Master Thesis Report

Semantic analysis of web archive historical data

the 1983 "Marche pour l'égalité et contre le racisme"

Erasmus Mundus Joint Master's Degree in Big Data Management and Analytics (BDMA)

Davide Rendina

davide.rendina@student-cs.fr

Advisor:	Mathieu Génois	-	CPT, AMU	mathieu.genois@univ-amu.fr
Advisor:	Sophie Gebeil	-	TELEMMe, AMU	sophie.gebeil@univ-amu.fr
Advisor:	Patrice Bellot	-	LIS, CNRS	patrice.bellot@univ-amu.fr

August 10, 2023

Acknowledgments

This is the end of a journey that started two years ago and that with its ups and downs brought me invaluable lessons and contributed to make me who I am now. I would like to take this opportunity to acknowledge everyone that, with their support, encouragement, and guidance, has been instrumental in helping me navigate the challenges and obstacles along the way.

First, I would like to express my gratitude to my supervisors: to Mathieu Génois for his continuous support and patience throughout the whole duration of my internship; to Sophie Gebeil for her feedback and help in achieving this project and for giving me the opportunity to present my work at the RESAW conference in Marseille; to Patrice Bellot for his invaluable advices and guidance; without them this project would have not been possible. In addition, I would like to thank INA and its director, Claude Mussou, for providing the data used in this project.

I would like to express my heartfelt gratitude to my family for their unwavering support throughout my academic journey. Their understanding and encouragement have allowed me to embark on this enriching experience abroad, which started with my Bachelor's and continued with this Master's. Although I may not always express it, their love and silent presence means the world to me. Thank you for everything. A special thanks to my grandfather, that with his passion he showed interest and did his best to support me in completing my work.

Thanks to my friends, my second family, that although physically far they never felt closer than these past years and never failed to be there for me whenever I needed it. Because as the saying goes "when winter comes the lone wolf dies while the pack survives", and I am incredibly grateful to have such a strong pack of friends.

Finally, thanks to every special person I met in these past two years that enriched me and made this experience unique.

List of Figures

Contents

Abstract: Web archives represent an invaluable source for historians and researchers in digital humanities to study the relation between a society and its past. However, the surplus of digitalised resources makes it difficult to deal with these sources in terms of navigation, exploration and analysis. To effectively address this matter, it is crucial to adopt an interdisciplinary approach that integrates the expertise and perspectives of both historians and computer scientists.
After a careful review of related work, it becomes clear that the assistance provided by Natural Language Processing (NLP) and network analysis tools is significant in this particular context.
In this thesis, we use the data obtained from the web archive of INA, which includes webpages referencing the 1983 " Marche pour l'egalité et contre le racisme ". This data serves as the foundation for implementing a framework that combines Named Entity Recognition (NER), Topic Modeling, and network analysis. The objective is to provide historians with a navigational tool that facilitates their exploration of extensive corpora. Promising results have been achieved, paving the way for future enhancements to the framework.

Keywords: *Web Archive, Data Analysis, Natural Language Processing, Network Analysis, Named Entity Recognition, Topic Modelling, Historical Data*

. . . . . .

Résumé: Les archives web représentent une source inestimable pour les historiens et les chercheurs en humanités numériques qui étudient la relation entre une société et son passé. Cependant, le surplus de ressources numérisées rend difficile le traitement de ces sources en termes de navigation, d'exploration et d'analyse. Pour traiter efficacement cette question, il est crucial d'adopter une approche interdisciplinaire qui intègre l'expertise et les perspectives des historiens et des informaticiens.
Après un examen minutieux des travaux connexes, il apparaît clairement que l'aide apportée par les outils de traitement du langage naturel (NLP) et d'analyse des réseaux est significative dans ce contexte particulier.
Dans cette thèse, nous utilisons les données obtenues à partir des archives web de l'INA, qui comprennent des pages web faisant référence à la " Marche pour l'egalité et contre le racisme " de 1983. Ces données servent de base à la mise en œuvre d'un cadre qui combine la Named Entity Recognition (NER), Topic Modelling et l'analyse des réseaux. L'objectif est de fournir aux historiens un outil de navigation qui facilite leur exploration de vastes corpus. Des résultats prometteurs ont été obtenus, ouvrant la voie à de futures améliorations du cadre.

Mots clés: *Archives Web, analyse de données, traitement du langage naturel, analyse de réseaux, Named Entity Recognition, Topic Modelling, données historiques*

Chapter 1

Introduction

The advent of digital humanities has facilitated the advancement of text and network analysis techniques that are specifically applied to digitally native documents sourced from the contemporary web. For historians of the present time, the living web is not sufficient to study recent phenomena as the classical historian method requires both a stable corpus and ruins, which in the context of the web would consist in old versions of sites, no longer accessible on the living web.

Following the example of the Internet Archive created in the United States in 1996, the Bibliothèque nationale de France (BnF, National Library of France) [5] and the Institut National de l'Audiovisuel (INA, National Audiovisual Institute) have been archiving the national web since 2006 thanks to the web legal deposit law. As new essential material for historians of the current era, the web archives bring together multidisciplinary research communities, computer science researchers in the humanities, and archivists, like the European RESAW collective [32] founded in 2011.

In France, web archives have been used from a historical perspective mainly to study the relations between a society and its past since the end of the 1990s, for example concerning the First World War [4] or memories of discriminated populations such as those of North African immigration [17]. In this context, modern historians confront a surplus of digitized and digitally originated resources. For historians and researchers in digital humanities alike, grappling with these sources in terms of navigation, exploration, and analysis can pose a significant hurdle in their research endeavours [7, 27]. In general, we want to emphasize that addressing the issue requires a strongly interdisciplinary approach, combining the expertise and perspectives of historians and computer scientists.

This research induces a reflection on the creation of innovative methodology associating history and computer science, as it is about inventing new methods and tools to study data from web archives, which are more complex than those from the live web.

1.1 Objective and Motivation

This thesis is part of the Polyvocal Interpretations of Contested Colonial Heritage (PICCH) [30] European project and offers an opportunity to continue this innovative research. Head by Daniela Petrelli (Sheffield University), the PICCH explores how archived material created in a colonial mindset can be re-appropriated and re-interpreted to become a compelling source for decolonization and the basis for a future inclusive society. Led by Sophie Gebeil, the French team is engaged in studying the evolution of the online representations of the memories of the 1983 "March for Equality and against Racism", a landmark event in the history of postcolonial anti-racism.

The main goal of the thesis is to explore the possibilities that computational methods for text analysis offer for the study of large corpora of historical data from the archived

web. In particular, for this thesis, we focus on the way the memory of a historical event is built through the recounting of web media. By utilizing techniques from Natural Language Processing (NLP) and Machine Learning, our aim is to extract semantic attributes, such as Named Entities and topics, from historical texts within the corpora of web pages archived by INA that contain mention of the 1983 "March for Equality and against racism" (later referred to as the '*Marche*'). Additionally, we seek to uncover various discourses surrounding this event. Finally, by conducting a network analysis of these attributes, we aim to explore the interconnectedness and organization of the texts that reference the event.

1.2 Chapter List

Chapter 2 Related Work. This chapter provides an overview and analysis of existing literature and research relevant to the thesis topic.
Chapter 3 Background. Focusing on the background information necessary for understanding the thesis topic, this chapter establishes a foundation of knowledge and provides the necessary background context to comprehend the following chapter.
Chapter 4 Methodology and Approach. This chapter details the specific methodologies, approaches, and techniques utilized in the research. It outlines the steps taken to analyze the large corpora of historical data.
Chapter 5 Analysis and Results. In this chapter, the thesis presents the experimental findings and outcomes. It discusses the results of applying NLP techniques to extract named entities (NE) and topics from the historical text in the web page corpus. Finally, the results of applying network analysis on the NE and topics are reported and discussed.
Chapter 6 Conclusions and Perspectives. This concluding chapter summarizes the key findings of the thesis and presents the main conclusions drawn from the research. Finally, possible future work is suggested.

Chapter 2

Related Work

2.1 Overview

This chapter provides an in-depth review of the use of computational methods for text analysis in investigating extensive collections of historical data obtained from archived web sources. The discussion starts by providing an overview of the challenges that historians and researchers in digital humanities encounter when it comes to navigating, exploring, and analyzing these sources. Afterwards, various methodologies applied for the study of web archive historical data are introduced and evaluated. The objective is to assess their potential implications for this project.

2.2 Literature Review

Contemporary historians encounter an overwhelming abundance of sources that have been digitized or originated in a digital format. As noted by [34], "Surely, the injunction of traditional historians to look at 'everything' cannot survive in a digital era in which 'everything' has survived."

In traditional literary analysis, researchers engage in the laborious task of manually examining a text or collection of texts to identify terms and phrases related to specific topics or themes. While this approach can yield insightful analyses, it becomes tedious or unfeasible when dealing with extensive corpora such as those of the archived web. Moreover, human researchers are prone to unintentional mistakes during their readings, resulting in incomplete analyses.

In addition, navigating and exploring the vast quantity of data present in web archives presents a significant challenge for researchers. The sheer volume of archived web pages can be overwhelming, making it difficult to locate and access relevant materials.

To address these challenges, numerous computational techniques have emerged since the mid-20th century to enable researchers to conduct more extensive investigations into literary texts. Among these techniques is a text analysis method known as topic modelling, which utilizes algorithms to automatically identify recurring themes or topics within a collection of texts. In essence, topic modelling serves as a 'framework' that assists the researcher in conducting subjective analysis as it offers a quick way to get a sense of what might be happening in a large body of text.

In a quantitative study, [22] used Latent Dirichlet Allocation (LDA), a popular topic modelling algorithm, to categorize corpora from a large digital news archive into distinct topics based on the underlying themes they cover. The authors demonstrate the usefulness of LDA by applying it to a dataset comprising thousands of newspaper articles.

The study showed that LDA proves to be an effective tool for swiftly examining trends and patterns within extensive digital news archives, aiding in the analysis of news content. However, the authors acknowledge that domain knowledge and subject expertise are crucial in interpreting and validating the generated topics. Human intervention and domain knowledge can help refine the topics, identify misclassified articles, and ensure the accuracy and meaningfulness of the results.

Similarly, [6] provides an overview of various algorithms, specifically probabilistic topic models such as LDA, that offer solutions for effectively managing large document archives. The application of probabilistic topic models in different scenarios is explored, including document clustering, information retrieval, and recommendation systems. The article discusses the challenges associated with modelling topics, such as the determination of the number of topics and the interpretation of the resulting topics. Finally, the author stresses the importance of teaming computer scientists with other scholars to use topic models to help explore, visualize, and draw hypotheses from their data.

Another valuable computational technique for literary analysis is Named-Entity Recognition (NER). This tool identifies and labels entities, primarily proper nouns, within a text. The entities are identified and assigned tags, such as 'location', 'organization', or 'person'. With fully tagged entities, the researcher can leverage their structured data to efficiently carry out literary analysis.

In this context, [24] presents a study focused on the application of NER to the Trove collection in the National Library of Australia which holds a large collection of digitised newspapers dating back to 1803. The primary objective is to provide valuable data to researchers in the Humanities field who utilize the HuNI Virtual Laboratory. The project utilized a standard Named Entity Recognition (NER) model to extract 27 million person name mentions from 17 million articles in the archive. The study demonstrates how this enriched data can be utilized, including the identification of individuals through clustering techniques and the presentation of the dataset as linked data over the web.

A research conducted by [1] aimed to explore the 52,000 pages of the British Journal of Psychiatry (BJP) and other historical psychiatric journals. The project combined different Natural Language Processing (NLP) tools, such as topic modelling and NER, with new visual interfaces to foster the exploration of the corpus. One of the main findings is that the combination of NLP tools and visual interfaces can help historians to explore large-scale corpora and discover new insights for historical research. In terms of limitations, the author notes that the interpretation of the page content remains the historian's task and cannot be replaced by the machine. Algorithms and visualisations in Histograph remain only a tool for facilitating research by pointing out potentially interesting instances to the historian. Additionally, the paper acknowledges that the dataset used in the study is limited to historical psychiatric journals and that further research is needed to explore the potential of these tools in other domains.

In a similar study, [9] presents an analysis of how the European press covered Covid-19 vaccination news in 2020-2021. The study collected a dataset of over 50,000 online articles published by 19 newspapers from five European countries over 22 months. The authors combined NER, topic modelling and sentiment analysis to identify main actors, subtopics,

and tone, and to compare trends across countries.

From a different perspective, [4] aimed to map the network of websites dedicated to the Great War and analyze the links between them using web archives from the Bibliothèque nationale de France. The study showed how network analysis can be a useful tool for exploring and navigating large corpora of web archives. By mapping the links between websites, network analysis can reveal patterns of connectivity and identify clusters of related content. This can help researchers to identify key actors, themes, and trends in the corpus, and to explore the evolution of the network over time.

2.3 Conclusions

The proliferation of digital sources and the vast quantity of data available in web archives pose significant challenges for contemporary historians. Traditional manual approaches to text analysis are laborious and prone to incomplete analyses. To overcome these challenges, computational techniques such as topic modelling, NER and network analysis have emerged as valuable tools for literary analysis. In particular, topic modelling proved to be an effective method for analysing trends and patterns within vast documents in digital archives, while, NER has been successfully used to identify main actors and facilitate the navigation of large corpora of texts. Furthermore, network analysis has been applied to map the relationships and connections between web pages, uncovering patterns of connectivity and identifying clusters of related content.

Employing NER and topic modelling in conjunction with network analysis, this projects aim to gain a comprehensive understanding of what are the main actors and topics involved when mentioning the *Marche* within the corpora extracted from INA's web archive and how the different texts associated with the event are organised and related based on the topics and Named Entities extracted. To address the limitations associated with topic labelling and named entity evaluation, the project will involve domain knowledge experts who will actively participate in the process. Their expertise will be instrumental in refining the topic labels, identifying misclassified entities, and ensuring the accuracy and meaningfulness of the results. By having domain knowledge experts in the project, the aim is to overcome these limitations and enhance the overall quality and reliability of the analysis.

Chapter 3

Background

3.1 Overview

This chapter provides an overview of the algorithms and methods needed to understand the following chapter. First, the different methods are introduced and explained. Afterwards, the choice of the models for the different methodologies is motivated and explained.

3.2 Named Entity Recognition

As defined by [13], NER is a sequence labelling task in which a system aims to assign labels (Named Entity classes) to a sequence of tokens. The system's objective is to learn identification and classification patterns from a set of labelled examples, observing the word-label correspondences and their distinguishing features. This learning process enables the system to infer labels for new, unseen sequences of tokens.

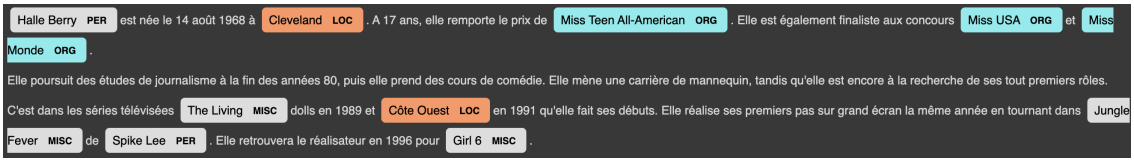


Figure 3.1: Named Entities extracted from text

As depicted in Figure 3.1, a model typically assigns each sequence of tokens to predefined semantic categories, such as Person (PER), Location (LOC), Organization (ORG) and Miscellaneous (MISC). These categories serve as the basis for labelling and organizing the identified entities within the sequence of tokens. NER is widely used in many downstream tasks, such as question answering, machine translation, information retrieval, text summarization, text understanding and entity linking, among others [36].

3.2.1 NER methods

Similarly to other NLP tasks, NER systems are developed based on different algorithmic approaches, namely: dictionary-based, rule-based, traditional machine learning methods, and deep learning methods. These four families of algorithms provide different techniques and frameworks for building NER systems, each with its own strengths and characteristics.

3.2.1.1 Dictionary-based

This approach utilizes a dictionary that contains a collection of predefined terms. Simple string-matching algorithms are employed to check if entities in the text match any entries

in the dictionary. However, this method has limitations as the dictionary needs to be continuously updated and maintained. In addition, this approach faces challenges with word ambiguities or polysemy, where a single word may have multiple meanings depending on the context, requiring more advanced techniques like contextual analysis and word sense disambiguation algorithms to accurately identify the intended entities in the text.

3.2.1.2 Rule-based

In this approach, a set of predefined rules is used for information extraction. Two types of rules are commonly employed: pattern-based rules, which rely on morphological patterns of words, and context-based rules, which depend on the context in which words appear in the text. For example, a context-based rule could be "If a person's title is followed by a proper noun, then that proper noun is the name of a person." These systems offer the benefits of not needing training data and being readily interpretable. However, they require significant time and expertise for their design and implementation [13].

3.2.1.3 Machine learning-based

These systems utilize statistical models to detect entity names. Machine learning models create feature-based representations of observed data, which helps in recognizing existing entity names even with slight spelling variations. Feature engineering could involve extracting features such as word embeddings, part-of-speech tags, syntactic dependencies, capitalization patterns, or contextual information surrounding the target words. These engineered features are then fed into the machine-learning model as inputs.

ML-based solutions for NER typically involve two phases, which are illustrated in Figure 3.2.

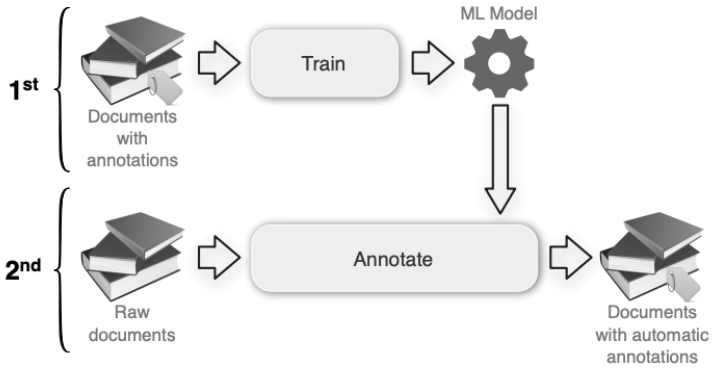


Figure 3.2: Development process of ML-based NER [8]

First, the model is trained on annotated documents, with the training time varying based on the complexity of the model. In the second phase, the trained model can be used to annotate raw documents (Inférence).

3.2.1.4 Deep learning methods

In recent years, the field of Natural Language Processing (NLP) has witnessed a remarkable surge in the popularity and effectiveness of deep learning-based models, particularly in the context of developing advanced Named Entity Recognition (NER) systems. These models have demonstrated significant advantages over traditional approaches, revolutionizing the way NER tasks are tackled.

One of the primary strengths of deep learning-based NER systems lies in their ability to transform input data into a non-linear representation. This transformation enables the models to capture and understand intricate relationships and dependencies within the data, which may not be apparent in the original feature space. By doing so, deep learning models can discern complex patterns and associations, leading to more accurate and contextually aware entity recognition. Another compelling aspect of deep learning models is their capacity to automatically learn meaningful representations from raw data. Unlike earlier approaches that heavily relied on manual feature engineering, deep learning algorithms excel at automatically extracting high-level features and representations from the data.

However, it's worth noting that deep learning-based NER systems also come with their own set of challenges. They typically require a considerable amount of labelled training data to achieve optimal performance and training these models can be computationally intensive, necessitating access to powerful hardware or cloud-based resources.

3.2.2 Deep Learning models

Among the aforementioned deep learning methods, following the advent of transformers models such as BERT (Bidirectional Encoder Representations from Transformers), a multitude of pre-trained models have been developed specifically for NER tasks. Pre-trained models refer to models that are trained on large-scale datasets, often using unsupervised learning techniques, to learn general language representations and capture linguistic patterns. These models, like BERT, are trained on vast amounts of text data and gain a deep understanding of language, making them highly effective at capturing context and semantics. These pre-trained models leverage the power of transformer architectures to capture contextual information and semantic relationships between words in a text. The issue with these models is the scarcity of large, high-quality multilingual training datasets, based on NER because deep learning models require large amounts of data to be trained effectively. To overcome this issue, [36] developed Babelscape, a multilingual pre-trained model which was trained on a dataset generated using WikiNEuRal for NER in 9 languages (German, English, Spanish, French, Italian, Dutch, Polish, Portuguese, Russian).

3.2.2.1 Babelscape's WikiNEuRal dataset

The WikiNEuRal dataset is created by combining Wikipedia and BabelNet. Specifically, the authors exploit the hyperlinked texts of Wikipedia articles to create a large-scale multilingual corpus of sentences, which are then automatically annotated with named entity tags using BabelNet. BabelNet is a multilingual lexical-semantic network that integrates information from various lexical resources, such as WordNet, Wikipedia, and

Wiktionary, among others [28]. The authors use BabelNet to automatically type tags by leveraging its rich semantic information, which includes synsets, definitions, translations, and semantic relations. They also use BERT-based models to discern entities from non-entity tags and a domain adaptation technique to produce NER training data for arbitrary domains. The process of creating the dataset is illustrated in Figure 3.3.

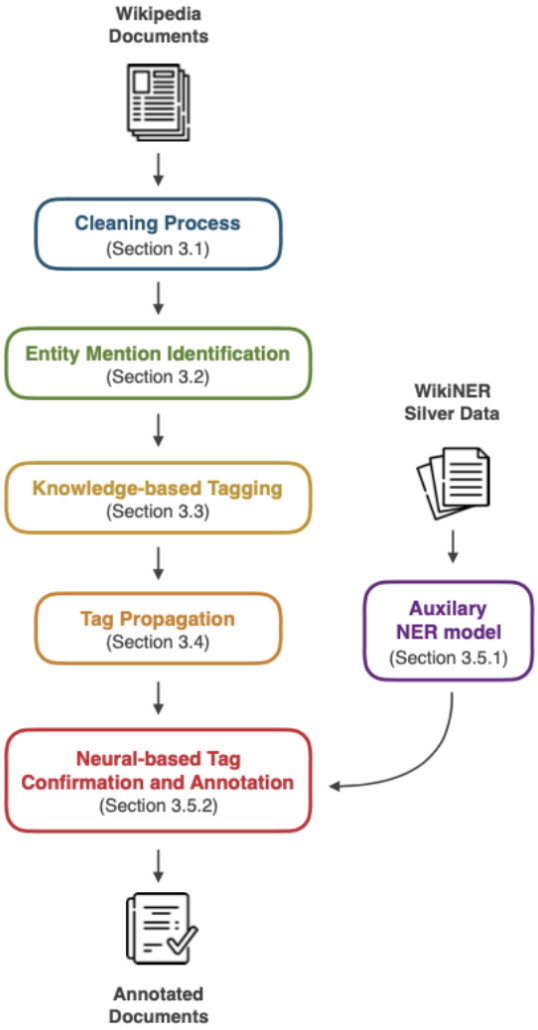


Figure 3.3: WikiNEuRal annotation pipeline [36]

The WikiNEuRal dataset is created through the following processes:

1. **Pre-processing Wikipedia articles**: The text of Wikipedia articles is cleaned up by removing sections with the 10 most frequent titles, as well as other elements that introduce noise, such as lists, tables, templates, formulas, etc.

