

École polytechnique de Louvain

Textual analysis of belgian newspapers with topic models

Author: **Stéphane LEBLANC**

Supervisor: **Pierre SCHAU**

Readers: **Antonin DESCAMPE, Harold KIOSSOU**

Academic year 2022–2023

Master [120] in Data Science: Information technology

Contents

1	Introduction	3
2	Datasets description	4
2.1	Motivation	4
2.2	Composition	4
2.3	Use cases	5
2.4	Collection process	6
2.5	Distribution	6
3	Latent Dirichlet Allocation	7
3.1	Introduction	7
3.2	Notations	7
3.3	Assumption of LDA	8
3.4	Dirichlet distribution	8
3.5	LDA model : a generative process	10
3.6	Inference and estimation of parameters	12
3.6.1	Variational inference	12
3.6.2	Parameter estimation	14
3.7	Toy example	14
3.8	Online learning	15
4	Text normalization	16
4.1	Punctuation removal	16
4.2	Lowercasing	17
4.3	Tokenization	17
4.4	Stopwords removal	19
4.5	Infrequent terms removal	19
4.6	Lemmatization	19

5	Evaluating topic models	21
5.1	Coherence measure framework	24
5.1.1	Segmentation of word subsets	24
5.1.2	Probability estimation	25
5.1.3	Confirmation measure	25
5.1.4	Aggregation	26
5.1.5	Best results	26
6	Visualizing topic models	28
7	Automatic labelling	34
8	Analysis of newspapers	35
8.1	Comparison of topics	35
8.2	Dominant topic by article	37
8.3	Topics distribution over time	44
8.4	Named entity recognition in topics	46
9	Other use cases of topics models	48
9.1	Dimensionality reduction	48
9.2	Information retrieval	50
10	Conclusion	53
A	Topic models	57
B	Topics distribution over time - heatmaps	68
C	Named entity recognition plots	84

Chapter 1

Introduction

In today's fast-paced world, we find ourselves bombarded with an overwhelming flood of information. As the volume of available data continues to grow exponentially, it becomes increasingly difficult for humans alone to effectively analyze and make sense of it all. This is where the power of artificial intelligence (AI) comes into play. By harnessing the capabilities of AI, we gain the ability to efficiently analyze large text corpora within a relatively short span of time. The objective of this master's thesis is to conduct a comprehensive textual analysis of different Belgian newspapers, with the aim of comparing and contrasting their content. While there exists a myriad of AI models specifically designed for text analysis, our research focused specifically on the utilization of Latent Dirichlet Allocation (LDA). LDA is an unsupervised learning technique capable of deciphering complex patterns within corpora by generating topics, which can be interpreted as categories for articles. By employing LDA, we were able to discern meaningful topics from the newspaper corpora, enabling us to explore the distribution and evolution of specific subjects over time or any other relevant criteria. With the foundation set, this master's thesis starts by providing a detailed description of the datasets used and the methodology employed to collect them. Furthermore, we outline the preprocessing techniques applied to the datasets to ensure optimal training of the models. Subsequently, we dive into the details of the Latent Dirichlet Allocation model itself. Additionally, we present techniques for evaluating and visualizing the LDA model, enabling a comprehensive analysis of the newspapers in the next chapter. Finally, we explore other potential use cases for topic models such as dimensionality reduction and information retrieval. The code utilized for this project is available on GitHub. ¹

¹<https://github.com/Balzard/memoire-topic-modeling>

Chapter 2

Datasets description

To the best of our knowledge, there is currently no existing dataset available for Belgian newspapers, or at least none that we are aware of. Consequently, we took the initiative to create individual datasets for each Belgian newspaper we focused on in our research. This chapter aims to provide a comprehensive description of these datasets and the underlying motivation behind their creation.

2.1 Motivation

The data itself plays a pivotal role in the creation of a machine learning model. A poorly constructed dataset or one that is inappropriately applied to the intended application will inevitably yield subpar results. Therefore, it is crucial to document a dataset to ensure that potential users are fully informed about its strengths, weaknesses, and appropriate contexts for its usage. To address this concern, we adopted the approach proposed by [8]. This involves providing a datasheet for each dataset, which outlines the rationale behind its creation, its composition, the methodology employed for data collection, recommended use cases, and other pertinent information. The datasets utilized in this master's thesis were created for the analysis of Belgian newspapers.

2.2 Composition

The datasets exclusively consist of unprocessed text data, which introduces some noise. While such occurrences are rare, there are instances where web pages deviate from the typical structure, resulting in missing or additional noisy data. For example, there might be extraneous text following an article that is unrelated to its content. Furthermore, the datasets may lack article introductions, as some articles solely consist of core text, or the newspaper does not differentiate between the

introduction and the core text. Similarly, article categories can occasionally be missing. Moreover, there is a possibility of some articles being absent, as certain entries were found to be absent during the collection process, rendering the datasets incomplete. Additionally, it is important to note that only articles published on the primary websites of each newspaper were collected. For instance, LeSoir has a separate platform called Geeko, and articles found on LeSoir that redirect to the Geeko website were not included in the datasets.

Regarding the time coverage of the datasets, they encompass articles within a specific time interval, spanning from 2020 to april 2023. Although a broader time range would have been desirable, web scraping limitations posed challenges. While certain newspapers, such as RTBF, maintain webpages listing all articles published by day from 2008 to the present, other newspapers do not provide such comprehensive archives. As a result, the collection of articles is constrained by the available time period.

Below are the respective numbers of documents/articles contained in each dataset:

	DH	Lavenir	RTBF	LeSoir	LaLibre
#documents	256 363	231 486	118 489	99 422	123 007

The following is an example instance extracted from the RTBF dataset, providing a representative snapshot of its content:

Title	Coronavirus en Belgique : le récapitulatif de toutes les mesures ...
Content	Vous avez un petit peu perdu le fil suite à l'annonce des dernières mesures ? Nous avons décidé de répertorier ...
Date of publication	2020-10-10
Premium	False
Categories (as given by the newspaper itself)	[Bon à savoir, Suggestion de la rédac, Coronavirus, Info, Coronavirus, Covid19]

2.3 Use cases

The datasets could be used for any unsupervised learning algorithm. It's also well suited for classification algorithms in order to classify articles by category.

2.4 Collection process

The collection process of the datasets was entirely automated and did not rely on human assistance. Python, along with the BeautifulSoup library, was employed to scrape each document from the respective corpora. The process was optimized using multithreading techniques to enhance efficiency. To provide insight into the methodology, a code template was utilized to generate the corpus for each newspaper. A dataframe was saved for every processed URL, and subsequently, all the dataframes were merged to form a comprehensive dataset for each newspaper, encompassing the collected articles.

```
import concurrent.futures, requests
from bs4 import BeautifulSoup

# save in a dataframe all articles from an url
def process_url(url: str) -> None:
    ...
    df.to_csv("path")

urls = collect_all_urls_needed()

with concurrent.futures.ThreadPoolExecutor(max_workers=30) as executor:
    for _ in executor.map(process_url, urls):
        pass
```

2.5 Distribution

All datasets are planned to be available on this platform : <https://dataverse.uclouvain.be/>

Chapter 3

Latent Dirichlet Allocation

3.1 Introduction

In this chapter, we dive into the explanation of the Latent Dirichlet Allocation (LDA) model. LDA is a probabilistic model widely utilized for uncovering hidden topics within document collections. Essentially, a topic can be understood as a cluster, where each topic is characterized by a multinomial distribution over the words in the dataset's vocabulary. Consequently, certain words within a topic are more likely to occur compared to others, effectively representing the essence of that topic. For instance, consider a topic with words like "Trump," "White House," and "Biden" as the three most probable words. It can be inferred that this topic revolves around USA policy. A document can be associated with a high probability for a specific topic if it contains words that are likely to occur within that topic, such as "Trump" and "Biden" in our example. However, it's important to note that a document can belong to multiple topics, albeit with varying probabilities. A document is essentially represented as a combination of finite mixtures of topics, represented by multinomial distributions. The strength of LDA lies in its ability to autonomously learn the topics within a document collection, without the need for manual annotation or hand-labeling of the corpus. It can effectively cluster together similar documents, allowing for unsupervised topic discovery.

3.2 Notations

Classical notations for *natural language processing* are used:

- A *vocabulary* is a set of words with size denoted as V .
- A *word* is an element of a vocabulary. It's represented by a very known *one hot encoded vector* with 1 at the position of the word in the vocabulary and

0 everywhere else.

- A *document* is a sequence of N words written as $\mathbf{w} = (w_1, w_2, \dots, w_N)$
- A *corpus* is a collection of M documents written as $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$

3.3 Assumption of LDA

- Some words belong to many topics but with different probability.
- The topics probability distribution of a document is more likely to contain some topics than others.
- The words probability distribution of a topic is more likely to contain some words than others.

Exchangeability

The LDA model operates under the "bag-of-words" assumption, which means that the order of words within a document is considered irrelevant. Similarly, this assumption extends to the documents themselves, implying that their ordering or sequence does not hold significance in the model.

3.4 Dirichlet distribution

A Dirichlet random variable θ of dimension k can take any value in the $k-1$ simplex ($\theta_i \geq 0, \sum_{i=1}^k \theta_i = 1$) and has the following probability density on the simplex:

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = p(\theta|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (3.1)$$

where $\boldsymbol{\alpha}$ is a vector of size k with each components $\alpha_i > 0$ and Γ is the gamma function. The Dirichlet distribution can be considered as a multivariate extension of the Beta distribution which can be viewed as a probability distribution of probabilities, specifically as a probability distribution for the success of a Bernoulli distribution. It allows us to quantify the fairness of a coin, for instance. The distribution is characterized by two shape parameters, α and β , which can be regarded as the results of many coin flips (e.g: $\alpha = 100, \beta = 100 \rightarrow 100$ heads and 100 tails). Here is the probability density function where $B(\alpha, \beta)$ is a normalization constant :

$$f(\theta; \alpha, \beta) = \frac{\theta^{(\alpha-1)}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \quad (3.2)$$

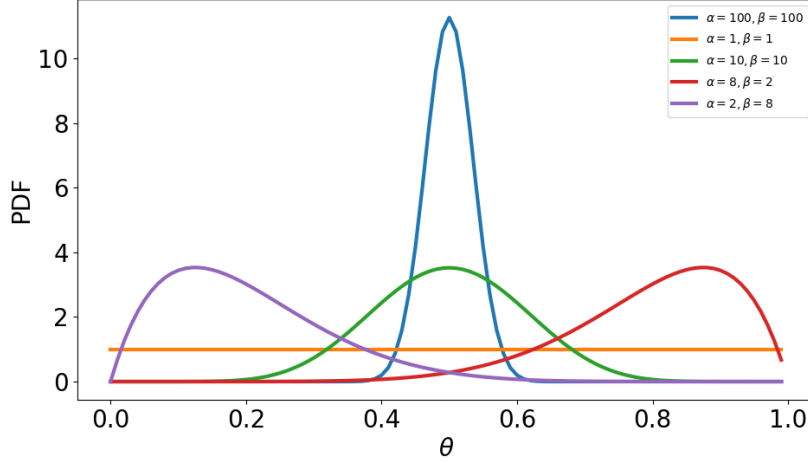


Figure 3.1: Beta distribution

When considering the Beta distribution, it becomes evident that greater values for α and β indicate a higher amount of information available, resulting in increased confidence in the fairness of a coin. Conversely, if α and β differ from each other, it implies the presence of a bias in the coin. This same principle can be extended to the α_i parameters in the Dirichlet distribution. Assigning a value of 1 to the parameters signifies a lack of prior knowledge or information. Conversely, when certain α_i values are greater than others, it signifies a larger proportion assigned to those specific α_i values. This relationship is visually represented in Figure 3.2, highlighting the varying proportions captured by the different α_i values.

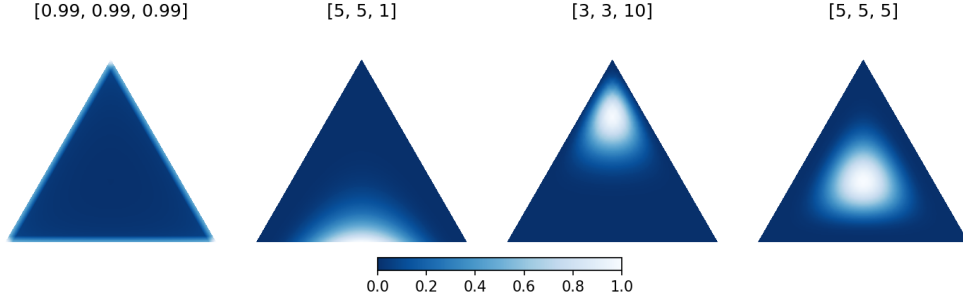


Figure 3.2: Dirichlet distributions from [9]. The parameters above each plot represent $\alpha_1, \alpha_2, \alpha_3$ respectively. Color indicates the probability density function.

Additionally, it is noteworthy that the Dirichlet distribution serves as a conjugate prior for the multinomial distribution. Consequently, the posterior distribution of the multinomial distribution exhibits the same form as the prior distribution, i.e., the Dirichlet distribution. This property is particularly advantageous for inference and parameter estimation of LDA, as it enables the estimation of the posterior distribution based on a known distribution.

3.5 LDA model : a generative process

LDA assumes that each document $\mathbf{w}$ in a corpus D was generated as follow:

1. Choose $\theta \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For each word w_n of document $\mathbf{w}$
 - (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - (b) Choose a word w_n from $p(w_n|z_n, \mathbf{B})$, a multinomial probability.

The size K of the vector $\boldsymbol{\alpha}$, and so the dimensionality of Dirichlet distribution θ is assumed to be fixed (it's a hyperparameter of the model i.e. the number of topics). Assuming this generative process for documents, we want to reverse engineering it. That is to say, find the latent variables of the model :

- θ_m is the probability distribution of a document $\mathbf{w}_m$ over the K topics.
- $\mathbf{B}$ is a $K \times V$ matrix where each row is a topic and each column is a word so each cell is the word probability for a topic k i.e. $\mathbf{B}_{ij} = p(w_j|z_i)$.
- $z_{m,n}$ is the topic assignment of the n -th word of document $\mathbf{w}_m$.

Figure 3.3 is a graphical example of this generative process for a document :

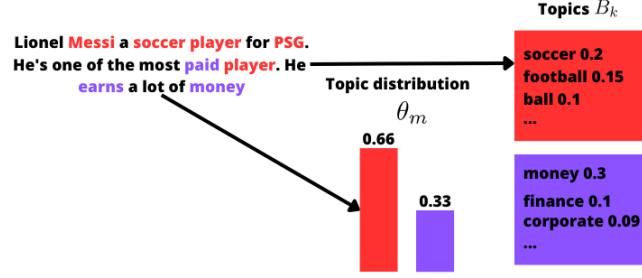


Figure 3.3: LDA generative process

There are three different levels in the generative process.

- Firstly, the parameters α and B are sampled once in generative the process.
- Secondly, the variables θ_m are sampled once for each of the M documents.
- Finally, the variables $z_{m,n}$ and $w_{m,n}$ are sampled once for each words of the M documents.

Figure 3.4 shows the geometry of the latent space of a document. LDA operates in the space of distributions over the words. Each one can be viewed as a point on the word simplex. The three points of the topic simplex are three different distributions over the words.

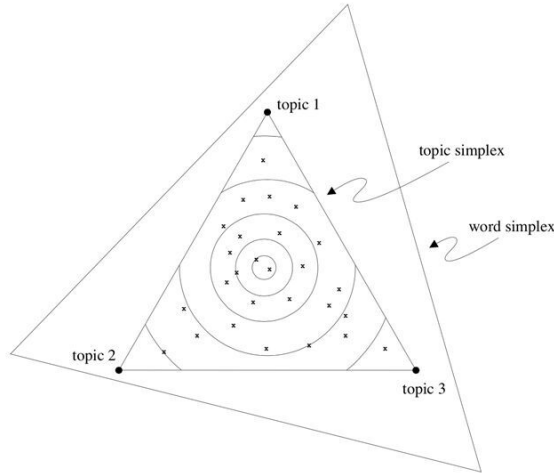


Figure 3.4: simplex for 3 topics embedded in the words simplex for 3 words from [2]

3.6 Inference and estimation of parameters

Given α and $\mathbf{B}$, the joint distribution of a topic mixture θ , a set of N topics $\mathbf{z}$ and a set of N words $\mathbf{w}$:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \mathbf{B}) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \mathbf{B}) \quad (3.3)$$

Marginal distribution of a document $\mathbf{w}$:

$$p(\mathbf{w} | \alpha, \mathbf{B}) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \mathbf{B}) \right) d\theta \quad (3.4)$$

The main challenge in LDA is the computation of the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \mathbf{B}) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \mathbf{B})}{p(\mathbf{w} | \alpha, \mathbf{B})} \quad (3.5)$$

Sadly, the distribution involved in LDA is intractable, meaning that it cannot be analytically computed. Consequently, we need to resort to approximation techniques. Several algorithms are available for this purpose, such as Laplace approximation, variational inference, and Markov Chain Monte Carlo. However, in this context, we will focus on variational inference, as it was the method originally employed in the original paper on LDA [2], as well as in the library we utilize for modeling the newspapers, called Gensim [20].

3.6.1 Variational inference

Variational inference aims to provide lower bounds, indexed by a collection of variational parameters, to approximate the log likelihood of a distribution. The authors of [2] derived this distribution as lower bound :

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta, \gamma) \prod_{n=1}^N q(z_n, \phi_n) \quad (3.6)$$

where the Dirichlet parameter ψ and the multinomial parameters $(\phi_1, \dots, \phi_n)$ are the free variational parameters. The subsequent step involves the optimization of the variational parameters. This is accomplished by minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior distribution $p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \mathbf{B})$. The Kullback-Leibler divergence serves as a metric to quantify the dissimilarity or distance between two distributions. The computation of this divergence is as follows:

$$D_{KL}(q||p) = q(\theta, \mathbf{z}|\gamma, \phi) E_q \left[\log \left(\frac{q(\theta, \mathbf{z}|\gamma, \phi)}{p(\theta, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \mathbf{B})} \right) \right] \quad (3.7)$$

Intuitively, 2 distributions are close if their sample results are similar. And we want to minimize this divergence to obtain the best variational parameters:

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D_{KL}(q||p) \quad (3.8)$$

[2] shows that this can be achieved via an iterative fixed point method. They obtain the following update equations for the variational parameters:

$$\phi_{ni} \propto \mathbf{B}_{i w_n} \exp(E_q[\log(\theta_i)|\gamma]) \quad (3.9)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni} \quad (3.10)$$

The expectation part of equation 3.9 can be computed as :

$$\exp[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right) \quad (3.11)$$

where Ψ is the Digamma function. It's worth noting that the variational distribution varies as a function of $\mathbf{w}$ and then the variational parameters (γ^*, ϕ^*) are document-specific. Moreover, $\gamma^*(\mathbf{w})$ can be viewed as a representation of a document in the topic simplex.

