorithms are applied to the test data sample.

***Chapter 5**: Conclusions and Future Work*

Chapter 5 concludes the results obtained to validate all objectives and future works related to this project will be discussed.

1.10 CHAPTER SUMMARY

In this chapter, the research project is depicted in a simple flow starting with the study's background, followed by the problem statement, project objectives, methodology, project scope, the significance of the project, project schedule, project outcome, and the project outline. The following chapter will touch on the literature review related to this research project.

CHAPTER 2: LITERATURE REVIEW

2.1 OVERVIEW

Chapter 2 includes the background study and literature review of this project. The purpose of this background research and literature review is to learn about different types of data and the comparison between the data types. This study also seeks to conduct a review on existing similar works and the different approaches in data visualisation.

2.2 DATA STRUCTURE

2.2.1 Unstructured Text

Unstructured text refers to text data that does not have a predefined structure or format. It is not organized in a specific way and does not conform to a specific schema or set of rules. Examples of unstructured text include:

- Social media posts: These often contain informal language, emojis, and hashtags, and have no predefined structure.
- News articles: They may have a headline and a byline, but the structure of the article itself is not predefined.
- Emails: Emails can contain a variety of information, including text, images, and attachments, and have no predefined structure.
- Customer feedback: Customer feedback may be in the form of text, voice or video, and has no predefined structure.
- Blogs and forums: They have no predefined structure and can include a variety of information, including text, images, and links.

Unstructured text is usually harder to analyze than structured text because it does not conform to a predefined schema. However, with the help of NLP techniques it can be transformed into structured data and analyzed.

2.2.2 Structured Text

Place Structured text refers to text data that has a predefined structure or format. It is organized in a specific way and conforms to a specific schema or set of rules. Examples of structured text include:

- Database records: These are typically organized into fields, such as name, address, and phone number, and have a predefined structure.
- Spreadsheets: They have a grid-like structure, with rows and columns, and the data is organized into cells.
- HTML documents: They have a predefined structure, with elements such as headings, paragraphs, and lists that can be identified and parsed.
- XML and JSON files: They have a predefined structure, with elements that can be identified and parsed.
- CSV files: They are plain-text files that store tabular data, and the data is organized into rows and columns.

Structured text is generally easier to analyze than unstructured text because it conforms to a predefined schema. The structure makes it easier to extract specific information and perform various types of analysis. However, the process of transforming unstructured text into structured text can be challenging and require NLP techniques.

2.2.3 Comparison of Unstructured and Structured Texts

Unstructured text and structured text are two different types of text data that have different characteristics and are used for different purposes. A comparison of the two is as follows:

Table 2.1 Comparison of unstructured and structured data

	Structured data	Unstructured data
Analysis	Quantitative	Qualitative
Schema Creation	Schema-on-write	Schema-on-read
Searching	Easy using SQL-based methods	May need special tools
Format	Predefined, using alphanumeric characters	Typically non-character-oriented digital representations
Storage	May require more storage to accommodate defined data structures	Some forms require less storage; others have large file formats, requiring more storage
Storage Format	Relational database management systems, data warehouses	Applications, NoSQL, databases, data lakes

2.3 REVIEW ON EXISTING SIMILAR WORKS

Geospatial text analysis of historical documents is an area of research that combines natural language processing and geospatial information to extract meaningful insights from historical texts that are spatially referenced. Some examples of existing works in this field include:

1. **Geocoding historical documents (Mertel, A. et al., 2021):** Researchers have used geocoding techniques to map the locations mentioned in historical documents and visualize patterns of historical events.

2. **Temporal and spatial analysis of historical texts (Yuan, Y., 2014):** Researchers have used natural language processing techniques to analyze the temporal and spatial aspects of historical texts, such as identifying the time and place of events described in the texts.
3. **Named Entity recognition and geolocation of historical texts (Tambuscio & Andrews, 2021):** Researchers have used named entities recognition techniques to extract the names of people, organizations, and locations from historical texts and geolocate them on the map.
4. **Geospatial sentiment analysis of historical texts (Rayson et al., 2017**): Researchers have used natural language processing techniques to analyze the sentiment of historical texts that are spatially referenced, such as diaries or letters.
5. **Geospatial topic modeling of historical texts (Gavin, M., & Gidal, E., 2016):** Researchers have used topic modeling techniques to uncover latent topics in historical texts and map them to specific geographic regions.

