Historic Events Search Engine

Information Processing and Retrieval - FEUP

Dinis Sousa
FEUP
Porto, Portugal
up202006303@edu.fe.up.pt

Gabriel Ferreira
FEUP
Porto, Portugal
up201906072@edu.fe.up.pt

João Matos
FEUP
Porto, Portugal
up202006280@edu.fe.up.pt

João Pinheiro
FEUP
Porto, Portugal
up202008133@edu.fe.up.pt

ABSTRACT

This article describes the first of three milestones of creating a search engine for historic events. Dataset extraction, preparation, and refinement will be at focus, as well as data analysis to further understand the needs of our future search engine.

CCS CONCEPTS

• **Information systems → Web searching and information discovery**; *Data management systems*; *Information retrieval query processing*.

KEYWORDS

historic events, search engine, dataset, data preparation, data analysis, information, retrieval, processing

1 INTRODUCTION

This paper was developed during the course of Information Processing and Retrieval, a Curricular Unit of the Master in Informatics and Computing Engineering course at FEUP.

During this course, the students were asked to make use of the theoretical knowledge from our Information Processing and Retrieval classes, building a search system on a topic of our choice.

We opted to use Wikidata [4] as our primary source of data, choosing historical events as the theme for our project (a historical event is defined by Wikidata as a "particular incident in history that brings about a historical change" [5]).

2 DATA SELECTION

Our initial step is to choose the datasets we will use. The only requirement was that the selected data should have both structured data and unstructured data.

We decided to explore datasets on wars and conflicts.

Some datasets, such as the GDELT (The GDELT Project) [2] or the UPPSALA Conflict Data Program (UCDP) [3], had too many entries and little to no rich text to explore.

We settled with using data from Wikidata, selecting the wanted data using the Wikidata Query Service, and using Wikipedia[6] articles to get the rich text.

After analyzing the different Wikidata parameters, we settled on using the following query to retrieve the main dataset used (containing 9134 results):

```
SELECT ?event (SAMPLE(?date_) as ?date) ?label (SAMPLE(?
    image_) as ?image) ?article  WHERE {

  ?event (wdt:P31/(wdt:P279*)) wd:Q13418847;
    wdt:P585 ?date_.
  ?event rdfs:label ?label.
  FILTER((LANG(?label)) = "en")
  OPTIONAL {?event wdt:P18 ?image_.}
    ?article schema:about ?event;
    schema:isPartOf <https://en.wikipedia.org/>.
}
GROUP BY ?event ?label ?article
LIMIT 10000
```

Listing 1: SPARQL query to retrieve historical events

This dataset has the event id, the event date, a label (name of the event), an image and the URL for the Wikipedia page.

The dataset is in a *.json* file format.

3 DATA PROCESSING

3.1 Data Collection

With the primary dataset in hand, we completed and treated the information as described in the annexed pipeline (5).

Iterating through the events dataset, we complete each event's information with: 1) a Wikipedia summary, and 2) other statements that Wikimedia had on the event.

Other statements include: list of participants, location, coordinate location, part of (if the historical event is a part of a bigger event, ex.: a battle is part of a war), time period, number of deaths, number of injured, etc...

After filtering out events that had no summary we ended up with 8998 events, mainly armed conflicts.

3.2 Data Refinement

When it comes to the Data Refinement process, the first step was to exclude all columns with too many missing values, more than 85%, with the most incomplete columns having only a dozen of values.

Everything described up to this point is represented in the data pipeline, show in Figure 5.

3.3 Conceptual Model

In our conceptual model (see Figure 1), we define several key entities that form the foundation of our historical event dataset:

- **Complex Conflict:** This entity represents a bigger conflict, such as a war. A Complex Conflict exists as a standalone entity and incorporates various Historical Conflicts that are part of it. It serves as a high-level container for historical conflict data.
- **Historical Conflict:** Historical Conflicts are components of Complex Conflicts. These represent individual conflicts within the broader context of a Complex Conflict. Historical Conflicts provide detailed information about specific events, including their participants, locations, and dates.
- **Place:** The Place entity represents various location types, including countries, regions, cities, rivers and more. It serves

as a critical element for specifying where historical events occurred. Each historical event is associated with a Place to indicate its geographical context.

- **Participant:** Participants in historical events are classified into different types, including individuals, groups of people, and organizations. This entity helps categorize and identify the key actors involved in Historical Conflicts, providing insights into the composition of these events.

