When news make history
The Pacific chapter of World War II

Joseph Vavalà
Epfl

Foivos Anagnou
Epfl

Lkham Nyambuu
Epfl

Abstract

The information we have access to defines the way we perceive the world. News media and journals in particular have contributed, since their very inception, to the spread of knowledge, influencing public opinion and setting the foundations for what we call history today.
This project focuses on the Asian-Pacific theatre of World War II and aims at correlating the frequency of reporting on events spanning from 1937 to 1945 and their main entities to their importance using newspapers archives in order to enliven crucial historic events.
The picks of frequency do indeed redirect to the turning points of the conflict that redefined the boundaries of the world.

1 Introduction

World War II has been the conflict that affected the highest number of people in human history and its consequences redefined the boundaries of most of the involved nations. The knowledge of the events that took place between 1939 and 1945, as many others, in most cases is acquired in school and through books that generally follows content that the each nation's education Ministry has previously agreed upon.
It's legitimate to suppose that this filtering can cause a certain degree of bias and make people around the world learn quite different version of the same events. As history books, newspapers and media keep influencing our perception of facts according to their the political leaning and the national policies. Still they provide a view that is different and more direct than books, providing an account of the events as they unfolded. The point of view we observe is the Swiss one, which has a long tradition of neutrality and despite the necessary mobilization managed to keep its independence throughout the conflict, specifically using the articles from the *Journal de Genève* and the *Gazette de Lausanne*.

2 Dataset

2.1 Data Collection

The sources are gently provided by *Le Temps*, that was founded by the merging of the aforementioned newspapers with *Le Nouveau Quotidien* in 1998 and represents today the only nationwide non specialized daily newspaper in Switzerland.

Being the analysis focused on the Pacific scenario, we decided to gather the data covering from 1937 to 1945, because despite being 1939 the year that marks the beginning of World War II, and 1941 the official beginning of the Pacific War according to most historians, the Asian situation was completely shocked by 1937, when the Empire of Japan, one of the main belligerents and already occupying Manchuria since 1931, started the full scale invasion of China, one of the widest theatre of the atrocities that followed. Thus, regarding the data collection, our dataset should include only a subset of the whole dataset stored in the cluster. The total size of this subset was just a few Gigabytes and therefore we could download it and work with it on our host machines without having to use any cluster-computing framework such as *Spark*.

We considered that it would be interesting and important to gather data form another news source as well, which would enable us to make comparisons and get meaningful results from the possible differences. Thus, we decided to access the New York Times archive of all the New York Times issues covering from 1851 to 2002. The TimesMachine dataset is accessed through the accessed through the API provided by the New York Times Developer Network. Initially we requested access

and then being provided an API key, we implemented a script which downloads all the files referring to our scenario. We realized that only a small small part of the articles was available. Unfortunately, after communication with the people in charge of the archive collection, we did not achieve to have access to such an old dataset in a suitable format; only through scanned newspaper pages.

2.2 Dataset Description

The dataset is made of 216 *xml* files; we got one file per month for this 9-year period and the newspaper sources were two.

Analyzing each file we find that the structure is the following: The file includes article objects and each of them entities. Inside an article more than an entity can be found. By observing the entities objects we can assume that it includes enough information to be considered an article instance. The most useful tag fields describing an such an object are: id, issue date, name (title), origin and full text. There are more tags describing each entity but either they do not provide meaningful information or we observed that they are not reliable enough. The

3 Methodology

Next, we describe our method for analyzing the corpus of news reports about conflict-related events among countries involved in the Pacific Theater scenario form *Le Temps* dataset. We proceed into filtering the documents according to our research (Sec. 3.1) and then into doing text analysis in these articles (Sec. 3.2).

3.1 Data filtering

To filter the articles related to our scenario we follow a three step approach. First we get only the articles that refer to regions concerning the countries involved and then we focus only on these documents which include conflict-related mentions in their texts. In the end, we discard the documents that despite seeming related to the chosen scenario contain references to many other external ones, such as those regarding the European scenario and Switzerland, because those include a considerable amount of irrelevant information.