Overall, the existing works demonstrate the potential of geospatial text analysis to uncover valuable insights from historical texts that can help us to understand the past and the context of the events that took place. These techniques could be applied to various fields such as history, sociology, archaeology, and more.

CHAPTER 3: METHODOLOGY

3.1 DIRECTION OF PROPOSED WORK

The proposed work for this project will involve the following steps: data preparation, geospatial text analysis, model training, and map visualisation.

3.1.1 Data preparation

The processed data will be separated into two sets – test set and training set for the model training. The test set will contain 5 samples of location and event name entities from Sarawak Gazette Volume 77, Issue No. 1125, December 1951 to test the accuracy and precision of the training model.

3.1.2 Geospatial text analysis

The processed data will then be analysed using natural language processing and geographic information systems techniques. This will involve extracting geographical entities and locations mentioned in the Sarawak Gazette, conducting sentiment analysis, and extracting key themes and patterns related to the colonial administration and society in Sarawak.

3.1.3 Model training

The extracted information will then be used to train a machine learning model and SpaCy that can automatically extract geographical information and insights from the Sarawak Gazette's text data.

3.1.4 Map visualization

The extracted geographical information will be mapped to specific locations in Sarawak to provide a visual representation of the Sarawak Gazette's text data.

3.2 PROJECT WORKFLOW

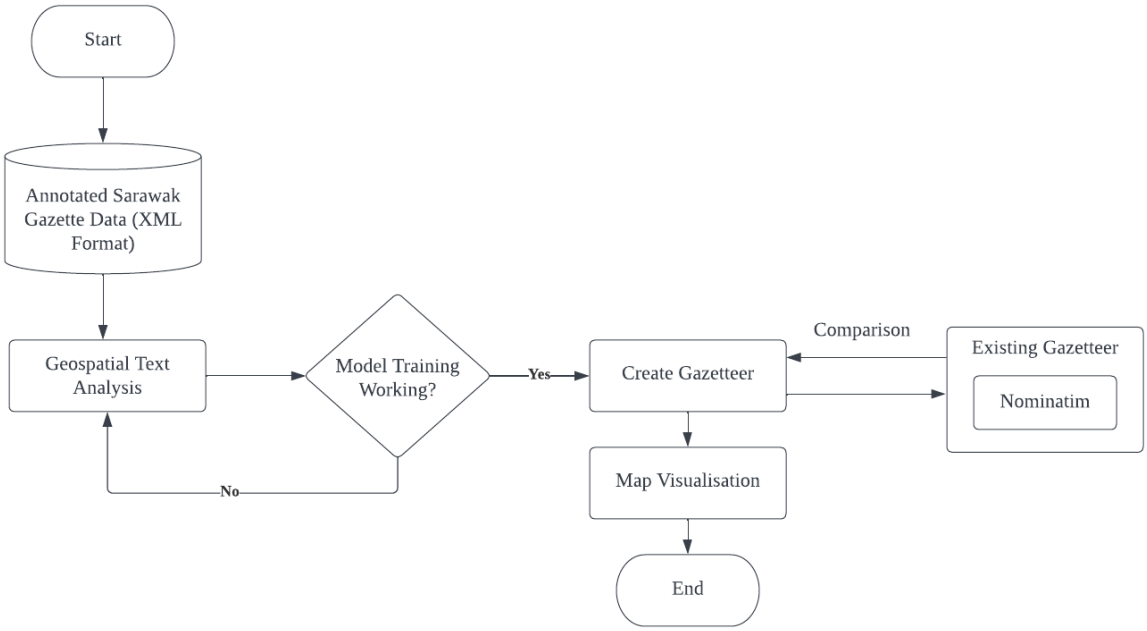

Figure 3.1 Project workflow

As shown in Figure 3.1, the project will use currently available processed Sarawak Gazette data in XML (Extensible Markup Language) format as the input data to conduct the model training using spaCy and Python in addition to creating a custom gazetteer for future references and works related to Sarawak Gazette. The training model should be able to output high-accuracy results in terms of data context and structure by training two sets of data – a test set and a training set – which assist in understanding the relationship between extracted location and event name entities. The Geographical Text Analysis (GTA) method will be adapted to achieve the objectives stated for this project.