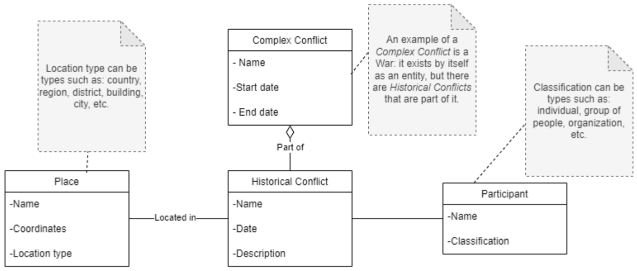
Location type can be types such as: country, region, district, building, city, etc.
Complex Conflict
- Name
-Start date
- End date
An example of a Complex Conflict is a War: it exists by itself as an entity, but there are Historical Conflicts that are part of it.
Classification can be types such as: individual, group of people, organization, etc.
Part of
Place
-Name
-Coordinates
-Location type
Located in
Historical Conflict
-Name
-Date
-Description
Participant
-Name
-Classification

Figure 1: Conceptual Model

The conceptual model defines the fundamental entities and their relationships within our historical event dataset, enabling a structured and comprehensive representation of historical conflicts and their contextual information.

4 DATA CHARACTERIZATION

After collecting and processing all the necessary data, our next step was to conduct a comprehensive analysis of the data.

4.1 Descriptive Analysis

We assessed the dataset for specific characteristics such as the number of events with available images and the number of events that included details about participants. These statistics are crucial for understanding the richness and completeness of our dataset.

Our analysis revealed that out of the total events, 5664 included images, providing visual context for a significant portion of the dataset. Additionally, 3240 events contained information about participants, indicating the extent to which historical records specify the key actors in these events.

4.2 Spatial Distribution Analysis

To gain insights into the geographic distribution of these events, we utilized an interactive map (see Figure 2) to visualize their global locations. The geographical coordinates of each event were extracted using the "coordinate location" statement, and we employed the Python module 'folium' [1] to plot these events on a world map.

Our initial hypothesis was the distribution of historical events would exhibit an asymmetric pattern, with a concentration of events in Europe. This hypothesis was based on historical records and expectations regarding the frequency of events across different regions.

The visualization on the interactive map allowed us to confirm and quantify this hypothesis, providing a clear representation of the distribution of historical events across the world.

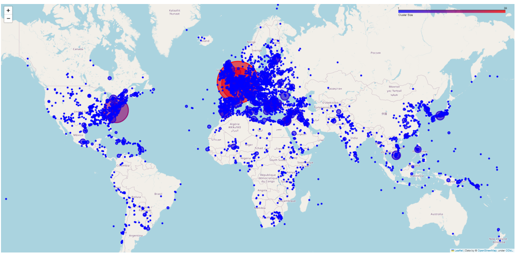

Figure 2: Geographic Distribution of Historical Events

4.3 Analysis of Word Clouds

To acquire a better understanding of the key terms and themes within our dataset of historic events, we generated word clouds for both event titles and summaries. These word clouds visually represent the frequency of words, with larger words indicating higher frequency. Our analysis reveals essential patterns and recurrent themes in the dataset.

4.3.1 Historic Event Titles Word Cloud, see Figure 3. The word cloud for historic event titles provides a snapshot of the most frequently occurring terms. Some initial observations include:

- The word 'Battle' prominently stands out, indicating its high frequency in event titles.
- Additional terms such as First and Second also feature prominently, suggesting a focus on particular historical themes like the worls wars.

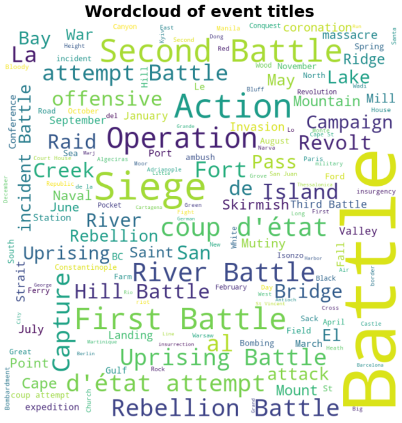

Figure 3: Word Cloud for Historic Event Titles

4.3.2 Historic Event Summaries Word Cloud, see Figure 4. The word cloud for historic event summaries offers a deeper insight into the content of these events. Key takeaways from the summary word cloud analysis include:

- The word 'Battle' remains a dominant term, reflecting its significance in historical narratives.
- The summary word cloud exhibits a broader diversity of terms, including european countries, types of locations, and terms within the same word family as 'war', indicating a range of themes and historical contexts.