With respect to step 1, we manually created a dictionary of words that includes all names of countries involved in the Pacific Theater as well as the names of cities with a significant importance in this scenario.
The dictionary contains only the ”root” of words, i.e. terms like japon.* and chin.*. This format was the preferred one because we wanted to detect all possible combinations, words such as japonais/e or chinois/e. The step 2, concerns the filtering based on manually built dictionary of conflict-related terms. This second phase was necessary because through the first one it was incredibly easy to retrieve documents not pertaining to our scenario simply because they included one mention to any of the Southeast Asia locations. After a more careful observation on the remaining documents we realized that there are many articles succinctly mentioning a variety of topics (including those related to the analysis scenario and at the same time to the European scenario). We dropped such articles the aim was to process texts with information exclusively on the Asian scenario, we acknowledge that with better filtering and selection methods the information retrieved could have been definitely relevant.

3.2 Information Extraction

After the articles identification step we started to analyze the documents to extract meaningful information out of them.

We consider that identifying the coverage of the newspapers on this topic would give important information concerning how relevant were international news at that time. We decided to group this information per month because an event that appears at a specific time, may have important impacts and be reported effectively during the days following it or in case of meetings and agreements it might be discussed some time in advance.

Value	Quantity
Total number of articles	395926
Untitled articles	119212
Empty articles	11857
Articles with misplaced body	2
Average text length (chars)	3392,97

Table 1: Dataset statistics.

3.3 Most frequent terms

To retrieve the most frequent terms per year and per journal we applied the basic steps of text analysis, that is we first cleaned as much as possi-

ble the input from unnecessary symbols, then tokenized each word in the documents, then applied an additional cleaning phase filtering out terms shorter than 5 characters in order not to include stopwords which could not be in the imported stopword list.

3.4 Entity extraction

In order to understand how the

4 Visualization

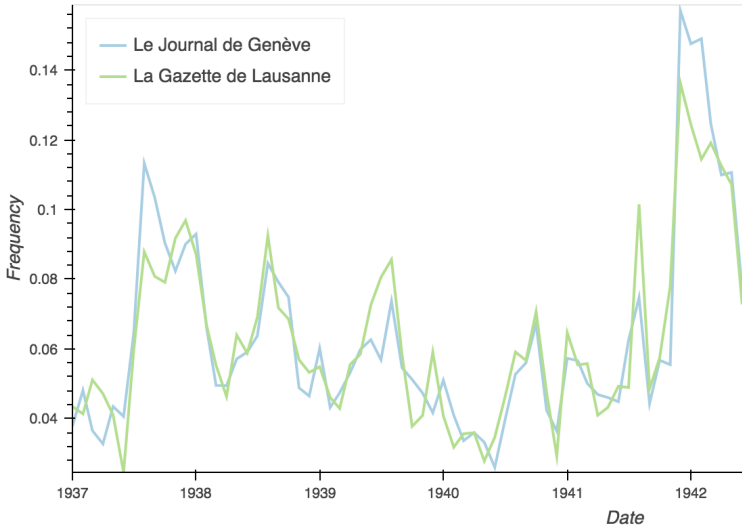
5 Results and Findings

Analyzing the coverage time series for the two newspapers we notice that the spikes occur reasonably in the same timespans and that looking closely the most evident ones happen:

- from the second half of 1937
- at the end of 1941
- between July and September 1945

These findings seem reasonable taking into consideration that those periods are correlated to the Empire of Japan invading China in July 1937, the attack on Pearl Harbor in December 1941 and the closing of the conflict following the atomic bombings in August 1945.

With the final filtering we noticed that the *Gazette de Lausanne* peaks are evidently curbed more than those of *Journal de Genève*, this might indicate that articles related to the scenario are generally more mixed with terms referring to the European events.



6 Conclusion

Gathering meaningful data from text always poses great challenges, especially when the corpus is not in English, the language for which most and the best Natural Language processing tools have been developed, and the digitalization process through OCR techniques is not totally effective, thus leading to the need to clean the data and sometimes the impossibility to retrieve particular entities.

Data analysis applied to newspapers and media could have an enormous impact on society and provide people with the instruments to have a better understanding of the reality they live in, not only because they could have access to a more direct interpretation of the past, but also because information could be collected from thousands of sources and dozens of countries contemporarily and then processed, hence reducing the bias and the noise coming from each of them. Achieving such a goal is way beyond the scope of this project and our means, still it gave us the opportunity to approach and explore the text mining field on non structured data (the thousands articles).