2. **Identifying entity mentions**: The remaining elements are Wikilinks, which provide potential entity mentions. The authors identify these mentions by leveraging the hyperlinked structure of Wikipedia articles.

3. **Tagging named entity links through synsets**: The authors use BabelNet, a multilingual lexical-semantic network, to automatically annotate the identified entity mentions with named entity tags. They do this by leveraging the rich semantic information in BabelNet, which includes synsets, definitions, translations, and semantic relations.

4. **Named entity tag propagation**: The authors propagate named entity tags to other mentions of the same entity in the same sentence, using heuristics based on the context of the sentence.

5. **Confirming and augmenting annotations**: The authors use a complementary approach to confirm and augment sentences with entity tags. They put forward a domain adaptation technique which can produce NER training data for arbitrary domains. They also use BERT-based models to discern entities from non-entity tags.

Through these processes, the authors are able to create a large-scale multilingual corpus of sentences that are automatically annotated with named entity tags. In particular, the dataset includes 4 named entity semantic categories: Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MISC). As shown in their report, models trained on the WikiNEuRal dataset achieved better results in comparison to those obtained by the two competitors (WikiANN and WikiNER) on different test sets (i.e., WikiGold, OntoNotes, and BSNLP).

Overall, when considering the various techniques for NER, including dictionary-based, rule-based, machine learning, and deep learning, the choice of a deep learning model holds significant advantages. These models, by leveraging the power of transformers are able to capture semantic relationships and contextual information in texts without the need for manual feature engineering require in more traditional approaches. In particular, the Babelscape pre-trained model was chosen for its performance, multilingual support and easiness of integration.

3.3 Topic Modelling

Topic modelling is an unsupervised method used to extract clusters of words from sets of documents. Unlike human readers who consider grammar and syntax, topic modelling algorithms view a corpus as mere 'buckets of words' disregarding sentences and paragraphs. Consequently, the topics generated by such algorithms appear as lists of seemingly related words without any discernible order, as follows:

- What human topic looks like → `Cinema`
- What topic modeling returns as a topic → `film_actor_release_theater`

This approach is rooted in the semantic theory that words derive meaning through their relations to other words, prioritizing relationality over syntax. The Swiss linguist Ferdinand de Saussure's concept of the relational model of linguistics supports this notion, emphasizing that words only hold meaning in relation to one another [35]. Thus,

the analysis of meaningful conversations relies on examining clusters of words and their relationships. Although individual words lack inherent meaning to topic modelling algorithms, meaningful results can be achieved by quantitatively grouping words into topics. Researchers can then subjectively derive meaning from these groupings and assign labels that reflect the most prevalent words within each topic.

3.3.1 Challenges

The main challenge of topic modelling is the subjective interpretation and validation of the generated topics. While topic modelling algorithms can automatically identify recurring themes or topics within a collection of texts, the resulting topics may not always align with the researcher's expectations or understanding of the data. It requires domain knowledge and subject expertise to accurately interpret and label the topics generated by the algorithm. Human intervention is crucial to validate the topics and ensure the accuracy and meaningfulness of the results.

Nevertheless, metrics have been developed to have an indication of the quality of the topics extracted. Among these, topic coherence is a metric that measures the degree of semantic similarity between high-scoring words in a topic [33]. It is used to evaluate the quality of the topics generated by a topic modelling algorithm. The coherence of a topic is calculated by comparing the co-occurrence statistics of the words in the topic against a reference corpus. The idea is that if the words in a topic tend to co-occur more frequently in the reference corpus than by chance, then the topic is considered to be more coherent. By using topic coherence, researchers can get an indication of how well the algorithm has identified recurring themes or topics within a collection of texts [19].

3.3.2 Methods

Various algorithms and methods have been developed to tackle the topic modelling task, offering different approaches and capabilities. These algorithms can be broadly categorized into two main categories: traditional topic modelling methods and deep learning-based methods.

3.3.2.1 Traditional topic modelling methods

Traditional approaches include well-known algorithms such as Latent Dirichlet Allocation (LDA), which leverages statistical techniques to extract topics from text data. LDA is a probabilistic model that uses Bayesian inference to estimate each topic's distribution and word distributions. It represents each document as a mixture of topics, where each topic is a probability distribution over words. It assumes that each word in a document is generated by one of the topics, and the goal is to infer the underlying topic structure that generated the observed documents [26]. The result of this approach, as shown in Figure 3.4, is a set of topic distributions for each document d in the dataset. These topic distributions represent the likelihood of each topic being present in a particular document. Each document is represented as a mixture of topics, where the proportions of each topic indicate the importance or relevance of that topic within the document. In addition to the document-topic distributions, LDA also provides the word distributions for each topic

k. These word distributions represent the probability of each word occurring within a specific topic. By examining the topic-word distributions, one can gain insights into the most representative words for each topic.

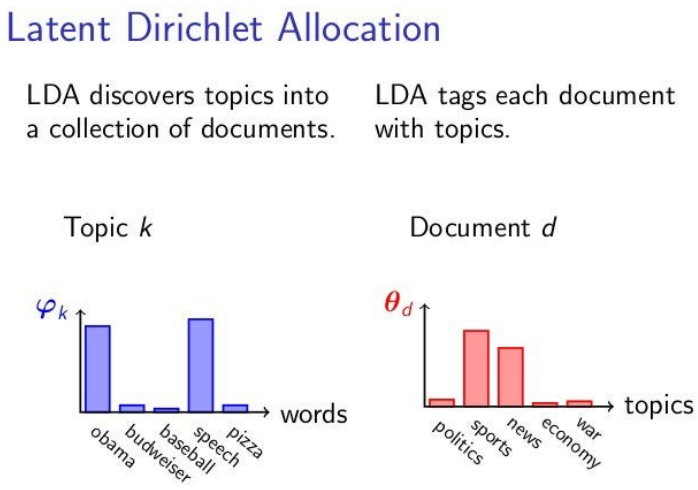


Figure 3.4: LDA topic and document distributions example [3]

The main advantage of this approach is that it is a probabilistic framework, providing a solid statistical foundation for topic modelling. In addition, this is a soft clustering method, which means that each document can belong to multiple topics with different probabilities, which is often the case in heterogenous texts such as blogs or forums. However, as stressed by [12], there are several limitations of statistical approaches such as LDA. Firstly, the model requires careful tuning of hyperparameters, such as the number of topics, which needs to be defined in advance. Secondly, LDA is sensitive to the quality and quantity of the input data, and it may not perform well on noisy or sparse datasets. Finally, LDA does not take into account the context or meaning of words, which can limit its ability to capture the semantic structure of the text

3.3.2.2 Deep Learning-based methods

Deep learning-based methods, such as BERTopic and Top2Vec, harness the power of neural networks to learn the underlying structure of the text data, which can overcome some of the limitations of statistical approaches like LDA.

One characteristic of deep learning approaches is that they can capture the semantic relationships between words and phrases, which can help to identify more meaningful and coherent topics. For example, BERTopic uses a pre-trained transformer model to encode the text data into high-dimensional vectors, which are then clustered using a graph-based algorithm to identify the topics. This approach can capture the contextual meaning of words and can identify topics that are more closely related to the underlying themes of the text. In addition, by leveraging transformers, deep learning models have the advantage of not requiring extensive text preprocessing tasks such as stopwords removal and stemming.

This is because they are capable of automatically learning and extracting meaningful features directly from raw text data.

Another advantage of deep learning approaches is that they have the ability to automatically detect the number of topics within a given corpus, which often is difficult to know a priori. This is achieved through techniques such as variational inference or latent Dirichlet allocation (LDA), where the model learns to assign probabilities to different topics for each document. By iteratively optimizing these probabilities, the model can discover the optimal number of topics that best represent the content of the corpus.

3.3.2.3 BERTopic

BERTopic is a neural topic modelling tool that leverages transformers and a class-based TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions [19]. BERTopic follows a default workflow, which is illustrated in Figure 3.5, that involves executing sentence transformers, UMAP, HDBScan, and c-TF-IDF in a sequential manner.

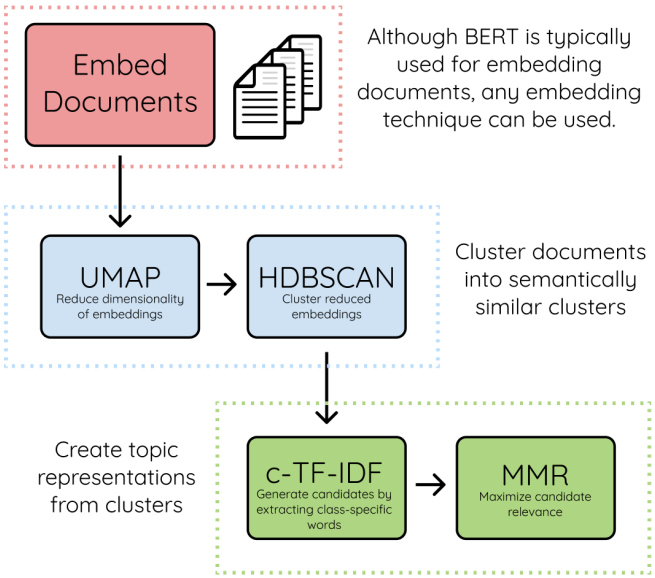


Figure 3.5: BERTopic steps [19]

BERTopic algorithm has five independent steps:

1. **Embed Documents**: BERTopic uses pre-trained language models to embed documents into dense vector representations that can be compared semantically. This step is performed using the Sentence-BERT (SBERT) framework, which is known for achieving state-of-the-art performance on various sentence embedding tasks.

2. **Dimensionality Reduction**: Before clustering the embeddings, the dimensionality of the resulting embeddings is reduced to optimize the clustering process. This is done using a technique called UMAP (Uniform Manifold Approximation & Projection), which is a non-linear dimensionality reduction algorithm that preserves the global structure of the data.

3. **Cluster Documents**: The embeddings are clustered using Hierarchical Density-Based Spatial Clustering (HDBScan), a density-based clustering algorithm that can automatically determine the number of clusters in the data. This step groups similar documents together based on their semantic content.

4. **Topic Representation**: From the clusters of documents, topic representations are extracted using a custom class-based variation of TF-IDF, a metric used to indicate the significance of a word in a document, which is modified to capture the relevance of a term to a topic instead. This step involves calculating the importance of each word in each cluster and using these weights to represent the topics. The c-TF-IDF formula is as follows:

$$W_{t,c} = \mathrm{tf}_{t,c} \cdot \log\left(1 + \frac{A}{\mathrm{tf}_t}\right) \tag{3.1}$$

   where $\mathrm{tf}_{t,c}$ is the frequency of term t in cluster c, and A is the average number of words per cluster. The term frequency models the frequency of term t in class c, which is the collection of documents concatenated into a single document for each cluster. The inverse class frequency measures how much information a term provides to a class and is calculated by taking the logarithm of the average number of words per class A divided by the frequency of term tf_t across all classes. To output only positive values, a one is added to the division within the logarithm.

5. **Fine-tuning Topics (Optional)**: The topics can be fine-tuned by adjusting the number of topics or by merging or splitting topics. This step allows users to refine the topic representations to better fit their specific needs.

Nonetheless, BERTopic is designed to be modular, allowing for some level of independence between these steps. For instance, different pre-trained language models for document embedding, such as BERT, RoBERTa, or DistilBERT, can be chosen depending on the specific task or dataset.

3.3.2.4 Top2Vec

Similarly to BERTopic, Top2Vec is a deep learning model that is capable of detecting automatically topics from the text by using pre-trained word vectors and creating meaningful embedded topics, documents and word vectors [2]. Top2Vec follows a similar approach to BERTopic in terms of embedding documents, reducing dimensionality, and clustering documents. However, it uses a different approach to generate topic vectors and extract topics, which relies on the cluster's centroid and not on c-TF-IDF. The steps are as follows:

1. **Embed Documents**: Top2Vec, similarly to BERTopic, embeds documents into dense vectors, which aims to ensure that similar documents are in closer proximity to each other in the embedding space, while dissimilar documents are positioned farther apart. This is achieved by default with Doc2Vec, however, sentence transformers such as SBERT can also be used.

2. **Dimensionality Reduction**: Similarly to BERTopic, UMAP is used to reduce the dimensionality of the resulting embeddings to optimize the clustering process.

3. **Cluster Documents**: As in BERTopic, to identify dense regions of similar documents, the density-based clustering method called HDBScan is employed. HDBScan categorizes each document as either noise, if it does not belong to a dense cluster, or assigns a label if it is part of a dense area.

4. **Topic Representation**: in Top2Vec, the topic representation step involves finding the nearest word vectors to the topic vector. This means that for each topic, as shown in Figure 3.6, we identify the words that are most closely related to that topic based on their vector representations.

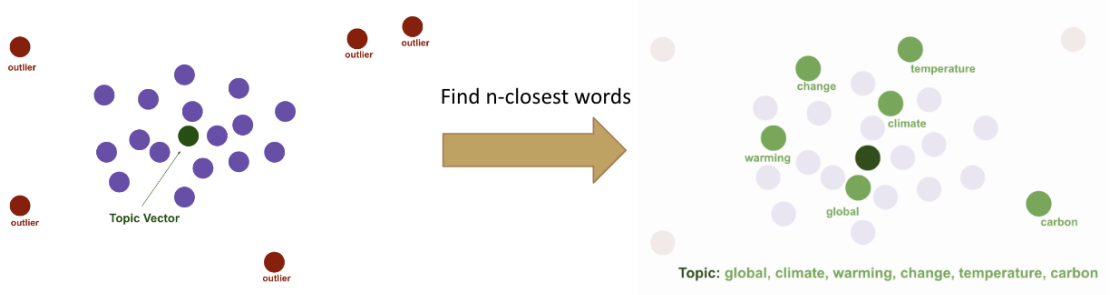


Figure 3.6: Top2Vec topic representation [2]

These words are then used to represent the topic as a weighted set of words, where the weights correspond to the importance of each word in the topic. This differs from the BERTopic approach in topic representation: while BERTopic emphasizes the discriminative power of words within topics based on their rarity in the corpus, Top2Vec focuses on semantic similarity between words and topics.

All in all, both BERTopic and Top2Vec share several similarities, such as automatically determining the number of topics, minimal reliance on pre-processing, utilizing UMAP for dimensionality reduction of document embeddings, and employing HDBScan for modelling these reduced embeddings. However, they differ significantly in how they assign topics to documents.

Top2Vec generates topic representations by identifying words that are in close proximity to a cluster's centroid. In contrast, BERTopic does not consider the cluster's centroid. Instead, it treats all documents within a cluster as a single entity and extracts topic representations using a class-based adaptation of TF-IDF. Ultimately, the choice between Top2Vec and BERTopic depends on the specific needs and the characteristics of the data and task at hand. As mentioned by [12], it may be beneficial to experiment with both methods and evaluate their performance based on your specific criteria, such as topic coherence and interpretability.

To sum up, the choice between traditional topic modelling approaches and deep learning-based methods depends on the specific requirements of the tasks at hand, as each of the algorithms comes with its strengths and weaknesses.

Traditional algorithms such as LDA have been widely used and studied in the field of topic modelling [26, 12, 23] and provide a well-established framework for estimating topic distributions and word distributions. In addition, they can identify mixed membership topics, where each document can belong to multiple topics with different probabilities. This is useful for analyzing complex or heterogeneous datasets, where each text may be represented by more than one topic. However, traditional approaches like LDA rely on statistical patterns and do not capture the semantic relationships between words and documents, plus, hyperparameters, such as the number of topics, need to be tuned carefully.

On the other hand, deep learning methods have been gaining ground as a valid alternative to traditional approaches due to their ability to capture the semantic relationships between words and documents, which allows them to better understand the contextual meaning of words and capture more nuanced and coherent topics. Due to this and to their ability to automatically find the number of topics and the absence of necessity for text preprocessing tasks such as removing stopwords and stemming, deep learning methods are often preferred to traditional topic modelling approaches such as LDA [12] and in particular for this project, Top2Vec and BERTopic will be tested.

3.4 Network Analysis

In NER and topic modelling, another valuable approach in text analysis is network analysis. This encompasses various tasks aimed at extracting meaningful insights and understanding the structure, dynamics, and relationships within complex networks. Some common network analysis tasks include Community detection, centrality measures, and influence maximization.

In a study conducted by [18], the authors provide a new perspective on identifying topical structures in text corpora by relating it to the problem of finding communities in complex networks. Community detection is used to detect topics by representing the word-document matrix as a bipartite network. In this network, the problem of inferring topics becomes a problem of inferring communities.

In our case, the nodes are either documents, topics, or entities. The links in our networks are established based on co-occurrences, indicating that if two nodes (documents, topics, or entities) share a significant number of common words or entities, a link is created between them. Finally, network analysis tools will be used after NER and Topic Modelling to provide another layer of information.

Overall, this approach opens up new possibilities for understanding and analyzing text corpora, including incorporating network analysis algorithms such as community detection.

3.4.1 Community detection

Community detection, also known as graph clustering, is the process of identifying groups of nodes in a network that are more densely connected to each other than to the rest of the network. These groups are called communities or clusters, and they represent subsets of nodes that are more similar to each other in terms of their connectivity patterns than to nodes in other communities. The goal of community detection is to reveal the underlying structure of a network and to provide insights into the organization and function of complex

systems, such as social networks, biological networks, and technological networks [15]. The Louvain algorithm was chosen for community detection due to its effectiveness in identifying communities in large-scale networks. It is known for its scalability and ability to handle networks with millions of nodes and edges [15]. We also chose the Girvan-Newman algorithm [10], to eventually verify the robustness of the communities detected by the Louvain algorithm.

3.4.1.1 Louvain algorithm

The Louvain algorithm, also known as Louvain modularity optimization, is a popular and widely used algorithm for community detection in graphs.

The Louvain algorithm is a hierarchical and iterative algorithm that optimizes the modularity of a network. Modularity is a quality function that measures the strength of the division of a network into communities. The Louvain algorithm starts with each node in its own community and iteratively merges communities to maximize modularity. This process continues until no further improvement in modularity can be achieved.

This algorithm is known for its efficiency and scalability, making it suitable for analyzing large-scale networks. It has been applied to various real-world networks, including social networks, biological networks, and technological networks, and has demonstrated good performance in detecting communities [15].

3.4.1.2 Girvan-Newman algorithm

The Girvan-Newman algorithm is a hierarchical algorithm that iteratively removes edges with high betweenness centrality to identify communities in a network. Betweenness centrality measures the extent to which a node or an edge lies on the shortest paths between pairs of other nodes in the network. By removing edges with high betweenness centrality, the algorithm gradually breaks down the network into smaller components, which correspond to communities.

The Girvan-Newman algorithm does not explicitly define the number of communities in advance but instead generates a hierarchical structure of communities. At each step, the algorithm calculates the edge betweenness centrality for all edges and removes the edge with the highest betweenness centrality. This process is repeated until all edges are removed, resulting in a hierarchical structure of communities.

3.5 Conclusions

This chapter provided the foundation needed to understand the methodology and approach implemented in this project. Looking at the different algorithms and models, by comparing their advantages and disadvantages the following were chosen to be implemented:

- **Named Entity Recognition**: Babelscape pre-trained model [36] was chosen for its performance and multilingual support.
- **Topic Modelling**: BERTopic and Top2Vec were selected over more traditional methods for their ability to capture semantic relationships in texts and the absence of

necessity for text pre-processing. These two models will be compared and evaluated based on their individual features and topic coherence score.

- **Network Analysis**: Community detection algorithms namely Louvain and Girvan-Newman will be used to discover how the different texts are organised in terms of topics and entities, identifying clusters and outliers.

It is also important to acknowledge the challenges discussed in this chapter, which lie in the need for human evaluation and often domain knowledge intervention to assess the quality of the results.

Chapter 4

Methodology and Approach

4.1 Overview

This chapter provides an overview of the methodology used in this project. First, a summary of the pipeline is provided with a mention of the different components. Afterwards, each component of the pipeline is explained in detail with reference to notations introduced in Chapter 3, with a particular focus on how the different choices were made to fit our approach.

4.2 Pipeline

As shown in Figure 4.1, after extracting the relevant data from the `JSON` files containing information regarding the data scraped from each web page, a sample of which can be seen in Figure A.1 in Appendix, the workflow is divided into two parallel streams:

1. **Topic Extraction**: In this step, the topics are extracted for each text, and instead of using a pre-defined closed label list, the labeling process is performed by domain knowledge experts. The experts assign relevant labels to each text based on their expertise and understanding of the content, allowing for a more flexible and contextually accurate labeling approach.

2. **Named Entities Extraction**: In this step, NE are extracted for each text. Domain knowledge intervention is then required to evaluate the quality and relevance of the extracted NE and to normalise and filter them.

The final lists of NE and Topics extracted are then added to the final dataset, which will be used to analyse the different results and possibly answer relevant questions from a historical perspective. Finally, this dataset will also be used to perform the network analysis, which will offer a final layer of analysis that will allow to analyse how the corpora of texts extracted that mentioned the event are related and organised based on the topics and NE extracted.

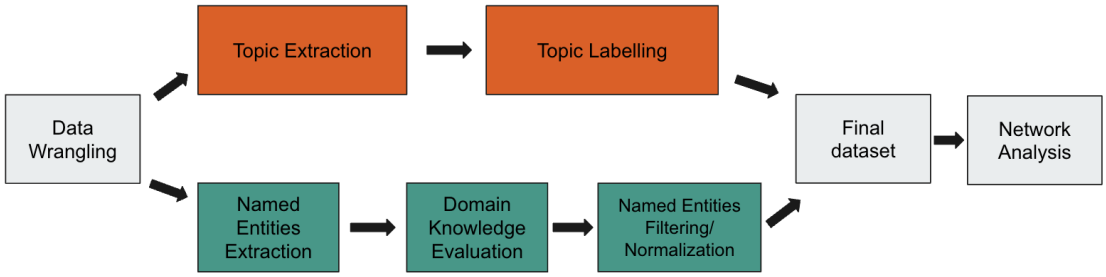


Figure 4.1: Project pipeline

4.3 Topic Modelling

Topics are extracted from the texts, and the models discussed in Chapter 3 are first deployed and compared on the basis of the coherence score and the features offered by the corresponding libraries. Both BERTopic and Top2Vec were deployed on the data presented in the next chapter with the following configurations:

- **Embedding**: a sentence BERT [31] was used to embed the texts. This is a sentence-transformers model that maps sentences and paragraphs to a 512-dimensional dense vector space.
- **Dimensionality Reduction**: UMAP was used for performing dimensionality reduction and embedding data points into a lower-dimensional space with the following configuration:

```
config = {
    'n_neighbors': 15,
    'n_components': 10,
    'min_dist': 0.0,
    'metric': 'cosine',
    'random_state': 100
}
```

  Which are defined as follows:

  - `n_neighbors`: Number of nearest neighbors to consider when constructing the fuzzy topological representation. In this case, it is set to 15.
  - `n_components`: The number of dimensions in the lower-dimensional space. The data points will be embedded into this space. Here, it is set to 10.
  - `min_dist`: The minimum distance between embedded points. A lower value results in a more compact embedding. In this case, it is set to 0.0, meaning no minimum distance constraint
  - `metric`: The distance metric used to measure distances between data points. Here, the 'cosine' distance metric is used.
  - `random_state`: Seed for random number generation to ensure reproducibility. It is set to 100.