Algorithm 1 Variational inference algorithm for LDA

```

init  $\phi_{ni}^0 := 1/k$  for all i and n
init  $\gamma_i := \boldsymbol{\alpha} + N/k$  for all i
while not converged do
  for  $n = 1$  to  $N$  do
    for  $i = 1$  to  $k$  do
       $\phi_{ni}^{t+1} := \mathbf{B}_{i w_n} \exp(\Psi(\gamma_i^t))$ 
    end for
    normalize  $\phi_n^{t+1}$  to sum to 1
  end for
   $\gamma^{t+1} := \boldsymbol{\alpha} + \sum_{n=1}^N \phi_n^{t+1}$ 
end while

```

3.6.2 Parameter estimation

Given a corpus D , we want to find the parameters α and B that maximize the log likelihood of the data:

$$l(\alpha, \beta) = \sum_{d=1}^M \log p(\mathbf{w}_d | \alpha, B) \quad (3.12)$$

As a reminder, $p(\mathbf{w} | \alpha, B)$ is untractable. Fortunately, variational inference gives us a tractable lower bound that we can maximize with respect to α and B . The procedure used to approximate the true posterior distribution is Expectation Maximization. It alternates between maximizing the lower bound with respect to γ and ϕ and, for fixed value of these two parameters, maximizing the lower bound with respect to α and B .

1. (E-step) For each document $\mathbf{w}_d$, find the optimized variational parameters $\{\gamma_d^*, \phi_d^* : d \in D\}$
2. (M-step) Maximize the lower bound on the log likelihood with respect to α and B .

This is repeated until the lower bound on the log likelihood converges.

3.7 Toy example

In this section, we aim to provide a toy example of the EM algorithm, specifically on the E-step with the variational inference algorithm detailed in Algorithm 1. Here is a toy example with 3 topics and 5 words in the vocabulary:

$$B = \begin{bmatrix} covid & test & football & but & film \\ 0.4 & 0.2 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.4 & 0.2 & 0.1 \\ 0.2 & 0.4 & 0.2 & 0.2 & 0.1 \end{bmatrix} \quad (3.13)$$

Suppose the first document being : **covid test covid** with $\alpha = (0.1, 0.1, 0.1)$. Then $\gamma = (1.1, 1.1, 1.1)$ following line 2 of Algorithm 1. We begin with the first word of our document, **covid** :

$$\phi(0) = 0.4 \times \exp(\Psi(1.1) - \Psi(1.1 + 1.1 + 1.1)) = 0.093 \quad (3.14)$$

$$\phi(1) = 0.2 \times \exp(\Psi(1.1) - \Psi(1.1 + 1.1 + 1.1)) = 0.046 \quad (3.15)$$

$$\phi(2) = 0.2 \times \exp(\Psi(1.1) - \Psi(1.1 + 1.1 + 1.1)) = 0.046 \quad (3.16)$$

After normalization : $\phi_0 = (0.5, 0.25, 0.25)$. Then do the same for **test** and **covid** which gives :

	ϕ_0	ϕ_1	ϕ_2
covid	0.5	0.25	0.25
test	0.25	0.25	0.5
covid	0.5	0.25	0.25

We need then to update γ :

$$\gamma_0 = 0.1 + 0.5 + 0.25 + 0.5 = 1.35 \quad (3.17)$$

$$\gamma_1 = 0.1 + 0.25 + 0.25 + 0.25 = 0.85 \quad (3.18)$$

$$\gamma_2 = 0.1 + 0.25 + 0.5 + 0.25 = 1.1 \quad (3.19)$$

We do all these steps until convergence for all documents.

The next step (M-step) is to maximise the lower bound with the optimised variational parameters on the log likelihood with respect to α and β . This step for β corresponds to :

$$B_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j \quad (3.20)$$

The update for α can be implemented using a Newton-Raphson method where :

$$\frac{\partial L}{\partial \alpha_i} = M \left(\Psi \left(\sum_{j=1}^k \alpha_j \right) - \Psi(\alpha_i) \right) + \sum_{d=1}^M \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^k \gamma_{dj} \right) \right) \quad (3.21)$$

3.8 Online learning

The LDA algortihm implemented in Gensim is the online version of LDA. It has been introduced by [10]. It is different from the standard batch learning where parameters are updated after learning the whole dataset. Online learning updates parameters for each new data. It's then more suitable for large datasets and constant incoming data. The authors showed that their algorithm approximate the posterior as well as the original variational algorithm but in a fraction of time.

Chapter 4

Text normalization

In the realm of Natural Language Processing (NLP), it is customary to begin by preprocessing the text before training a model. This crucial step involves transforming the text into a more suitable format that enhances model performance. Fortunately, the corpora utilized in this context are inherently less noisy, with the exception of noise introduced during the collection process. Since journalists typically adhere to proper language conventions, it is unlikely to encounter slang, spelling mistakes, emojis, and similar artifacts within the articles. This inherent cleanliness of the data simplifies our task, as an excessive number of spelling mistakes would complicate word recognition. Nevertheless, even though our documents exhibit less noise compared to sources such as tweets, it remains essential to clean any existing noise. This process serves to improve model performance, enhance interpretability, minimize complexity, and mitigate the risk of overfitting. To achieve this, we employed the Spacy Python library [11] and utilized the `fr_core_news_md` pipeline. This pipeline facilitated tokenization of the texts and filtering of the tokens based on specific categories such as nouns, proper nouns, adjectives, verbs, numbers, and adverbs. By employing these preprocessing steps, we ensure that the data is in an optimal state for subsequent modeling and analysis, ultimately leading to improved performance and more reliable insights. In the rest of this chapter, we list all the transformations used to normalize the text.

4.1 Punctuation removal

The question of whether to retain or remove punctuation in documents is one of the primary considerations in text processing. In the field of computer science, the answer to this question, as with many others, is often *it depends*. The appropriate course of action may vary based on the specific context and requirements of the task at hand. For instance, in an N-gram language model aiming to predict the

next word in a given sentence, retaining punctuation marks such as periods is crucial, as they indicate the end of a sentence. In the case of analyzing tweets, it may be desirable to preserve symbols like the # symbol. However, in our scenario, where punctuations are not of significant importance, the emphasis lies on selecting meaningful words to represent the corpus effectively. That said, there are instances where it is beneficial to retain internal punctuation. Take, for example, the word "Covid-19". In such cases, preserving the internal punctuation allows for the accurate representation of these specialized terms.

4.2 Lowercasing

Lowercasing is a common preprocessing step in many NLP applications. Its purpose is to ensure that the capitalization of the first letter of a word does not affect its meaning in most cases. For instance, both "Football" and "football" refer to the same sport, so we treat them as identical words. While there are exceptions to this rule, they are relatively rare. An example would be the word "rose" (referring to the flower) and "Rose" (referring to the first name) which have distinct meanings but are generally insignificant in the context of newspaper articles. Detecting such exceptions can be challenging, and it is generally more beneficial to convert all words to lowercase. Otherwise, our vocabulary would be larger, leading to a more complex model that may unnecessarily duplicate words in topics.

4.3 Tokenization

A classical approach which is almost used before training of every language models, is tokenization. It consists of segmenting a text into smaller units called token according to a set of rules. It can be either words, characters, subwords or sentences.

Word tokenization

Word tokenization splits a text into a set of words. For example, the sentence : "Hello, how are you ?".

The simplest word tokenizer would basically segment this sentence by using the white spaces between words. It's quite straightforward and is then subject to approximations. The first default is that it does not take punctuation into account. Here, the tokenized sentence would be :

["Hello," , "how" , "are" ,"you?"]

We would probably like to separate *you* from ? and so not having them as one token. Therefor, our tokenizer should take punctuation into account. Otherwise,

our model would learn different representations of a same word if it is often followed by a punctuation symbol. The vocabulary size would then grow. Moreover, we would certainly like to get elements like *New York* as a single token, and it would be nice to separate *don't* as ["do", "n't"]. This introduces more complexity to the tokenization rules. It's worth noting that such tokenization can be challenging for large text corpora, resulting in a substantial vocabulary size if the corpus contains many unique words. Another challenge is dealing with out-of-vocabulary words, which are words appearing in the test set but not in the training set. The common solution is to map these unknown tokens to the special "UNK" token, although this approach leads to the loss of specific word information.

Character tokenization

Character tokenization involves splitting a text into individual characters. For example, the sentence "Hello there !" would be tokenized as:

['H', 'e', 'l', 'l', 'o', ' ', 't', 'h', 'e', 'r', 'e', ' ', '!', '']

This approach overcomes the issue of out-of-vocabulary words because even if a word is not in the training set, it can still be represented in terms of its constituent characters. It also limits the size of the vocabulary as it consists of single characters. However, character tokenization increases the input sentence length significantly, making it more challenging for the model to learn the relationships between characters to form meaningful words.

Subword Tokenization

Subword tokenization involves dividing a text into meaningful subwords. It operates on the principle that common words should remain intact, while rare words can be decomposed into meaningful subword units. For example, the word "annoyingly" can be decomposed into "annoying" and "ly". By representing the word using these subwords, which appear more frequently, the meaning is preserved. This technique allows for a manageable vocabulary size while capturing word relationships effectively. Additionally, it enables the model to process unknown words by breaking them down into subwords it already knows.

In this work, we employ word tokenization. Our datasets have reasonable sizes, so the vocabulary does not become excessively large. Furthermore, we remove stopwords and infrequent terms to further refine the tokenization process.

4.4 Stopwords removal

Once the text has been tokenized, there may still be words that contribute little information, commonly known as "stopwords." In the French language, examples of stopwords include "le," "la," "de," and so on. These words do not hold significant meaning for most NLP applications, so it is considered good practice to exclude them. While there are numerous stopwords lists available for various languages, there isn't a definitive standard list to use. Although words like "la" can often be removed without consequence, the importance of certain stopwords may vary depending on the specific domain or context in which we are working.

4.5 Infrequent terms removal

In addition to removing stopwords, it is common practice to eliminate infrequent words. This is because words that appear very rarely do not contribute significant information when computing document similarity. Furthermore, removing infrequent words helps reduce the size of the vocabulary, resulting in faster computations and decreased complexity. One straightforward approach to achieve this is to discard all words that appear below a certain frequency threshold. Although this method can be effective, choosing the appropriate threshold is not straightforward, and there is no universal rule of thumb. To determine the threshold, an analysis of the token frequencies and the tokens themselves can be performed. For instance, tokens that appear only eight times or less may predominantly represent noise introduced during the data collection process. Therefore, removing all words that appear less than eight times could be a suitable approach for a given dataset. This kind of process has been done for removing infrequent terms.

4.6 Lemmatization

Lemmatization is a linguistic process to reduce words to their base or canonical form, known as the lemma. The lemma represents the dictionary or base form of a word, from which all inflected or derived forms can be generated. The purpose of lemmatization is to normalize words so that variations of a word with different inflections or derived forms are treated as the same word. For example, the words "run," "running," and "ran" would all be reduced to the lemma "run" through lemmatization.

Lemmatization takes into account the grammatical aspects of a word and applies morphological analysis to determine the lemma. It considers factors such as the word's part of speech (noun, verb, adjective, etc.), tense, number, and gender. By reducing words to their lemmas, lemmatization enables better analysis and

understanding of text by reducing the vocabulary size and grouping together words with the same semantic meaning.

Compared to stemming, which simply removes suffixes to achieve word normalization, lemmatization produces more accurate and meaningful lemmas by considering the word's linguistic context and rules. Indeed, stemming is subject to approximations:

1. under-stemming: failing to get the same word stems for words with same conceptual meaning.
Example: divide $\rightarrow$ divid and division $\rightarrow$ divis.
2. over-stemming: producing same word stems for words with different conceptual meaning.
Example: wand $\rightarrow$ wand and wander $\rightarrow$ wand.

However, lemmatization is generally computationally more expensive than stemming due to its linguistic analysis and reliance on language-specific rules. It still remains more accurate so we have used the lemmatization process of the `fr_core_news_md` pipeline of Spacy [11].

Chapter 5

Evaluating topic models

When evaluating topic models, one of the primary objectives is to assess their interpretability, as these models do not inherently guarantee interpretability. Inadequate preprocessing, for example, can result in topics where words like "the," "have," and other less meaningful words are identified as the most probable words for a given topic. Consequently, such topics become challenging to interpret, and it becomes unclear what their core subject is. Evaluating the semantic coherence of topics is therefore crucial.

The research literature has proposed numerous methodologies to evaluate topic models, but the question of evaluation remains open. One such approach, suggested by [4], is the word intrusion method. This approach involves randomly introducing "intruder words" into topics and asking human users to identify them. The assumption is that coherent topics will make it easier to identify intruder words compared to incoherent topics. Thus, the interpretability of a topic can be assessed by measuring the ease of identifying intruder words. However, this approach relies on manual annotation and lacks automation. Consequently, there is a need to develop automated methods for evaluating topic models.

A classical and straightforward approach to automatically evaluate language models is by using likelihood. This method involves splitting the dataset, as is often done, into a training set and a testing set. The testing set comprises unseen documents $\mathbf{w}_m$. By calculating the log likelihood of the test set, we can compare different models. The log likelihood measures the probability of the entire test set according to the model. Therefore, a higher likelihood indicates a better-performing model.

$$L(\mathbf{W}) = \log_2 p(\mathbf{W}|\alpha, \beta) = \sum_{i=1}^M \log_2 p(w_i|\alpha, \beta) \quad (5.1)$$

But usually, we use the perplexity measure, which is the inverse of the geometric mean per-word likelihood.

$$\sqrt[M]{\prod_{i=1}^M \frac{1}{p(w_i|\alpha, \beta)}} \quad (5.2)$$

Perplexity is a useful metric for model selection and adjusting hyperparameters, such as the number of topics, in topic modeling. However, its main limitation is that it does not consider semantic information. Therefore, even if we choose the model with the highest perplexity, there is no guarantee of topic interpretability. In fact, studies have shown a negative correlation between perplexity and human interpretability [4].

To address this issue, researchers have proposed coherence measures that evaluate topics based on their coherence to human understanding. In a study by [19], a coherence measure based on pointwise mutual information was introduced. Among various metrics, this measure was found to be the most similar to human judgment. It operates on the intuition that similar words tend to co-occur frequently. It quantifies how often two events (words w_i and w_j here) occur, compared with what we would expect if they were independent:

$$\text{PMI}(w_i, w_j) = \max \left(\log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}, 0 \right) \quad (5.3)$$

Probabilities are estimated by treating the corpus as a single meta-document and using a sliding window of 10 words. Single word probabilities are estimated as follow:

$$p(w_i) = \frac{C(w_i)}{M} \quad (5.4)$$

where M stands for the number of words in the corpus and $C(w_i)$ for the number of times the word w_i appears in the corpus. Co-occurring words probabilities are estimated based on word co-occurrence counts. ϵ is a smoothing factor to avoid to compute the log of 0.

$$p(w_i, w_j) = \frac{C(w_i, w_j) + \epsilon}{\sum_{i=1}^V \sum_{j=1}^C (C(w_i, w_j) + \epsilon)} \quad (5.5)$$

For a topic $\mathbf{t}$, they computed its PMI score as did [18]:

$$\text{PMI-score}(\mathbf{t}) = \text{median}\{\text{PMI}(w_i, w_j), i, j \in 1...10\} \quad (5.6)$$

So for the 10 most probable words of a topic, they compute the median of the PMI for all pair words, which is less affected by outliers. It's also known as the UCI

metric.

Another coherence measure called Umass was proposed by different researchers [17]. In their approach, experts were asked to assess the quality of each topic in a topic model, categorizing them as good, intermediate, or bad. For topics identified as intermediate or bad, the researchers identified the following issues:

- **Chained:** every word is connected to every other words but not all word pairs make sense.
- **Intruded:** there are some intruder words or unrelated sets of words.
- **Random:** no clear semantic connections between more than a few pair of words.
- **Unbalanced:** top words are semantically connected to each other but there are also very general terms that give less understandability.

The researchers proposed a method to identify chained, intruded, and random bad topics using a metric based on word co-occurrence, similar to the UCI metric. They observed that in chained and intruded bad topics, pairs of words belonging to the same concept tend to co-occur within a single document, while word pairs from different concepts do not. On the other hand, in random topics, there is a low probability of word co-occurrence. Taking these observations into account, they developed the Umass coherence metric, which calculates probabilities based on document frequency:

$$C_{Umass}(\mathbf{t}) = \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \quad (5.7)$$

So it's very similar to UCI metric except that $P(w_i)$ is not used. The reason is that this metric is more close to their expert annotations than PMI. Note that Umass is not symmetric.

[14] proposed an approach to automate the word intrusion technique created by [4]. It's based on the methodology proposed by [13] where we can learn the best topic word that summarises the semantic of a topic. From that, the word intrusion task, so detecting the least representative word of a topic, becomes the converse of the best topic word task. It works as follow: for each topic, compute the word association features for each of the N top words of a topic, and then combine the features in a ranking support vector regression model to learn the intruder words. From 4 words association measures, they found PMI as the most correlated with the human score in [4]:

$$\text{PMI}(w_i) = \sum_j^{N-1} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \quad (5.8)$$

Note that this technique can also be adapted to a topic level by grouping all topics for a dataset (without distinguishing the models that produce them) and calculate the correlation with human judgment but the result is worse for the model level.

5.1 Coherence measure framework

For this study, we will utilize the framework introduced [21], which is already implemented in Gensim [20]. The authors present a framework that enables the construction of word-based coherence measures by combining various elementary components according to our preferences. With this approach, we have the flexibility to create both existing coherence measures and novel ones. The framework is structured into four distinct parts

5.1.1 Segmentation of word subsets

Coherence of a set of words tries to evaluate how well single words or subsets of words fit together. The initial step in the framework involves determining the desired approach for segmenting the word set associated with a particular topic. While several coherence measures, like UCI coherence, evaluate pairwise relationships between individual words, alternative segmentation methods can also be employed. Here are a few examples of possible segmentations for a given word set, denoted as W :

$$S_{one}^{one} = \{(W', W*) | W' = \{w_i\}; W* = \{w_j\}; w_i, w_j \in W; i \neq j\} \quad (5.9)$$

$$S_{all}^{one} = \{(W', W*) | W' = \{w_i\}; w_i \in W; W* = W \setminus \{w_i\}\} \quad (5.10)$$

$$S_{any}^{one} = \{(W', W*) | W' = \{w_i\}; w_i \in W; W* \subseteq W \setminus \{w_i\}\} \quad (5.11)$$

$$S_{set}^{one} = \{(W', W*) | W' = \{w_i\}; w_i \in W; W* = W\} \quad (5.12)$$

S_{one}^{one} is the segmentation of W where each word is compared with every other word. S_{all}^{one} compares every single word to all other words of the word set. S_{any}^{one} extends S_{all}^{one} by using every subset as condition. S_{set}^{one} compares words to the complete word set W using words context vectors

5.1.2 Probability estimation

This section focuses on determining the approach for estimating word probabilities based on our corpus. Multiple methods can be employed for this purpose. For instance, in the case of the UCI measure, word counts can be calculated using a sliding window technique that moves over the documents with a step size of one word. Alternatively, the *Boolean document* method can be utilized, where the probability of a single word is estimated by dividing the number of documents in which the word occurs by the total number of documents. The joint probability of two words is then estimated by dividing the number of documents containing both words by the total number of documents. Similarly, the *Boolean paragraph* and *Boolean sentence* methods operate on the same principle, but consider paragraphs and sentences as individual documents, respectively.

5.1.3 Confirmation measure

In this section, we determine the approach for calculating the similarity between words or word subsets, denoted as $S_i = (W', W^*)$, based on their respective probabilities. This can be achieved either directly or indirectly.

