ARI2201 Individual Assigned Practical Task

Location Chronicles

Matthew Kenely

matthew.kenely.21@um.edu.mt

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

The influx of data in recent years has contributed to the quality of news platforms significantly, with the provision of personalised news content based on users' locations and preferences becoming more accurate by the day. These platforms often focus on larger geographical regions such as countries or cities. For users in smaller regions such as Malta, however, there is a lack of localised news content that caters to their specific interests and preferences on a per-village basis.

Malta has a unique landscape with distinct villages, each with its own local events, issues, and news. For the Maltese population, having access to news articles that are specifically relevant to their local villages can greatly enhance their news consumption experience and keep them informed about happenings in their immediate surroundings.

The aim of this project is to provide Maltese users with a platform which facilitates the recognition of where news articles took place geographically, as well as the ability to filter a database of news articles based on geographical locations, either tailored to a user's location or to whichever village they select. This entails the creation of a news article database, as well as the training of a machine learning model to recognise locations in news articles.

The task of recognising locations in news articles is a Natural Language Processing (NLP) problem, and more specifically, a Named Entity Recognition (NER) problem. NLP is a field of computer science which deals with the interaction between computers and human (natural) languages. NER is a subtask of NLP which deals with the identification of named entities (subjects of interest) in text, such as people, locations and organisations [1]. In this project, the task of the developed NER algorithm will be to identify the locations (specifically, Maltese villages) in which news articles took place.

To successfully create this platform, the

following tasks shall be carried out:

1. The creation of a **dataset** of news articles which take place in (and ideally mention) Maltese villages.

2. The training of a **Named Entity Recognition model** on this dataset, with the goal of it being able to recognise where Maltese news articles took place.

3. The development of a **web application** on which the following features will be hosted:

    - An article location detector which utilises the trained model to identify locations in news articles provided by the user.
    - A news article database lookup which allows the user to filter said articles in the created dataset based on where they took place in Malta, either tailored to their location (given their consent) or to whichever village they select.

By creating a Maltese village location detector, the proposed platform aims to bridge this gap and provide a tailored news experience to Maltese users. Furthermore, the development of such a platform aligns with the growing trend of hyper-localized news consumption, where users seek information that is highly relevant and specific to their immediate surroundings. By focusing on Maltese villages, the platform can cater to the unique needs and interests of the Maltese population, ensuring that they stay informed about events and developments in their local communities.

2 Background

In recent years, the tendency has been for people to consume news on online platforms [2] [3], offering users a vast amount of news articles from various sources. One popular example of such a platform is Google News, a news aggregator that collects headlines and snippets from news sources worldwide, allowing users to access a wide range of news articles on different topics. One key feature of Google News relevant to this project is its ability to tailor news articles to users' locations. This feature works well when it comes to filtering news articles on a per-country basis (including Malta), and while the option to filter on a per-village basis is available, it is not refined, with Google News often returning any articles which took place in Malta rather than in a specific village.

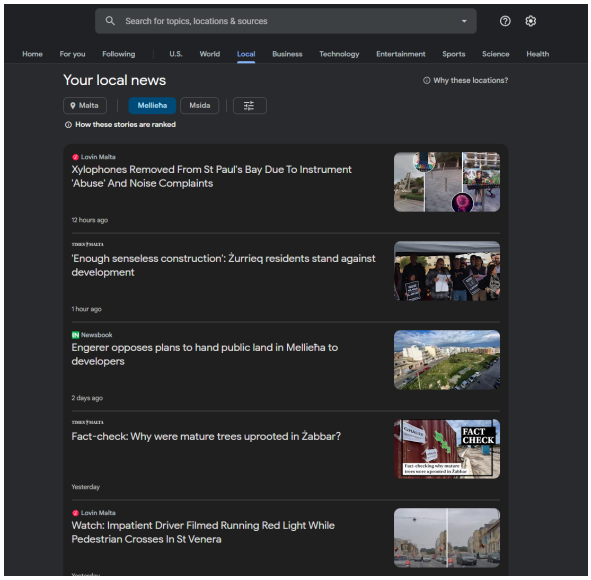

Figure 1: Google News filtering by village (Mellieha) displaying articles which took place in St. Paul's Bay, Żurrieq and Żabbar

BBC News also offers this feature to a limited extent in its World section, allowing users to filter news articles on a per-continent basis. However, this feature is not available for smaller geographical regions such as, and as such, cannot tailor news to a user based on the country they live in, let alone the specific city/village.