- **Clustering algorithm**: For clustering HDBScan was used with following configuration:

```
config = HDBSCAN(
    min_cluster_size=30,
    metric='euclidean',
    prediction_data=True
)
```

  Which are defined as follows:

  - `min_cluster_size`: The minimum number of data points required to form a cluster. Clusters with fewer points than this will be considered noise. In this case, it is set to 30.
  - `metric`: The distance metric used for clustering. Here, the 'euclidean' distance metric is used.
  - `prediction_data`: A flag indicating whether to return the data used for clustering during prediction. It is set to True.

The topic representation, as detailed in the previous chapter, differs between the two models. Since both models utilize a transformer (i.e. BERT), there is no requirement for text pre-processing, and the raw text is directly input into the models. In addition, an important consideration to make is that the average length of the texts is approximately 2,000 tokens with some texts exceeding 10,000. Given that the model is based on BERT [11], we had to acknowledge the limitation of 512 tokens. Given that the model is based on BERT, the limitation of 512 tokens stems from the architecture's design. BERT, like other transformer-based models, relies on self-attention mechanisms to process the input text. However, the self-attention mechanism has a quadratic time and space complexity with respect to the input sequence length. As a result, processing very long sequences becomes computationally expensive and memory-intensive. To strike a balance between computational efficiency and the ability to handle long documents, BERT was originally designed with a maximum sequence length of 512 tokens. Longer texts would need to be truncated or split into smaller segments to fit within this limit. This constraint ensures that BERT remains feasible for practical use without excessively sacrificing performance on shorter texts.

To overcome this, different approaches have been used such as selecting a part of the original text instance (i.e. first n tokens) or dividing the text into chunks that fit a model with a 'standard' 512 limit of tokens per instance, deploying the model on each part separately, join the resulting vector representation [14].

In this case, given the high heterogeneity of the data selecting only a part of the text would not be representative of the whole. Furthermore, we assume that texts may be assigned more than one topic, as this often shifts significantly in forums or blogs. Therefore, we decided to split each texts into chunks of 512 token length and feed this to the models. The result is a list of n topics assigned to each text, one for each chunk.

Comparing the coherence score, as suggested by [33], achieved by the two models, BERTopic obtained a coherence score of 0.73, on the other hand, Top2Vec achieved a

score of 0.67. While this could be an indication that the topics generated by BERTopic are more semantically coherent or interpretable than those generated by Top2Vec, it is important to note that coherence scores alone do not provide a comprehensive evaluation of topic modelling models. Ultimately, BERTopic was selected for extracting the topics for its flexibility in the model construction, with different embedding models compatible including transformers, and the ability to perform dynamic topic modelling [19].

4.3.1 Topic Labelling

The final topics extracted by the selected model, were then given to domain knowledge experts for evaluation and labelling. As shown in the Figure in Appendix C.2 the topics were labelled by two different people and, in addition, ChatGPT was asked to label the set of words representing each topic, the prompt used can be found in Appendix C.2. In doing so, we aimed to combine the domain knowledge of the event, which would be useful for interpreting the topics where words refer to historical events or people, with the power and efficiency of AI. Finally, a final label was assigned to each topic based on the 3 different labels which will be used for the subsequent analysis.

4.4 Named Entity extraction

Named Entities (NE) are extracted using the Babelscape pre-trained model presented in Chapter 3. Since the model leverages a transformer, no text pre-processing is needed and the raw texts are directly fed into the model. In addition, as acknowledged in the topic labelling, we decided to split each text into chunks of 512 token lengths and feed this to the model that will extract a list of NE for each chunk, which will finally be merged into a single list of NE for each text.

4.4.1 Domain knowledge evaluation

After extracting the NE, there are two main issues due to the nature of the data and limitations of the model:

- **Non-relevant Entities**: due to heterogeneous contents, may be extracted NE that are not related at all to the event that is being analysed (e.g. ‘Halle Berry (PER)’ or ‘Batman (PER)’)
- **Ambiguous entities**: these can either be due to mistakes in detecting a NE, which leads to truncated words that are not recognisable (e.g. ‘Ba (PER)’ or ‘Eri (PER)’ or multiple NE which have been detected as a single NE (e.g. ‘Zemmour Naulleau (PER)’).

To mitigate these, there are two possible solutions: (1) Fine-tune the NER model, (2) Human-evaluation and filtering. Both solutions require human intervention, to ensure more accurate results the choice is to go with domain knowledge evaluation for manual filtering and normalising the NE extracted.

4.4.2 Named Entities filtering and normalization

At this point, a dictionary containing the NE extracted is provided to domain knowledge experts, whose task is to flag entities that are not interpretable with a 'no', non-relevant entities with an 'oi' and other entities with an 'ok'. In addition, when deemed possible NE representing the same entity (e.g. "Marche Des Beurs" and "Marche des Beurs") will be normalised with a single representation. The dictionary is shown in Appendix B.

This will then be used to filter out all the NE flagged with a 'no', which will be left out from the final dataset, and, in addition, the NE that were normalised will be now considered the same NE.

4.5 Network Analysis

The final layer of analysis is provided by the network analysis. In this stage, the NE and topics extracted are leveraged to build various networks which can then be analysed using community detection algorithms to uncover how the different texts are related based on the topics and NE.

4.5.1 Topic networks

To examine the interconnections and organization of topics, as well as the interrelations and organization of texts with respect to the topics they contain, two networks are defined:

- **Topic-Topic network**: In this network, each node represents a topic and a link between two nodes is drawn whenever the two topics (nodes) share co-appear in at least 1 text.

- **Document-Document (topics) network**: In this network, each node represents a text document and a link between two nodes is drawn whenever the two texts (nodes) share at least one topic.

4.5.2 Named Entity networks

Similarly, two networks are established to investigate the relationships and structure of entities, along with the correlations and organization of texts in relation to the entities they contain:

- **Entity-Entity network**: In this network, each node represents a topic and a link between two nodes is drawn whenever the two topics (nodes) share co-appear in at least one text.

- **Document-Document (entities) network**: In this network, each node represents a text document and a link between two nodes is drawn whenever the two texts (nodes) share at least one NE.

4.6 Conclusions

All in all, following the approaches reviewed in Chapter 2, our approach aims at incorporating NER, Topic Modelling and Network Analysis to analyse the text corpora extracted from INA's web archive.

A two-stream pipeline is developed where: on one side Topic Modelling is used to extract topics from the text corpora collected from the web archive, on the other, NER is applied to extract NE present in the texts. These methodologies will offer two layers of analysis. The first one is given by the topics and NE, which can be used by historians to efficiently organise and search through the corpora and answer specific historical questions; the second layer is offered by the network analysis of the corpora in relation to the topics and NE, which will give an idea of how the different texts mentioning the event are organised and related.

Chapter 5

Analysis and Results

5.1 Overview

This chapter covers the analysis carried out following the methodology explained in the previous chapter. First, the data used is presented, including its provenance and how the relevant dataset used for the following experiments was obtained. Afterwards, the different layers of analysis proposed in the approach presented in Chapter 4 are presented and the results are discussed.

5.2 Hardware Setup

The experiments were conducted on a main workstation (MacBook Pro) featuring the following specifications:

- CPU Model: Apple M1;
- GPU Model: Apple M1 (Built-in GPU);
- RAM Capacity: 8 GB;
- RAM Type: LPDDR4X.

In addition, to speed up more computationally expensive steps, such as training the BERTopic models, Google Colab with enabled Tesla V100-SXM2-16GB GPU was used.

Finally, a secondary workstation, provided by the laboratory, with 64 GB RAM was utilized for Network Analysis tasks that required higher RAM, namely plotting larger graphs.

5.3 Data Overview and Specifications

As mentioned in Chapter 1, this project is part of the PICCH, which aims to reframe and reinterpret colonial material held across the web in Europe in order to contribute to decolonization efforts and the creation of a more inclusive society. Therefore, the focus of the analysis for this project is on the "March for Equality and against Racism" that happened in 1983 in France that aimed to address issues of racism and immigration in France, align with the broader objectives of decolonization and fostering a more inclusive society that PICCH seeks to achieve

The data was provided by INA in collaboration with PICCH. INA is responsible for the legal deposit of the French web in connection with audiovisual communication. They started collecting web pages in 2009 before the Internet Archive provided that few websites.

INA is responsible for archiving: (1) Sites from audiovisual media services (public and private channels), including on-demand audiovisual media services (SMAD); (2) Web TV and web radio (including podcasts from Radio France); (3) Sites mainly devoted to radio and TV programs (sites devoted to streaming programs, series, fan sites); (4) Sites of professional and institutional organizations in the audiovisual communication sector

In order to retrieve the data related to the event, the web archive was queried for: `‘‘Marche des Beurs’’ OR ‘‘Marche pour l’égalité et contre le racisme’’ OR ‘‘Marche de 1983’’`. The webpages were scraped using Boilerpipe, which is an algorithm that processes the HTML file and returns the main content, which implies ignoring the text related to the navigation links of the site, and other parasite elements. It is important to note that since a website is archived every time single byte changes, the same webpage may be present multiple times in the archive. Therefore, a deduplication was performed and only the first (chronologically) URL of each webpage was scraped. The resulting JSON files were extracted, a sample of which is shown in Appendix A.1, and provided for analysis. For the purpose of this project, only the relevant data was kept and extracted in a CSV file, this included:

- `id`: the id of the webpages, which would be used as unique identifier.
- `url`: the URL of the webapges.
- `title`: the HTML title of the webpages.
- `date`: the extraction date of the webpage. From which the `year` was derived, which would be used to allow a diachronic analysis of the corpora as well.
- `webpage_text`: the text scraped by the boilerpipe algorithm from the webpage.

A sample of the resulting data frame is shown in Appendix A.2. In total, the dataframe contains 12,688 entries, whose distribution across the years can be seen in Figure 5.1 below.

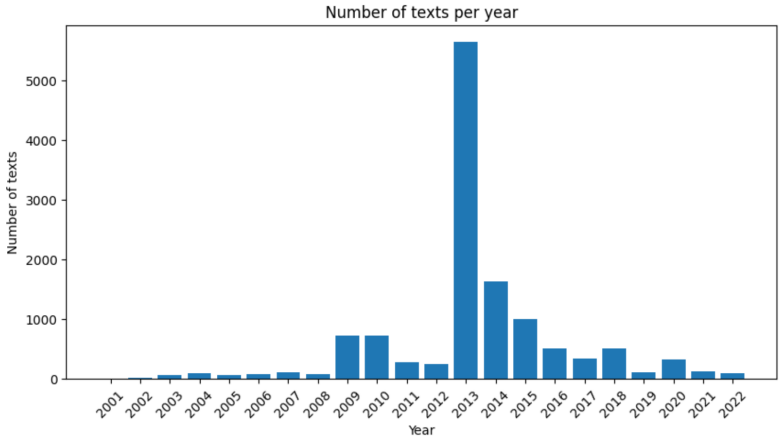


Figure 5.1: Webpage distribution across the years

As we can see, the distribution is highly skewed, with 44% of the web pages dated 2013. This could be due to the fact that in that year the 30th anniversary of the *Marche*

and a movie came out as well as a documentary, which could explain the rise in mentions of the event in that year. This is important to consider if we want to analyse the data diachronically as it would be a limitation not having enough data for each year. In addition, the data includes 558 different domains, from radio (franceinfo.fr) and television (non-stop-people.fr) websites including blogs and forums. Therefore, we can expect the data to be heterogeneous in both content and form.

5.4 Results

Following the methodology and approach explained in Chapter 4, two layers of analysis were performed on the data: the first one given by the NE and topics extracted, the second one given by the network analysis performed on these features, which aims to create a map of the corpora.

5.4.1 NE and Topics

The first layer of analysis is provided by the Named Entities and topics extracted. This can be used by historians and researchers in digital humanities to filter for texts that mention a particular person or organisation or that are related to a specific topic. In addition, these can also allow answering relevant questions about the discourse mentioning the *Marche*, such as which topics are discussed by different websites or how these changes through the years considering milestone events (i.e. the release of the movie about the *Marche*).

5.4.1.1 Named Entities

After deploying the Babelscape model [36] as detailed in Chapter 4, a list of NE was extracted from each text and 127,195 unique Named Entities were obtained and divided in the semantic categories shown in Table 5.1 below:

Entity type	Count
PER	50,016
MISC	37,743
ORG	21,858
LOC	17,578
Total	127,195

Table 5.1: Unique Named Entities extracted from data

This, however, as mentioned in the previous chapter contain NE that are not interpretable due to errors of the model (i.e. '##eté Terre Vo (MISC)') or different representations of the same entity due to the heterogeneity of the data (i.e. 'Marche des Beurs (MISC)' and 'Marche Des Beurs (MISC)'). Therefore, we proceeded to involve domain knowledge experts to filter and normalise the NE extracted by providing them with the dictionary of entities that can be found in Appendix C.1.

As mentioned in the previous chapter, the following flags were defined:

- Flag 'ok': which was assigned to entities that were deemed relevant to the *Marche*.

- Flag ‘oi’: which was assigned to entities that are correctly detected, but not strictly relevant to the event.

- Flag ‘no’: which was assigned to entities that are not interpretable.

- Flag ‘mu’: which was assigned to entities that contain more than one entity together (e.g. ‘Zemmour Naulleau (PER)’)

After flagging the entities present in the dictionary the following were flagged:

Entity type	Flag ‘ok’	Flag ‘oi’	Flag ‘no’	Flag ‘mu’
PER	781	23,155	26,000	80
MISC	109	3,317	34317	0
ORG	73	2,857	18927	1
LOC	1,171	3,184	13223	0

Table 5.2: Unique Named Entities extracted from data

The entities flagged with a ‘no’ or a ‘mu’ were filtered out from the dataset, leaving only the ‘ok’ and ’oi’ in the list of NE for each text. Leaving with a total of 34,728 unique NE.

In Figure 5.2 we can see the frequency of the top 20 NE entities belonging to the PER category, while in Figure 5.3 we can see a WordCloud representing all the different PER NE extracted, with the size of the entity proportional to its count. The other results can be found in Appendix B.1.

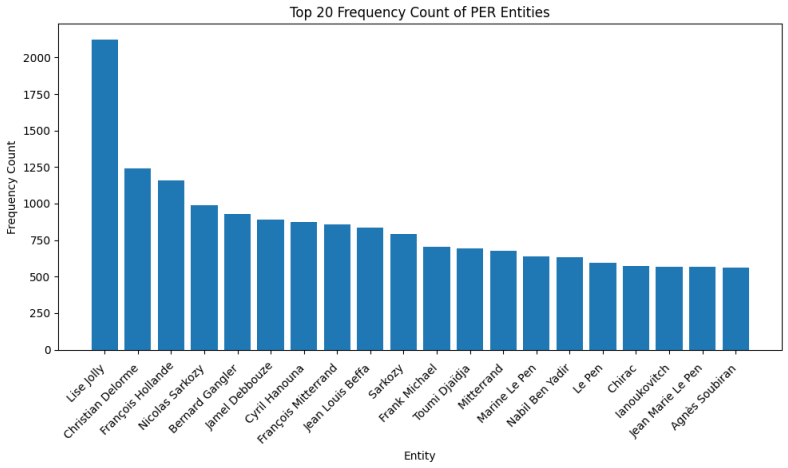


Figure 5.2: Frequency of top 20 PER NE extracted

Figure 5.3: WordCloud of PER NE extracted

This allows us to have a first look at which are the main characters or organisations involved when mentioning the *Marche*. For instance, it is evident that politics are often involved in the discourse around the event, which can be seen looking at figures such as François Holland, Nicolas Sarkozy, François Mitterand or organisations like Front National, PS and UMP. This may be due to the fact that the March was an event that emerged against tensions between different political parties, and discriminatory policies and aimed to demand equal rights, an end to injustice, and social inequality in France. Therefore, politics and political factors are integral to the overall narrative related to the event.

Finally, using the Named Entities list associated with each webpages' text it is also possible to rapidly filter and search through the corpora for specific entities mentioned.

5.4.1.2 Topics

Similarly to the NE, topics provide a way to efficiently explore the corpus. As mentioned in Chapter 4, BERTopic was used to extract topics from the corpus. In addition to the configuration detailed in the previous chapter, a diversification threshold was used to obtain different representative words per topic as follows: `MaximalMarginalRelevance(diversity=0.45)` . This would allow to diversify the topic representation, which might include words that are too similar (i.e. ' ' ' 'musulmane' and 'musulman') that do not offer a broad description of the topic.

Furthermore, it is important to note that as seen in Figure B.4 in Appendix B.2 topics extracted may contain stop words. This is because, as mentioned also by [19], occasionally stop words can inadvertently appear in topic representations, which is something we generally wish to prevent since they add minimal value to the interpretation of the topics. However, it is not recommended to remove stop words as a preprocessing step because the transformer-based embedding models we utilize require the complete context to generate precise embeddings. Therefore, a list of French and English stop words as well as frequent words occurring in HTML elements (i.e. 'class', 'div', 'php') was input into a `CountVectorizer` that was used to remove these from the topic representations. In combination with this, use also used the `ClassTfidfTransformer` to reduce the impact of frequent words.

In Figure 5.4 can be seen the top 20 topics are extracted with their first 5 representative

words.

Figure 5.4: Top 20 topics extracted after refining

Finally, 106 topics were extracted and the full list of topics with their corresponding representative words can be found in Table B.1 in Appendix B.2.

At this point, the labelling process described in the previous chapter was applied, obtaining the labelled topics shown in Table B.2 in Appendix B.2. By looking at the

final labelled topics, it is evident how among the most frequent topics there are multiple about Television and Films (Topic 2 and 3) probably related to the movie about the *Marche*. In addition, as it was noted from the NE, also numerous topics are related to Politics, in particular in the contexts of debates and comments (i.e. topics labelled 'Political Commentary', 'TV Guests and Political Entertainment', 'Politics and Criticism') and Political Elections (i.e. topics labelled 'Political Propaganda', 'Election Campaign' and 'Holland and Elections').

5.4.1.3 Conclusions

All in all, NE and topics offer a first layer to analyse the corpora, this enables digital historians to effectively explore the corpora and provide relevant answers to their inquiries. By identifying and extracting named entities like individuals, organizations, and places, it is possible to gain valuable insights into important entities and geographical information in the text. Furthermore, by utilizing topic modelling techniques, it is possible to unveil prominent themes and subjects, facilitating effortless navigation through large corpora of texts. Altogether, these analytical approaches play an indispensable role in our endeavour to efficiently explore corpora and derive significant conclusions to pertinent questions.

To show the potential of these we built a Dashboard using Tableau, which can be accessed in the footnote [1]. The first step in building the Dashboard was data preparation. The data extracted from the Topic Modelling and NER steps was pivoted in order to analyse the Topics and Named Entities diachronically and display their trends. In addition, from the 'url' the domain was extracted in order to provide an overview of the different topics present in the various web pages domains.

In Figure 5.5 below a screenshot of the aforementioned dashboard shows how it is possible to select a specific domain from a dropdown menu and visualise through a bar chart the distribution of the different topics extracted from the corpora. In the figure, it can be seen how the ccme.org, which is the website for the "Conseil de la communauté marocaine à l'étrange", predominantly deals with the topic labelled "Immigration and Integration", which is not surprising given that the CCME is responsible for monitoring and evaluating the Kingdom's public policies towards its nationals abroad. Additional screenshots of the different pages offered in the Dashboard can be found in Appendix B.3.

[1] NE & Topics Dashboard

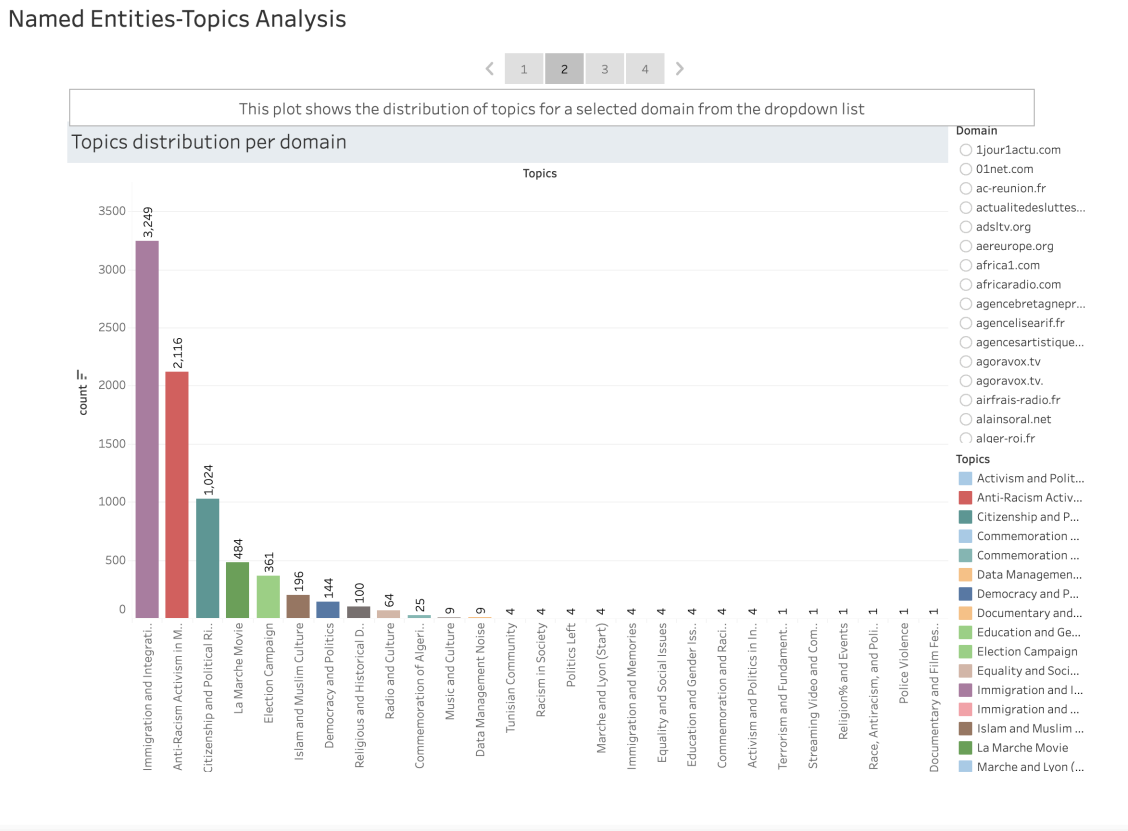


Figure 5.5: Dashboard exploring domain's topics

This Dashboard is an example of how Named Entities and topics can be exploited to rapidly navigate large corpora, looking for specific topics

5.4.2 Network Analysis

The network analysis serves as the final layer of analysis. During this stage, as detailed in Chapter 4, the extracted Named Entities and topics are utilized to construct multiple networks. These networks can then be examined using community detection algorithms to reveal the interconnectedness of texts based on the topics and NE. Additionally, this network analysis functions as a kind of map for the corpora.