Direct confirmation

These types of measures directly calculate the similarity for an individual pair S_i . For example, in the case of UCI coherence, the PMI measure can be utilized. Alternatively, other measures such as the log-conditional probability measure used in Umass can also be employed:

$$score(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W^*)} \quad (5.13)$$

Another measure, known as the Jaccard measure, can also be used:

$$score(S_i) = \frac{P(W', W^*)}{P(W' \vee W^*)} \quad (5.14)$$

These are just a few examples, and there are numerous other measures available, although we won't list them all here.

Indirect confirmation

Indirect confirmation operates under the assumption that when considering a word in set W , the direct confirmation measures of words in set W' should be similar to the direct confirmation measures of words in set W^* for that particular word. This implies that although two words, such as z and x may have low joint probability due

to infrequent co-occurrence, their confirmation measures can still exhibit correlation with respect to other words in the corpus.

For example, consider two competing car brands that are semantically related but rarely appear together. Despite the low joint probability, their confirmation measures can strongly correlate with other words in the corpus, such as "road" or "wheel." In this case, indirect confirmation captures the semantic similarity between the two words, which might be missed by direct confirmation.

This can be done by representing the word sets W' and W^* as vectors with dimension equal to the total size of W . Such vectors can be computed with any confirmation measure. If W' is a single word, the vector elements correspond to the direct confirmation measures with every other words of W . If W' is a set of words, the vector elements are the sum of the direct confirmation measures of the individual words.

$$\vec{v}_{m,\gamma}(W') = \left\{ \sum_{w_i \in W'} \text{score}(w_i, w_j)^\gamma \right\}_{j=1, \dots, |W|} \quad (5.15)$$

Given two vectors $\vec{u} = \vec{v}(W')$ and $\vec{w} = \vec{v}(W^*)$ for the pair $S_i = (W', W^*)$, we can compute indirect confirmation as vector similarity. Below are the cosine and jaccard similarity:

$$s_{\cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \quad (5.16)$$

$$s_{jac}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} \min(u_i, w_i)}{\sum_{i=1}^{|W|} \max(u_i, w_i)} \quad (5.17)$$

5.1.4 Aggregation

Given confirmations of all pairs S_i , we aggregate them as a single coherence score. The median σ_m , as in UCI coherence, and arithmetic mean σ_a are used in the literature but we can use other aggregation measure such as geometric mean σ_g or quadratic mean σ_q .

5.1.5 Best results

So, for example, if we wanted to represent the UCI measure with the framework, it would look to something like this:

$$C_{UCI} = (S_{one}^{one}, P_{sw}(10), \text{PMI}, \sigma_m) \quad (5.18)$$

$$C_{U_{mass}} = (S_{pre}^{one}, P_{booldoc}, \text{Umass score}, \sigma_a) \quad (5.19)$$

The authors conducted numerous benchmark tests to evaluate different combinations of coherence measures. Each test utilized a dataset accompanied by topics and human ratings for each topic. Specifically, for a word set representing the top words of a topic, human participants provided ratings based on their understandability. Additionally, each combination of coherence measures generated a ranking for a topic, which was then compared to the human ranking using Pearson correlation. A high correlation with human ratings indicated a good coherence measure. On average, the C_v measure demonstrated the best performance:

$$C_v = (S_{set}^{one}, P_{sw}(10), m_{cos}(\text{NPMI}, 1), \sigma_a) \quad (5.20)$$

This measure also yielded the best results when applied to the New York Times articles dataset, which is similar to our datasets. Therefore, we will adopt this coherence metric to evaluate our models.

Visualizing topic models

Although automated evaluation of topic models is possible, it is still beneficial to conduct a quick human check for validation purposes. One simple approach is to examine the list of most frequent terms for each topic. This method is effective as it allows users to quickly understand how words are organized within topics. Another popular method for visualizing word lists is through the use of word clouds. In addition to presenting the words themselves, word clouds also incorporate the size of words to reflect their probability within the topic. An example of a word cloud is shown in Figure 6.1, where "match" and "club" are the most probable words. However, it is important to note that word clouds lack of contextual information between words. Although there may appear to be connections between closely placed words, their arrangement in the cloud is random and does not indicate any inherent relationships



Figure 6.1: Word cloud

Alternatively, word associations can be used to establish connections between words, offering an alternative visualization method. One commonly used measure for assessing word associations is Pointwise Mutual Information (PMI). By calculating

the PMI for each term in the set of most probable words for a given topic, we can identify pairs of words with a PMI value above a certain threshold, effectively linking them together. An example of this visualization approach is depicted in Figure 6.2. In this visualization, the area of each word circle is proportional to its probability in the topic. Unlike word clouds, the visual representation is not affected by word length. However, it is important to note that a threshold needs to be defined in order to determine which word associations to include.

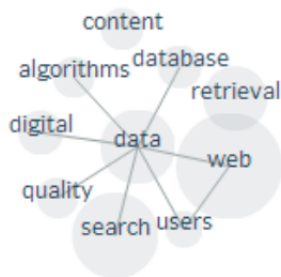


Figure 6.2: Word associations visualization from [3]

Anyway, these methods lack of comparison between topics themselves. Indeed, it would be nice to see how topics relate to each other. This can be achieved by the *Termite* [5] and *LDavis* [22] visualization.

Termite

Termite is an additional visual analysis tool that can be used to assess the quality of topic models. It employs a matrix layout to compare terms both within and across topics. Unlike other methods, Termite does not rely solely on word probabilities to select words for visualization. Instead, it utilizes a measure called *saliency* to identify more distinctive terms than the most probable words which can be generic.

The saliency measure is defined as follows : for a given word w , its conditional probability $P(T|w)$ is computed, representing the likelihood that word w was generated by topic T . This estimation is accomplished using G^2 statistics [7], leveraging the knowledge of $P(w|T)$. Consider this table :

	w	$\neg w$
T	$a = P(w T)$	$b = P(\neg w T)$
$\neg T$	$c = P(w \neg T)$	$d = P(\neg w \neg T)$

The G^2 statistic is then defined as :

$$G^2 = a \log \frac{a(c+d)}{c(a+b)} + b \log \frac{b(c+d)}{d(a+b)} = P(T|w) \quad (6.1)$$

Additionally, the likelihood $P(T)$ that any word w was generated by topic T is calculated. As a result, the *distinctiveness* of word w can be quantified by the Kullback-Leibler divergence between $P(T|w)$ and $P(T)$:

$$\text{distinctiveness}(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)} \quad (6.2)$$

This defines *how informative the specific term w is for determining the generating topic versus a random word w'* . If a word w occurs in all topics, it would have a low distinctiveness score. Consequently, the *saliency* of a term can be defined as follows:

$$\text{saliency}(w) = P(w) \times \text{distinctiveness}(w) \quad (6.3)$$

Additionally, Termite offers a seriation feature, which allows for the reordering of terms within the matrix. This permutation of term order can unveil clustering structures, making it easier to identify concepts within topics and identify shared properties among them. The provided visualization, represented in Figure 6.3, illustrates a Termite visualization. The horizontal axis represents words, while the vertical axis represents topics. It is worth noting that the sizes of the circles in the visualization remain proportional to the word probabilities.

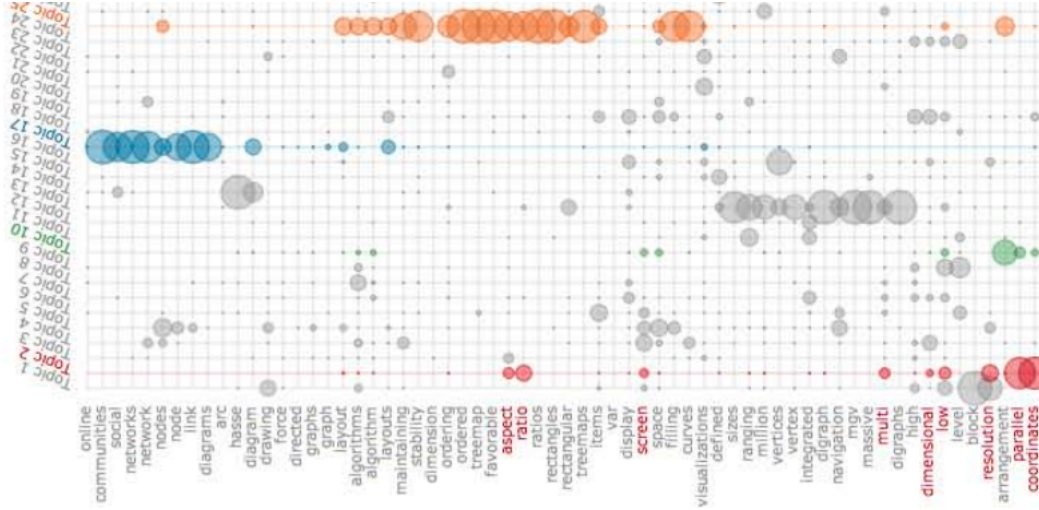


Figure 6.3: Termite matrix visualization from [5]

LDavis

LDavis is a visualization tool specifically designed for exploring topics generated by Latent Dirichlet Allocation (LDA) models. It offers both a global view of the topics and individual views for each topic, allowing users to visualize the most important terms within each topic. Like Termite, LDavis employs a different metric for selecting the terms to be presented in a topic. The authors of LDavis found that ranking terms solely by their probability is suboptimal for topic interpretation. One limitation of Termite is that it primarily provides a global view of the model and lacks the ability to thoroughly examine each individual topic. Consequently, we have chosen to utilize LDavis for visualizing our models, as it offers a more comprehensive visualization and presents words within each topic in a highly coherent manner.

Relevance of terms

In this section, we define the *relevance* method for ranking terms in a topic. Let ϕ_{kw} be the probability of term w in topic k and p_w the probability of term w in the corpus. Then the *relevance* of a term w to k given a weight parameter λ ($0 \leq \lambda \leq 1$) is defined as:

$$r(w, k|\lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right) \quad (6.4)$$

By setting $\lambda = 1$, the terms are ranked solely based on their probability for topic k . Conversely, when $\lambda = 0$, the terms are ranked using the lift measure. In order to determine the optimal value for λ , the authors conducted a user study based on a 50-topic model trained on the 20 Newsgroups dataset. They enlisted 29 subjects from their colleagues who were asked to review a list of the top 5 terms for each topic, ranked by relevance according to randomly sampled λ values. The subjects then had to identify the topic from a list of three possible topics, with one being the correct match. Through the study, the authors discovered that an optimal value of $\lambda = 0.6$ yielded a 70% probability of correct topic identification. This finding was based on the subjects ability to interpret the topics using the lists of terms generated. In our own experience, we have found that this value of λ indeed enhances the coherence and interpretability of the topics, surpassing the use of word probabilities alone.

In Figure 6.4, we present the LDavis visualization of a 30-topic model trained on RTBF data. The left panel offers a comprehensive overview of the 30 topics. The size of each circle represents the topic’s distribution within the corpus, with the selected topic being the most prevalent in this particular dataset. In this case, the selected topic prominently revolves around the COVID epidemic. To determine the

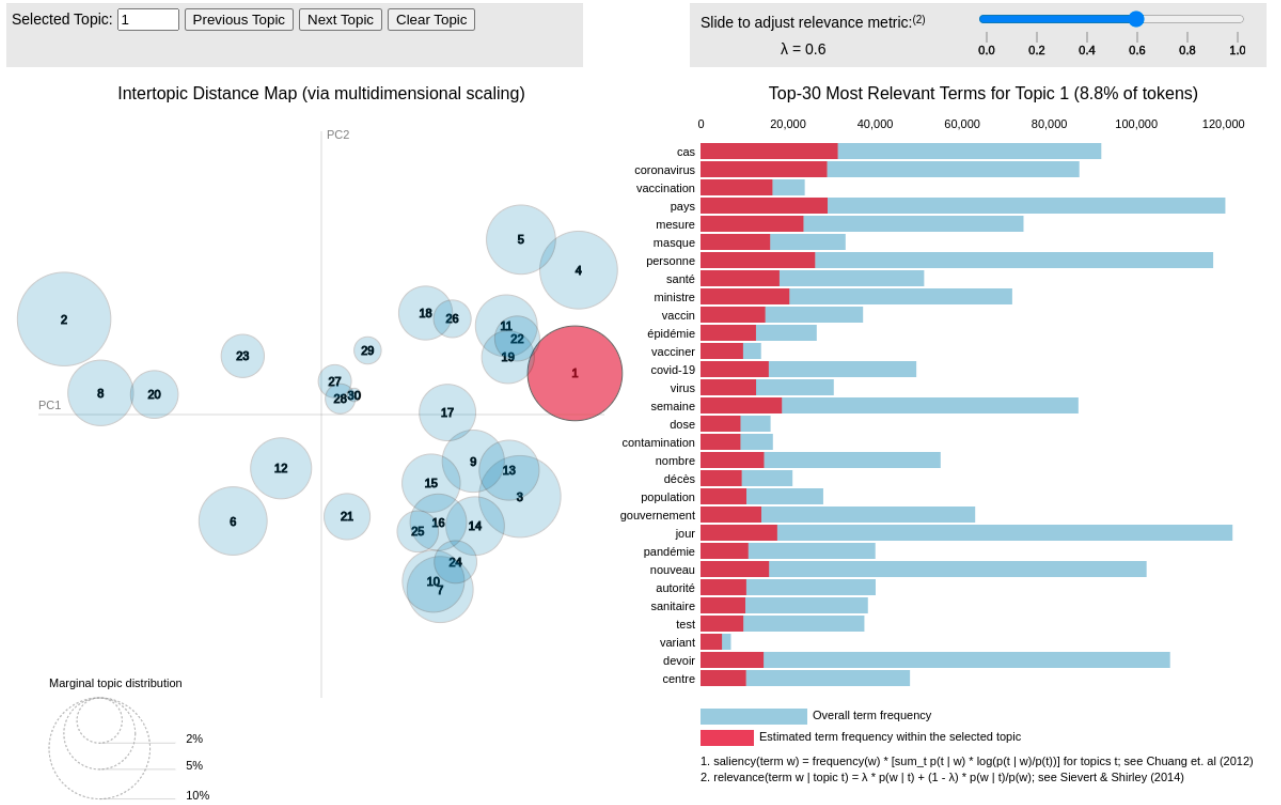


Figure 6.4: LDAvis of the RTBF model

distances between topics, the Jensen-Shannon divergence is employed. It calculates a weighted sum of Kullback-Leibler divergences, ensuring symmetry:

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}\left(p||\frac{p+q}{2}\right) + \frac{1}{2}D_{KL}\left(q||\frac{p+q}{2}\right) \quad (6.5)$$

To map these distances onto a two-dimensional space, Principal Components Analysis (PCA) is utilized. This transformation allows topics that are close in the 2D space to be considered similar to each other.

In Figure 6.5 and Figure 6.6, we present the most representative words for topic 8, which is the second topic from the left in the LDAvis global view. In Figure 6.6, where $\lambda = 1$, the displayed words undoubtedly pertain to sports. Specifically, it appears to be related to racing sports, such as running, Formula 1, and cycling. Although these words are the most probable within the topic, their estimated term frequencies (indicated by the red bars) are considerably lower than the overall term frequencies (indicated by the blue bars). This makes it challenging to distinguish this topic from others. However, by changing the value of λ to 0.6 (as seen in Figure 6.5), the most representative words become more discriminative. These words are predominantly present in this particular topic, making it easier to determine that the topic is specifically discussing cycling.

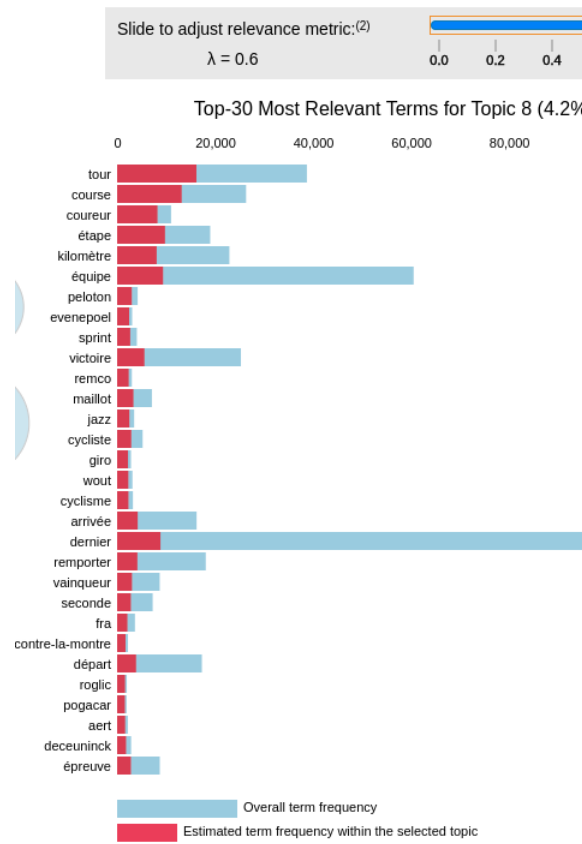


Figure 6.5: Most representative words of topics 2 of RTBF model with $\lambda = 0.6$

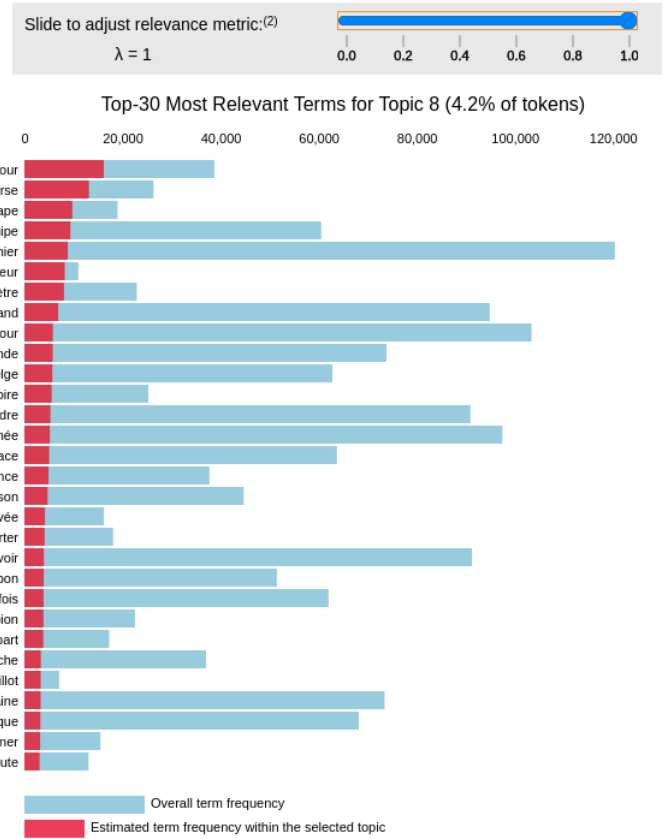


Figure 6.6: Most representative words of topics 2 of RTBF model with $\lambda = 1$

Chapter 7

Automatic labelling

One of the main limitations of Latent Dirichlet Allocation is that topics generated by the model lack of clear labels and require manual review by humans. However, automating the process of topic labeling is desirable. Several papers have attempted to address this challenge using complex techniques. Initially, our intuition was to leverage Word2Vec, a word embedding model, to identify the most similar terms to the top terms of a topic. One such paper, [1], employed this approach. However, their technique is computationally intensive, involving training Word2Vec and Doc2Vec models on a substantial Wikipedia dataset, which weighs 7 GB. They then computed the most similar article titles to the top 10 terms and ranked them using another measure. While effective, we opted for a simpler approach. In our case, we trained a Doc2Vec model on the articles from the five newspapers. Since certain websites, such as RTBF, provide multiple labels/categories for each article, we utilized the trained Doc2Vec model to identify the most similar labels to the top N terms of a topic. Given a set of labels L and the top N terms of a topic T , we computed the best label as follows:

$$\max_{l \in L} \left(\frac{\sum_{t \in T} \cos_sim(w, l)}{length(T)} \right) \quad (7.1)$$

The approach we adopted for topic labeling is relatively straightforward and yields satisfactory results. In our perspective, the most effective way to identify a topic was by utilizing the two best labels we could find for each topic. Using only one label often resulted in a vague representation, while incorporating two labels provided more informative and sufficient labeling. Another parameter to consider is the number of top terms used to compute the best label. We conducted tests with different settings, including 10, 15, and 20 terms. We found that using 20 terms often led to a focus on a specific topic, which may be overly narrow. On the other hand, using 10 and 15 terms yielded more general and improved results. Ultimately, we settled on using 15 terms as it consistently produced better outcomes.