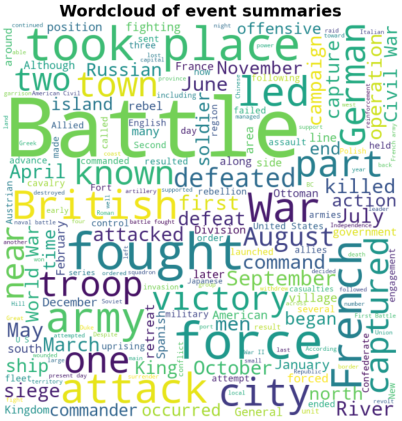

Figure 4: Word Cloud for Historic Event Summaries

These word clouds offer valuable visual representations of the most frequent terms in event titles and summaries, highlighting key themes and recurring words. The prominence of 'Battle' in both titles and summaries underscores its historical significance. The summary word cloud, in particular, showcases the diversity of terms surrounding historical events, providing a nuanced understanding of the dataset.

The spatial distribution analysis we have seen before, along with the characterization of dataset attributes, forms the foundation for our search engine's capabilities in providing users with a rich and diverse set of historical information. It enables users to explore and understand the geographical and contextual aspects of historical events in an interactive and informative manner.

5 INFORMATION NEEDS

In order to develop a search engine for historical events, it is essential to identify the information needs of potential users. Here, we outline four specific information needs related to historical events, each of which corresponds to a different aspect of the search engine's functionality:

1. **Participation in 19th Century Battles:** Users may require information about individuals, military units, or nations that participated in the major battles of the 19th century. This information can help researchers and history enthusiasts understand the key players in historical conflicts during this period.
2. **Deadliest Battles in Europe during World War I:** Users may seek data on the deadliest battles that occurred in Europe during World War I. Such information is crucial for those interested in the impact and scale of warfare in this region during the early 20th century.
3. **Portuguese Participation in Battles (1300-1800):** Users interested in Portuguese military history may want to know about the battles in which Portugal participated between the 14th and 18th centuries. This historical context can help researchers explore the country's military involvement during this period.
4. **Consequences of the Spanish Civil War:** The consequences of historical events are often of great interest. Users may seek detailed information on the consequences of the Spanish Civil War, including its impact on Spain and the world. This information is valuable for understanding the broader implications of historical conflicts.

To address these information needs, our search engine utilizes a dataset sourced from Wikidata and Wikipedia, which contains structured data, rich text, and images related to historical events. The data has been processed and organized to facilitate effective information retrieval.

6 SEARCH DOCUMENT

In our search result document, we present a comprehensive overview of a historical event, including the following elements:

- **Event Name:** The title or name of the historical event, which provides a concise identifier for the event.
- **Truncated Summary:** A brief summary or description of the event, providing key details and context. The summary may be truncated for readability.
- **Image:** An image or visual representation associated with the historical event, offering visual context and insights.
- **Map:** A geographical map displaying the location of the event. The map helps users understand where the event took place and its spatial context. Location information may include country, region, district, or city.
- **Timeline:** A timeline indicating the date and period during which the historical event occurred. It provides a temporal context, allowing users to understand when the event took place in history.

The search result document combines these elements to provide users with a comprehensive and informative representation of the historical event they are interested in. Users can gain insights into the event's name, summary, visual aspects, location, and temporal context, facilitating a deeper understanding of the historical narrative.

7 CONCLUSION

Main takeaways from the project

In the course of developing our rich-text, coherent dataset of historical events, several key takeaways have emerged:

1. Utilizing Wikidata as a data source has proven to be a valuable approach. The combination of structured data and rich text from Wikipedia articles has allowed us to create a robust dataset for historical events.
2. The spatial distribution analysis of historical events has provided insights into their geographic concentration, with a notable focus on Europe. This information is essential for understanding the global distribution of historical conflicts.
3. The generation of word clouds for event titles and summaries has revealed significant recurring themes and terms, such as the prominence of 'Battle.' This analysis provides users with an immediate visual representation of key historical themes.
4. Our dataset's inclusion of images and participant information is critical for enhancing users' understanding of historical events. Images offer visual context, while participant details help identify the key actors involved.