5.4.2.1 Topic Networks

As described in the previous chapter, two networks are defined to explore the interconnections and structure of topics, as well as the relationships and arrangement of texts based on the topics they encompass.

These networks are the Topic-Topic and Document-Document (topics).

Topic-Topic Network

This network provides an overview of how the topics extracted are related to each other, the goal is to detect whether there is a structure in the organisation of the topics and what

type of communities emerge. First of all, looking at the degree distribution, which can be seen in Figure B.8 in Appendix B.4, we can see that the topic-topic network is homogenous which means that there is no node having significantly more connections than others. In addition, we can also see, in Figure B.9 in Appendix B.4, by plotting the relationship between the node size, given by the number of chunks in which a topic appears, and the degree of the node that, intuitively, the nodes with higher appearances have also higher degree on average. Finally, the graph has a density of 40% which indicates a high level of connectivity among its nodes, as there are many edges connecting the nodes.

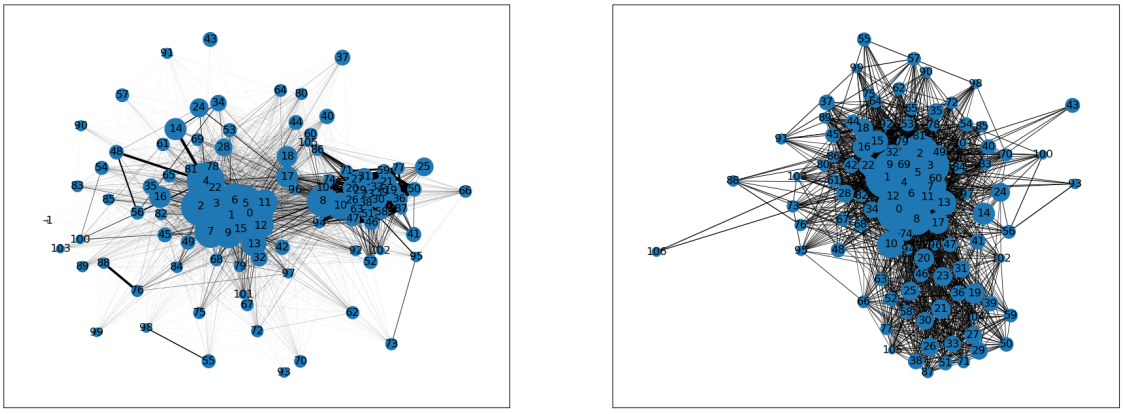
In order to assess that the communities identified do not depend on the definition of the edge's weight, we define two edges:

- **Co-occurrence**: This is simply the number of documents in which two given topics co-occur.
- **Pointwise Mutual Unformation (pmi)**: This is based on the PMI score. The idea of PMI is that we want to quantify the likelihood of the co-occurrence of two words, taking into account the fact that it might be caused by the frequency of the single words. Hence, the algorithm computes the (log) probability of co-occurrence scaled by the product of the single probability of occurrence. The formula is given as follows:

$$\text{PMI}(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}\right) \tag{5.1}$$

  In this formula, w_1 and w_2 represent two topics, $P(w_1, w_2)$ is the joint probability of w_1 and w_2 occurring together, and $P(w_1)$ and $P(w_2)$ are the individual probabilities of w_1 and w_2 occurring respectively. In doing so, we take into consideration the fact that some topics may be highly more frequent than others.

Figure 5.6 shows the topic-topic network, with two edge weights defined. The networks are displayed using the spring layout, which is commonly used to visualize networks and it is based on the concept of simulating a physical system where nodes repel each other while edges act as springs that pull connected nodes closer together.

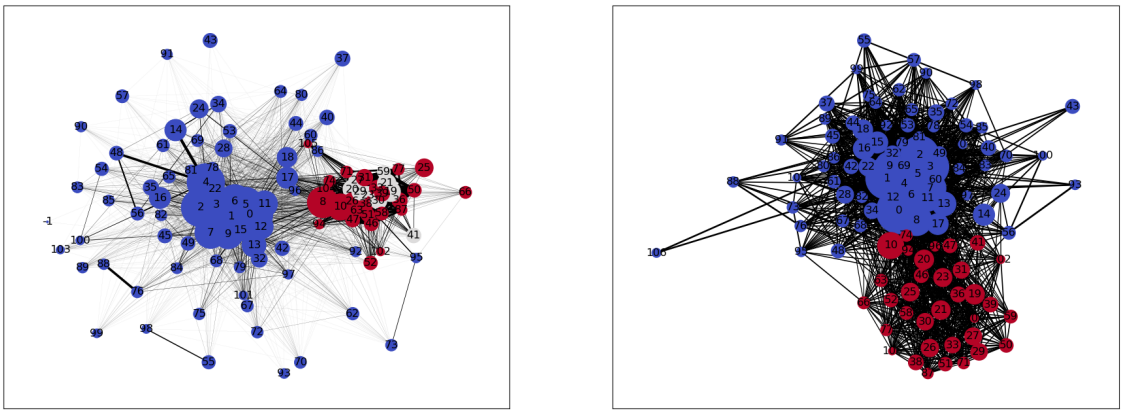


(a) Co-occurrance edge weight

(b) PMI edge weight

Figure 5.6: Topic-Topic network

Looking at the two network representations we can note that there may be two distinct communities. After applying the Louvain algorithm we detect the communities displayed in Figure 5.7.



(a) Co-occurrance edge weight

(b) PMI edge weight

Figure 5.7: Topic-Topic network Louvain communities

We can see that on the co-occurrence network, it finds 2 communities, while in the PMI network, it finds 3 communities. In order to see how different these communities are, we use Jaccard similarity and we find that actually communities 1 and 2 of the PMI score merged together have a similarity of 95% with community 1 found in the co-occurrence network, as it can be seen in the Table B.3 in Appendix B.4. This proves that the communities detected by Louvain are independent of the edge's weight definition and there are two main communities in the network.

To further explore the effectiveness of community detection algorithms and check whether different communities are detected, we decided to apply the Girvan-Newman algorithm as well.

As a result of applying the original version of Girvan-Newman, we can see in Figure B.10 in Appendix B.4 that two communities are detected: one with only 1 node and the other one with all the rest. This can be explained by the high density of the network. In fact, Girvan-Newman uses betweenness centrality to find communities, however, this in a highly dense network is irrelevant as many edges will have the same betweenness centrality.

For this reason, we try to apply a modified version of Girvan-Newman, proposed by [25], in which the criteria to cut an edge is the edge's weight. In practice, the edge's weights are ordered from lower to higher and iteratively remove the edge with the lowest weight is. However, even using this version we are only able to find one community which has 8 nodes and the others are split into one node per community. We can look at the average edge's weight inter and intra-community, reported in Table B.4 in Appendix B.4, and see how this is very similar which explains why the modified Girvan-Newman does not work in this case as there is no high discrimination between weights. However, these two communities, whose topics are listed in Table B.5 in Appendix B.4, have been detected by Louvain and as shown in Figure B.11 in Appendix B.4, generating a random network, removing any correlation and keeping only the number of edges and weight's

distribution as detailed by [16], we would not be able to see any structure or detect defined communities.

Document-Document Network (topics)

Similarly, as described in Chapter 4, we obtain a document-document network where each node represents a webpage text and a link is drawn whenever two nodes share at least one topic. This time, for simplicity we represent the edge's weight simply by the number of shared topics. However, different ways could also be tested such as normalising the weight considering the length (in terms of the number of chunks) of the documents.

In Figure 5.8 can be seen the network which is represented with the Kamada-Kawai layout, whose goal is to arrange the nodes of a graph in a way that represents their connectivity and proximity accurately. This is done by considering the pairwise distances between nodes in the graph and attempting to find an arrangement that minimizes the total energy of the system. The energy is calculated based on the distances between connected nodes, with shorter distances indicating stronger connections.

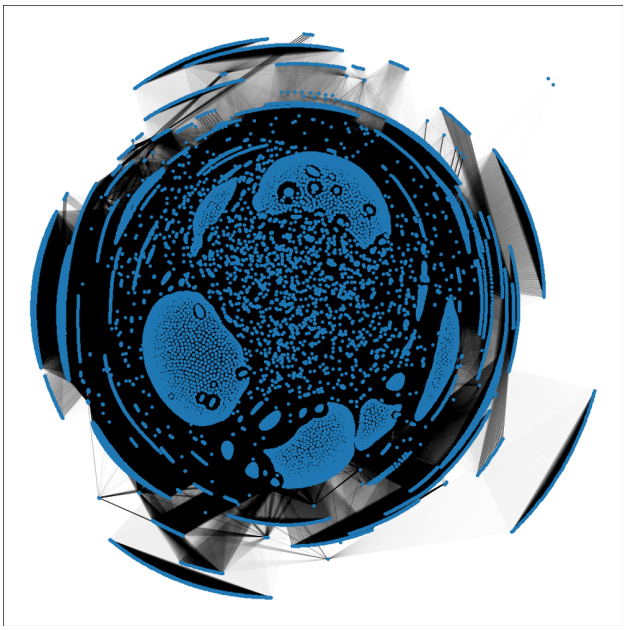


Figure 5.8: Document-Document (topics) Network

As we can see from the Figure, it is immediately evident that there is a structure in the network with peripherical nodes organised in groups as well as a few communities in the core of the network.

Also in this case we apply the Louvain algorithm, whose communities can be seen in Figure 5.9 below.

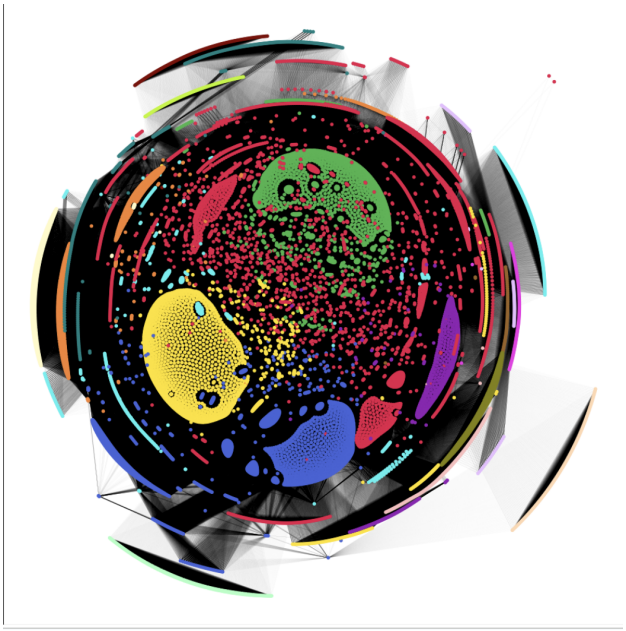


Figure 5.9: Document-Document (topics) Network's Louvain communities

As a result, 17 communities have been detected, which would not be the same case in the case a random network is drawn with the same characteristics (i.e. the same number of nodes and the number of edges).

Therefore, we can conclude that indeed there is a structure in the corpora extracted from INA's web archive mentioning the "Marche pour l'égalité et contre le racisme" as the webpages are organised into communities based on the topics extracted. These can serve as a valuable resource for analyzing individual communities and facilitating the retrieval of similar texts. This allows for efficient text retrieval by narrowing down the search space to specific communities rather than the entire corpus. Additionally, it provides a means to explore related texts and discover connections between different documents.

5.4.2.2 Entities Networks

As detailed in Chapter 4, also, in this case, two networks are established with the intention of investigating the interconnections and structure of entities, along with examining the relationships and organization of texts in relation to these entities.

These networks are Entity-Entity and Document-Document (entities).

Entity-Entity Network

This network provides an overview of how the Named Entities extracted are related to each other, the goal is to detect whether there is a structure in the organisation of the NE and what type of communities emerge.

First of all, looking at the degree distribution, which can be seen in Figure B.13 in Appendix B.4, we can see that the topic-topic network is heterogeneous, which means that there are few nodes having significantly more connections than others. This is expected given that some entities such as 'France (LOC)' or 'Paris (LOC)', as shown in Appendix

B.2, appear significantly more times than the rest.

By applying Louvain algorithm we obtain the communities shown in Figure 5.10 below:

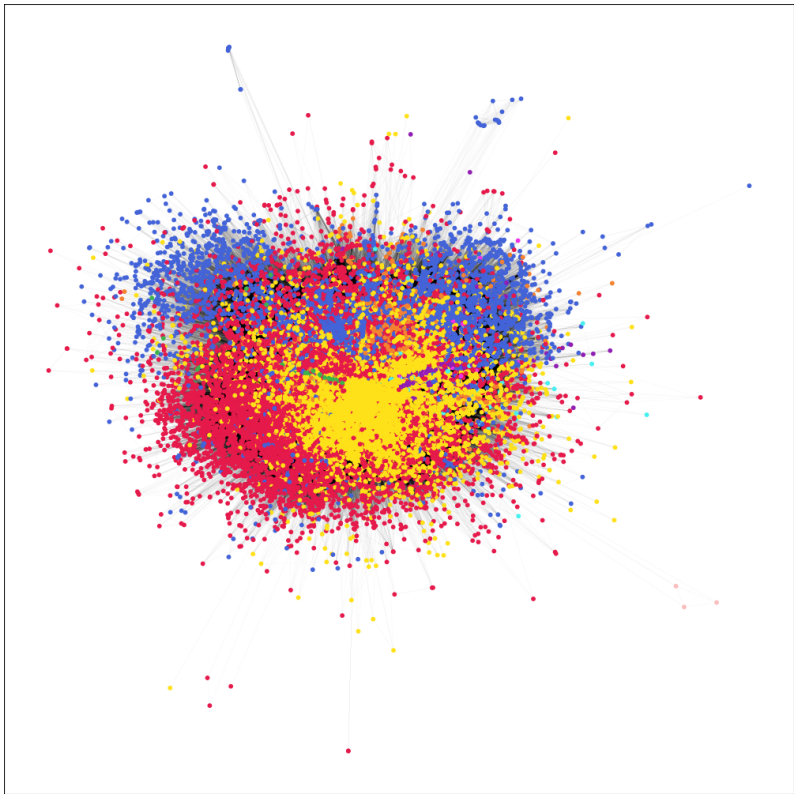


Figure 5.10: Document-Document (entities) Network's Louvain communities

The Louvain algorithm detected 10 communities and, as we can see from the Figure, there is a structure in the network that is even more evident when comparing it to a random network. However, the communities detected in this network are not well-defined. A possible explanation could be attributed to the highly heterogeneous degree distribution (shown in Figure B.13, where some nodes have significantly more connections than others, which means that some entities appear in almost every text. These high-degree nodes potentially serve as critical connectors that hold the network together. To investigate the significance of these nodes in maintaining network integrity, we can examine the resilience of the network by observing how the connected components evolve when we progressively remove nodes starting from those with high degrees and moving towards those with lower degrees. This analysis will provide insights into the importance of these high-degree nodes in sustaining the overall connectivity and structure of the network.

By removing the nodes with high degrees, it is possible that the network's communities will become more clearly defined. This refinement of communities can facilitate the rapid analysis of corpora grouped within these communities. Once the high-degree nodes, acting as network hubs, are removed, the resulting communities may exhibit more

cohesive and distinct patterns of connectivity among their constituent nodes. These refined communities can then be leveraged to facilitate the analysis of corpora associated with each community.

Document-Document Network (entities)
Similarly to the document-document network presented before, this network provides an overview of how the texts of the corpora are organised in relation to the Named Entities extracted. Each node represents a text and an edge is drawn between two nodes when these share at least one NE. For simplicity, the edge's weight is given by the number of shared NE.

In order to obtain a less dense network, we decide to filter out those entities that appear in more than 10% of the texts as they would generate links between the majority of nodes, as well as those that appear in less than 15 texts. After that, we apply the Louvain algorithm and detect the communities shown in Figure 5.11 below.

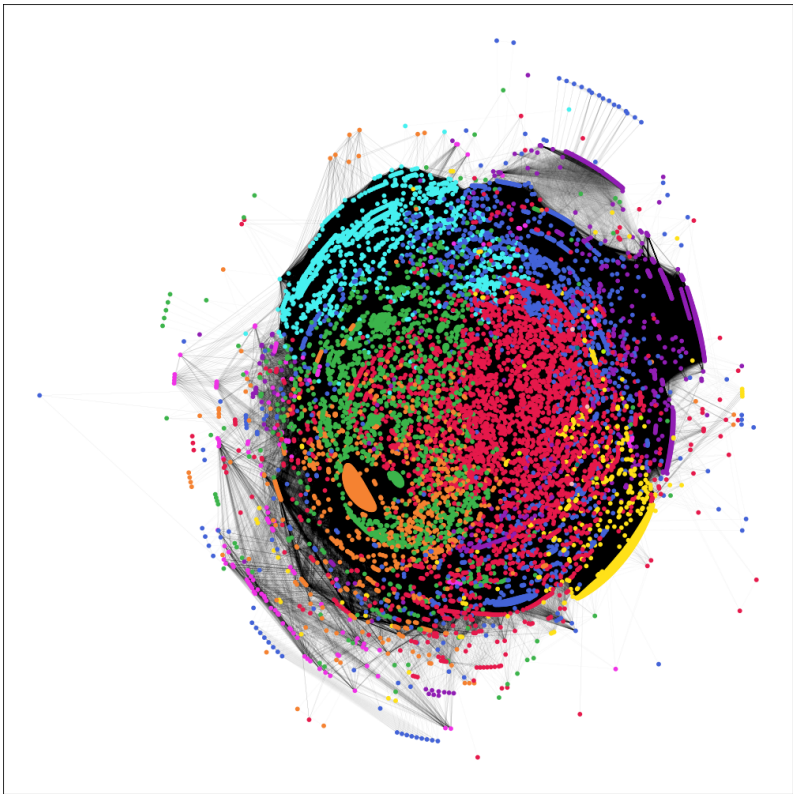


Figure 5.11: Document-Document (entities) Network's Louvain communities

The figure displays the 10 communities detected by the Louvain algorithm in the network. This time, contrary to the document-document network generated using the topics, the overall structure of the network and the communities detected are less evident. However, compared with the structure obtained by generating a random network with the same number of edges and the same weight distribution, as shown in Figure B.12 in Appendix B.4 it is clear that there is a signal of a structure in the organisation of the texts in relation to the Named Entities extracted. This can be used to navigate the corpora by

communities, isolating groups of texts that share Named Entities and further analysing them.

5.4.2.3 Conclusions

To sum up, network analysis provides a further layer of analysis on the corpora mentioning the Marche extracted from INA's web archive. Four different networks have been generated leveraging the topics and Named Entities extracted from the texts. From a first analysis of these, it is immediately clear that there is a structure on how the corpora are organised in relation to the semantic features extracted. Therefore, network analysis offers a unique perspective on the organization of corpora, revealing patterns of relationships and structural properties that may not be readily apparent through other methods.

5.5 Limitations

All in all, following the pipeline described in the previous chapter, we provided two layers of analysis: the first one, offered by the Named Entities and Topics extracted from the corpora which can be leveraged as shown in the Dashboard to rapidly gain insights about the discourse mentioning the "Marche pour l'égalité et contre le racisme" in the texts from INA's web archive; the second one, provided by the network analysis that uses the semantic features extracted from the texts to discover how the different texts are organised in terms of topics and entities. The results of these analyses represent a valid tool to explore the vast quantity of data available in web archives.

However, there are limitations to this work that need to be addressed:

- **Topic Modelling**: Topic modelling pertains to its relationship with present research in Computer Science (CS) communities. While topic modelling techniques have gained significant attention and adoption in various domains, including natural language processing and information retrieval, the field continues to evolve rapidly. Therefore, it is important to recognize that the findings and performance of the topic modelling techniques applied may be influenced by the specific framework in which they are applied.

- **Network Analysis**: It is crucial to acknowledge that network analysis, specifically in the context of community detection, faces several limitations. As highlighted by [29], a notable challenge lies in the absence of definitive ground truth. This issue was further discussed in [20], which emphasized that community detection is an inference problem, thus making it subject to the limitations inherent in statistical inference.

  Furthermore, a study referenced as [21] conducted a comparison between the outcomes of various community detection algorithms and ground truth data. The results revealed that these algorithms often fail to accurately identify communities within real-world networks. This finding stresses the need to employ different algorithms and compare their results to ensure the reliability and robustness of the detected communities across diverse networks.

- **Quantity of data**: For the scope of this project, the data was provided by INA and extracted from the web pages present in their web archive that mentioned the

Marche. This affects, for instance, the diachronic analysis of the corpora, since as noted in Figure 5.1 it is for the majority around 2013 and fewer data is present for the other years. In addition, the main property of web pages, namely links, is not fully exploited due to the absence of links in the extracted web pages. Therefore, the analysis does not consider the interconnectedness and referencing nature of web content through hyperlinks.

- **Evaluation**: When employing topic modelling and NER techniques, it is important to acknowledge the limitations in evaluation. One notable limitation in topic modelling is the absence of meaningful benchmarks to assess the quality and accuracy of the generated topics. Unlike some other machine learning tasks that have established evaluation metrics, topic modelling lacks universally accepted standards for evaluation. As a result, it becomes challenging to objectively measure the performance of different topic modelling algorithms or compare the results across different studies.

  On the other hand, in NER the lack of benchmark datasets makes it difficult to objectively assess and compare the performance of different NER models. In this case, we relied on domain knowledge experts to evaluate and filter the NE extracted, however, this still requires caution in interpreting the results.

Chapter 6

Conclusions and Perspectives

6.1 Main Contributions

In conclusion, this interdisciplinary project has demonstrated the power of integrating NLP techniques and network analysis in the exploration of large corpora extracted from INA's web archive. By leveraging the capabilities of NER, topic modelling, and network analysis, we have provided digital historians with valuable tools to navigate and extract meaningful insights from vast amounts of historical data.

Through the application of NER, we were able to extract named entities, such as people, organizations, and locations, from the web archive, enabling a better understanding of the key actors and entities involved in the historical context under study.

Furthermore, topic modelling allowed us to uncover latent themes and topics within the texts, providing a comprehensive view of the recurring patterns, events, and discourses. This automated approach facilitated the identification of important trends and allowed historians to delve deeper into specific topics of interest.

The subsequent application of network analysis techniques enhanced our understanding of the relationships between the different texts in relation to the entities and topics extracted. By applying community detection algorithms, we were able to identify clusters of texts.

The collaboration between computer science, network science, and historians played a crucial role throughout the project. The expertise of historians in evaluating named entities and labelling topics ensured the accuracy and relevance of the extracted information. Meanwhile, computer science and network science provided the necessary computational tools and techniques to process and analyze the large corpora, thus facilitating the exploration of the web archives.