Chapter 8

Analysis of newspapers

In this section, we will analyze the newspapers using Latent Dirichlet Allocation (LDA) models. Each newspaper's topic model can be found in Appendix A in a table format, showcasing the ten most relevant words for each topic. This provides a straightforward and intuitive visualization of the topic models in report format. Furthermore, we employ t-SNE dimensionality reduction to visualize the topic models. Each article is represented as a point in the plot, and the color of each point indicates its most dominant topic. This visualization allows for a visual comparison of topic distributions within each topic model. Additionally, we examine the temporal distribution of topics for each week in each newspaper. We create a heatmap where the x-axis represents the weeks of a year, and the y-axis represents the topics of the newspaper. This heatmap facilitates a straightforward comparison of the topic proportions over time for each newspaper and are listed in Appendix B. Lastly, we utilize a named entity recognition model to identify the entities mentioned in each article. This analysis enables us to explore the most cited entities within each topic, shedding light on the entities that are frequently discussed in the articles. These plots are available in Appendix C.

8.1 Comparison of topics

In this section, we conduct a comprehensive comparison of the topics of the newspapers.

In general, newspapers tend to cover similar major topics, although these topics are often broad and lack a specific subject definition. The dominant topics usually consist of common words such as "bien" (well), "chose" (thing), "beaucoup" (a lot), and others. These terms are frequently used in French texts, making it difficult to assign a precise label to these topics. While they appear in many articles, they do

not define a clear subject due to their general nature.

Football is also a significant topic across all newspapers. The most dominant football topic focuses on football terms without providing specific information about a particular league or subject. Terms such as "match" (match), "but" (goal), "club" (club), and others are relevant to any document discussing football. Lavenir and DH are newspapers with a more Belgium-oriented focus, having a topic specifically dedicated to regional football. Other football topics include the Pro League in Belgium, the Belgian national football team (Red Devils), and European football. The Pro League is covered by all newspapers, although RTBF and Lalibre have less coverage of football. European football and the Red Devils have dedicated topics in other newspapers.

In addition to football, newspapers also cover topics related to tennis, cycling, and racing sports. These sports are popular in Belgium, so their inclusion in newspapers is not surprising. Another sports topic that appears in every newspaper is the Olympic Games in Tokyo.

One subject that cannot be overlooked in recent years is the COVID-19 pandemic, which has affected people worldwide for several months or even years. Therefore, it is logical that all newspapers cover this topic. Common COVID-19 topics include vaccination, the pandemic situation and measures taken, flight restrictions, and the closure of the catering industry. Lalibre and RTBF also have a topic dedicated to health in general.

The American elections, which received significant global coverage in 2020, are also covered by each newspaper. While there are slight differences in the marginal distribution, the topics related to the American elections are generally the same across newspapers, with Lesoir covering it slightly less in percentage.

Culture, art, cinema, and music are also covered by all newspapers. Music, festivals, and concerts are topics found in each newspaper, and they all have a topic dedicated to cinema. Art is also covered by all newspapers, and Lalibre even has a topic about novels and reading, which is unique to that newspaper.

Another major and ongoing global event is the war in Ukraine, which is covered by every newspaper. Lalibre and RTBF also have a topic about wars outside Ukraine, including Afghanistan, Israel, Syria, and others. Migration, often a consequence of war, is a topic for RTBF and Lesoir. However, it may also be discussed in the context of climate, which is a topic for Lalibre, Lavenir, and RTBF. Energy

consumption, an area for improvement in climate change efforts, is a topic covered by every newspaper, likely due to recent rises in energy prices.

Politics is an essential subject covered by all newspapers, although the specific areas they focus on may vary. All newspapers generally cover topics related to the Belgian government, courts/jails, and police interventions. Lalibre, Lesoir, and RTBF provide more coverage of international politics, including European politics and international conflicts. These three newspapers also have topics related to finance. Lalibre covers finance in the most detail, with topics about financial markets, Wall Street, and banks. Additionally, a related topic is companies/jobs, which is covered by Lavenir, Lalibre, Lesoir, and RTBF.

The remaining topics are generally similar across different newspapers. All newspapers have a topic dedicated to weather reports, and most include a separate topic for fires/floods, except for Lalibre, where these topics may be combined. They also cover topics related to education, mobility/accidents, royalty, and consumption.

8.2 Dominant topic by article

This section displays the t-SNE dimensionality reduction plots of the document-topic matrices. Each document is represented by a color corresponding to its most likely topic. The topic labels above each plot are also color-coded to match the corresponding colors in the plot. To ensure readability, some labels have a black background for contrast, as reading them on a white background would be challenging.

• vivre ici, finlande le pays du bonheur • pro league challenger, pro league • droit social, service citoyen • arbitre, hors-jeu • tribunal correctionnel de mons, tribunal correctionnel • parlement flamand, commission parlementaire • aménagement du temps de travail, rénovation • commerce local, centre culturel • délit de fuite, comportement suspect • reporter, annulation • princesse stephanie de monaco, princesse elisabeth • covid-19 long, contamination • euro istanbul, euro u21 2023 • ligue 1 - football, coupe de la ligue • twitter, démocrate • championnats de belgique, championnat de belgique de rugby • sprint course, sprint • patients covid-19, patients non vaccinés • voirie, fermeture voirie • pop rock, spectacle musical • compagnie aérienne, avion • invasion russe, soldat ukrainien • écrans, look • mont ventoux, denivele challenge trail • manchester city, swansea city • installations électriques, vélo électrique • art contemporain, exposition - musées • grand chelem, tennis - roger federer • pompiers, incendie de voiture • accueillir la vie sauvage au jardin, ramassage des déchets • bio, consommateur • province de brabant wallon, hôpital • max hurdebroucq, rallye de spa • scolaire, enseignement secondaire • prévisions météo, thermomètre • jean paul gaultier, jean dujardin • réalisateur, festival du film documentaire • blessure, genou • maxime richard, citadelle de namur • anne-françoise rouche, premier café liégeois • logement, appartement

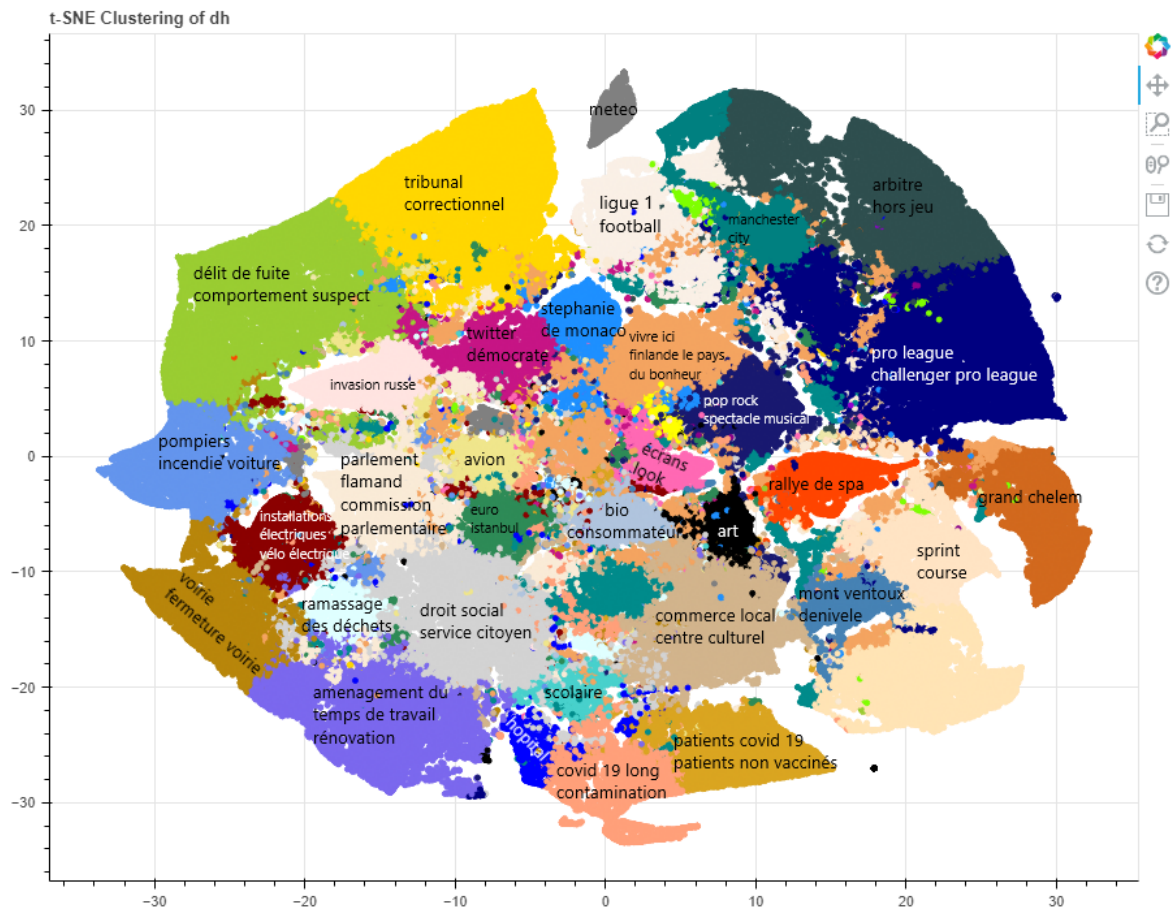


Figure 8.1: t-SNE dimension reduction of DH

• femme de vie, famille • champions league, champions league - foot • super 8, demain • salaires, indexation des salaires • patients non vaccinés, covid-19 long • pop rock, single • parlement flamand, parlement wallon • art contemporain, musée autoworld • poutine, chine de mao • bouclier aérien, espace aérien • incendie de voiture, violence armée • cinéaste, festival du film documentaire • euro de sofia, dollar • procureur général, tribunal correctionnel de mons • agriculture, plantes • républicain, démocrate • crise climatique, réchauffement climatique • application smartphone, web • journée internationale contre le racisme et la discrimination, harcèlement sexuel • harcèlement scolaire, enseignement secondaire • princesse astrid, prince william • cancer de la peau, maladies cardiaques • chrono, course sprint • grand chelem de paris, grand chelem • tram, bus • finale coupe de belgique basket, championnat de belgique de cyclisme • régions brabant wallon, brabant wallon • horeca wallon, restaurants • prévisions météo, thermomètre • dune, citadelle de namur • beurre de cacao, chocolat au lait • piano, orchestre philharmonique de liege • le cactus, grand cactus • criminal, animal max purcell • compagnie maritime, belge navire • boisson, vin sans alcool • inondation, inondations 2021 • le pape françois, pape • scène - accueil, asile • bitcoin, cryptomonnaie • benjamin liennot, tcheque • conflit israélo-palestinien, palestinien

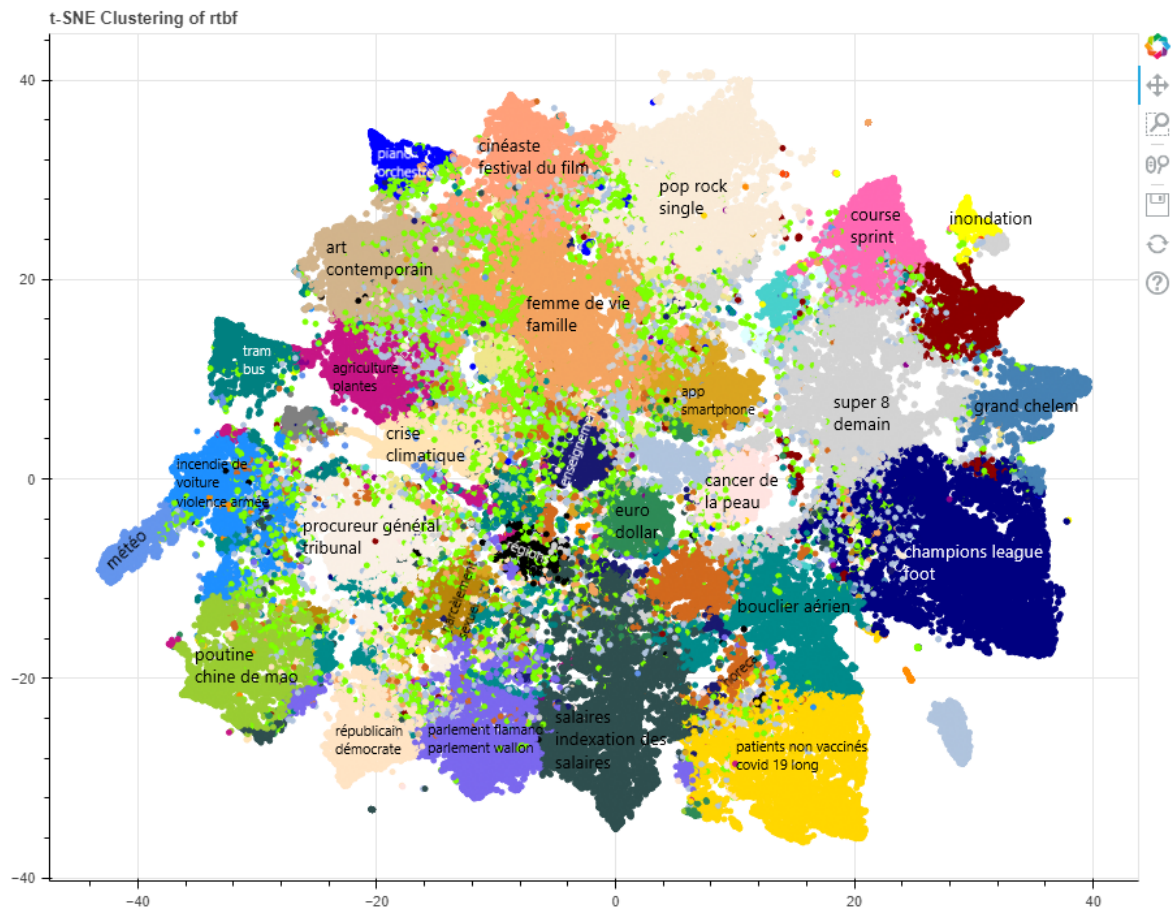


Figure 8.2: t-SNE dimension reduction of RTBF

•maman, dans quel monde on vit •application smartphone, applications •bye-bye la démocratie, démocratie en questions •musée des arts contemporains, art contemporain •soignant, personnel soignant •terroristes palestiniens, groupe terroriste etat islamique •invasion de, invasion russe •réalisateur, réalisatrice •expert scientifique, chercheur •parlement flamand, parti socialiste •impôt, impôt sur la fortune •covid-19, contamination •derby, coupe davis 2021 •énergies fossiles, carbone •scrutin, parti démocrate •immunité parlementaire, commission parlementaire •euro istanbul, inflation •attentat de nice, viol •bruxelles environnement, le fonds du logement à bruxelles •tribunal correctionnel de tournai, tribunal correctionnel •roman noir, roman policier •pluies, températures •new delhi, wall street •réouverture lieux culturels, bar •voiture électrique, conducteur de voiture •ursula von der leyen, ingrid von wantoch rekowski •enseignement supérieur, enseignement secondaire •droit de grève, fin de grève •finale coupe de belgique, finale coupe de belgique basket •marine le pen, marine le pen •delta, infection respiratoire •vaccins, injection •sprint, course sprint •vendre, consommateur •chine, chinois •grand chelem, tournoi •tartare aux légumes, salade boulgour •avion belge, avion •festival musical, eurovision jeune musicien •tir sportif, occasion •real, ligue 1 •fréquences radio, émission radio •reine mathilde, prince de galles •dhaese racing team, venturi racing •lauréat, récompense

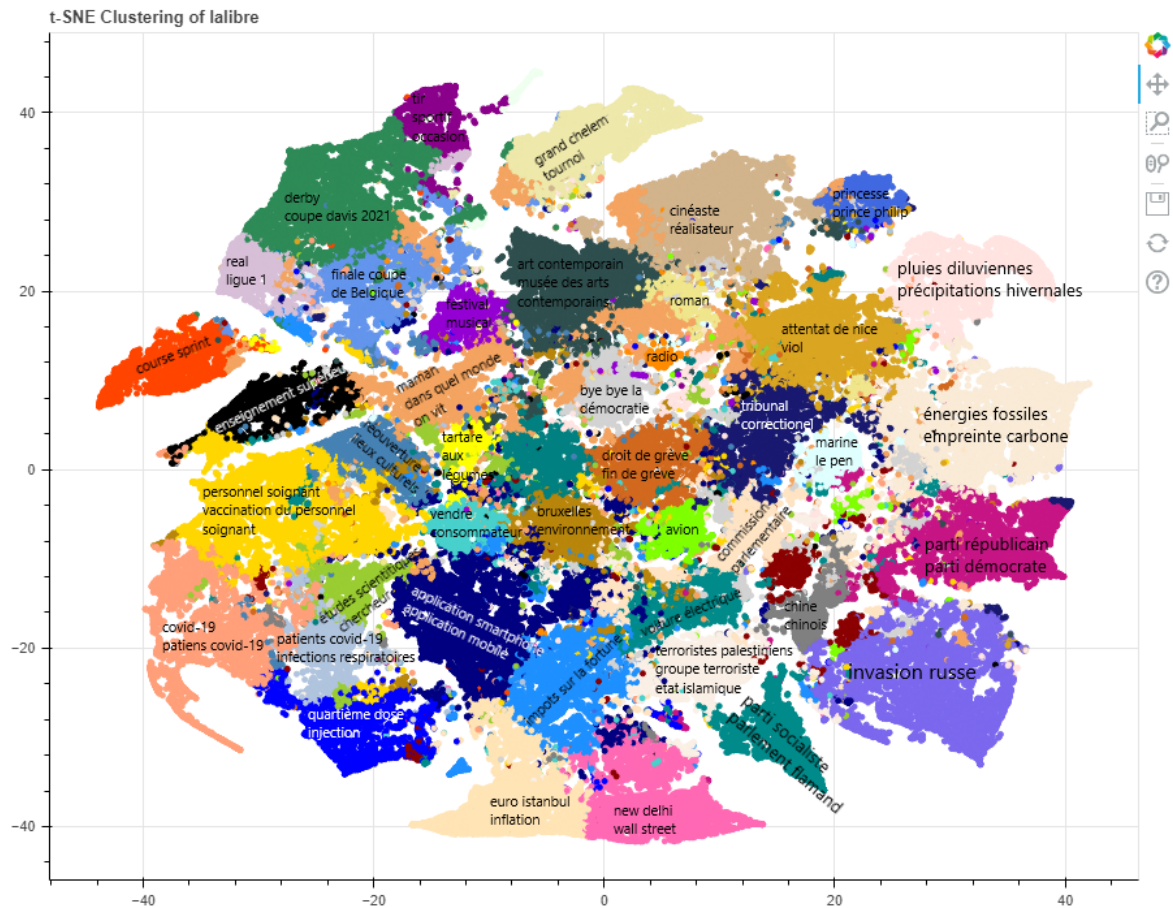


Figure 8.3: t-SNE dimension reduction of Lalibre

To begin with, an evident observation is that topics sharing the same dominant theme tend to cluster together effectively. Additionally, for DH, RTBF, and Lavenir, the clusters located on the sides of the visual representation are larger compared to those in the center. In contrast, Lalibre and Lesoir exhibit more uniform cluster sizes. This suggests that DH, RTBF, and Lavenir are more focused on specific dominant topics, whereas Lalibre and Lesoir cover dominant topics more evenly. Furthermore, an examination of topic positions reveals that related topics are situated in close proximity to one another. For instance, in the case of DH, all sport-related topics are positioned on the right side, while topics related to Covid-19 are found at the bottom. Other subjects such as politics, government, and justice are situated on the left side. This arrangement reflects the semantic closeness of the topics, aligning with our expectations.