3 Methodology

3.1 Dataset

All source files for work related to this section can be found in the `docs/data` *di-*

rectory

To create a dataset of Maltese news articles, I created an algorithm (found in `scraper.ipynb`) which scrapes news articles in the Newsbook English local archive (found here). I was able to scrape the following content of 5797 articles: URL, title, date, and article text. This dataset can be found in `all_articles.csv`. I proceeded to clean this dataset (specifically removing duplicates) in Google Sheets, followed by a filtering process to only include articles which mention the names of Maltese villages (found in `villages.txt`) in their text, as well as the start and end index of these villages within the text (this is required to train the NER model), resulting in a dataset of 2454 articles (found in `location_articles.csv`). The code for this filtering process can be found in `location_extractor.ipynb`. A final pass through of the dataset was carried out to get the URL of each articles' corresponding image (if any) and add it to the dataset (this was also done in `location_extractor.ipynb`), resulting in the final dataset of 2454 articles (found in `location_articles_images.csv`).

3.2 Model

All source files for work related to this section can be found in the `docs/data` *directory*

The Named Entity Recognition model for this project was trained using the spaCy library in Python. spaCy is an open-source library for NLP written in Python, which provides a wide range of NLP tools and features, including the ability to train custom NER models.

I used the blank English model provided by spaCy as a base model (with the intention of creating a model targetted towards the recognition of the names of Maltese villages), and proceeded to train it on the `location_articles_images` dataset, taking the news article text and corresponding village name indices and entity labels (scraped in the Dataset subsection) as inputs. The model was trained for 10 epochs, with a dropout rate of 0.3.

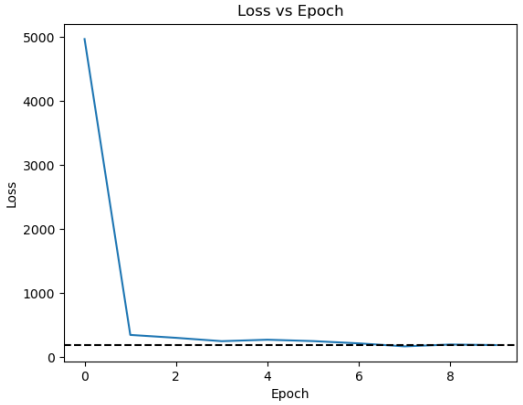

Figure 2: Loss of the NER model over 10 epochs

The code used to train the model can be found in `train_model.ipynb`. The model was saved in the `model` directory.

3.3 Web Application

All source files for work related to this section can be found in the `docs` *directory*

The web application for this project was developed using PHP, CSS, JavaScript and Python. The application must be hosted (this can be done locally using open-source software such as XAMPP).

The application is mobile responsive, allowing users to access it from whichever device they prefer.

The application makes use of the OpenStreetMap API to display a map of Malta on the home page, as well find the coordinates of the villages identified by the Article Location Detector and subsequently display said villages on the map. The Navigator Geolocation API is used on the the News page to ask the user for permission to access their location, which, if provided, is used to automatically filter the news articles in the database based on the village the user is currently in.

3.3.1 Home Page

The home page contains a form (called the Article Location Detector) which allows

the user to input a news article URL, which is then passed to the NER model to attempt to identify the Maltese village in which the article took place.

3.3.2 News Page

The news page contains a form (called the Database Lookup) which allows the user to either provide their current location or select a village manually, which is then passed to a python script which filters the `location_articles_images.csv` dataset, and retrieves news articles which took place in that village. Where an article took place is determined by checking which village name occurred most frequently in the article text, and if there is a tie, the village name which occurred first is chosen. The articles are then displayed on a grid of cards, each containing the article title, date, image (if any), and a link to the original article. The cards are sorted by date, with the most recent articles appearing first.

4 Results

4.1 Dataset

4.2 Model

4.3 Web Application

5 Challenges and Limitations

6 Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[2] S. Bennett, K. Maton, and L. Kervin, "The 'digital natives' debate: A critical review of the evidence," *British journal of educational technology*, vol. 39, no. 5, pp. 775–786, 2008.

[3] A. C. Ripollés, "Beyond newspapers: News consumption among young people in the digital era," *Comunicar. Media Education Research Journal*, vol. 20, no. 2, 2012.