6.2 Reception by Domain Experts

The outcomes of this research project were met with exceptional enthusiasm from the involved domain experts in the field.

Moreover, the opportunity to showcase our research at the RESAW conference in Marseille provided an excellent platform to disseminate our results to a broader audience. The presentation was met with keen interest and engagement from fellow researchers and academics in attendance. We received encouraging feedback and constructive suggestions, affirming the relevance and novelty of our work. The positive feedback and valuable insights not only validated the significance of our findings but also encouraged further exploration into this domain.

In addition, the interdisciplinary project's success in combining NLP techniques and network analysis to explore INA's web archive demonstrates its applicability beyond historical data. This approach can be extended to other domains, such as academic research papers, where vast amounts of literature exist with limited navigation tools. By employing NER, topic modelling, and network analysis, researchers can uncover key entities, themes, and relationships within research papers. This integration opens up new possibilities for gaining deeper insights, identifying trends, and facilitating cross-disciplinary research, empowering scholars to efficiently navigate and interpret a vast sea of academic knowledge.

6.3 Future Work

In addition to the achievements and contributions made in this thesis, there are several potential avenues for future work that can further enhance the capabilities of the proposed framework. Firstly, incorporating additional data sources, such as tweets archived from INA and potentially from the Bibliothèque nationale de France (BnF) web archive, would provide a richer and more diverse corpus for analysis. In addition, Twitter data could also offer valuable insights into contemporary discussions and events, complementing the existing data.

Secondly, exploring and testing different algorithms for community detection within the network analysis component would be an important step towards improving the identification and understanding of cohesive groups or clusters within the data. By considering alternative algorithms, such as that suggested by [29], and comparing their performance, we can refine the analysis and potentially uncover more nuanced relationships and patterns among the entities and topics. In addition, it could be also interesting to dive deeper into the communities detected by the Louvain algorithm, by for example re-applying the community detection algorithm to detect smaller communities within the existing communities.

Lastly, it would be valuable to apply the framework to a different dataset, particularly one consisting of more heterogeneous texts such as news articles. By examining how the framework performs on different types of data, we can assess its generalizability and robustness. News articles, for example, typically exhibit a different writing style and structure compared to blogs and forums, which could yield different insights and potentially lead to improved results.

By pursuing these future directions, we can continue to advance the field of digital history and facilitate the exploration of large corpora of text. Incorporating additional data sources, refining network analysis techniques, and testing the framework on different datasets will contribute to a more comprehensive understanding of historical contexts and enable historians to delve deeper into the complexities of the past.

Bibliography

[1] Andersen, E., Biryukov, M., Kalyakin, R., and Wieneke, L. How to read the 52.000 pages of the british journal of psychiatry? A collaborative approach to source exploration. J. Data Min. Digit. Humanit. 2020 (2020). (Cited on page 6.)

[2] Angelov, D. Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470 (2020). (Cited on pages iii, 17 and 18.)

[3] Bansal, H. Latent dirichlet allocation, March 2020. (Cited on pages iii and 15.)

[4] Beaudouin, V., and Pehlivan, Z. Cartographie de la Grande Guerre sur le Web. Research report, Bibliothèque nationale de France ; Bibliothèque de documentation internationale contemporaine ; Télécom ParisTech, Jan. 2017. (Cited on pages 3 and 7.)

[5] Bibliothèque nationale de France. Bibliothèque nationale de France. https://www.bnf.fr/fr, Accessed: 20/06/2023. (Cited on page 3.)

[6] Blei, D. M. Probabilistic topic models. Commun. ACM 55, 4 (2012), 77–84. (Cited on page 6.)

[7] Brügger, N. The archived web: doing history in the digital age. MIT Press, 2018. (Cited on page 3.)

[8] Campos, D., Matos, S., and Oliveira, J. Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. 11 2012, pp. 175–195. (Cited on pages iii and 10.)

[9] del Barrio, D. A., and Gática-Pérez, D. How did europe's press cover covid-19 vaccination news? A five-country analysis. In MAD@ICMR 2022: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, June 27 - 30, 2022 (2022), B. Ionescu, G. Kordopatis-Zilos, S. Papadopoulos, A. Popescu, and L. Cuccovillo, Eds., ACM, pp. 35–43. (Cited on page 6.)

[10] Despalatović, L., Vojković, T., and Vukicević, D. Community structure in networks: Girvan-newman algorithm improvement. In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (2014), pp. 997–1002. (Cited on page 20.)

[11] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers) (2019), J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, pp. 4171–4186. (Cited on page 25.)

[12] Egger, R., and Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology 7 (2022). (Cited on pages 15, 18 and 19.)

[13] Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A. Named entity recognition and classification on historical documents: A survey. CoRR abs/2109.11406 (2021). (Cited on pages 9 and 10.)

[14] Fiok, K., Karwowski, W., Gutierrez, E., Davahli, M. R., Wilamowski, M., and Ahram, T. Revisiting text guide, a truncation method for long text classification. Applied Sciences 11, 18 (2021), 8554. (Cited on page 25.)

[15] Fortunato, S. Community detection in graphs. CoRR abs/0906.0612 (2009). (Cited on page 20.)

[16] Gauvin, L., Génois, M., Karsai, M., Kivelä, M., Takaguchi, T., Valdano, E., and Vestergaard, C. L. Randomized reference models for temporal networks. SIAM Rev. 64, 4 (2022), 763–830. (Cited on page 39.)

[17] Gebeil, S. Website story. Histoire, mémoires et archives du Web. Ina Éditions, Aug. 2021. (Cited on page 3.)

[18] Gerlach, M., Peixoto, T. P., and Altmann, E. G. A network approach to topic models. CoRR abs/1708.01677 (2017). (Cited on page 19.)

[19] Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022). (Cited on pages iii, 14, 16, 26 and 33.)

[20] Hastings, M. B. Community detection as an inference problem. Physical Review E 74, 3 (2006), 035102. (Cited on page 43.)

[21] Hric, D., Darst, R. K., and Fortunato, S. Community detection in networks: Structural communities versus ground truth. CoRR abs/1406.0146 (2014). (Cited on page 43.)

[22] Jacobi, C., Van Atteveldt, W., and Welbers, K. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital journalism 4, 1 (2016), 89–106. (Cited on page 5.)

[23] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multim. Tools Appl. 78, 11 (2019), 15169–15211. (Cited on page 19.)

[24] Kim, S. M., and Cassidy, S. Finding names in trove: Named entity recognition for australian historical newspapers. In Proceedings of the Australasian Language Technology Association Workshop, ALTA 2015, Parramatta, Australia, December 8 - 9, 2015 (2015), B. Hachey and K. Webster, Eds., ACL, pp. 57–65. (Cited on page 6.)

[25] Kontro, I., and Génois, M. Combining surveys and sensors to explore student behaviour. CoRR abs/2003.04137 (2020). (Cited on page 38.)

[26] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures 12, 2-3 (2018), 93–118. (Cited on pages 14 and 19.)

[27] Milligan, I. History in the Age of Abundance?: How the Web Is Transforming Historical Research. McGill-Queen's Press-MQUP, 2019. (Cited on page 3.)

[28] Navigli, R., and Ponzetto, S. P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence 193 (2012), 217–250. (Cited on page 12.)

[29] Peixoto, T. P. Bayesian stochastic blockmodeling. Advances in network clustering and blockmodeling (2019), 289–332. (Cited on pages 43 and 46.)

[30] PICCH. Polyvocal interpretation of contested colonial heritage. https://picch-project.org/, Accessed: 02/06/2023. (Cited on page 3.)

[31] Reimers, N., and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (11 2019), Association for Computational Linguistics. (Cited on page 24.)

[32] RESAW. RESAW collective. https://cc.au.dk/en/resaw, Accessed: 04/06/2023. (Cited on page 3.)

[33] Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015 (2015), X. Cheng, H. Li, E. Gabrilovich, and J. Tang, Eds., ACM, pp. 399–408. (Cited on pages 14 and 25.)

[34] Rosenzweig, R. Scarcity or abundance? preserving the past in a digital era. The American historical review 108, 3 (2003), 735–762. (Cited on page 5.)

[35] Saussure, F. d. Course in general linguistics. In The Norton Anthology of Theory and Criticism, V. B. Leitch and et al., Eds., 3rd ed. W.W. Norton & Company, 1916, pp. 820–840. (Cited on page 13.)

[36] Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. Wikineural: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021 (2021), M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Association for Computational Linguistics, pp. 2521–2533. (Cited on pages iii, 9, 11, 12, 20 and 31.)

Appendix A

About the data

A.1 Original data

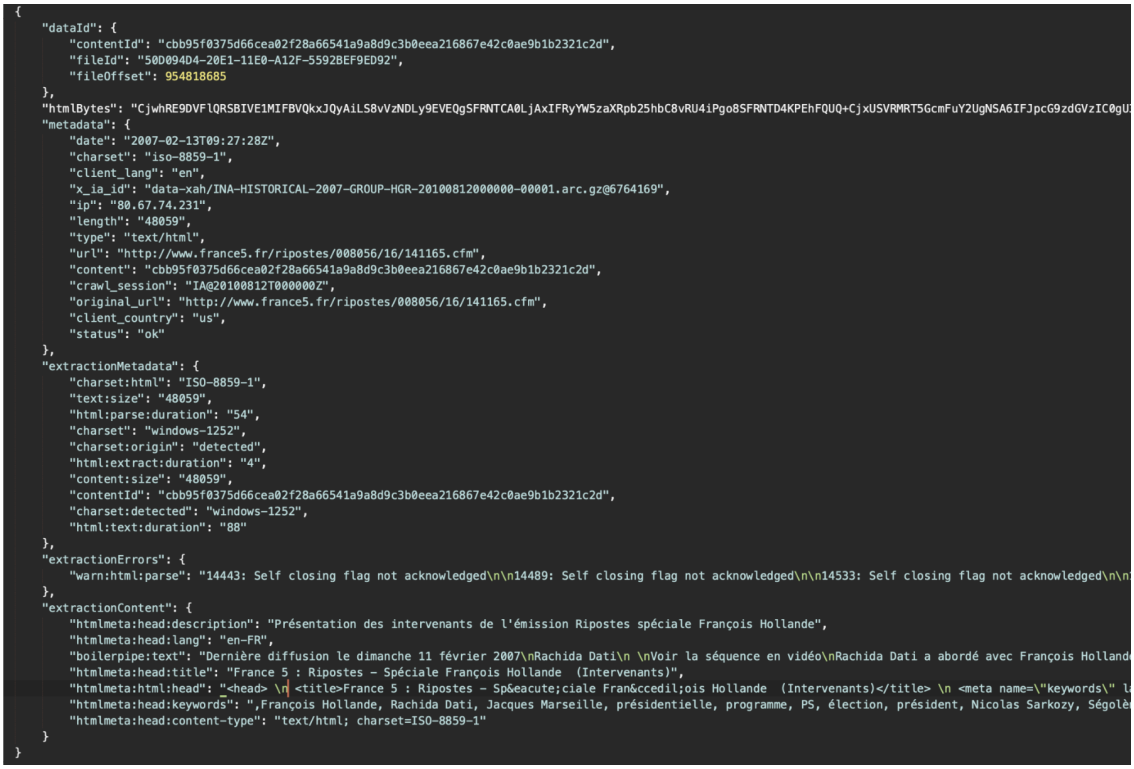


Figure A.1: JSON record from original data extracted from INA's web archive

A.2 Dataset extracted

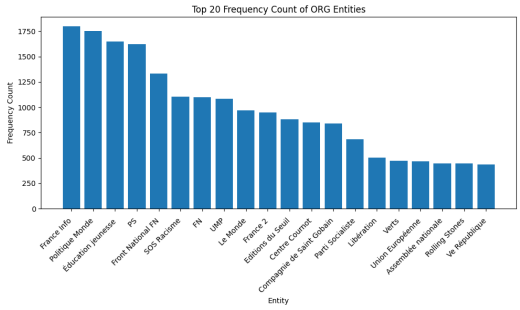
	id	url	title	webpage_text	year
0	238d63ae7714c91726b87d9be1527b47ab6a7dd527a943...	http://faea.skyblog.com/2.html	SkyBlog : C'est MY BLOG !!	Description du Blog:\nCoucou à toi !! J'espère...	2003
1	ea5e9277a5b4a323affe4d2118924dcb1e8c0a2225cca0...	http://www-org.france5.fr/cdanslair/006055/69/...	France 5 : C dans l'air - La grande déception ...	Prochaine émission lundi à 17h50\nDernière dif...	2004
2	62f4c6e7dd958ba2a51d5c987a8d7ef89a3326cd0ce232...	http://pr1cess-maleka.skyblog.com/35.html	Skyblog de pr1cess-maleka : KIFFANCE a 3000%	Description du Skyblog :\n1 blog parmis tant d...	2004
3	7fc59ddce9be2aaec9bd528972bfac706d8bbda2482b9d...	http://machaallah.skyblog.com/2.html	Skyblog de machaallah : Réalité Anonyme~~~~~~>...	Description du Skyblog :\nBienvenus à toutes e...	2004
4	07a59b03205107689784ac2108b3f3e7cb50154b6b9b95...	http://stefano147.skyblog.com/	Skyblog de stefano147 : Stephan	Date de création :\nDu 31 juillet au 4 août 20...	2004

Figure A.2: Sample dataframe extracted from original data

Appendix B

Results

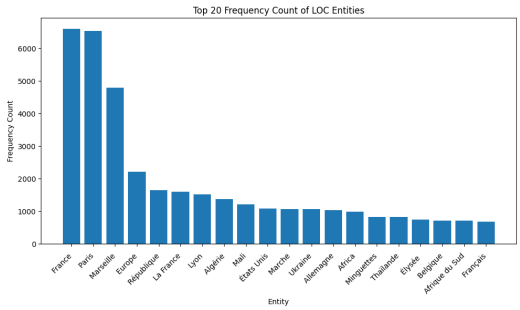
B.1 NER



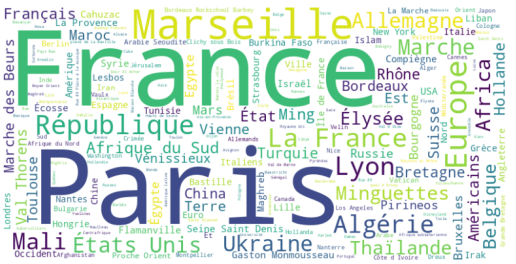
(a) Frequency of top 20 ORG NE extracted