DH, Lavenir, and Lesoir demonstrate a strong inclination towards sports, with nearly one-third of their topics being sports-related. These three newspapers occupy the entirety of the right side of the plot, emphasizing their focus on sports. Unsurprisingly, the largest topics within this category revolve around football, with Lavenir and Lesoir featuring a significant majority of football-related topics. In DH, the distribution is slightly more balanced across various sports. Cycling, tennis, and motorsport, when combined, encompass a size comparable to football. On the other hand, sports topics constitute a considerably smaller portion of the plot for the other two newspapers, with football continuing to dominate even in those cases.

Following sports, other prominent topics revolve around Covid-19 and health. Lalibre and RTBF exhibit a significant cluster of topics related to health, with a majority of them specifically addressing the Covid-19 pandemic. Lesoir also dedicates a substantial portion to health-related subjects, primarily focusing on Covid-19. In contrast, DH and Lavenir cover these topics to a lesser extent, with a more concentrated emphasis on Covid-19 rather than a broader range of health-related subjects.

Culture plays a significant role in articles, with RTBF, Lalibre, and Lesoir dedicating a larger portion of their content to cultural topics compared to DH and Lavenir. The cultural coverage primarily revolves around cinema, music, and art. This emphasis is relatively consistent for Lesoir and RTBF, although Lalibre focuses slightly less on music-related subjects.

Articles related to politics, war, and law hold significant importance in the realm of information. DH and Lavenir allocate a substantial portion of their content to

such topics, including accidents/mobility, courts/murders, regional politics, and unions. These subjects predominantly revolve around factual events in Belgium and constitute approximately one-third of the overall topics. However, international topics like the Ukraine war, USA politics, and European politics cover a comparatively smaller portion of their content. In contrast, RTBF, Lalibre, and L'esoir exhibit a more uniform distribution of these types of topics. They dedicate a larger proportion of their content to political, war, and law-related articles compared to DH and L'avenir. Furthermore, these three publications give more prominence to international news. Additionally, all three newspapers allocate a significant portion of their content to world finances, with Lalibre featuring three consequent topics related to this subject.

8.3 Topics distribution over time

Analyzing the distribution of topics over time in various newspapers offers valuable insights into their editorial focus and evolving coverage. By examining the fluctuations and patterns within these publications, we can gain a deeper understanding of their priorities and the changing landscape of news. This section delves into the temporal distribution of topics across the 5 newspapers, shedding light on the emphasis given to specific subjects and potentially uncovering underlying trends and shifts in journalistic coverage. Through this analysis, we aim to explore how newspapers allocate their attention to various themes and how these distributions may evolve over time. In order to accomplish this analysis, we have created heatmap plots for each year of each newspaper which are available in Appendix B. These plots present the distribution of topics over time, with the X-axis representing the weeks of the year and the Y-axis representing the various topics covered by the newspapers. Each cell in the heatmap is color-coded to represent the percentage of coverage dedicated to a specific topic within a given week. This visual representation allows us to observe and compare the fluctuations and trends in topic coverage throughout the weeks and years for each newspaper.

2020

In 2020, a significant starting event unfolded, namely the COVID pandemic, which had a profound impact on news coverage. This event becomes evident in the newspapers around the 10th week, coinciding with the Belgian government's announcement of strict containment measures. The historical significance of this event is reflected in its substantial presence within the topics covered throughout the year. Furthermore, analyzing other topics reveals the recurrence of certain events. In the realm of sports, annual competitions play a prominent role. For

example, the renowned cycling competition, Le Tour de France, typically occurs around the 30th week, as visible in DH, L'Espresso, and Libération. Similarly, in the world of tennis, the Open d'Australie serves as a major tournament taking place towards the end of January. Its impact can be observed in the 4th week in DH, Libération, and L'Espresso. While Libération and L'Espresso exhibit notable coverage of events involving Belgian governments, it is challenging to determine the exact nature of these events without further context or specific information.

2021

In 2021, unfortunately, the Covid pandemic continued to dominate the news, but we can explore other significant events. For instance, the American elections occurred towards the end of 2020, resulting in Joe Biden becoming the official president of the USA on January 20, 2021. This event received notable coverage in Libération, L'Espresso, and RTBF during the second and third weeks. However, DH covered it to a lesser extent, despite having a dedicated topic for it. Additionally, the football European Cup was initially scheduled for 2020 but was affected by the Covid pandemic. Its coverage was prominent in DH, L'Espresso, Libération, and L'avenir around the 23rd and 24th weeks. Surprisingly, RTBF, which typically covers football extensively, showed limited coverage of this event. The reason for this discrepancy is unclear, considering that RTBF is known for reporting on the European Cup. Another significant sporting event in 2021 was the Olympic Games in Tokyo, which took place around the 29th to 31st weeks. All newspapers extensively covered this event. In Belgium, there was also a major flooding event in 2021, visible in all newspapers around the 28th and 29th weeks. Libération stands out once again by featuring a notable event during the 33rd week, specifically within their war/terrorism coverage. This event revolves around the Taliban's military offensive against the Afghan government. It highlights Libération's strong focus on international affairs and reinforces its reputation as a newspaper with a significant emphasis on global news.

2022

In 2022, there was a noticeable decrease in the coverage of the Covid pandemic, reflecting its better containment. However, an ongoing and somber event that remained in the headlines was the war in Ukraine. This conflict was extensively covered in every newspaper around the 8th week, except for L'avenir, where it received comparatively less attention. Another significant event in 2022 was the football World Cup in Qatar, which took place towards the end of the year. This event garnered coverage in all newspapers, except for RTBF, due to the issue mentioned in the 2021 section. The French presidential elections also unfolded in

2022, receiving coverage around the 14th to 16th weeks in RTBF, Lalibre, and Lesoir. Finally, on the 8th of September 2022, the Queen of England passed away. Lesoir provided the most extensive coverage of this event, while RTBF and Lalibre covered it to a lesser extent during the 36th week. These three newspapers once again demonstrated their international orientation, extending their focus beyond Belgium, whereas DH and Lavenir had a comparatively more localized perspective.

8.4 Named entity recognition in topics

Named Entity Recognition (NER) algorithms offer a powerful tool for understanding the mentions and prominence of specific entities within different newspapers. By employing NER algorithms, we can analyze and identify the most cited entities within various topics, providing valuable insights into the sources and individuals that shape the news narratives. This section delves into the analysis of the most cited entities by topics in the 5 newspapers. The `fr_core_news_md` pipeline from Spacy has been used to detect entities in each article.

Football topic

In the Football topic, it is not surprising that the most frequently mentioned teams are Belgian. Across all five newspapers, Standard and Anderlecht emerge as the most cited teams. Historically, these two teams have been regarded as the top contenders in Belgian football, although their current standing may not necessarily reflect that. Lesoir predominantly covers Belgian teams, while RTBF, Lalibre, and DH also provide coverage of non-Belgian teams in their football-related content. This indicates a broader scope of coverage by RTBF, Lalibre, and DH, extending beyond the boundaries of Belgian football. Notably, Lavenir primarily focuses on regional football, encompassing amateur categories such as P2 and P3.

Tennis topic

In the tennis topic, the Women's Tennis Association (WTA) is the most frequently mentioned entity across all five newspapers. This comes as no surprise, considering the talent of Elise Mertens, a Belgian tennis player, who is highly regarded on the global stage. Belgian newspapers extensively cover her achievements and the WTA due to her performances. Following the WTA, the next most cited entities are the Grand Slam tournaments, including Roland Garros, Wimbledon, and the US Open. Additionally, David Goffin, another Belgian tennis player, garners considerable

attention in the newspapers. Lastly, it is unsurprising that renowned players like Novak Djokovic and Rafael Nadal receive significant coverage as well.

Political topic

When analyzing the political aspect of this topic, it is evident that the most frequently cited political parties across the five newspapers are the PS followed by the MR. These parties enjoy significant popularity in Belgium and are considered to be on opposite sides of the political spectrum, with PS leaning towards the left and MR towards the right. Alexander De Croo, the Prime Minister of Belgium, is prominently cited in all newspapers. This is understandable as he holds a prominent position in the government and received extensive media coverage during the Covid pandemic. N-VA and CD&V are Flemish political parties that also receive considerable attention in the newspapers. However, the coverage for other political parties such as Ecolo and PTB is slightly less extensive, although they are still frequently cited in the publications.

Covid vaccine topic

The selection of Covid vaccines has been a significant topic in the recent past. Among the five newspapers, Astrazeneca is the most frequently mentioned vaccine in DH, Lalibre, and Lenoir, while Pfizer receives the most citations in RTBF and Lenoir. These two vaccines, Astrazeneca and Pfizer, have been extensively used and widely discussed in the context of Covid vaccination. Furthermore, other Covid vaccines such as Moderna and Johnson & Johnson also receive considerable attention in DH, Lalibre, and Lenoir. These newspapers frequently mention these additional vaccine options in their coverage.

Chapter 9

Other use cases of topics models

Topic models, such as Latent Dirichlet Allocation (LDA), provide a valuable tool for comprehending large document collections. By mapping documents to topics, which are further associated with the vocabulary of the corpus, topic models enable us to determine the affinity of a document towards specific topics. This capability facilitates answering questions about the content of the corpus in an accessible manner. However, the utility of Latent Dirichlet Allocation extends beyond topic identification. The mapping of documents to topics and words offered by LDA can be leveraged for a range of other applications, including dimensionality reduction, information retrieval, machine translation, and part-of-speech tagging. In this chapter, we will focus specifically on exploring dimensionality reduction and information retrieval.

9.1 Dimensionality reduction

Dimensionality reduction is a crucial process in machine learning, aimed at transforming high-dimensional data into a lower-dimensional space while preserving the essential properties of the original data. This is done to address the issue of model performance degradation that arises when dealing with an excessive number of dimensions, often referred to as the "curse of dimensionality." In high-dimensional spaces, data points become sparser, necessitating larger amounts of data to obtain representative samples.

When working with text data, the Bag of Words method is a simple representation of a corpus that directly maps each document to all the tokens in the corpus, resulting in a document-term matrix. However, this approach often leads to sparse matrices with numerous zero entries, resulting in high-dimensional data, with one dimension for each token in the vocabulary. Dimensionality reduction can be

applied to group words with similar semantic meanings together, thus obtaining a low-dimensional representation of documents that reflects concepts rather than individual words. In Latent Dirichlet Allocation (LDA), the modeling of documents involves utilizing topics as a bridge between the tokens of the vocabulary. Instead of representing each document with all the tokens, LDA assigns documents to topics, providing a more meaningful and compact representation.

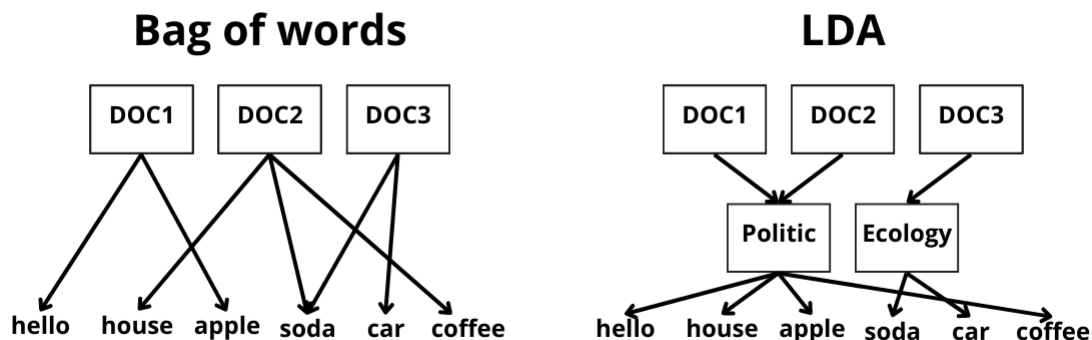


Figure 9.1: BOW vs LDA

Given N documents and K topics, we can represent each document as a vector of length K that signifies the probability distribution of the document across the topics. Combining all these vectors results in an $N \times K$ feature matrix that represents the corpus. This matrix can be utilized as input for various purposes, including training a supervised model.

Test classification accuracy

The test involves using a previous preprocessed RTBF dataset with a reduced number of documents. A decision tree classifier is employed to categorize the articles based on the newspaper's assigned categories. After removing irrelevant categories, the dataset consists of 95,648 articles for 30 categories. The vocabulary size is 28,231 words. To represent the words, TF-IDF embedding is utilized. This results in a matrix of dimensions $95,648 \times 28,231$, totaling 2,700,238,688 cells. The dataset is split in 75% for training and 25% for testing. To prevent overfitting, the only hyperparameter fine-tuned is the maximum depth of the decision tree. This is accomplished using 5-fold cross-validation, resulting in the best maximum depth of 20. It gives us a test set accuracy of 0.4972 by using the TF-IDF embedding. The figure 9.2 displays test set accuracies of a decision tree using document-topic matrix for different numbers of topics. The best test set accuracy is 0.54767 with 27 topics. It implies a matrix of dimensions $95,648 \times 27$ totaling

2, 582, 496 cells. This improvement in test set accuracy of 5% also leads to a reduction in dataset dimensions by over 99%, resulting in faster training times.

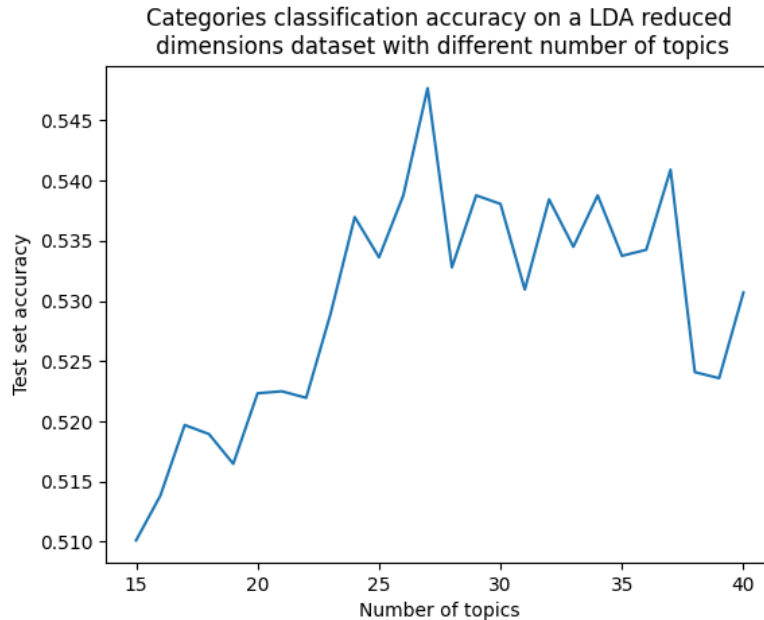


Figure 9.2: Classification accuracy for different number of topics

9.2 Information retrieval

Information retrieval involves the retrieval of pertinent documents based on a user's query. It differs from topic models that explore and summarize documents in the absence of any given context. Nevertheless, despite this distinction, there exist mathematical and conceptual connections between these two algorithms. Indeed, topic models can be viewed as an indexing method. It is possible to identify topics that assign a high probability to a query and subsequently locate the documents that have a high probability for these topics. In a basic information retrieval system, queries and documents are treated as collections of words, and retrieval occurs by evaluating the overlap between the two. However, this simple matching approach has a limitation in that it disregards semantics. In reality, words with similar meanings should be taken into consideration. Language modeling addresses this issue by capturing semantic relationships, making it a primary framework for utilizing topic models in information retrieval.

Language modeling

The concept of language modeling is to estimate the likelihood of a query by a probabilistic language model based on documents. Henceforth, the probability of generating a query q given a document d is :

$$p(q|d) = \prod_{w \in q} p(w|d) = \prod_{w \in q} \frac{n_{d,w}}{n_d} \quad (9.1)$$

where $n_{d,w}$ is the frequency of w in d , and n_d is the number of tokens in d .

To rank the documents, it is sufficient to consider their probabilities. However, if a query contains a word that is absent from a document d , the probability of generating that query given d becomes zero. This issue can be addressed through smoothing techniques or by employing topic models. Topic models offer a means to extract word probabilities from the corpus, which can be utilized for smoothing. In the subsequent discussion, we will outline two smoothing methods that can be incorporated within topic models.

Linear interpolation

Linear interpolation discounts the counts of seen words and distributes the extra to both seen and unseen words:

$$p(w|d) = (1 - \lambda)p(w|d) + \lambda p(w|C) \quad (9.2)$$

where C is the corpus. This approach effectively addresses the issue of sparsity. When a word w is not present in document d , its probability is adjusted to the corpus level probability denoted as $p(w|C)$.

With topic models, we can also compute this probability by marginalizing over the topics k :

$$p_{tm}(w|d) = \sum_k p(w|k)p(k|d) \quad (9.3)$$

[23] proposed to combine these two models through a linear interpolation :

$$p(w|d) = \lambda'((1 - \lambda)p(w|d) + \lambda p(w|C)) + (1 - \lambda')p_{tm}(w|d) \quad (9.4)$$

where λ' the LDA document model with the general smoothed model.

This approach use topic models to connect the query with document through the topics. The advantage is that the base language modeling approach could low rank

a document which doesn't contain some words of the queries while topic models could connect this document with topics that have connections with these missing words and then give a better result. Therefore, a better semantic relationship is learned.