Future directions

As we conclude this first milestone, we recognize potential directions for the future:

- Enhancing the search engine's query capabilities to support a wider range of user queries, including complex historical research questions.
- Expanding the dataset to include additional historical events, thus broadening the engine's coverage and historical context.
- Incorporating advanced natural language processing (NLP) techniques to improve the search engine's understanding of user queries and the relevance of search results.
- Further refining the user interface and interactivity of the search engine to provide a seamless and informative user experience.

In conclusion, this project represents an exploration of historical data retrieval and processing, leveraging Wikidata and Wikipedia to create a valuable resource for history enthusiasts, researchers, and educators.

REFERENCES

[1] Folium. 2023. Retrieved 2023-10-12 from https://python-visualization.github.io/folium/latest/
[2] GDELT. 2023. Retrieved 2023-10-12 from https://www.gdeltproject.org/
[3] Uppsala Universitet. 2023. Retrieved 2023-10-12 from https://ucdp.uu.se/
[4] Wikidata. 2023. Retrieved 2023-10-12 from https://www.wikidata.org/wiki/?variant=zh-tw
[5] Wikidata. 2023. Retrieved 2023-10-07 from https://www.wikidata.org/wiki/Q13418847
[6] Wikipedia. 2023. Retrieved 2023-10-12 from https://www.wikipedia.org/

A ANNEX

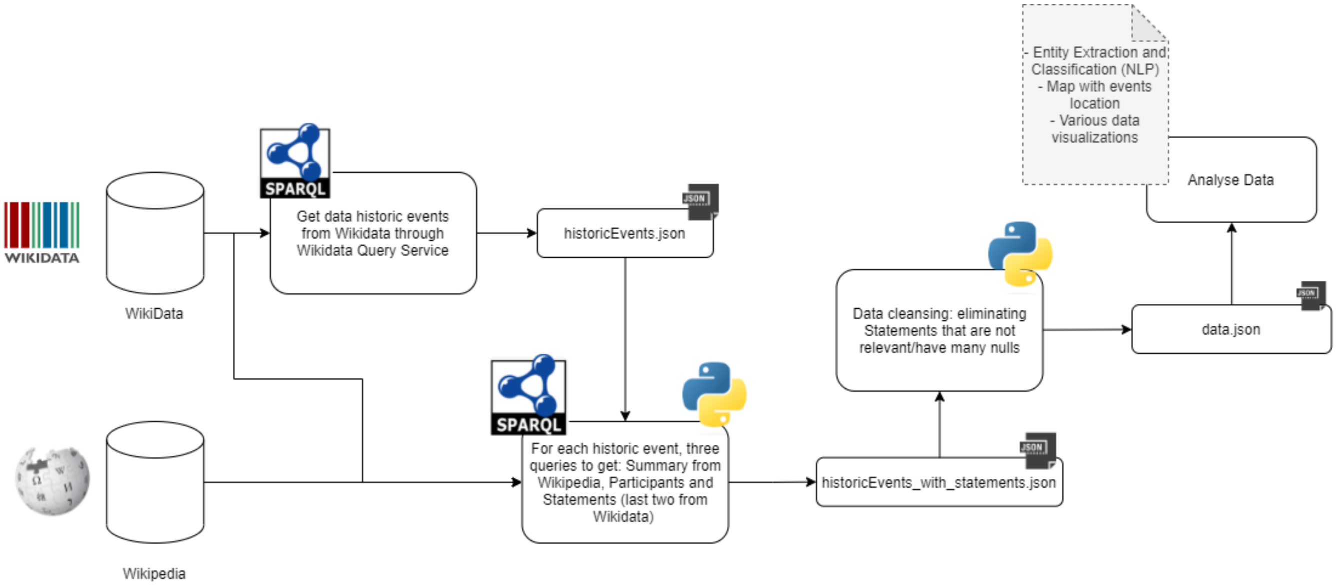
- Entity Extraction and Classification (NLP)
- Map with events location
- Various data visualizations
Analyse Data
WIKIDATA
WikiData
SPARQL
Get data historic events from Wikidata through Wikidata Query Service
JSON
historicEvents.json
Data cleansing: eliminating Statements that are not relevant/have many nulls
data.json
SPARQL
For each historic event, three queries to get: Summary from Wikipedia, Participants and Statements (last two from Wikidata)
historicEvents_with_statements.json
Wikipedia

Figure 5: Data pipeline