(b) WordCloud ORG NE extracted

Figure B.1: Comparison of ORG NE extraction results

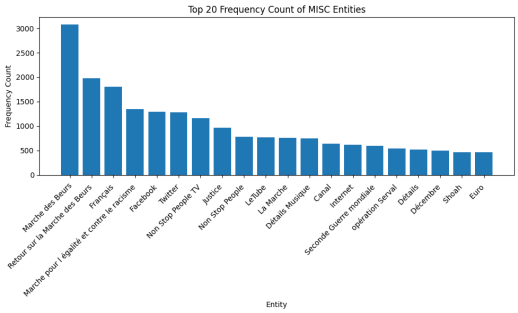


(a) Frequency of top 20 LOC NE extracted



(b) WordCloud LOC NE extracted

Figure B.2: Comparison of LOC NE extraction results



(a) Frequency of top 20 MISC NE extracted

(b) WordCloud MISC NE extracted

Figure B.3: Comparison of MISC NE extraction results

B.2 Topic Modelling

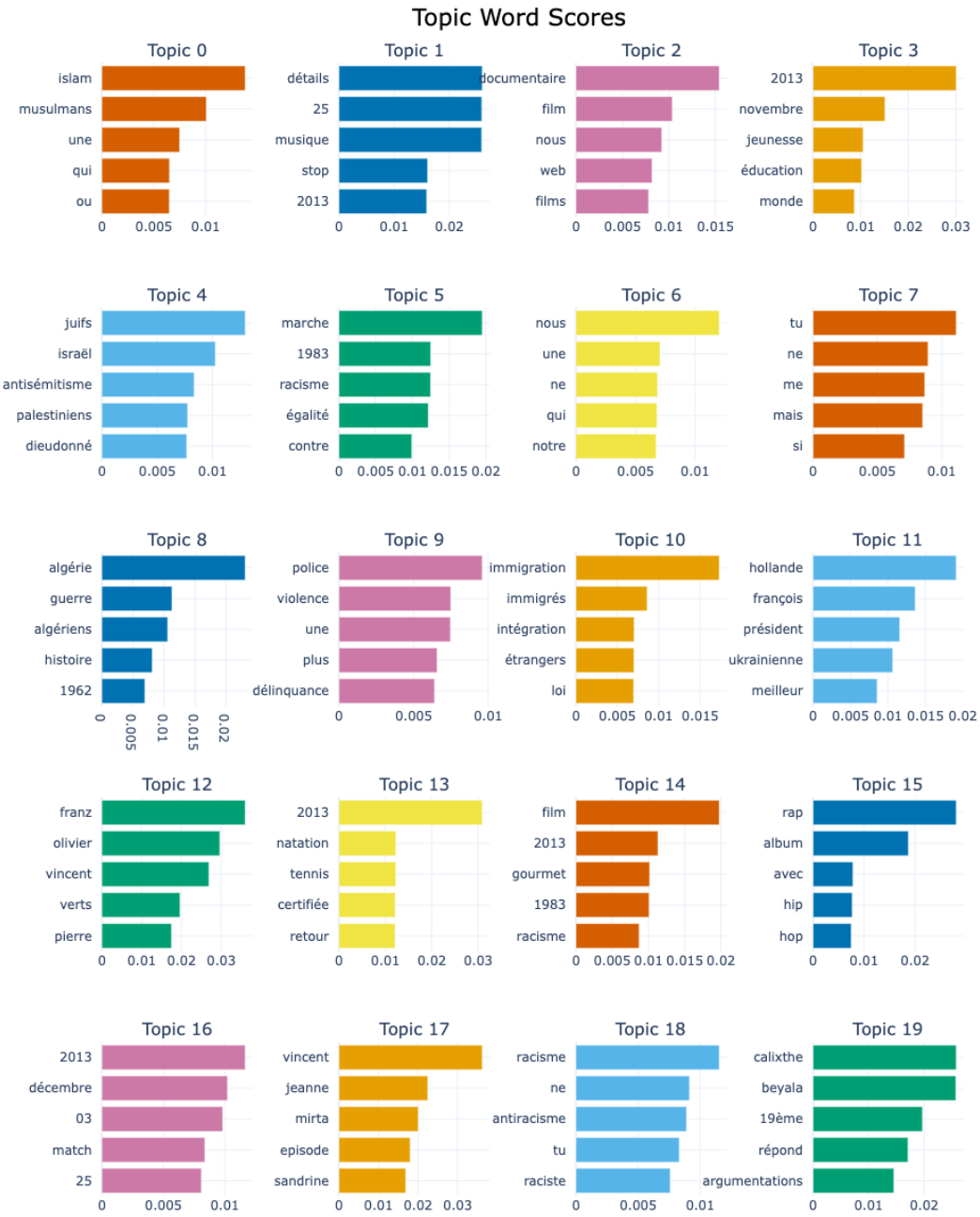


Figure B.4: Top 20 topics extracted before diversification and stopwords removal

Table B.1: Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
0	2517	[('islam', 0.2108), ('musulmane', 0.1578), ('ramadan', 0.1360), ('islamique', 0.1262), ('mosquée', 0.1216), ('signes', 0.1158), ('communauté', 0.1127), ('école', 0.1035), ('autres', 0.1031), ('islamophobie', 0.1025)]
1	2134	[('égalité', 0.1444), ('racisme', 0.1353), ('1983', 0.1351), ('octobre', 0.1210), ('jeunes', 0.1203), ('immigrés', 0.1191), ('marseille', 0.1152), ('collectif', 0.1146), ('mouvement', 0.1125), ('violences', 0.1124)]
2	1766	[('détails', 0.2738), ('musique', 0.2447), ('stop', 0.2320), ('25', 0.2277), ('légales', 0.2054), ('partenaires', 0.1938), ('twitter', 0.1933), ('hanouna', 0.1925), ('réagissez', 0.1909), ('adolescence', 0.1894)]
3	1719	[('documentaire', 0.1982), ('films', 0.1822), ('blog', 0.1517), ('commentez', 0.1415), ('festival', 0.1361), ('caméra', 0.1231), ('production', 0.1231), ('réel', 0.1229), ('email', 0.1229), ('changer', 0.1217)]
4	1699	[('jeunesse', 0.1739), ('indifférence', 0.1733), ('éducation', 0.1715), ('générale', 0.1659), ('publicité', 0.1649), ('changé', 0.1615), ('hôtellerie', 0.1614), ('sports', 0.1531), ('trente', 0.1505), ('guides', 0.1504)]
5	1443	[('démocratie', 0.1458), ('croissance', 0.1336), ('modèle', 0.1233), ('social', 0.1187), ('question', 0.1178), ('élus', 0.1173), ('travail', 0.1170), ('délégation', 0.1155), ('amendement', 0.1124), ('politiques', 0.1113)]
6	1393	[('police', 0.1699), ('article', 0.1404), ('violence', 0.1360), ('intérieur', 0.1279), ('amendement', 0.1264), ('justice', 0.1209), ('ministère', 0.1193), ('rumeur', 0.1155), ('jeunes', 0.1119), ('sécurité', 0.1095)]
7	1187	[('immigration', 0.1993), ('immigrés', 0.1780), ('intégration', 0.1484), ('travailleurs', 0.1457), ('immigré', 0.1226), ('migratoires', 0.1176), ('immigre', 0.1154), ('familles', 0.1144), ('population', 0.1143), ('africains', 0.1121)]
8	1144	[('juifs', 0.2230), ('antisémitisme', 0.1990), ('palestiniens', 0.1986), ('israéliens', 0.1523), ('humoriste', 0.1480), ('paix', 0.1440), ('antisémites', 0.1402), ('football', 0.1395), ('sionisme', 0.1394), ('allemands', 0.1347)]
9	1005	[('guerre', 0.1745), ('1962', 0.1726), ('algérien', 0.1665), ('indépendance', 0.1559), ('histoire', 0.1372), ('mémoire', 0.1309), ('peuple', 0.1253), ('colonial', 0.1240), ('française', 0.1181), ('nationalisme', 0.1174)]
10	826	[('ukrainienne', 0.2034), ('djihadistes', 0.2010), ('sauver', 0.1869), ('fondamentalistes', 0.1862), ('président', 0.1817), ('fiscale', 0.1800), ('charlie', 0.1778), ('françois', 0.1752), ('saoudiennes', 0.1683), ('inutile', 0.1646)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
11	818	[('jaune', 0.1671), ('ha', 0.1558), ('smiley', 0.1471), ('merci', 0.1382), ('meme', 0.1365), ('rien', 0.1290), ('blueman', 0.1222), ('message', 0.1177), ('peux', 0.1141), ('belghoul', 0.1099)]
12	729	[('raciste', 0.1832), ('antiracisme', 0.1758), ('blanc', 0.1745), ('vote', 0.1481), ('répondre', 0.1441), ('oui', 0.1315), ('micnet', 0.1254), ('africains', 0.1248), ('décolonial', 0.1219), ('authentique', 0.1163)]
13	693	[('inconnue', 0.2246), ('février', 0.1950), ('relation', 0.1920), ('royal', 0.1764), ('voter', 0.1737), ('12', 0.1561), ('gauche', 0.1471), ('boycotté', 0.1392), ('ségolène', 0.1318), ('bien', 0.1161)]
14	550	[('hôteliers', 0.2768), ('clauses', 0.2759), ('réservations', 0.2756), ('centrales', 0.2726), ('nuits', 0.2669), ('miniatures', 0.2498), ('parfum', 0.2474), ('flacons', 0.2467), ('jolly', 0.2461), ('1800', 0.2461)]
15	542	[('rap', 0.3036), ('album', 0.2778), ('feat', 0.1911), ('soul', 0.1601), ('arabian', 0.1576), ('video', 0.1568), ('mixtape', 0.1553), ('jazz', 0.1538), ('label', 0.1494), ('solo', 0.1483)]
16	540	[('yadir', 0.2945), ('gourmet', 0.2389), ('film', 0.2254), ('charlotte', 0.1828), ('flux', 0.1539), ('récupérez', 0.1510), ('podcasts', 0.1484), ('égalité', 0.1481), ('1000', 0.1456), ('marcheurs', 0.1451)]
17	537	[('match', 0.2102), ('hier', 0.2038), ('ecologie', 0.1712), ('week', 0.1657), ('5min', 0.1602), ('field', 0.1523), ('38s', 0.1516), ('monaco', 0.1504), ('footballeur', 0.1489), ('suspendu', 0.1454)]
18	470	[('immigration', 0.2183), ('sélectionné', 0.1941), ('suivantes', 0.1908), ('génériques', 0.1499), ('écouter', 0.1477), ('17h00', 0.1437), ('mémoires', 0.1370), ('documentaire', 0.1331), ('université', 0.1329), ('chabani', 0.1327)]
19	469	[('tristitude', 0.4817), ('réinterprète', 0.4021), ('chocolat', 0.3638), ('plateau', 0.3510), ('commentant', 0.2874), ('guests', 0.2872), ('caricaturer', 0.2855), ('lâchez', 0.2795), ('instrumentalisé', 0.2758), ('statuts', 0.2739)]
20	442	[('franz', 0.4372), ('vincent', 0.3805), ('olivier', 0.3526), ('khan', 0.3178), ('particulièrement', 0.2884), ('antigouvernemental', 0.2863), ('couleuvre', 0.2859), ('explicable', 0.2793), ('leonarda', 0.2758), ('manipule', 0.2750)]
21	418	[('définition', 0.3937), ('sceptique', 0.3775), ('pape', 0.3505), ('compiègne', 0.3502), ('diouf', 0.3496), ('propagandiste', 0.3481), ('conscient', 0.3479), ('interrompu', 0.3451), ('développée', 0.3403), ('remarques', 0.3342)]
22	402	[('plaque', 0.3040), ('racisme', 0.2147), ('colère', 0.1884), ('1986', 0.1807), ('commémorative', 0.1723), ('vingt', 0.1639), ('julien', 0.1606), ('étudiant', 0.1551), ('madame', 0.1527), ('association', 0.1497)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
23	399	[('reconnaissiez', 0.3101), ('fraternités', 0.3071), ('comparaison', 0.3070), ('désespérée', 0.3061), ('kiri', 0.3043), ('poubelle', 0.2989), ('samuel', 0.2946), ('socialisme', 0.2900), ('pouvoirs', 0.2894), ('aristocratie', 0.2801)]
24	396	[('manifestants', 0.2312), ('francophones', 0.2272), ('triplent', 0.2271), ('certifiée', 0.2246), ('tennis', 0.2246), ('blogs', 0.2242), ('average', 0.2237), ('opération', 0.2184), ('rugby', 0.2169), ('changé', 0.2153)]
25	383	[('étapes', 0.3704), ('lyonnais', 0.3678), ('capitale', 0.3445), ('32', 0.3314), ('filmera', 0.3246), ('koné', 0.3232), ('monmousseau', 0.3229), ('décida', 0.3222), ('passeront', 0.3204), ('connexions', 0.3193)]
26	367	[('thiriez', 0.4225), ('frédéric', 0.3809), ('football', 0.3512), ('vague', 0.3384), ('basketteurs', 0.3073), ('sport', 0.3055), ('millionnaires', 0.3012), ('clientélisme', 0.2984), ('accomplissements', 0.2981), ('rivale', 0.2936)]
27	353	[('debré', 0.5023), ('rouges', 0.3690), ('contradiction', 0.3659), ('tendances', 0.3269), ('instantané', 0.3094), ('orateurs', 0.3053), ('déficits', 0.3045), ('tyrannie', 0.3044), ('breton', 0.2914), ('reviennent', 0.2829)]
28	349	[('cookies', 0.3381), ('merah', 0.2971), ('navigation', 0.2736), ('acceptez', 0.2715), ('inondations', 0.2139), ('paramètres', 0.2123), ('publicités', 0.1966), ('services', 0.1912), ('statistiques', 0.1793), ('optimale', 0.1781)]
29	335	[('romero', 0.4328), ('lobby', 0.3797), ('chinois', 0.3502), ('différencié', 0.3216), ('anticonceptionnel', 0.3178), ('euthanasie', 0.3061), ('ironisant', 0.3060), ('juge', 0.3047), ('américains', 0.3046), ('manifestant', 0.2964)]
30	333	[('adjectif', 0.4311), ('bruno', 0.3973), ('maire', 0.3653), ('présidence', 0.3232), ('amusé', 0.3064), ('reproches', 0.3033), ('commence', 0.2988), ('admiratif', 0.2951), ('tueurs', 0.2943), ('hollande', 0.2896)]
31	324	[('calixthe', 0.4193), ('19ème', 0.3413), ('argumentations', 0.3370), ('fallacieuses', 0.2681), ('sondage', 0.2678), ('perroquet', 0.2664), ('vôtres', 0.2652), ('moralité', 0.2597), ('apprennent', 0.2555), ('judiciaire', 0.2521)]
32	303	[('candidats', 0.2252), ('circonscriptions', 0.2119), ('présidentielle', 0.2110), ('diversité', 0.2011), ('secrétaire', 0.1833), ('législatives', 0.1820), ('tour', 0.1807), ('socialiste', 0.1702), ('royal', 0.1649), ('campagne', 0.1613)]
34	299	[('certifiée', 0.5780), ('natation', 0.5082), ('rugby', 0.5018), ('blogs', 0.5006), ('tennis', 0.5005), ('sports', 0.4813), ('foot', 0.4545), ('26', 0.3621), ('biologie', 0.2251), ('silence', 0.2114)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
33	299	[('style', 0.3428), ('christine', 0.3317), ('bobard', 0.3204), ('complètes', 0.3199), ('bonds', 0.3196), ('racontez', 0.3180), ('germanophobie', 0.3133), ('pardonnez', 0.3125), ('antiquité', 0.3125), ('chinoise', 0.3079)]
35	285	[('insert', 0.3537), ('number', 0.3506), ('cb', 0.3493), ('delivery', 0.3486), ('programme', 0.3420), ('data', 0.3382), ('blaster', 0.3121), ('left', 0.3101), ('border', 0.2876), ('blank', 0.2868)]
36	276	[('opposition', 0.3272), ('even', 0.3248), ('résumé', 0.3179), ('christopher', 0.3088), ('métaphore', 0.3071), ('idiots', 0.3064), ('légale', 0.3039), ('futures', 0.3033), ('féminisme', 0.2987), ('parisiens', 0.2949)]
37	270	[('chansons', 0.3217), ('album', 0.3101), ('séchan', 0.2433), ('romane', 0.2381), ('gagnant', 0.2021), ('enfer', 0.1928), ('mistral', 0.1917), ('polydor', 0.1841), ('lolita', 0.1830), ('blouson', 0.1799)]
38	267	[('opposition', 0.3272), ('even', 0.3248), ('résumé', 0.3179), ('christopher', 0.3088), ('métaphore', 0.3071), ('idiots', 0.3064), ('légale', 0.3039), ('futures', 0.3033), ('féminisme', 0.2987), ('parisiens', 0.2949)]
39	265	[('henri', 0.3509), ('hollandais', 0.2924), ('républiques', 0.2896), ('meurent', 0.2840), ('amar', 0.2587), ('euro', 0.2532), ('ironique', 0.2408), ('multi', 0.2386), ('partiraient', 0.2343), ('éléctorale', 0.2343)]
40	257	[('jeanne', 0.3822), ('vincent', 0.3747), ('sandrine', 0.3453), ('episode', 0.3333), ('babeth', 0.3286), ('frémont', 0.2863), ('patrick', 0.2541), ('casteygnac', 0.2530), ('convainc', 0.2487), ('refuse', 0.2451)]
41	255	[('rockschool', 0.5334), ('rock', 0.4790), ('barbey', 0.4709), ('marchons', 0.4613), ('électro', 0.4579), ('concert', 0.4109), ('jubilation', 0.3685), ('certes', 0.3654), ('jeudi', 0.3618), ('tubes', 0.3560)]
42	247	[('rock', 0.2515), ('burger', 0.2205), ('punk', 0.2128), ('khaled', 0.2060), ('clash', 0.2056), ('reggae', 0.1996), ('album', 0.1903), ('techno', 0.1825), ('musique', 0.1750), ('soleils', 0.1733)]
43	240	[('détails', 0.4254), ('stop', 0.4214), ('musique', 0.3845), ('légales', 0.3327), ('depardieu', 0.3276), ('contacter', 0.3212), ('z3m3b2whsf', 0.3145), ('taylorswift', 0.3145), ('20h15', 0.3094), ('adolescence', 0.3091)]
45	237	[('1904', 0.2687), ('exposition', 0.2215), ('carole', 0.1987), ('malentendants', 0.1985), ('musulmans', 0.1950), ('guerre', 0.1944), ('reconstruction', 0.1882), ('africains', 0.1851), ('cristallisation', 0.1789), ('sélectionné', 0.1785)]
44	237	[('minutes', 0.2943), ('christianisme', 0.2531), ('diversifiés', 0.2396), ('apocalypse', 0.2376), ('911', 0.2242), ('documentaire', 0.2207), ('francophone', 0.1974), ('gratuitement', 0.1971), ('pétrole', 0.1927), ('bible', 0.1896)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
46	236	[('votation', 0.4734), ('suisse', 0.3816), ('henri', 0.3772), ('légué', 0.3337), ('indisciplinées', 0.3306), ('assertion', 0.3289), ('erronée', 0.3256), ('privilégiée', 0.3063), ('barbares', 0.3058), ('fonctionné', 0.2938)]
47	234	[('nicolas', 0.4000), ('lobbies', 0.3921), ('améliorer', 0.3644), ('antinationale', 0.3373), ('frustre', 0.3367), ('décrivez', 0.3343), ('campe', 0.3204), ('endormir', 0.3183), ('travailleurs', 0.3172), ('interrogée', 0.3132)]
49	229	[('id', 0.4066), ('chaine', 0.3239), ('libelle', 0.2800), ('hierarchie', 0.2794), ('dureeensecondes', 0.2790), ('magazine', 0.2372), ('court', 0.1979), ('episode', 0.1915), ('24mn', 0.1894), ('diffusion', 0.1858)]
48	229	[('washkansky', 0.4525), ('christiaan', 0.4517), ('organe', 0.4426), ('opération', 0.4085), ('1967', 0.4052), ('accident', 0.3918), ('55', 0.3642), ('hôtels', 0.3611), ('étoiles', 0.3528), ('complexe', 0.3398)]
50	226	[('gonzague', 0.4880), ('oh', 0.4670), ('aristocratie', 0.4305), ('terreur', 0.3777), ('bourgeois', 0.3449), ('parlant', 0.3424), ('applaudir', 0.3143), ('ballons', 0.3105), ('gâteau', 0.3101), ('zidane', 0.3075)]
51	222	[('corbière', 0.3691), ('démissionne', 0.3682), ('francis', 0.3488), ('aimez', 0.3227), ('socialisme', 0.3160), ('incompétentes', 0.3004), ('fiscaliste', 0.2971), ('philanthropiques', 0.2964), ('parodie', 0.2951), ('banquiers', 0.2941)]
52	220	[('loldf', 0.5624), ('emissions', 0.5140), ('oreilles', 0.4359), ('front', 0.3241), ('radiophonique', 0.3183), ('sociologue', 0.3155), ('optionnel', 0.3148), ('tactikollectif', 0.2859), ('mobilisations', 0.2842), ('administrateur', 0.2838)]
53	216	[('papiers', 0.3764), ('nice', 0.3436), ('régularisation', 0.2671), ('gymnase', 0.2484), ('neuville', 0.2253), ('maire', 0.2123), ('fontaines', 0.1925), ('déjeuner', 0.1810), ('entendrons', 0.1772), ('élus', 0.1766)]
54	214	[('ajouté', 0.4981), ('assassinats', 0.4191), ('89', 0.4110), ('playlist', 0.3865), ('seul', 0.3426), ('impossible', 0.3230), ('12h33', 0.3176), ('médias', 0.3127), ('partenariat', 0.2712), ('défi', 0.2532)]
55	212	[('matins', 0.6658), ('session', 0.6262), ('vidéo', 0.5221), ('voir', 0.4801), ('culture', 0.4153), ('moondog', 0.2908), ('intégralité', 0.2534), ('2h31', 0.2410), ('contemporain', 0.2404), ('love', 0.2345)]
56	206	[('triplent', 0.3626), ('sommeil', 0.3539), ('accidents', 0.2829), ('inquiétant', 0.2817), ('ukrainien', 0.2798), ('protester', 0.2781), ('opération', 0.2780), ('dessins', 0.2753), ('scientifique', 0.2654), ('masque', 0.2644)]
57	202	[('journal', 0.4369), ('d8', 0.4059), ('rentrée', 0.3702), ('canal', 0.3472), ('audiences', 0.3397), ('idol', 0.3154), ('paye', 0.3140), ('télé', 0.3115), ('singles', 0.3086), ('crochets', 0.3082)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
58	201	[('remplies', 0.3427), ('compliquées', 0.3374), ('financements', 0.3231), ('responsabilité', 0.3197), ('programmaient', 0.3182), ('nu', 0.3171), ('lesbiens', 0.3154), ('bouteille', 0.3149), ('hypocrisie', 0.3132), ('totalitaire', 0.3119)]
59	200	[('geoffroy', 0.4716), ('propositions', 0.3886), ('victime', 0.3194), ('gadgets', 0.3104), ('virez', 0.3104), ('mittal', 0.3104), ('internationaliste', 0.3101), ('homos', 0.3094), ('discriminé', 0.3092), ('disqualifier', 0.3077)]
60	189	[('théorie', 0.3068), ('belghoul', 0.3066), ('retrait', 0.2642), ('enseignants', 0.2322), ('parents', 0.2246), ('education', 0.2081), ('garçons', 0.2020), ('rumeurs', 0.1904), ('masturbation', 0.1880), ('février', 0.1774)]
61	188	[('citizenship', 0.3611), ('political', 0.3422), ('right', 0.2870), ('poet', 0.2665), ('years', 0.2630), ('arabic', 0.2576), ('new', 0.2549), ('access', 0.2546), ('values', 0.2545), ('two', 0.2524)]
62	184	[('forumdesimages', 0.5570), ('goldbronn', 0.5564), ('1451xjt', 0.5515), ('masterclass', 0.5505), ('interactives', 0.5501), ('archive', 0.5232), ('publications', 0.5210), ('aider', 0.4486), ('gagner', 0.4481), ('partenariat', 0.4462)]
63	183	[('opposition', 0.3426), ('résumé', 0.3284), ('geoffroy', 0.3280), ('philippe', 0.3273), ('chanterait', 0.2997), ('cholestérol', 0.2988), ('disqualifier', 0.2977), ('pharmaceutique', 0.2947), ('nullité', 0.2940), ('pen', 0.2936)]
65	176	[('auteures', 0.3043), ('clermont', 0.2526), ('femmes', 0.2106), ('entretien', 0.1896), ('présidente', 0.1884), ('2003', 0.1872), ('association', 0.1842), ('brûlée', 0.1805), ('chauffard', 0.1784), ('sabeg', 0.1744)]
64	176	[('description', 0.2702), ('programme', 0.2590), ('99', 0.2347), ('jazz', 0.2255), ('type', 0.2125), ('vendredi', 0.2122), ('ghetto', 0.2072), ('festival', 0.2053), ('20h30', 0.2026), ('electro', 0.1975)]
67	174	[('musqua', 0.3765), ('courriel', 0.3252), ('masser', 0.3140), ('enthousiaste', 0.3012), ('impossibles', 0.2924), ('défis', 0.2759), ('judith', 0.2596), ('reconstituer', 0.2590), ('oublié', 0.2531), ('matinaux', 0.2193)]
66	174	[('youtube', 0.3966), ('orkut', 0.3847), ('chargement', 0.3015), ('fonctionnalité', 0.2993), ('lyonnais', 0.2745), ('cliquer', 0.2742), ('étapes', 0.2726), ('capitale', 0.2537), ('32', 0.2501), ('monmousseau', 0.2456)]
68	173	[('charlie', 0.3607), ('nekfeu', 0.2365), ('rap', 0.2331), ('censure', 0.1914), ('satirique', 0.1907), ('kilo', 0.1831), ('islam', 0.1782), ('caviar', 0.1727), ('liberté', 0.1599), ('approuvé', 0.1591)]
69	172	[('christiane', 0.3457), ('minute', 0.2333), ('associations', 0.1956), ('villes', 0.1876), ('raciste', 0.1818), ('baromètre', 0.1802), ('opinionway', 0.1782), ('insultes', 0.1687), ('bananes', 0.1624), ('sondage', 0.1603)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
70	169	[('commentaires0', 0.7749), ('streaming', 0.6456), ('direct', 0.6016), ('replay', 0.4063), ('voir', 0.3709), ('déconnecté', 0.3194), ('feel', 0.3072), ('37k', 0.2852), ('movie', 0.2765), ('loto', 0.2313)]
71	164	[('lesbienne', 0.4231), ('pensants', 0.4185), ('aimable', 0.4150), ('edouard', 0.3929), ('adolescent', 0.3823), ('confesser', 0.3376), ('scepticisme', 0.3326), ('martin', 0.3314), ('paradoxal', 0.3263), ('vatican', 0.3218)]
72	162	[('email', 0.4900), ('vérification', 0.3971), ('blog', 0.3928), ('confidentielle', 0.3914), ('échoué', 0.3794), ('goldbronn', 0.3700), ('icône', 0.3684), ('interactives', 0.3644), ('masterclass', 0.3631), ('1451xjt', 0.3626)]
73	161	[('migrants', 0.3835), ('gentlemen', 0.3802), ('plane', 0.3770), ('expose', 0.3569), ('sanctions', 0.3419), ('turquie', 0.3397), ('lesbos', 0.3272), ('législatives', 0.3221), ('mur', 0.3154), ('externes', 0.3043)]
74	160	[('antérieure', 0.6558), ('canope', 0.6529), ('camp', 0.6076), ('inscrire', 0.5778), ('2nd', 0.5557), ('autorisés', 0.5122), ('publicité', 0.4897), ('académie', 0.4608), ('connectant', 0.4469), ('enseignant', 0.4397)]
76	158	[('nova', 0.4143), ('pacifiste', 0.4094), ('mélange', 0.3729), ('radio', 0.3475), ('enregistrais', 0.3469), ('magnéto', 0.3469), ('toxico', 0.3461), ('épatant', 0.3450), ('créatives', 0.3417), ('gymnases', 0.3406)]
75	158	[('marocains', 0.3429), ('casa', 0.1923), ('sam92', 0.1744), ('ghayat', 0.1629), ('royaume', 0.1618), ('lobbying', 0.1617), ('communiste', 0.1613), ('origine', 0.1514), ('inconnu', 0.1498), ('émigrés', 0.1486)]
77	157	[('rabbins', 0.3953), ('flamanville', 0.3950), ('belgique', 0.3803), ('geoffroy', 0.3412), ('propositions', 0.3327), ('victime', 0.3279), ('zoophiles', 0.3207), ('étonnerait', 0.3188), ('internationaliste', 0.3184), ('gadgets', 0.3176)]
78	152	[('clip', 0.2073), ('activé', 0.1874), ('strange', 0.1853), ('npns', 0.1770), ('prostitution', 0.1745), ('tumblr', 0.1655), ('féministes', 0.1648), ('écouter', 0.1593), ('clique', 0.1591), ('podcast', 0.1537)]
80	150	[('a45', 0.2773), ('oules', 0.2249), ('dossiers', 0.2131), ('ppp', 0.2120), ('infrastructure', 0.2070), ('georg', 0.2061), ('autrichienne', 0.1998), ('prison', 0.1980), ('28', 0.1975), ('autoroutes', 0.1953)]
79	150	[('tions', 0.2424), ('tique', 0.2076), ('men', 0.2047), ('ti', 0.2037), ('und', 0.2030), ('prési', 0.1937), ('nation', 0.1881), ('dent', 0.1840), ('cles', 0.1837), ('ver', 0.1826)]
81	146	[('email', 0.3764), ('value', 0.3681), ('code', 0.3595), ('swf', 0.3164), ('message', 0.3111), ('repaire', 0.2869), ('5316bee15e5b', 0.2861), ('playerkey', 0.2861), ('oubliez', 0.2848), ('flash', 0.2806)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
82	144	[('minutes', 0.2525), ('planète', 0.2265), ('1939', 0.2247), ('concentration', 0.2122), ('nazisme', 0.2096), ('pétrole', 0.2056), ('documentaire', 0.2046), ('maggio', 0.2035), ('guerre', 0.2029), ('technologie', 0.1897)]
83	144	[('clés', 0.6029), ('production', 0.5299), ('date', 0.5175), ('lyonenfrance', 0.2565), ('strasbourg', 0.2299), ('20', 0.2123), ('classés', 0.2015), ('brève', 0.1960), ('cinéma', 0.1942), ('marseille', 0.1939)]
84	143	[('cuddy', 0.1999), ('house', 0.1926), ('walt', 0.1877), ('heisenberg', 0.1863), ('commandant', 0.1749), ('marines', 0.1713), ('episode', 0.1679), ('aventures', 0.1652), ('bad', 0.1635), ('avril', 0.1621)]
85	143	[('fpf', 0.3837), ('naissance', 0.3015), ('prêtre', 0.2286), ('benzine', 0.2143), ('protestante', 0.2003), ('newsletters', 0.1720), ('1982', 0.1668), ('antipolis', 0.1659), ('mort', 0.1645), ('écouter', 0.1601)]
86	138	[('plane', 0.2766), ('migrants', 0.2680), ('expose', 0.2618), ('turquie', 0.2541), ('sanctions', 0.2539), ('ei', 0.2498), ('mur', 0.2364), ('législatives', 0.2359), ('modification', 0.2348), ('humanité', 0.2052)]
87	136	[('soviétique', 0.3564), ('censure', 0.3485), ('cuirassé', 0.3279), ('cruche', 0.3269), ('payre', 0.3259), ('personnaliser', 0.3193), ('critiquer', 0.3127), ('valérie', 0.3104), ('débuter', 0.3090), ('curieusement', 0.3083)]
88	135	[('rendons', 0.6877), ('investis', 0.6810), ('barons', 0.6587), ('victimisation', 0.6510), ('sombre', 0.5945), ('égalité', 0.5300), ('41', 0.5223), ('acteurs', 0.4221), ('least', 0.4219), ('plan', 0.3920)]
89	133	[('martin', 0.2649), ('1963', 0.2530), ('freedom', 0.2523), ('droits', 0.2116), ('dream', 0.2067), ('cinquante', 0.2039), ('retrouverons', 0.2014), ('américains', 0.1966), ('telephant', 0.1946), ('organisaient', 0.1920)]
90	128	[('recommandations', 0.7219), ('synopsis', 0.6649), ('commentaires0', 0.6107), ('3ème', 0.4816), ('mauront', 0.4000), ('commémorer', 0.3467), ('biologi', 0.3095), ('atelier', 0.2905), ('reporters', 0.2877), ('jardiniers', 0.2874)]
91	126	[('92', 0.4798), ('27500', 0.4291), ('annemasse', 0.4252), ('amman', 0.4245), ('alexandrie', 0.4128), ('awards', 0.4079), ('jordanie', 0.4019), ('fréquences', 0.3934), ('normale', 0.3830), ('médiamétrie', 0.3266)]
92	123	[('tunisie', 0.3460), ('tunisiens', 0.2668), ('tunisienne', 0.2472), ('tunisien', 0.2165), ('tunisia', 0.2124)]
93	114	[('galite', 0.7024), ('nova', 0.5172), ('refait', 0.5014), ('nadia', 0.4628), ('citoyenne', 0.4570), ('tchatcheur', 0.4418), ('83e', 0.4418), ('dialoguiste', 0.4399), ('saltimbanque', 0.4362), ('hume', 0.4345)]