Application to RTBF model

We conducted tests using linear interpolation and observed the following results. Initially, calculating the likelihood of a query on multiple documents using this method was found to be time-consuming. Specifically, for 957 documents, it took approximately 9 to 10 minutes to compute the likelihoods of a one-word query on an AMD Ryzen 5 3600XT processor with 10 processes. Due to the lengthy computation time, we did not test with a larger number of documents. However, despite the time required, the results were encouraging. We tested four queries: "lukaku," "trump," "covid," and "sprint." Each query yielded top-ranked documents that were semantically related. For example, the "lukaku" query generated documents related to the Belgian national football team, while "trump" produced documents about American elections or American politics. Similarly, the "covid" query resulted in documents containing information about the COVID situation, including the number of cases and deaths. Lastly, the "sprint" query returned documents about cycling competitions. Overall, the results were quite satisfactory in terms of semantic relevance, but the computational requirements were high.

Chapter 10

Conclusion

In conclusion, Latent Dirichlet Allocation (LDA) has proven to be a powerful model for analyzing Belgian newspapers. We started by collecting and cleaning data through web scraping, ensuring the construction of high-quality datasets for our analysis. We then proceeded to visualize and evaluate the LDA topic model, recognizing its importance in the training process. One significant step was the automatic labeling of topics generated by our LDA model, which allowed us to interpret and understand the underlying themes within the topics. The topics covered a wide range of subjects, including the Covid pandemic, sports, politics, and more. We gained insights into the varying coverage and prominence of specific topics across different newspapers. One key advantage of LDA is its unsupervised learning capability, which eliminated the need for manual annotation and provided us with a deeper understanding of the newspapers corpora's underlying structure. Furthermore, we explored additional use cases of LDA, such as dimensionality reduction and information retrieval, highlighting its versatility and applicability in various scenarios. Overall, Latent Dirichlet Allocation proved to be an effective and valuable tool for analyzing large collections of documents, extracting meaningful insights, and exploring patterns within the Belgian newspapers. Its ability to uncover latent topics and its unsupervised learning approach make it a powerful tool for data analysis in the field of natural language processing.

Bibliography

- [1] S. Bhatia, J. H. Lau, and T. Baldwin. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [3] J. Boyd-Graber, Y. Hu, and D. Minmo. 2017.
- [4] J. Chang, S. Gerrish, C. Wang, J. Boyd-graber, and D. Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.
- [5] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12*, page 74–77, New York, NY, USA, 2012. Association for Computing Machinery.
- [6] S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha. *Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond*, pages 129–161. Springer US, Boston, MA, 2012.
- [7] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, mar 1993.
- [8] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.
- [9] G. Gundersen. Understanding dirichlet–multinomial models.

- [10] M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.
- [11] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
- [12] D. Jurafsky and J. H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.
- [13] J. Lau, D. Newman, S. Karimi, and T. Baldwin. Best topic word selection for topic labelling. volume 2, pages 605–613, 01 2010.
- [14] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics.
- [15] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: An empirical study of plsa and lda. *Inf. Retr.*, 14:178–203, 04 2011.
- [16] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.
- [17] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.
- [18] D. Newman, S. Karimi, and L. Cavedon. External evaluation of topic models. *ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium*, 01 2011.
- [19] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California, June 2010. Association for Computational Linguistics.

- [20] R. Rehurek and P. Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.
- [21] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408, 02 2015.
- [22] C. Sievert and K. Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.
- [23] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 178–185, New York, NY, USA, 2006. Association for Computing Machinery.

Appendix A

Topic models

In this chapter, we present the topic models in a tabular format. The topics are listed in descending order of their marginal distribution, and for each topic, we provide the ten most relevant words. Additionally, we include the labels assigned to the topics using our automatic labelling technique.

RTBF	Top 10 words	Label	ist.
1	femme, vie, livre, homme, vivre, histoire, mot, chose, jeune, bien	femme de vie, famille	11.9%
2	match, club, but, joueur, league, football, diable, point, rencontre, minute	champions league, champions league - foot	11.0%
3	bien, beaucoup, chose, équipe, virement, moment, gagner, essayer, difficile, espérer	super 8, demain	4.6%
4	secteur, emploi, euro, syndicat, travailleur, travail, crise, entreprise, salaire, grève	salaires, indexation des salaires	4.4%
5	coronavirus, vaccin, covid-19, test, vaccination, cas, virus, santé, hôpital, patient	patients non vaccinés, covid-19 long	4.1%
6	album, chanson, chanteur, titre, loading, artiste, musique, groupe, morceau, concert	pop rock, single	4.1%
7	ministre, européen, commission, gouvernement, accord, flamand, parlement, fédéral, politique, proposition	parlement flamand, parlement wallon	4.0%
8	art, musée, exposition, œuvre, découvrir, siècle, artiste, patrimoine, collection, château	art contemporain, musée autoworld	3.8%
9	ukraine, russe, guerre, ukrainien, moscou, sanction, poutine, pays, européen	poutine, chine de mao	3.7%
10	mesure, coronavirus, aéroport, pays, sanitaire, vol, annuler, pandémie, restriction, voyage	bouclier aérien, espace aérien	3.3%
11	mort, incendie, tuer, armée, militaire, feu, force, ville, civil, afp	incendie de voiture, violence armée	2.5%
12	film, festival, cinéma, acteur, série, spectacle, réalisateur, documentaire, théâtre, scène	cinéaste, festival du film documentaire	2.4%
13	million, milliard, dollar, euro, vente, entreprise, marché, vendre, banque, prix	euro de sofia, dollar	2.3%
14	police, enquête, prison, justice, tribunal, policier, avocat, parquet, procès, cour	procureur général, tribunal correctionnel de mons	2.3%
15	eau, déchet, jardin, plante, naturel, forêt, nature, arbre, plastique, environnement	agriculture, plantes	2.2%
16	président, trump, parti, élection, donald, américain, biden, joe, politique, vote	républicain, démocrate	2.2%
17	étude, chercheur, climatique, scientifique, recherche, université, climat, science, réchauffement, changement	crise climatique, réchauffement climatique	2.1%
18	vidéo, réseau, application, facebook, plateforme, utilisateur, social, apple, internet, instagram	application smartphone, web	2.0%
19	manifestation, droit, violence, femme, manifestant, journaliste, loi, dénoncer, liberté, mouvement	journée internationale contre le racisme et la discrimination, harcèlement sexuel	1.8%
20	enfant, école, étudiant, élève, parent, enseignement, scolaire, jeune, wallonie, don	harcèlement scolaire, enseignement secondaire	1.7%
21	famille, père, mère, fils, roi, prince, fille, enfant, décéder, couple	princesse astrid, prince william	1.7%
22	maladie, traitement, santé, médecin, peau, cerveau, forme, cancer, mental, patient	cancer de la peau, maladies cardiaques	1.6%

23	tour, étape, kilomètre, coureur, course, sprint, evenepoel, peloton, victoire, fra	chrono, course sprint	1.6%
24	tournoi, finale, atp, set, tennis, open, roland, mertens, joueur, mondial	grand chelem de paris, grand chelem	1.6%
25	bruxelles, rue, véhicule, voiture, vélo, circulation, mobilité, gare, train, bus	tram, bus	1.5%
26	olympique, tokyo, médaille, sport, record, monde, belge, sportif, athlète, jeu	finale coupe de belgique basket, championnat de belgique de cyclisme	1.5%
27	commune, bourgmestre, communal, ville, wallon, province, liège, brabant, échevin, collège	régions brabant wallon, brabant wallon	1.4%
28	masque, port, magasin, rouvrir, réouverture, fermer, restaurant, iranien, commerce, client	horeca wallon, restaurants	1.4%
29	degré, température, pluie, vent, neige, tempête, chaleur, météo, canada, ciel	prévisions météo, thermomètre	1.3%
30	carnaval, pied, empereur, heure, petit, dune, fouille, citadelle, adrien, bout	dune, citadelle de namur	1.3%
31	gramme, huile, recette, lait, pâte, sucre, œuf, légume, ingrédient, pain	beurre de cacao, chocolat au lait	1.3%
32	musique, orchestre, compositeur, concert, piano, opéra, musicien, concours, reine, instrument	piano, orchestre philharmonique de liege	1.2%
33	-vou, connecter, contenu, gratuitementconnectez, cactus, marathon, irlandais, vousà, grimpeur, vousla	le cactus, grand cactus	1.2%
34	animal, pilote, hamilton, verstappen, max, formule, bull, mercedes, ferrari, leclerc	criminal, animal max purcell	1.1%
35	migrant, mer, bateau, spatial, bord, navire, nasa, marin, méditerranée, côte	compagnie maritime, belge navire	0.9%
36	noël, fête, bière, vin, alcool, boire, bouteille, café, boisson, verre	boisson, vin sans alcool	0.7%
37	inondation, thierry, rallye, neuville, sébastien, wrt, hyundai, toyota, evans, ogier	inondation, inondations 2021	0.6%
38	chien, chat, suède, avion, pape, george, floyd, corée, suédois, église	le pape français, pape	0.5%
39	asile, réfugié, demandeur, liban, motogp, assise, libanais, marquez, accueil, bangladesh	scène - accueil, asile	0.5%
40	turquie, turc, nintendo, syrien, syrie, erdogan, mario, soudan, switch, boogie	bitcoin, cryptomonnaie	0.3%
41	suisse, alexandre, biélorussie, tchèque, renard, benjamin, netanyahu, pologne, loukachenko, united	benjamin hennot, tchèque	0.3%
42	israélien, israël, palestinien, juif, mali, arabe, slovène, cyril, jérusalem, gaza	conflit israélo-palestinien, palestinien	0.1%

Lesoir	Top 10 words	Label	Dist
1	aujourd'hui, bien, question, problème, exemple, chose, monde, beaucoup, manière, sens	ma rue couche-toi là, mon plus beau village	5.2%
2	film, vie, cinéma, personnage, chose, petit, histoire, amour, vivre, femme	comédien, festival du film documentaire	5.1%
3	but, match, minute, face, jeu, ballon, rencontre, penalty, point, score	la vie en pause, le dernier des siens	4.7%
4	gouvernement, ministre, accord, fédéral, réforme, proposition, loi, mesure, pension, plan	pension minimum, contrôle budgétaire	4.3%
5	euro, million, banque, milliard, marché, financier, entrepreneur, dollar, vente, montant	dollar, euro	4.0%
6	projet, construction, bâtiment, espace, logement, technologie, développement, site, développer, environnement	elevage industriel, environnement	3.3%
7	politique, droit, démocratie, liberté, citoyen, mouvement, manifestation, démocratique, pays, action	militant, démocratie en questions	3.3%
8	reine, famille, père, hommage, fils, prince, elizabeth, cérémonie, nom, roi	reine mathilde, reine victoria	3.2%
9	standard, club, joueur, saison, charleroi, sporting, entraîneur, mercato, anderlecht, contrat	derby liégeois, mercato	3.2%
10	police, policier, enquête, parquet, suspect, fait, victime, homme, accident, violence	homicide, comportement suspect	3.1%
11	diable, joueur, manchester, rouge, psg, city, coupe, martinez, messi, football	swansea city, thunder oklahoma city	3.0%
12	masque, mesure, sanitaire, coronavirus, concertation, comité, horeca, obligatoire, port, pass	masque obligatoire, réouverture lieux culturels	3.0%
13	flamand, bien, paul, trop, magnette, flamand, wever, bart, dire, côté	flamand, budget flamand	2.6%
14	olympique, sport, jeu, tokyo, sportif, athlète, médaille, compétition, mondial, belge	les athlètes belges, athlète paralympique	2.5%
15	coronavirus, patient, hôpital, santé, soin, cas, maladie, nombre, infection, contamination	patients covid-19, infection respiratoire	2.5%
16	étudiant, travail, fédération, personnel, emploi, travailleur, congé, formation, université, métier	employeur, emploi et stage	2.2%
17	ukrainien, ukrainie, guerre, russe, armée, militaire, kiev, israël, soldat, israélien	soldat ukrainien, palestinien	2.1%
18	art, livre, musée, œuvre, exposition, collection, artiste, galerie, dessin, siècle	oeuvres art , musée des beaux-arts de tournai	2.1%
19	tour, étape, course, coureur, evenepoel, remco, kilomètre, wout, pogacar, peloton	cycliste, cyclisme	2.0%
20	chine, russie, otan, ukrainie, chinois, uni, sanction, diplomatie, russe, pays	mer de chine, royaume uni	2.0%
21	eau, incendie, inondation, pompier, kilomètre, chaleur, tempête, climatique, secours, dégât	pluies torrentielles, pluies diluviennes	2.0%
22	bruxelles, bruxellois, commune, bruxelle, bourgmestre, wallon, wallonie, ville, région, rue	etat des routes à BX, le fonds du logement à BX	2.0%
23	prix, gaz, énergie, nucléaire, électricité, énergétique, produit, consommation, centrale, consommation	transition énergétique, approvisionnement énergétique	1.9%

24	congo, parlement, cour, corruption, congolais, affaire, colonial, floyd, pape, ancien	parlementaire, commission parlementaire	1.8%
25	européen, pays, europe, union, asile, commission, frontière, migrant, belge, réfugié	crise migratoire, asile et migration	1.7%
26	club, bruges, league, union, match, supporter, pro, stade, football, mazzù	pro basket league, jupiler pro league	1.7%
27	expert, virus, scientifique, pandémie, vague, chine, population, covid, confinement, risque	nouvelle vague, vaccination covid	1.6%
28	classement, troisième, point, place, qualification, maroc, coupe, victoire, deuxième, groupe	coupe des nations, qualifications mondial 2022	1.6%
29	parti, député, ministre, libéral, gouvernement, bouche, droite, georges, majorité, socialiste	parti socialiste, coalition	1.6%
30	musique, album, festival, concert, artiste, chanson, chanteur, musical, disque, musicien	tournai jazz, spectacle musical	1.6%
31	atp, tournoi, tennis, wta, goffin, set, open, chelem, roland, djokovic	grand chelem de paris, grand chelem	1.5%
32	voyage, certificat, code, belge, orange, zone, amende, article, étranger, donnée	qr code, spatial	1.4%
33	train, magasin, snb, bus, grève, gare, stib, métro, chauffeur, delhaize	ferroviaire, chauffeur de bus	1.4%
34	américain, trump, biden, etats-unis, donald, joe, new, york, démocrate, républicain	parti démocrate, républicain	1.4%
35	procès, prison, attentat, avocat, tribunal, condamner, justice, cour, terroriste, peine	condamnation, tribunal correctionnel de tournai	1.4%
36	français, france, paris, juif, droite, eric, extrême, olivier, français, michel	olivier marchal, jean olivier collinet	1.2%
37	vaccin, vaccination, dose, vacciner, pfizer, médicament, astrazeneca, injection, vacciné, administrer	injection, vaccin	1.2%
38	enfant, école, élève, parent, enseignant, enseignement, scolaire, jeune, secondaire, classe	enseignants, enseignement secondaire	1.2%
39	test, aéroport, vol, tester, avion, compagnie, passager, aérien, positif, quarantaine	avion, avion belge	1.2%
40	madrid, real, ligue, champion, bayern, allemand, barcelone, munich, liverpool, dortmund	real valladolid, ligue 1 - football	1.1%
41	degré, vent, météo, averse, sec, pluie, nuage, prévision, ardennes, modérer	prévisions météo, averse	1.0%
42	animal, agricole, espèce, manger, maison, agriculteur, viande, asbl, poisson, plante	viande de veau, pain de viande	1.0%
43	émission, rtbf, boris, rtl, média, johnson, chaîne, climat, radio, journaliste	radio caroline, émission radio	1.0%
44	macron, emmanuel, candidat, iran, iranien, pen, marine, présidentielle, président, français	marine le pen, marine le pen	0.9%
45	pilote, formule, hamilton, mercedes, écurie, max, voiture, verstappen, ferrari, bull	max purcell, max hurdebroucq	0.9%
46	russe, poutine, turquie, russie, turc, vladimir, moscou, kremlin, Erdogan, istanbul	céréales, poutine	0.6%
47	nba, hockey, lions, james, vin, jeu, astérix, onana, binche, franchise	three lions, rap game	0.5%
48	stéphane, voo, nethy, notaire, grégoire, moreau, jet, basket, décerner, dimitri	classic 21, sports moteur classic 21	0.4%

Lalibre	Top 10 words	Label	Dist.
1	chose, bien, beaucoup, enfant, vie, vivre, jeune, temps, vraiment, dire	maman, dans quel monde on vit	5.7%
2	entreprise, société, service, emploi, plateforme, numérique, réseau, application, travail, start	application smartphone, applications	4.3%
3	politique, femme, liberté, droit, opinion, société, démocratie, débat, question, social	bye-bye la démocratie, démocratie en questions	4.1%
4	art, artiste, musée, exposition, œuvre, théâtre, livre, siècle, spectacle, dessin	musée des arts contemporains, art contemporain	3.5%
5	mesure, santé, soin, situation, ministre, concertation, comité, sanitaire, personnel, crise	soignant, personnel soignant	3.4%
6	iran, israélien, taliban, israël, iranien, palestinien, afghanistan, armée, attaque, militaire	terroristes palestiniens, groupe terroriste etat is-lamique	3.2%
7	russe, ukraine, russie, ukrainien, guerre, moscou, poutine, kiev, militaire, vladimir	invasion de, invasion russe	3.1%
8	film, acteur, série, personnage, cinéma, scène, réalisateur, comédien, documentaire, rôle	réalisateur, réalisatrice	3.1%
9	scientifique, étude, chercheur, risque, recherche, expert, cancer, analyse, professeur, facteur	expert scientifique, chercheur	3.1%
10	ministre, parti, gouvernement, flamand, libéral, fédéral, cdv, accord, socialiste, politique	parlement flamand, parti socialiste	3.0%
11	euro, million, milliard, montant, fonds, revenu, salaire, fiscal, payer, financier	impôt, impôt sur la fortune	2.9%
12	coronavirus, cas, test, nombre, contamination, décès, épidémie, covid-19, chiffre, moyenne	covid-19, contamination	2.8%
13	match, joueur, anderlecht, saison, jouer, équipe, but, bruges, coach, bruyne	derby, coupe davis 2021	2.7%
14	énergie, climatique, gaz, énergétique, électricité, production, climat, carbone, environnemental, transition	énergies fossiles, carbone	2.7%
15	trump, biden, président, donald, élection, joe, démocrate, républicain, américain, vote	scrutin, parti démocrate	2.7%
16	loi, commission, parlement, conseil, député, texte, droit, parlementaire, assemblée, membre	immunité parlementaire, commission parlementaire	2.6%
17	banque, taux, hausse, bourse, inflation, baisse, euro, croissance, point, marché	euro istanbul, inflation	2.6%
18	victime, sexuel, homme, attentat, spatial, suspect, police, viol, femme, princesse	attentat de nice, viol	2.4%
19	bruxellois, wallon, wallonie, commune, bruxelles, liège, région, logement, bourgmestre, communal	bruxelles environnement, le fonds du logement à bruxelles	2.4%
20	justice, tribunal, avocat, cour, juge, enquête, procès, judiciaire, dossier, prison	tribunal correctionnel de tournai, tribunal correctionnel	2.3%
21	roman, mort, noir, mourir, vie, livre, homme, récit, drame, écrire	roman noir, roman policier	2.2%