3.3 PHASE 1: DATA PREPARATION

Following the first phase of GTA method shown in Figure 1.3, a computer-readable corpus will be obtained. The data that will be used in this project is extracted from Sarawak Gazette Volume 77, Issue No. 1125, December 1951 which has been annotated and saved in the form of XML format. The name entities required for this project are location and event name entities, hence, the test set will consist of 5 data samples of location and event name entities while the training set will consist of the whole Sarawak Gazette Volume 77, Issue No. 1125, December 1951 data.

```
<LOCATION gate:gateId="95703">KUCHING</LOCATION>
```

Figure 3.2 Annotated text in Sarawak Gazette labelled as <LOCATION> name entity

3.4 PHASE 2: MODEL TRAINING

To start with the model training, BeautifulSoup can be used to parse the XML document to find location and event name entities. A pre-trained core language model from the spaCy library will be utilised to train the extracted data. Figure 3.3 shows visually that the subject of the sentence is the proper noun Gus and that it has a learn relationship with piano.

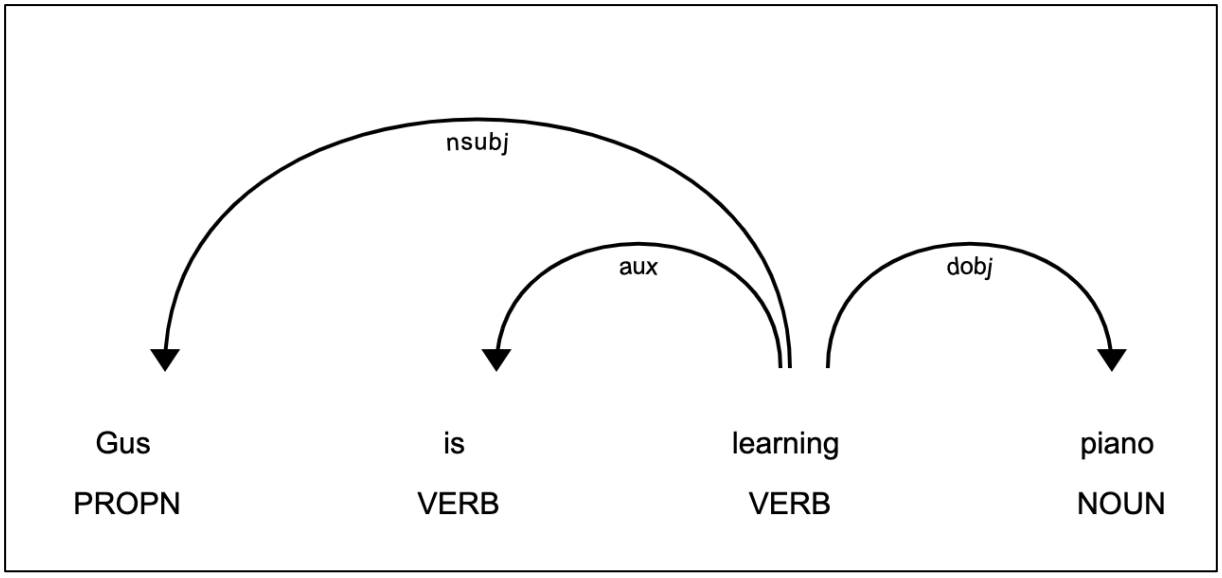

Figure 3.3 Taranjeet Singh (Author). (2023). displaCy: Dependency Parsing Demo [Digital Image]. https://realpython.com/natural-language-processing-spacy-python/