Continued on next page

Table B.1 Final topics extracted from corpora

Topic	Chunk Count	Representative Keywords
94	112	[('sophia', 0.2887), ('debré', 0.2851), ('maxence', 0.2659), ('connecté', 0.2582), ('fonctionnalité', 0.2202), ('pierre', 0.2103), ('contradiction', 0.2068), ('editions', 0.2054), ('challenge', 0.2012), ('euros', 0.2002)]
95	111	[('externes', 0.5050), ('humanitaire', 0.4301), ('référendum', 0.3707), ('blessée', 0.3650), ('seuls', 0.3598), ('migrants', 0.3588), ('32', 0.3066), ('militants', 0.2885), ('webradio', 0.2577), ('observateurs', 0.2527)]
96	110	[('guaino', 0.3337), ('meurent', 0.2385), ('totalitaire', 0.2353), ('euro', 0.2103), ('programmaient', 0.1952), ('lesbiens', 0.1935), ('partiraient', 0.1920), ('éléctorale', 0.1920), ('mariage', 0.1909), ('anthropologiques', 0.1899)]
97	104	[('editions', 0.4152), ('euros', 0.3594), ('sand', 0.2933), ('burgi', 0.2417), ('musée', 0.2400), ('farhad', 0.2383), ('khosrokhavar', 0.2332), ('lumières', 0.2232), ('exhibition', 0.2217), ('muros', 0.2211)]
98	94	[('myskreen', 0.6964), ('catalogues', 0.5413), ('programmes', 0.5316), ('horaires', 0.4806), ('playlists', 0.4530), ('innovant', 0.4488), ('meilleurs', 0.4350), ('replay', 0.4318), ('streaming', 0.4287), ('préférences', 0.3635)]
99	92	[('exclusives', 1.0204), ('recevez', 0.8944), ('newsletter', 0.7509), ('partenaires', 0.6693), ('alyson', 0.5064), ('d8', 0.4954), ('mother', 0.4932), ('vendredi', 0.4839), ('médias', 0.4682), ('boutons', 0.4174)]
100	91	[('faudel', 0.4966), ('barbès', 0.4289), ('sihem', 0.3819), ('mami', 0.3708), ('philipe', 0.3649), ('orchestre', 0.3400), ('chahim', 0.3237), ('25min50s', 0.3237), ('cafés', 0.3213), ('extraits', 0.3163)]
101	90	[('skyblog', 0.2952), ('rap', 0.2056), ('description', 0.1945), ('billal', 0.1864), ('tchatche', 0.1710), ('sky', 0.1619), ('date', 0.1613), ('francais', 0.1593), ('fé', 0.1544), ('roseraie', 0.1522)]
102	80	[('tv5', 0.4531), ('panafricaine', 0.3729), ('commençaient', 0.3600), ('antifrançais', 0.3577), ('autoritarisme', 0.3533), ('automobiles', 0.3372), ('industries', 0.3320), ('réchauffement', 0.3317), ('kremlin', 0.3140), ('francophone', 0.2898)]
103	79	[('parishad', 0.4216), ('janadesh', 0.3725), ('terre', 0.2824), ('indiens', 0.2822), ('agricole', 0.2791), ('solidaires', 0.2609), ('jaitapur', 0.2316), ('accès', 0.2289), ('hindoue', 0.2237), ('org', 0.2212)]
104	75	[('geoffroy', 0.3278), ('propositions', 0.2746), ('lobby', 0.2559), ('romero', 0.2547), ('gadgets', 0.2495), ('benguigui', 0.2482), ('élargir', 0.2412), ('tentative', 0.2366), ('philippe', 0.2312), ('êtes', 0.2291)]
105	67	[('oh', 0.3302), ('baroin', 0.3096), ('romero', 0.3094), ('lobby', 0.2758), ('anticonceptionnel', 0.2233), ('différencié', 0.2229), ('lesbienne', 0.2224), ('ironisant', 0.2223), ('salade', 0.2221), ('euthanasie', 0.2220)]

Table B.2: Labelled Topics

Topic	Final Label
0	Islam and Muslim Culture
1	Anti-Racism Activism in Marseille in 1983
2	Music, Television and Entertainment
3	Documentary and Film Festivals
4	Youth Education
5	Democracy and Politics
6	Police Violence
7	Immigration and Integration
8	Antisemitism
9	Commemoriation of Algerian War
10	Terrorism and Fundamentalism
11	Miscellaneous
12	Race, Antiracism, and Political Involvement
13	Politics Left
14	Hotel Industry and ads
15	Hip Hop and Music
16	La Marche movie
17	Football
18	Immigration and Memories
19	TV Guests and Political Entertainment
20	Politics
21	Poltical Propaganda
22	Commemoriation and Racism
23	Politics and Socialism
24	Current Events and News
25	Marche and Lyon (Start)
26	Football and Politics
27	Bonnets rouge (Red Caps movement)
28	Website Cookies
29	Lobby, Politics and Social Issues
30	Politics and Public Figures
31	Media and Politc controversary
32	Election Campaign
34	Sport
33	Political Commentary
35	Data Management noise
36	Politics Opposition
37	Music Song
38	Political Commentary
39	Holland and Elections
Continued on next page	

Table B.2 Labelled Topics

Topic	Final Label
40	Television Drama
41	Music Performance
42	Music and Culture
43	Music, Television and Entertainment
45	Historical Events and Exhibitions
44	Religious and Historical Documentariestaries
46	Swiss Politics and Social issues
47	Political Critique
49	Television Programming
48	Medical Breakthroughs
50	French Culture and Society
51	Politics and Criticism
52	Antiracism and Social Engagement
53	Immigration and Social Movements
54	Media and News
55	Video and Culture
56	Media and News
57	Television News
58	Homophobic speech
59	Political Statements and Controversial Remarks
60	Education and Gender issues
61	Citizenship and Political Rights
62	Film Archives and Interactive Events
63	Politics Far-right
65	Women Issues and Activism
64	Music Genres and Programme
67	TV Challenge and Commemoration of the Marche
66	Youtube video
68	Terrorism and Freedom of expression
69	Racism in Society
70	Streaming Video and Comments
71	Homophobic comments
72	Online Communication and Technology
73	Migration and Politics (2015)
74	Educational Resources
76	Radio and Culture
75	Moroccan Community in France
77	Political Controversial commentary
78	Video and Commentary
80	Infrastructure and Public Projects
79	Miscelaneous
Continued on next page	

Table B.2 Labelled Topics

Topic	Final Label
81	Webpage Prompts
82	History and Documentary War
83	Film production news
84	Television series and characters
85	Religion and events
86	Migration crisis and Politics (2015)
87	Censorship and Critiques
88	Victimisation and Activism
89	Civil rights and Social justice
90	Movie recommendations and film synopses
91	Broadcasting, radio frequencies, and awards
92	Tunisian community
93	Equality and social issues
94	Personalities
95	Humanitarian efforts and social issues
96	Poltical Commentary and Social issues
97	Art and culture
98	Streaming and video content
99	Newsletters and media
100	Musical artists and Immigration
101	Online blogs and Comments
102	Media and Politic ideologies
103	Activism and Politics in India
104	Political Proposals and Advocacy
105	Politics, Social Issues, and Satire
-1	Outliers

B.3 NE & Topics Dashboard

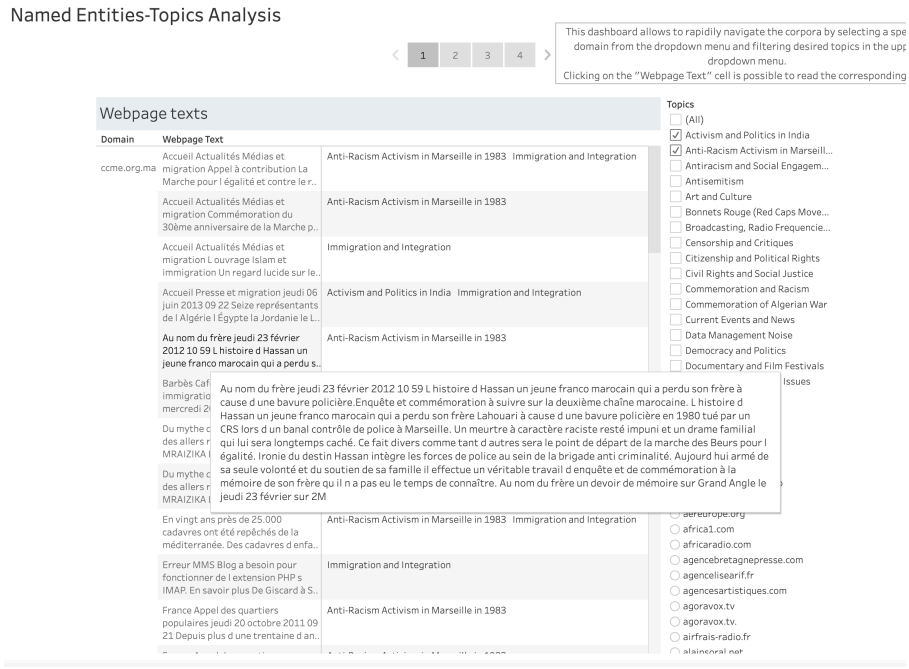


Figure B.5: Dashboard exploring corpora by topics and domain

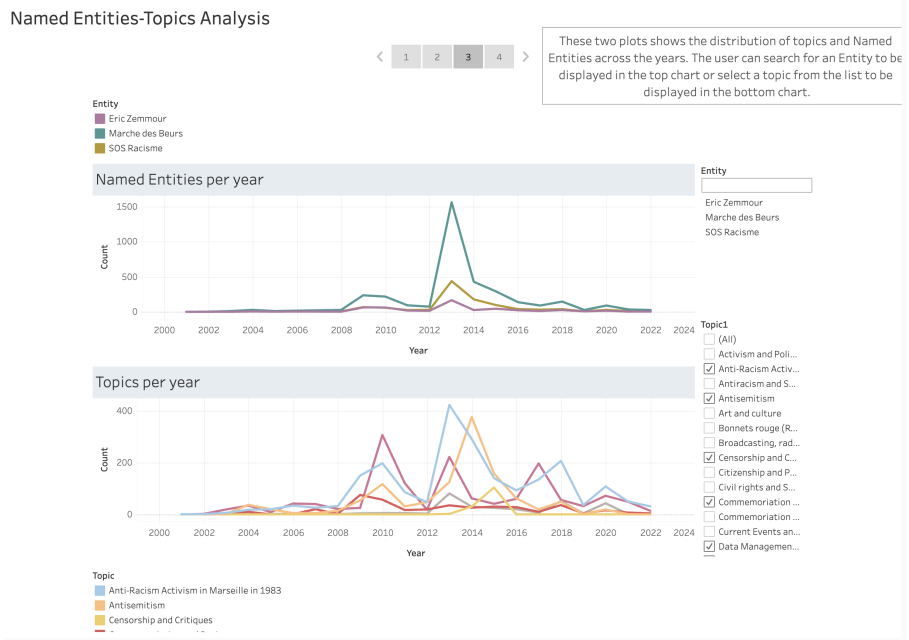


Figure B.6: Dashboard exploring corpora diachronically using Topics and NE

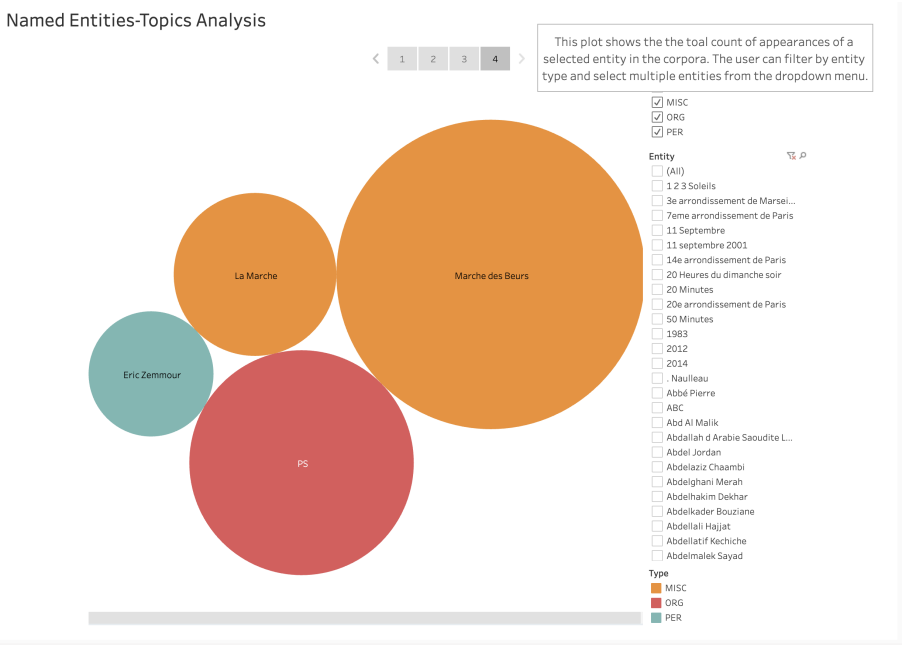


Figure B.7: Dashboard exploring enity total appearances

B.4 Network Analysis

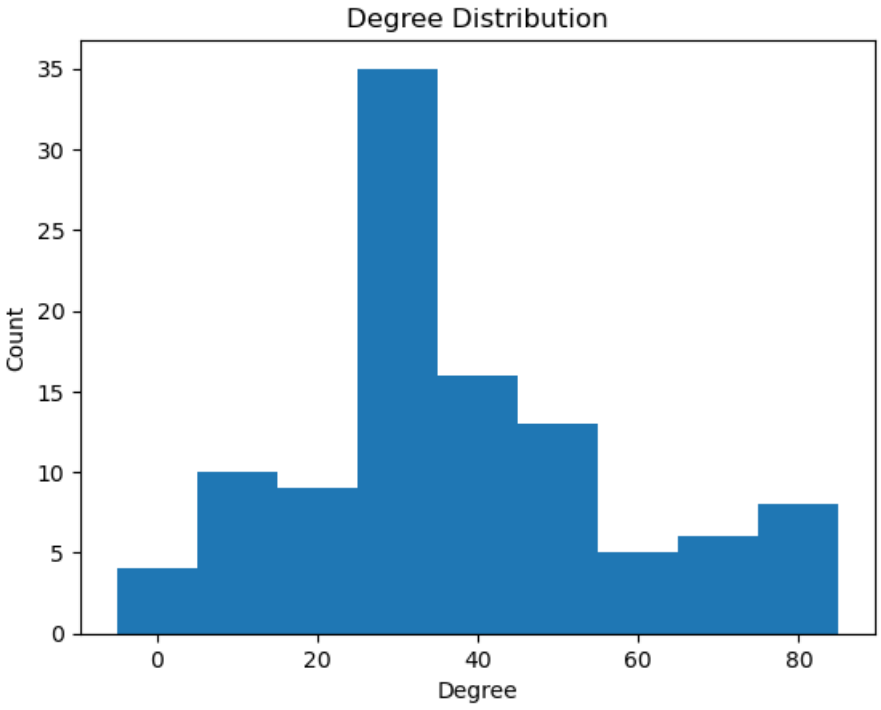


Figure B.8: Degree distribution of topic-topic Network

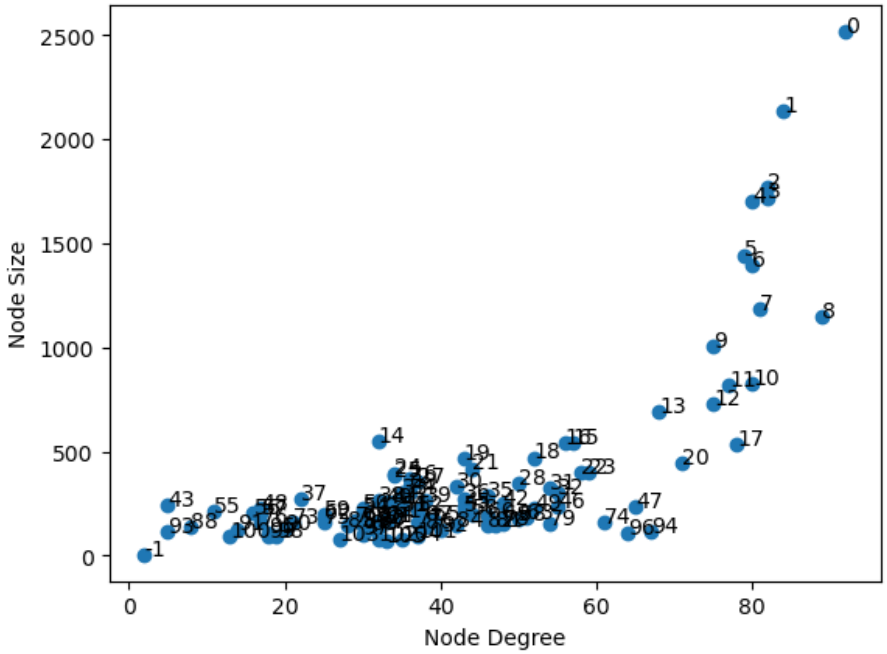


Figure B.9: Relationship between node size and degree in topic-topic network

Table B.3: Jaccard similarity between Louvain communities in topic-topic network

pmi\co-occurrance	Community 0	Community 1
Community 0	0.95	0.01
Community 1	0.00	0.29
Community 2	0.01	0.66
Community 1+2	0.01	0.94

Table B.4: Average edge weight inter and intra Louvain communities

	Community 0	Community 1
Community 0	0.42	0.36
Community 1	0.36	0.45

Table B.5: Communities of topics extracted by Louvain

	Topics
Community 0	['Commemoration of Algerian War', 'Immigration and Memories', 'TV Guests and Political Entertainment', 'Politics', 'Commemoration and Racism', 'Current Events and News', 'Football and Politics', 'Website Cookies', 'Lobby, Politics and Social Issues', 'Politics and Public Figures', 'Election Campaign', 'Data Management Noise', 'Music Song', 'Television Drama', 'Religious and Historical Documentaries', 'Swiss Politics and Social Issues', 'Medical Breakthroughs', 'French Culture and Society', 'Politics and Criticism', 'Television News', 'Homophobic Speech', 'Film Archives and Interactive Events', 'Migration and Politics (2015)', 'Equality and Social Issues', 'Humanitarian Efforts and Social Issues', 'Online Blogs and Comments', 'Activism and Politics in India', 'Marche and Lyon (Start)', 'Music Genres and Programme', 'Political Commentary', 'Moroccan Community in France', 'Migration Crisis and Politics (2015)', 'Political Proposals and Advocacy', 'Streaming Video and Comments']
Community 1	['Outliers', 'Islam and Muslim Culture', 'Anti-Racism Activism in Marseille in 1983', 'Music, Television and Entertainment', 'Documentary and Film Festivals', 'Youth Education', 'Democracy and Politics', 'Police Violence', 'Immigration and Integration', 'Antisemitism', 'Terrorism and Fundamentalism', 'Miscellaneous', 'Race, Antiracism, and Political Involvement', 'Politics Left', 'Hotel Industry and Ads', 'Hip Hop and Music', 'La Marche Movie', 'Football', 'Political Propaganda', 'Politics and Socialism', 'Bonnets Rouge (Red Caps Movement)', 'Media and Political Controversy', 'Sport', 'Political Commentary', 'Politics Opposition', 'Holland and Elections', 'Music Performance', 'Music, Television and Entertainment', 'Historical Events and Exhibitions', 'Political Critique', 'Television Programming', 'Antiracism and Social Engagement', 'Immigration and Social Movements', 'Media and News', 'Political Statements and Controversial Remarks', 'Education and Gender Issues', 'Citizenship and Political Rights', 'Politics Far-right', 'Women Issues and Activism', 'TV Challenge and Commemoration of the Marche', 'Youtube Video', 'Terrorism and Freedom of Expression', 'Racism in Society', 'Homophobic Comments', 'Online Communication and Technology', 'Educational Resources', 'Radio and Culture', 'Political Controversial Commentary', 'Video and Commentary', 'Infrastructure and Public Projects', 'Miscellaneous', 'Webpage Prompts', 'History and Documentary War', 'Film Production News', 'Television Series and Characters', 'Religion% and Events', 'Censorship and Critiques', 'Victimization and Activism', 'Civil Rights and Social Justice', 'Broadcasting, Radio Frequencies, and Awards', 'Personalities', 'Political Commentary and Social Issues', 'Art and Culture', 'Streaming and Video Content', 'Musical Artists and Immigration', 'Media and Political Ideologies', 'Media and News', 'Movie Recommendations and Film Synopses', 'Newsletters and Media', 'Music and Culture', 'Video and Culture', 'Tunisian Community', 'Politics, Social Issues, and Satire']