22	degré, eau, incendie, pompier, température, pluie, vent, ouest, météo, mer	pluies, températures	2.2%
23	dollar, indice, marché, bourse, américain, street, wall, fed, nasdaq, investisseur	new delhi, wall street	2.1%
24	masque, restaurant, rouvrir, port, fermer, réouverture, salle, sanitaire, fermeture, horeca	réouverture lieux culturels, bar	2.0%
25	voiture, animal, véhicule, vélo, train, kilomètre, parc, mobilité, route, circulation	voiture électrique, conducteur de voiture	1.9%
26	européen, europe, pays, allemagne, union, allemand, turquie, turc, grèce, suède	ursula von der leyen, ingrid von wantoch rekowski	1.9%
27	école, étudiant, élève, enseignement, enseignant, scolaire, université, classe, secondaire, cours	enseignement supérieur, enseignement secondaire	1.8%
28	police, syndicat, policier, grève, travailleur, direction, syndical, manifestation, détenu, travail	droit de grève, fin de grève	1.8%
29	coupe, football, club, sport, sportif, diable, équipe, belge, match, stade	finale coupe de belgique, finale coupe de belgique basket	1.7%
30	français, france, macron, emmanuel, paris, candidat, pen, marine, eric, mélenchon	marine le pen, marine le pen	1.6%
31	maladie, virus, variant, patient, oms, santé, covid-19, omicron, traitement, médicament	delta, infection respiratoire	1.6%
32	vaccin, dose, vaccination, royaume, johnson, uni, britannique, pfizer, vacciner, astrazeneca	vaccins, injection	1.6%
33	course, tour, étape, coureur, evenepoel, kilomètre, remco, wout, peloton, maillot	sprint, course sprint	1.5%
34	vente, magasin, prix, vendre, immobilier, consommateur, acheter, produit, client, enseigne	vendre, consommateur	1.5%
35	chine, chinois, pékin, corée, américain, etats-unis, hong, japon, kong, australie	chine, chinois	1.4%
36	tournoi, finale, olympique, atp, set, tennis, jeu, médaille, chelem, goffin	grand chelem, tournoi	1.4%
37	vin, cuisine, produit, bouteille, plastique, bio, hamilton, huile, plante, plat	tartare aux légumes, salade bolognais	1.3%
38	vol, avion, aéroport, compagnie, aérien, passager, migrant, asile, brussels, navire	avion belge, avion	1.3%
39	musique, chanson, concert, album, chanteur, église, festival, pape, musical, musicien	festival musical, eurovision jeune musicien	1.2%
40	minute, but, score, tir, jeu, point, poteau, ballon, mi-temps, penalty	tir sportif, occasion	1.1%
41	madrid, real, club, ligue, espagnol, messi, psg, barcelone, mbappé, manchester	real, ligue 1	1.1%
42	émission, laurent, rtl, olivier, rtbf, radio, julien, journaliste, animateur, chaîne	fréquences radio, émission radio	0.9%
43	prince, reine, royal, harry, roi, elizabeth, charles, meghan, britannique, william	reine mathilde, prince de galles	0.8%
44	racing, balle, boon, daring, watdicks, orée, léopold, dragons, charlier, cercle	dhaese racing team, venturi racing	0.6%
45	meilleur, prix, récompense, cérémonie, César, récompenser, catégorie, jury, décerner, doren	lauréat, récompense	0.4%

Laverir	Top 10 words	Label	Dist.
1	bien, chose, petit, vie, beaucoup, jamais, moment, temps, vivre, vraiment	combien de temps ?, souvenir	6.8%
2	bien, équipe, match, jouer, coach, gagner, beaucoup, mieux, semaine, saison	la touche jeune talent, victoire de donald trump	6.5%
3	artiste, festival, concert, art, musique, spectacle, album, musée, exposition, culturel	pop art, artiste belge	4.3%
4	communal, commune, projet, bourgmestre, ville, échevin, collège, aménagement, conseil, bâtiment	plan communal de mobilité, collège communal	4.2%
5	score, minute, point, but, mi-temps, pause, penalty, quart, second, visiteur	hors-jeu, le dernier des siens	3.9%
6	ville, enfant, centre, asbl, accueil, association, organiser, activité, accueillir, Noël	père Noël, asbl	3.8%
7	club, saison, joueur, équipe, jeune, entraîneur, sportif, football, foot, prochain	complexe sportif de rixensart, coach sportif	3.6%
8	euro, prix, million, énergie, vente, déchet, coût, montant, budget, énergétique	autonomie alimentaire et énergétique, prix des carburants	3.2%
9	décision, conseil, procédure, demander, situation, question, demande, comité, règlement, réunion	recours en annulation, règlement	3.2%
10	mesure, masque, magasin, confinement, sanitaire, coronavirus, fermer, mars, SPF, rouvrir	réouverture, masque obligatoire	3.1%
11	entreprise, emploi, travailleur, secteur, syndicat, crise, grève, personnel, travail, milliard	syndicats, travailleur	3.0%
12	match, coupe, point, finale, championnat, play, victoire, off, rencontre, face	free to play, coupe de la ligue	2.9%
13	école, nombre, élève, chiffre, taux, moyenne, enseignement, rapport, scolaire, hausse	enseignement secondaire, décrochage scolaire	2.6%
14	président, parti, ministre, politique, parlement, gouvernement, député, élection, vote, commission	immunité parlementaire, parlement	2.5%
15	prison, tribunal, avocat, prévenu, fait, peine, condamner, procès, cour, accuser	tribunal correctionnel de Tournai, tribunal correctionnel	2.5%
16	heure, dame, nicolas, senior, olivier, laurent, christophe, julien, thomas, antoine	jean marie lepen, jean olivier collinet	2.4%
17	coronavirus, test, pays, cas, covid-19, trump, américain, chine, pandémie, tester	covid-19 long, patients covid-19	2.4%
18	réseau, étudiant, vidéo, social, facebook, université, internet, plateforme, message, photo	regé-jean page, publication en 1782	2.2%
19	police, policier, enquête, parquet, suspect, interpellé, victime, individu, arme, homme	agression au couteau, meurtre	2.1%
20	film, acteur, hommage, cinéma, prince, reine, réalisateur, star, père, roi	prince william, prince	2.0%
21	pompier, incendie, eau, feu, secours, inondation, dégât, zone, évacuer, habitation	incendie, incendie de voiture	2.0%
22	tour, étape, coureur, peloton, course, sprint, maillot, evenepoel, kilomètre, remco	course sprint, sprint	1.9%
23	standard, anderlecht, bruges, genk, but, antwerp, eupen, match, malines, zulte	anne-françoise rousse, union européenne	1.9%

24	kilomètre, course, épreuve, challenge, rallye, parcours, esp, départ, organisateur, édition	rallye finlande, mont ventoux denivele challenge	1.9%
25	russe, ukrainien, guerre, ukrainien, russe, militaire, armée, moscou, poutine, kiev	soldat ukrainien, invasion de	1.9%
26	olympique, record, médaille, tokyo, athlète, bronze, jeu, champion, catégorie, compétition	athlétisme, athlète paralympique	1.8%
27	tournoi, set, atp, wta, finale, tennis, open, australie, chelem, djokovic	grand chelem, tennis de table	1.7%
28	climatique, forêt, espèce, nature, arbre, plante, jardin, île, naturel, terre	forêt, biodiversité	1.5%
29	véhicule, voiture, accident, circulation, rue, conducteur, gare, route, bus, train	conducteur de voiture, accident de bus	1.5%
30	vaccin, vaccination, santé, patient, hôpital, dose, médecin, soin, vacciner, médical	soins hospitaliers, patients non vaccinés	1.5%
31	belgique, france, belge, allemagne, pays, italie, suisse, espagne, qatar, portugal	choix goncourt de la belgique, coupe de belgique	1.3%
32	femme, attentat, violence, manifestation, manifestant, paris, droit, nucléaire, terroriste, taxi	groupe terroriste etat islamique, etat islamique	1.3%
33	huy, carte, jaune, arbitre, verlainne, hutois, seraing, hamoir, rebecq, warnant	d3 acff, d3a acff	1.3%
34	église, produit, œuf, chocolat, pomme, légume, spatial, huile, pape, pain	chocolat au lait, galette de pomme de terre	1.2%
35	league, ligue, contrat, manchester, city, club, pro, liverpool, united, milan	pro league, pro basket league	1.2%
36	degré, vent, température, sec, ouest, nba, pluie, météo, sen, averse	mercure, météo neige	1.1%
37	diable, real, hazard, madrid, lukaku, bruyne, eden, martinez, kevin, ronelu	real lessard, real sociedad	1.0%
38	vol, aéroport, avion, pont, compagnie, chantier, aérien, passager, brussels, stib	aérien, espace aérien	1.0%
39	namur, dinant, ciney, grez, namurois, namuroi, andenne, gembloux, tubize, loyer	anne loyer, loyer	0.9%
40	animal, parc, chien, carnaval, chat, cheval, cortège, oiseau, sauvage, loup	le carnaval des animaux, chat sauvage	0.9%
41	tournai, mouscron, ath, tournaisien, péruwelz, picard, bel, comines, templeuve, brunehaut	concours ixpé, patois gaumais	0.9%
42	liège, arlon, province, luxembourg, libramont, bastogne, liégeois, neufchâteau, pepinster, saint-hubert	gare liège guillemins, province de liège	0.7%
43	verviers, waremme, herve, hanut, aubel, malmedy, verviétois, battice, welkenraedt, hombourg	sport namurois, salle communale	0.7%
44	fra, pilote, ita, mercedes, hamilton, max, verstaappen, bull, ferrari, écurie	max hurdebroucq, one piece red	0.7%
45	braine, brabant, jodoigne, genappe, nivelles, Waterloo, ottignies, étoile, wallon, perwez	festival musiq3 brabant wallon, parquet du brabant wallon	0.5%
46	seh, sef, chastre, sem, forem, amélie, jacobs, jakobsen, oreye, fusil	coup de poing, armes de chasse	0.4%

DH	Top 10 words	Label	Dist.
1	bien, chose, beaucoup, jamais, dire, vraiment, moment, trop, mal, temps	vivre ici, finlande le pays du bonheur	8.0%
2	club, saison, joueur, équipe, coach, anderlecht, jouer, match, entraîneur, football	pro league challenger, pro league	5.4%
3	communal, social, travail, service, travailleur, action, citoyen, emploi, conseil, aide	droit social, service citoyen	4.8%
4	but, match, minute, point, face, ballon, score, rencontre, jeu, victoire	arbitre, hors-jeu	4.7%
5	prison, tribunal, fait, prévenu, avocat, peine, victime, condamner, sursis, correctionnel	tribunal correctionnel de mons, tribunal correctionnel	4.5%
6	ministre, politique, gouvernement, député, président, parti, parlement, commission, débat, fédéral	parlement flamand, commission parlementaire	3.9%
7	projet, bâtiment, ville, rénovation, site, construction, espace, quartier, communal, aménagement	aménagement du temps de travail, rénovation	3.8%
8	ville, proposer, culturel, Noël, commerçant, fête, événement, centre, organiser, commerce	commerce local, centre culturel	3.7%
9	police, policier, accident, véhicule, parquet, suspect, homme, voiture, conducteur, individu	délit de fuite, comportement suspect	3.5%
10	mesure, sanitaire, annuler, décision, prévoir, reporter, date, règle, comité, organisateur	reporter, annulation	3.4%
11	famille, père, prince, livre, roi, reine, fils, fille, femme, mère	princesse stephanie de monaco, princesse elisabeth	2.8%
12	masque, test, nombre, chiffre, coronavirus, cas, covid, positif, contamination, taux	covid-19 long, contamination	2.6%
13	euro, million, montant, payer, prix, financier, salaire, somme, milliard, budget	euro istanbul, euro u21 2023	2.6%
14	ligue, real, madrid, coupe, psg, messi, club, champion, espagne, barcelone	ligue 1 - football, coupe de la ligue	2.6%
15	américain, trump, afp, biden, donald, macron, joe, états-unis, twitter, emmanuel	twitter, démocrate	2.6%
16	olympique, record, médaille, athlète, tokyo, compétition, jeu, belge, champion, belge	championnats de belgique, championnat de belgique de rugby	2.4%
17	tour, étape, coureur, maillot, evenepoel, peloton, remco, wout, course, sprint	sprint course, sprint	2.4%
18	vaccin, vaccination, dose, santé, vacciner, maladie, variant, pfizer, médicament, cancer	patients covid-19, patients non vaccinés	2.4%
19	rue, circulation, chantier, pont, travail, avenue, voirie, parking, stationnement, piéton	voirie, fermeture voirie	2.3%
20	émission, artiste, chanteur, musique, festival, concert, album, chanson, spectacle, scène	pop rock, spectacle musical	2.3%
21	aéroport, belge, belge, avion, vol, voyage, pays, étranger, aérien, européen	compagnie aérienne, avion	2.1%
22	russe, ukrainien, ukrainien, russie, guerre, militaire, armée, moscou, poutine, kiev	invasion russe, soldat ukrainien	2.1%
23	modèle, version, marque, moteur, johnny, clip, tabac, mode, écran, james	écrans, look	1.9%

24	kilomètre, course, vélo, parcours, épreuve, courir, montagne, trail, marathon, pied	mont ventoux, denivele challenge trail	1.9%
25	diable, city, lukaku, martinez, manchester, inter, chelsea, bruyne, liverpool, roberto	manchester city, swansea city	1.9%
26	énergie, électrique, transport, train, voiture, électricité, gaz, énergétique, snob, véhicule	installations électriques, vélo électrique	1.8%
27	musée, art, exposition, cuisine, œuvre, siècle, chef, patrimoine, étoile, château	art contemporain, exposition - musées	1.8%
28	tournoi, tennis, atp, open, set, wta, finale, chelem, roland, australie	grand chelem, tennis - roger federer	1.8%
29	pompier, animal, incendie, inondation, secours, feu, chien, eau, dégât, évacuer	pompiers, incendie de voiture	1.7%
30	déchet, commune, arbre, sac, collecte, tournai, jardin, église, sauvage, soignées	accueillir la vie sauvage au jardin, ramassage des déchets	1.7%
31	produit, magasin, vin, vente, bière, client, alimentaire, bouteille, production, consommateur	bio, consommateur	1.6%
32	province, liège, brabant, hôpital, luxembourg, wallon, centre, résident, repos, provincial	province de brabant wallon, hôpital	1.6%
33	pilote, rallye, mercedes, ferrari, bull, hamilton, lewis, max, toyota, verstappen	max hurdebroucq, rallye de spa	1.4%
34	école, enfant, élève, étudiant, scolaire, parent, enseignement, jeune, enseignant, classe	scolaire, enseignement secondaire	1.3%
35	degré, vent, température, pluie, sec, météo, neige, averse, mer, chaleur	prévisions météo, thermomètre	1.2%
36	combat, disque, steve, boxe, lions, boxeur, raphaël, arthur, accusé, john	jean paul gaultier, jean dujardin	0.8%
37	film, acteur, cinéma, thierry, série, sébastien, réalisateur, personnage, neuville, tournage	réalisateur, festival du film documentaire	0.7%
38	louvre, blessure, visé, jonathan, genou, eden, julien, loup, hazard, entraînement	blessure, genou	0.7%
39	namur, braine, waterloo, namurois, alleud, maxime, namuroi, jodoigne, rochefort, citadelle	maxime richard, citadelle de namur	0.5%
40	standard, rouche, deila, liégeois, sclessin, amallah, ronny, ema, forain, google	anne-françoise rouche, premier café liégeois	0.5%
41	logement, dinant, luc, ciney, locataire, gembloux, christophe, couvin, vervoort, pepinster	logement, appartement	0.4%

Appendix B

Topics distribution over time - heatmaps

This chapter displays the heatmap plots for every year and every newspapers. The x-axis represents the weeks of a year, and the y-axis represents the topics of the newspaper. Therefor, each cell indicates the marginal probability of a topic in a week.