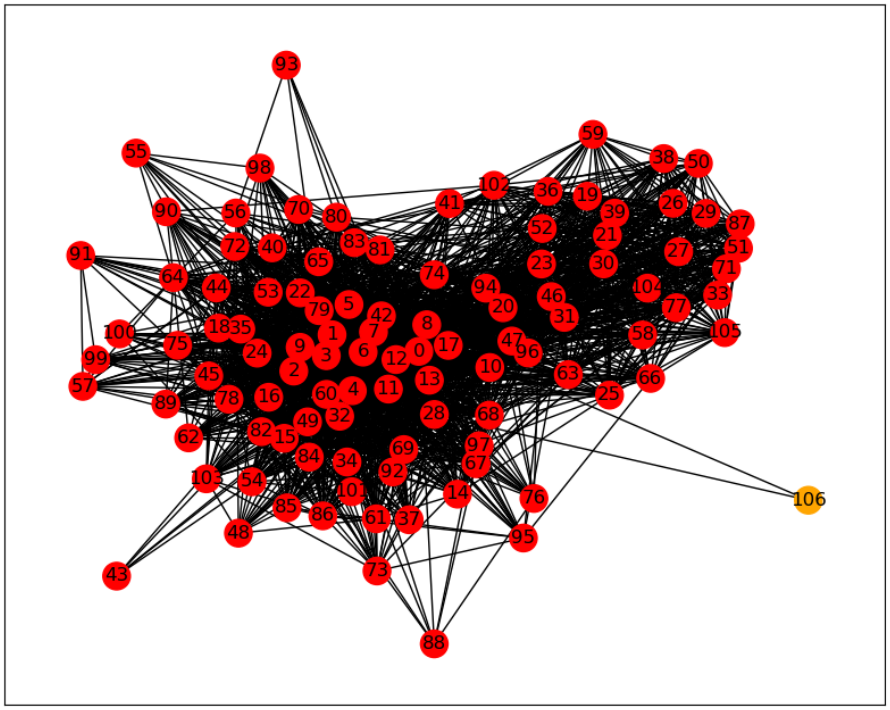


Figure B.10: Girvan-Newman communities in topic-topic Network

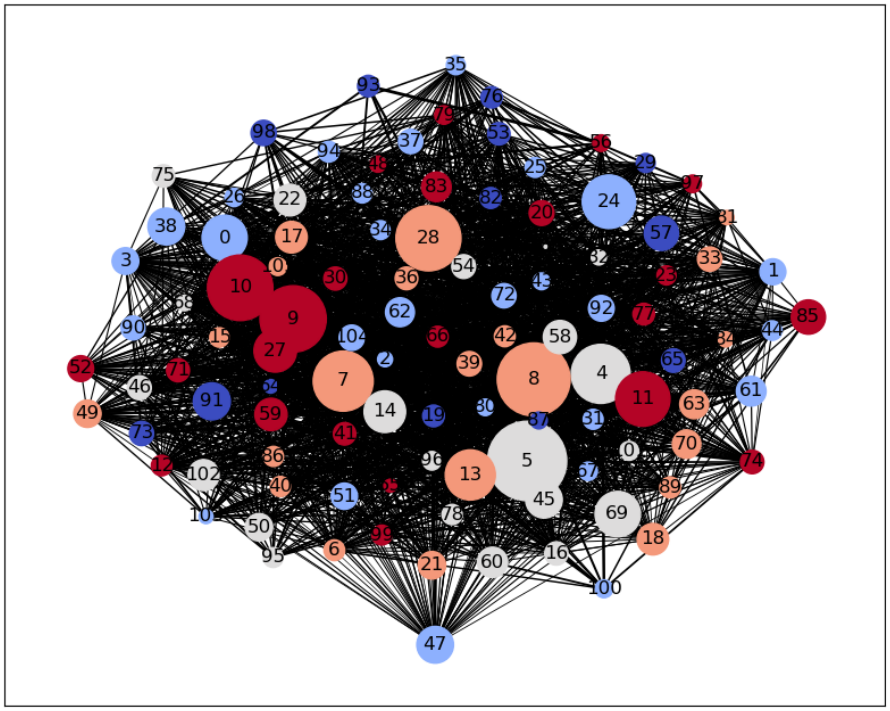


Figure B.11: Louvain communities in Topic-Topic Random Network

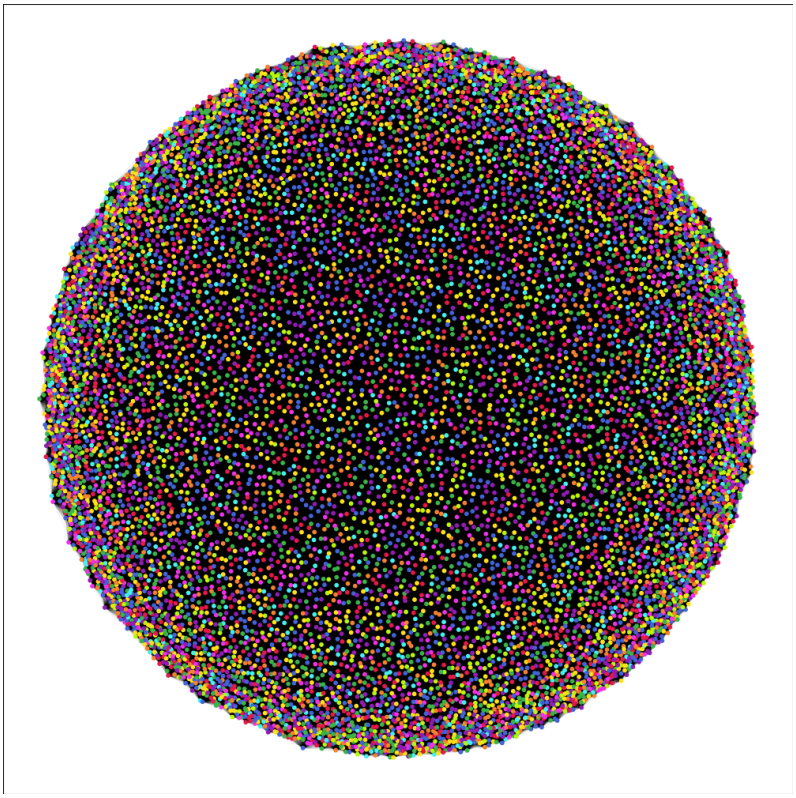


Figure B.12: Louvain communities in Document-Document Random Network

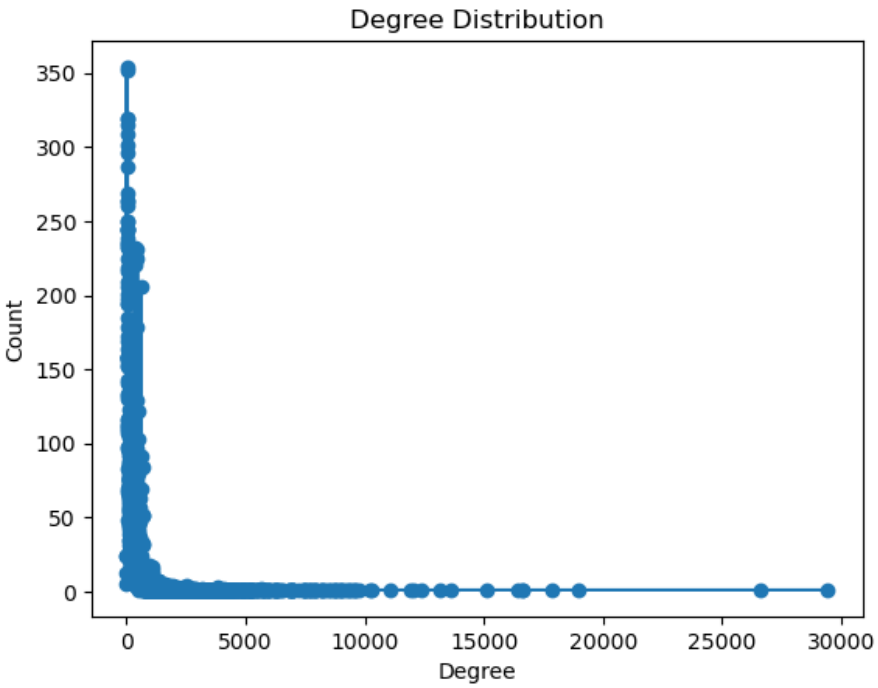


Figure B.13: Degree distribution of enity-entity Network

Appendix C

Domain Knowledge evaluation

C.1 NE dictionary

The dictionary showed in Figure C.1 was provided to domain knowledge experts for filtering and normalisation of the NE extracted by the Babelscape model.

Do NOT CHANGE			to validate		
entity original (dictionary)	cnt		entity noramlized (dictionary)	type	flag
'. I. A. Pentago (LOC)'	1		. I. A. Pentago	LOC	
'1 (LOC)'	1		1	LOC	
'10 (LOC)'	4		10	LOC	
'10e arrondissement de Paris (LOC)'	4		10e arrondissement de Paris	LOC	
'10ème arrondissement de Paris (LOC)'	1		10ème arrondissement de Paris	LOC	
'11 (LOC)'	2		11	LOC	
'11e arrondissement de Paris (LOC)'	2		11e arrondissement de Paris	LOC	
'11eme arrondissement de Paris (LOC)'	2		11eme arrondissement de Paris	LOC	
'11ème arrondissement de Paris (LOC)'	1		11ème arrondissement de Paris	LOC	
'11ème circonscription des Français de l Etranger (LOC)'	1		11ème circonscription des Français de l Etranger	LOC	
'12e arrondissement de Paris (LOC)'	1		12e arrondissement de Paris	LOC	
'13 (LOC)'	1		13	LOC	
'13 arrondissement de Paris (LOC)'	1		13 arrondissement de Paris	LOC	
'13 Nord (LOC)'	1		13 Nord	LOC	
'13e arrondissement de Paris (LOC)'	1		13e arrondissement de Paris	LOC	
'13e circonscription de Paris (LOC)'	1		13e circonscription de Paris	LOC	
'13e circonscription des Hauts de Seine (LOC)'	1		13e circonscription des Hauts de Seine	LOC	

Figure C.1: Named Entities dictionary provided for filtering and normalization

C.2 Topic Labelling dictionary

In Figure C.2 can be seen the excel sheet with the labels given by the two domain knowledge experts, the ones given by ChatGPT and the Final label assigned to each topic. The following prompt was used to ask ChatGPT:

```
"How would you label this topic:  [('corbière', 0.3691), ('démissionne',
0.3682), ('francis', 0.3488), ('aimez', 0.3227), ('socialisme', 0.3160),
('incompétentes', 0.3004), ('fiscaliste', 0.2971), ('philanthropiques',
0.2964), ('parodie', 0.2951), ('banquiers', 0.2941)]"
```

topic_info_reduced		new_topics_dict_reduced 4	Label	Label 2	Label ChatGPT	Final Label	Relevance
0_islam_musulmane_ramadan_islamique	2517	[('islam', 0.2108), ('musulmane', 0.1578), ('ramadan', 0.1360), ('islamique', 0.1262), ('mosquée', 0.1216), ('voiles', 0.1158), ('communauté', 0.1127), ('école', 0.1035), ('autres', 0.1031), ('islamophobie', 0.1025)]	Islam	Islam	Islam and Muslim Culture	Islam and Muslim Culture	yes
1_égalité_racisme_1983_octobre	2134	[('égalité', 0.1444), ('racisme', 0.1353), ('1983', 0.1351), ('octobre', 0.1210), ('jeunes', 0.1203), ('immigrés', 0.1191), ('marseille', 0.1152), ('collectif', 0.1146), ('mouvement', 0.1125), ('violences', 0.1124)]	Marche 1983	Marche 1983	Social Justice and Anti-Racism Activism in Marseille in 1983	Anti-Racism Activism in Marseille in 1983	yes
2_détails_musique_stop_25	1766	[('détails', 0.2738), ('musique', 0.2447), ('stop', 0.2320), ('25', 0.2277), ('légales', 0.2054), ('partenaires', 0.1938), ('twitter', 0.1933), ('hanoune', 0.1925), ('téqjsssc', 0.1909), ('adolescence', 0.1894)]	cinema	NonStopPeople_Tv	Music and Entertainment	Music, Television and Entertainment	no
3_documentaire_films_blog_commentez	1719	[('documentaire', 0.1982), ('films', 0.1822), ('blog', 0.1517), ('commentez', 0.1415), ('festival', 0.1361), ('caméra', 0.1231), ('production', 0.1231), ('réel', 0.1229), ('email', 0.1229), ('changer', 0.1217)]	Documentary	Documentary_Cinema	Documentary Films and Film Festivals	Documentary and Film Festivals	yes
4_jeunesse_indifférence_éducation_généi	1699	[('jeunesse', 0.1739), ('indifférence', 0.1733), ('éducation', 0.1715), ('générale', 0.1659), ('publicité', 0.1649), ('changé', 0.1615), ('hôtellerie', 0.1614), ('sports', 0.1531), ('mental', 0.1505), ('guides', 0.1504)]	Youth	Youth_Education	Youth Education	Youth Education	yes
5_démocratie_croissance_modèle_social	1443	[('démocratie', 0.1458), ('croissance', 0.1336), ('modèle', 0.1233), ('social', 0.1187), ('question', 0.1178), ('élus', 0.1173), ('travail', 0.1170), ('délégation', 0.1155), ('amendement', 0.1124), ('politiques', 0.1113)]	Democraty	Politics_Democracy	Democracy, Growth, and Social Models in Politics	Democracy and Politics	no
6_police_article_violence_intérieur	1393	[('police', 0.1699), ('article', 0.1404), ('violence', 0.1360), ('intérieur', 0.1279), ('amendement', 0.1264), ('justice', 0.1209), ('ministère', 0.1193), ('humeur', 0.1155), ('jeunes', 0.1119), ('sécurité', 0.1095)]	Police	Police_Violence	Police Violence and Justice in Law Enforcement	Police Violence	yes
7_immigration_immigrés_intégration_trav	1187	[('immigration', 0.1993), ('immigrés', 0.1780), ('intégration', 0.1484), ('travailleurs', 0.1457), ('immigré', 0.1226), ('migratoires', 0.1176), ('immigré', 0.1154), ('familles', 0.1144), ('population', 0.1143), ('africains', 0.1121)]	Immigration	Immigration	Immigration and Integration of Workers and Families.	Immigration and Integration	yes
8_juifs_antisémitisme_palestiniens_israé	1144	[('juifs', 0.2230), ('antisémitisme', 0.1990), ('palestiniens', 0.1986), ('israéliens', 0.1523), ('humoriste', 0.1480), ('juifs', 0.1440), ('antisémites', 0.1402), ('football', 0.1395), ('sionisme', 0.1394), ('allemands', 0.1347)]	Antisémitisme_Israël ?	Antisemitism	Antisemitism, Palestine, and Israel	Antisemitism	no
9_guerre_1962_algérien_indépendance	1005	[('guerre', 0.1745), ('1962', 0.1726), ('algérien', 0.1685), ('indépendance', 0.1559), ('histoire', 0.1372), ('mémoire', 0.1309), ('peuple', 0.1253), ('colonial', 0.1240), ('française', 0.1181), ('nationalisme', 0.1174)]	Commémoration Guerre d'Algérie	Commemoration_Algerian War	The Algerian War of Independence and its Historical and Cultural Signifi	Commemoration of Algerian War	yes
10_ukrainienne_djihadistes_sauver_fonda	826	[('ukrainienne', 0.2034), ('djihadistes', 0.2010), ('sauver', 0.1869), ('fondamentalistes', 0.1862), ('président', 0.1817), ('fiscale', 0.1800), ('charlie', 0.1778), ('français', 0.1752), ('saoudiennes', 0.1683), ('multile', 0.1646)]	Dieudonné	Fundamentalists	Inetnational affairs and Security	Terrorism and Fundamentalism	no
11_jaune_ha_smiley_merci	818	[('jaune', 0.1671), ('ha', 0.1558), ('smiley', 0.1471), ('merci', 0.1382), ('meme', 0.1365), ('lien', 0.1290), ('bluemari', 0.1222), ('message', 0.1177), ('peux', 0.1141), ('belghoul', 0.1099)]	Abbas	Miscellaneous	Internet Memes and Expressions	Miscellaneous	no
12_raciste_antiracisme_blanc_vote	729	[('raciste', 0.1832), ('antiracisme', 0.1758), ('blanc', 0.1745), ('vote', 0.1481), ('répondre', 0.1441), ('oui', 0.1315), ('micmet', 0.1254), ('africaine', 0.1248), ('décolonial', 0.1219), ('authentique', 0.1163)]	antiracisme	Racism	Race, Antiracism, and Political Involvement.	Race, Antiracism, and Political Involvement.	yes
13_inconnue_février_relation_royal	693	[('inconnue', 0.2246), ('février', 0.1950), ('relation', 0.1920), ('royal', 0.1784), ('voter', 0.1737), ('12', 0.1561), ('gauche', 0.1471), ('boycotté', 0.1392), ('ségolène', 0.1318), ('bien', 0.1161)]	Segolène Royal	Politics_Left	Unknown Person, Political Relations and Voting.	Politics Left	yes
14_hôteliers_clauses_réservations_centr	550	[('hôteliers', 0.2768), ('clauses', 0.2759), ('réservations', 0.2756), ('centrales', 0.2726), ('nuits', 0.2669), ('miniatures', 0.2498), ('parfum', 0.2474), ('flacons', 0.2467), ('jolly', 0.2461), ('1900', 0.2461)]	hôtelerie	Hôtelerie_Ads	Hotel Industry, Reservations, and Fragrances	Hotel Industry and ads	no
15_rap_album_feat_soul	542	[('rap', 0.3036), ('album', 0.2778), ('feat', 0.1911), ('soul', 0.1601), ('arabian', 0.1576), ('video', 0.1568), ('mixtape', 0.1553), ('jazz', 0.1538), ('label', 0.1494), ('solo', 0.1483)]	Rap	Music	Hip Hop and Music	Hip Hop and Music	no
16_yadir_gourmet_film_charlotte	540	[('yadir', 0.2945), ('gourmet', 0.2389), ('film', 0.2254), ('charlotte', 0.1828), ('flux', 0.1539), ('récupérez', 0.1510), ('podcasts', 0.1484), ('égalité', 0.1481), ('1100', 0.1456), ('marcheurs', 0.1451)]	film	Film	Unrelated or Random Words.	La Marche movie	yes
17_match_hier_ecologie_week	537	[('match', 0.2102), ('hier', 0.2038), ('écologie', 0.1712), ('week', 0.1657), ('5min', 0.1602), ('field', 0.1523), ('38s', 0.1516), ('monaco', 0.1504), ('footballeur', 0.1489), ('suspendu', 0.1454)]	football	Football	Sports and Environmentalism.	Football	no
18_immigration_sélectionné_suivantes_qi	470	[('immigration', 0.2183), ('sélectionné', 0.1941), ('suivantes', 0.1908), ('génériques', 0.1499), ('écouter', 0.1477), ('17h00', 0.1437), ('mémoires', 0.1370), ('documentaire', 0.1331), ('université', 0.1329), ('chabani', 0.1327)]	Mémoires de l'immigration	Immigration_Memories	Immigration and Media	Immigration and Memories	yes
19_1ristitude_réinterprète_chocolat_plateau	469	[('tristitude', 0.4817), ('réinterprète', 0.4021), ('chocolat', 0.3638), ('plateau', 0.3510), ('commentant', 0.2874), ('guests', 0.2872), ('caricaturer', 0.2855), ('lâcher', 0.2795), ('instrumentalisé', 0.2758), ('statuts', 0.2739)]	Zemmour	Commenting_Guests	Humour and Entertainment	TV Guests and Political Entertainment	no
20_franz_vincent_olivier_khan	442	[('franz', 0.4372), ('vincent', 0.3805), ('olivier', 0.3526), ('khan', 0.3178), ('particulièrement', 0.2884), ('antigouvernemental', 0.2863), ('couleuvre', 0.2859), ('explicable', 0.2793), ('leonarda', 0.2758), ('manipule', 0.2750)]	politique	Politics	Unrelated or Random Words.	Politics	no
21_définition_sceptique_pape_compiègne	418	[('définition', 0.3937), ('sceptique', 0.3775), ('pape', 0.3505), ('compiègne', 0.3502), ('dioul', 0.3496), ('propagandiste', 0.3481), ('conscient', 0.3479), ('interrompu', 0.3451), ('développée', 0.3403), ('remarques', 0.33423	Zemmour	Propaganda	Satirical Commentary	Polical Propaganda	no

Figure C.2: Topics file for labelling

Appendix D

Project Management

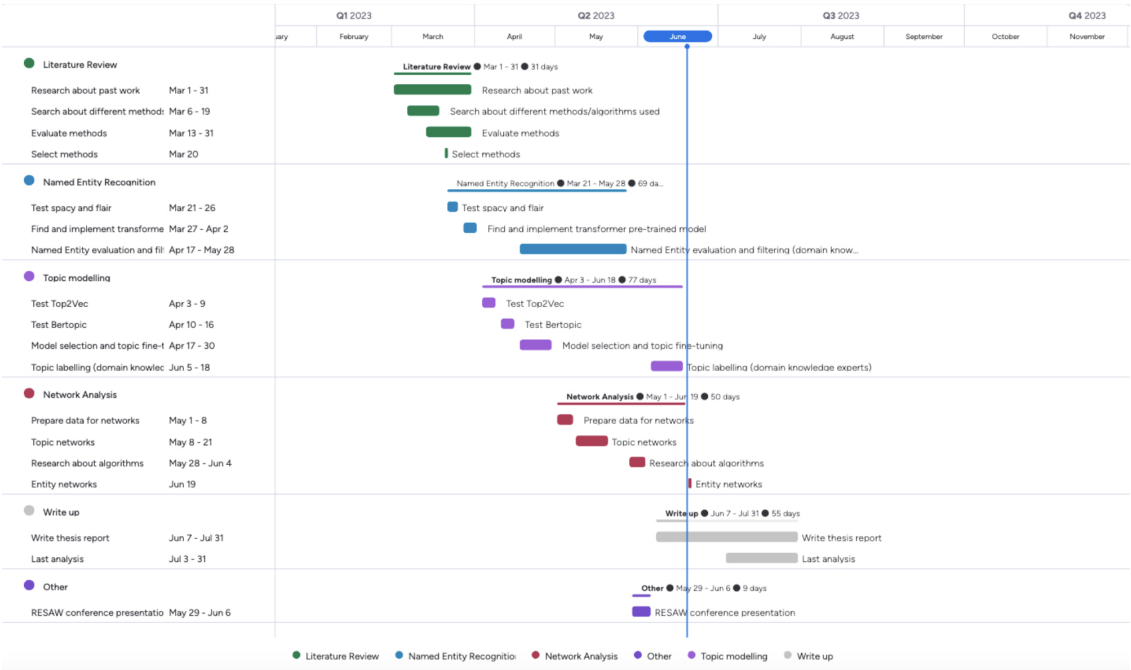


Figure D.1: Project gantt chart