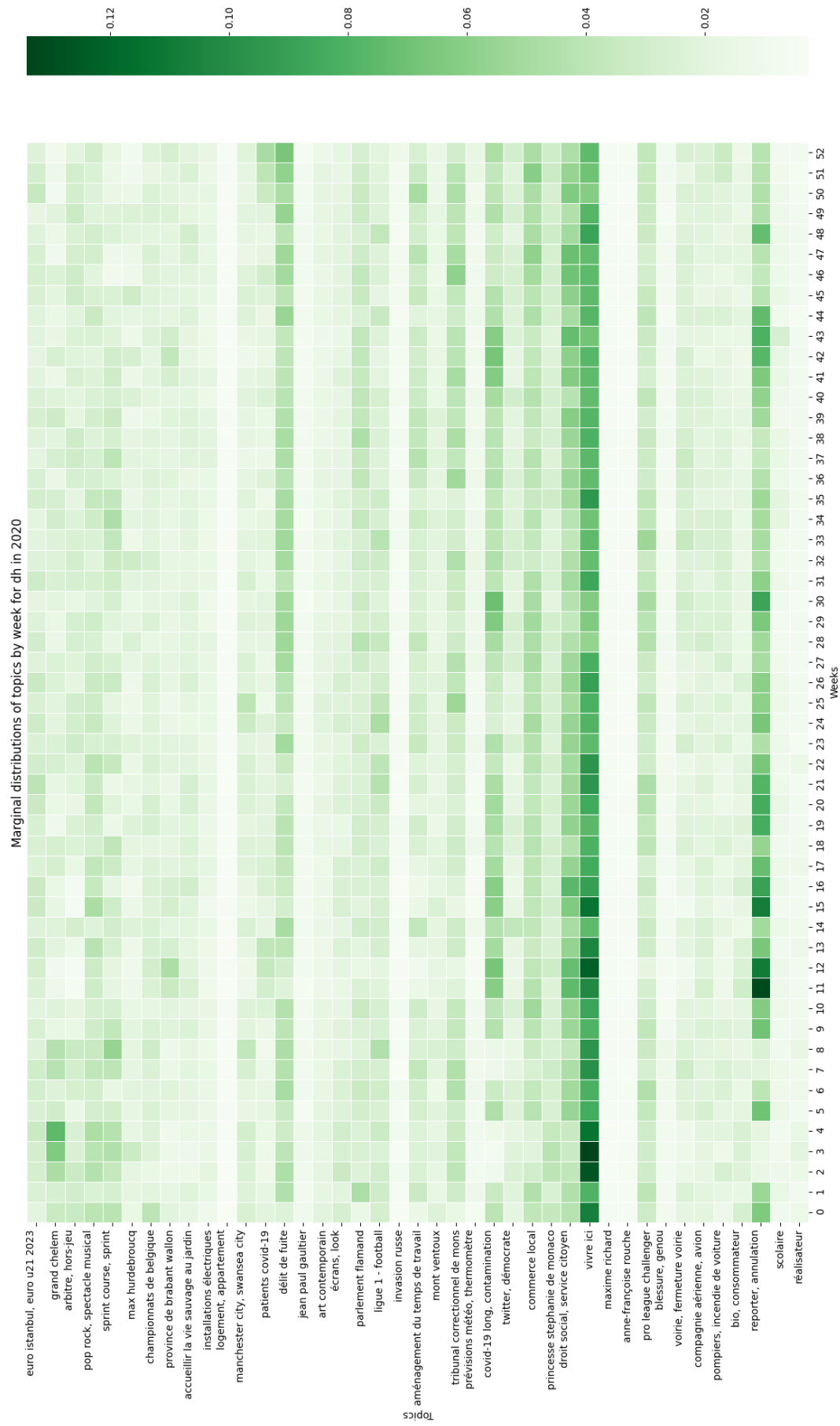


Figure B.1: Marginal distributions of topics by week for DH in 2020

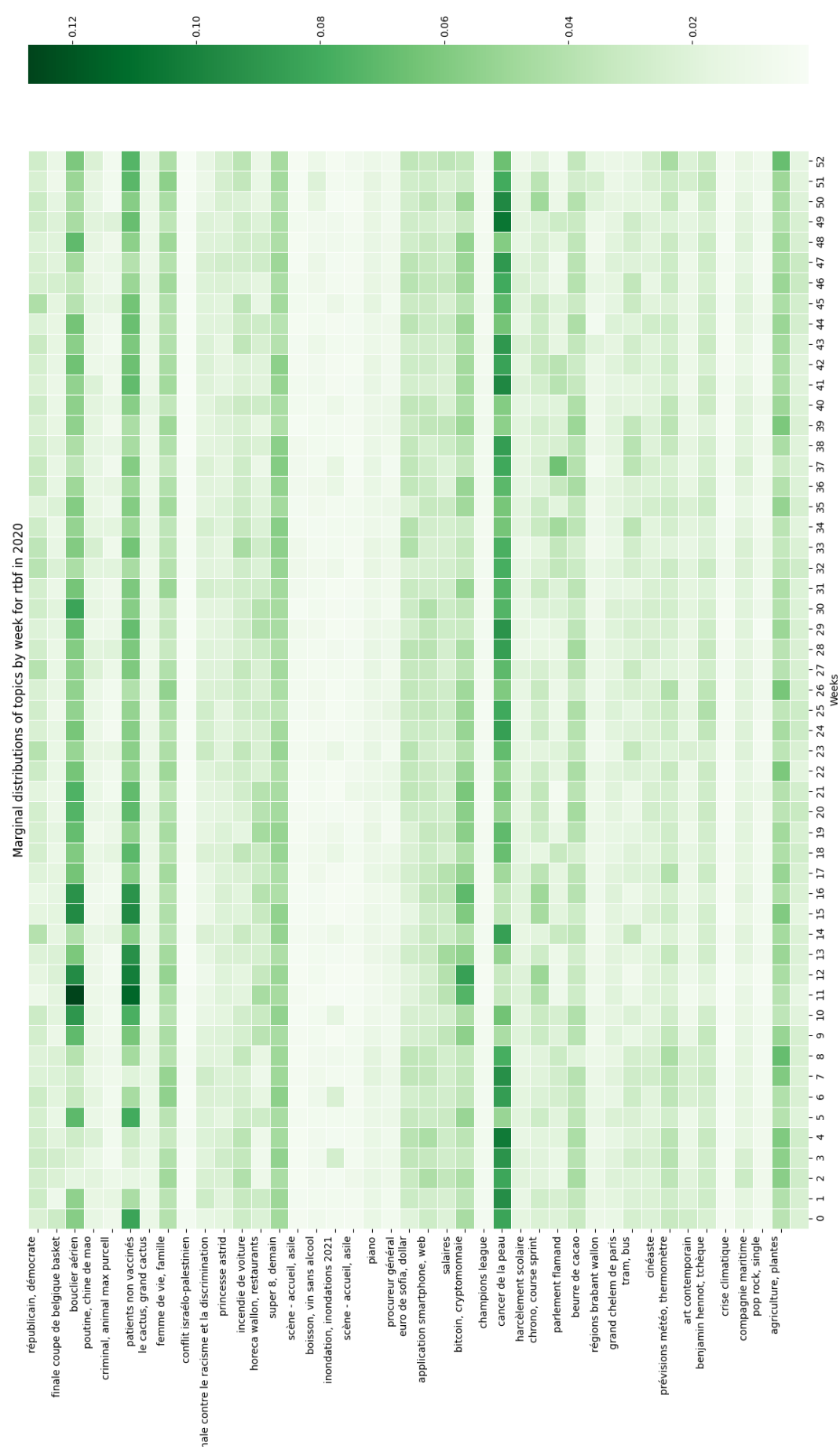


Figure B.2: Marginal distributions of topics by week for RTBF in 2020

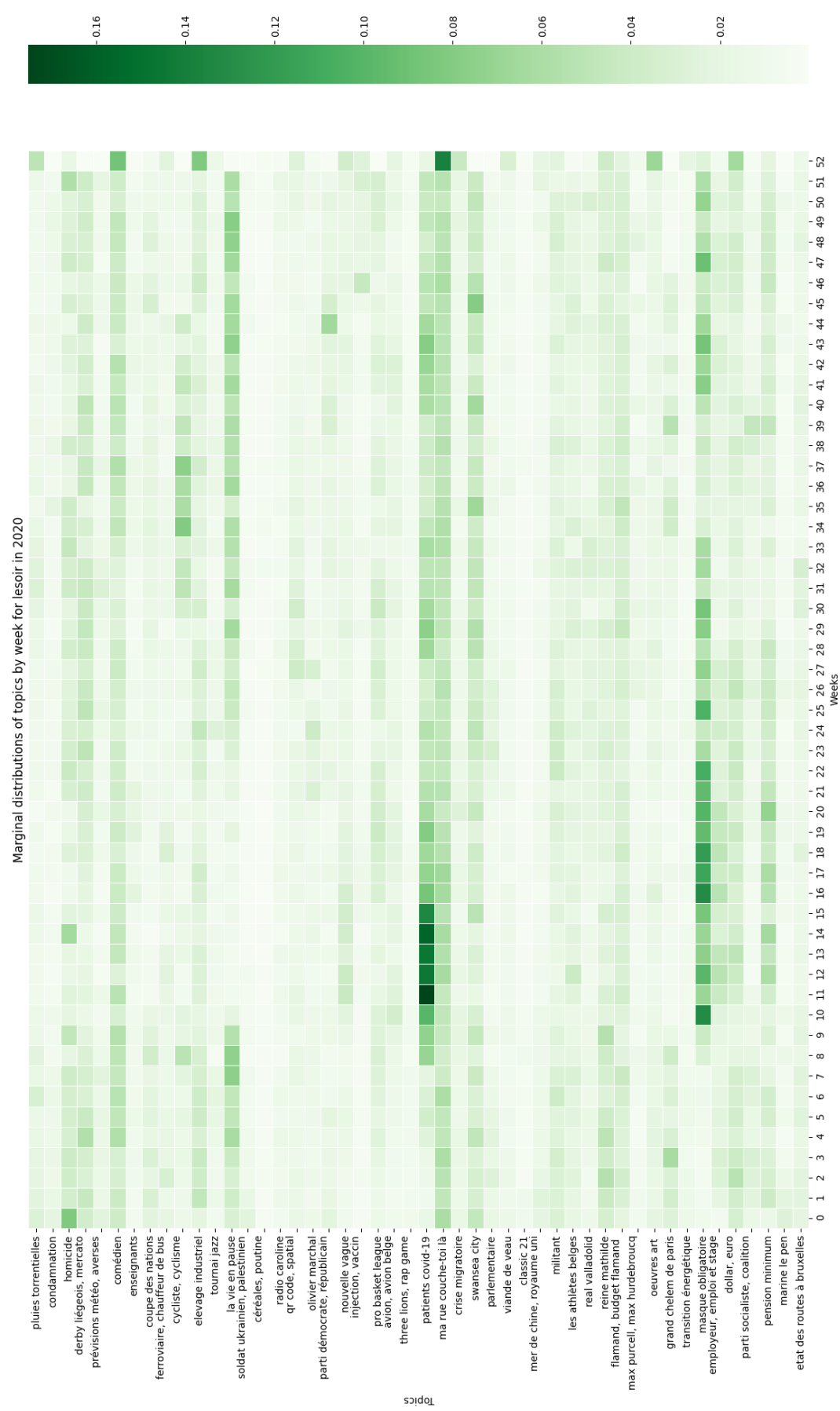


Figure B.3: Marginal distributions of topics by week for Lesoir in 2020

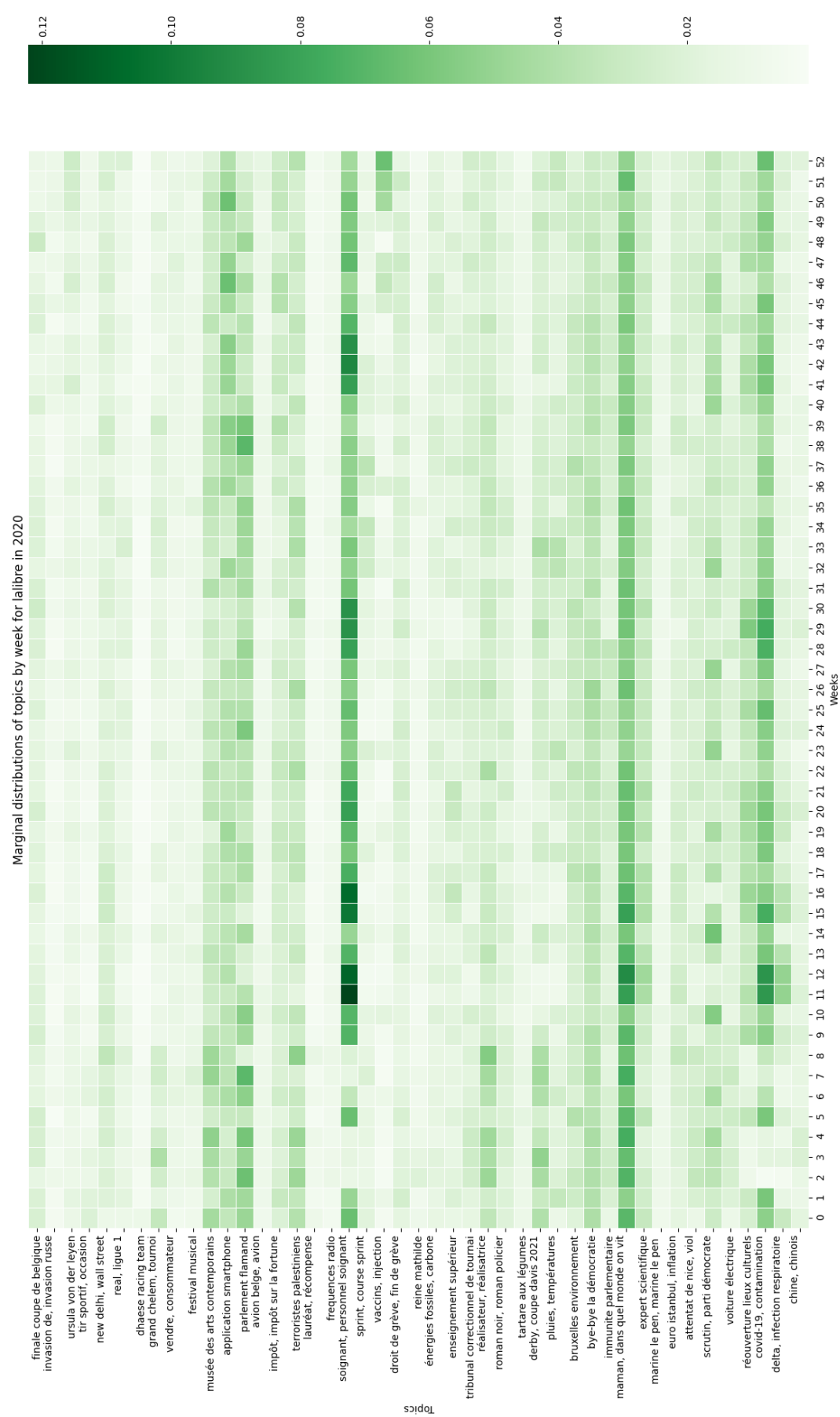


Figure B.4: Marginal distributions of topics by week for Lalibre in 2020

Marginal distributions of topics by week for lavenir in 2020

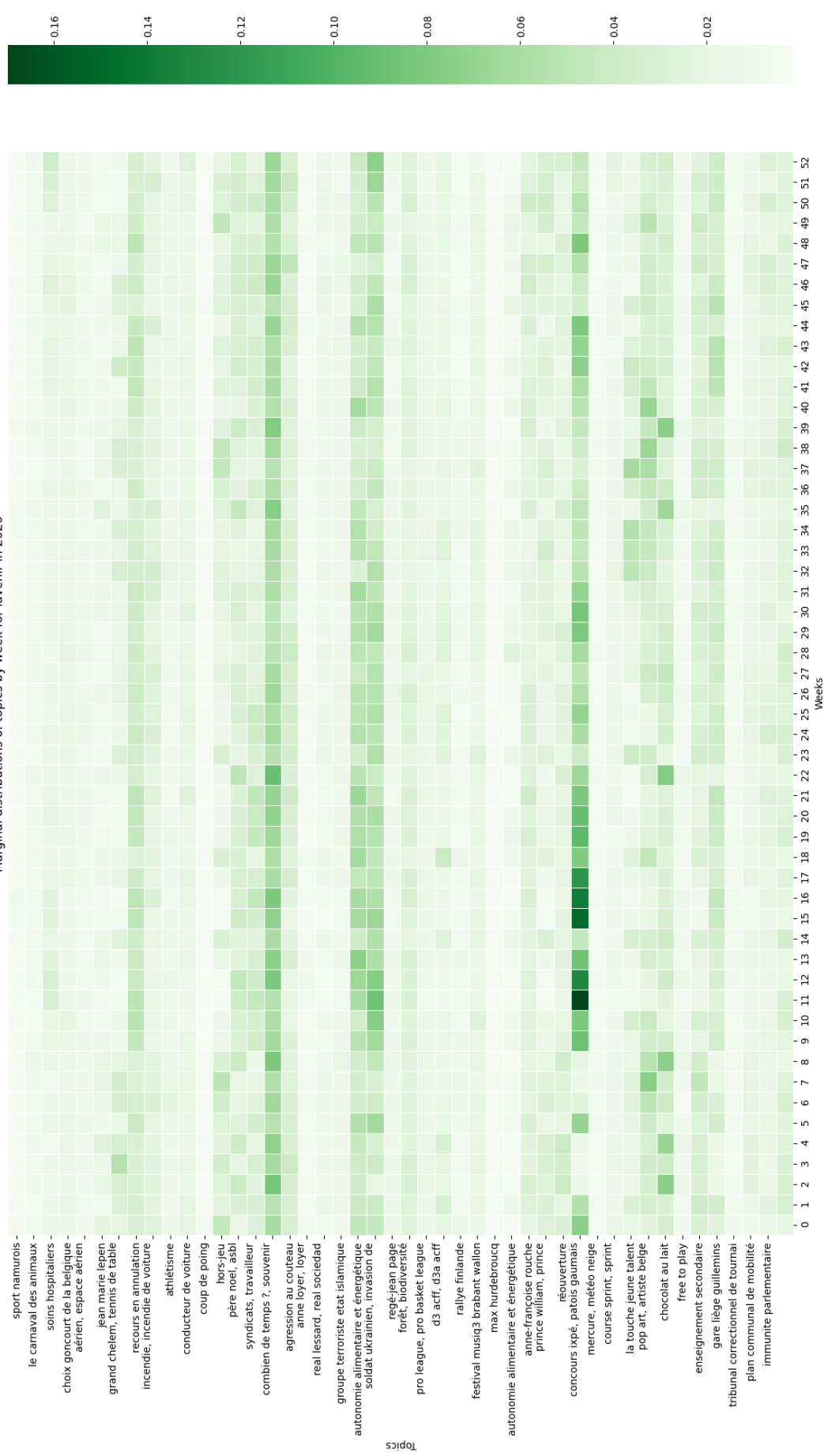


Figure B.5: Marginal distributions of topics by week for Lavenir in 2020

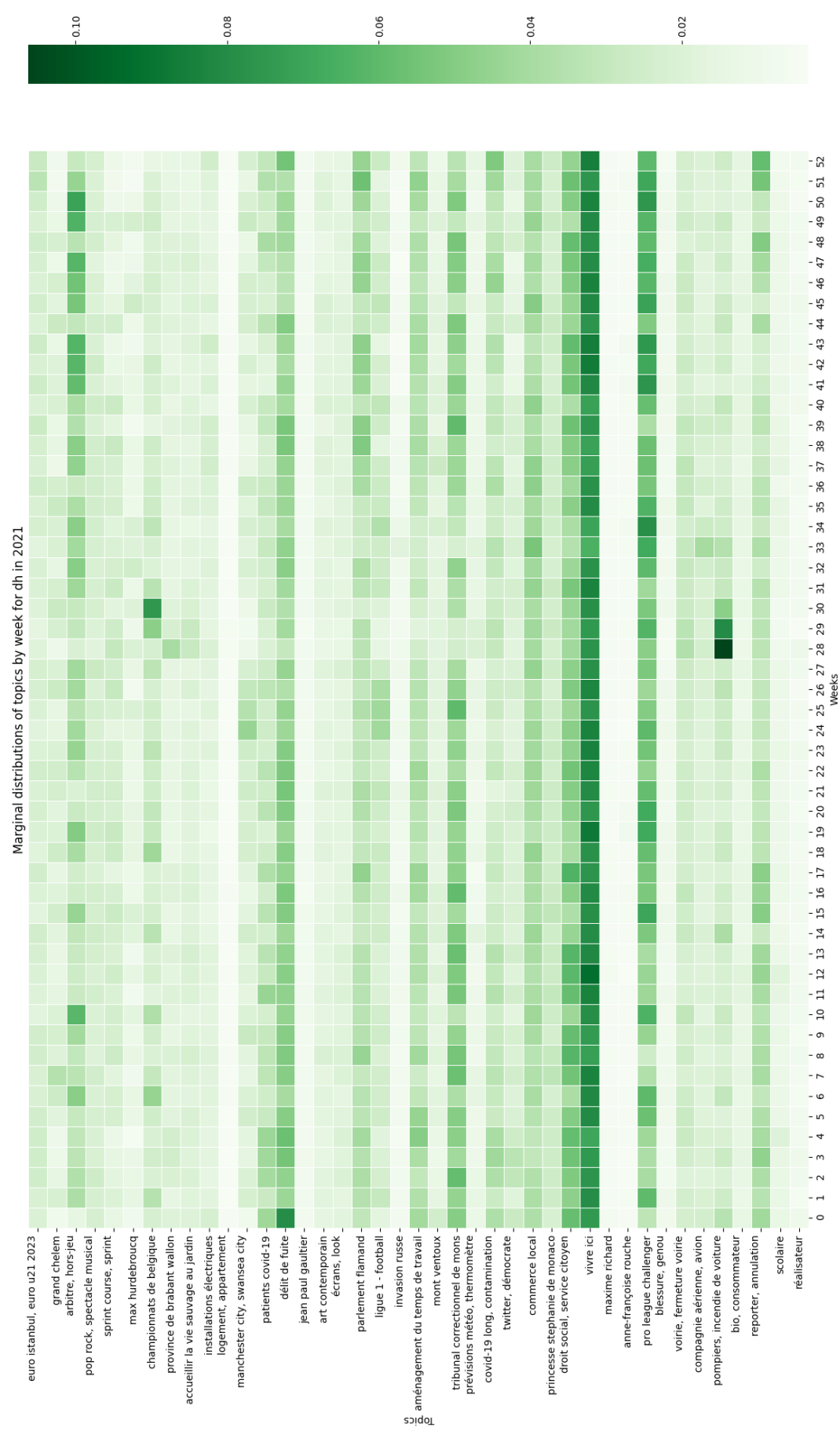


Figure B.6: Marginal distributions of topics by week for DH in 2021

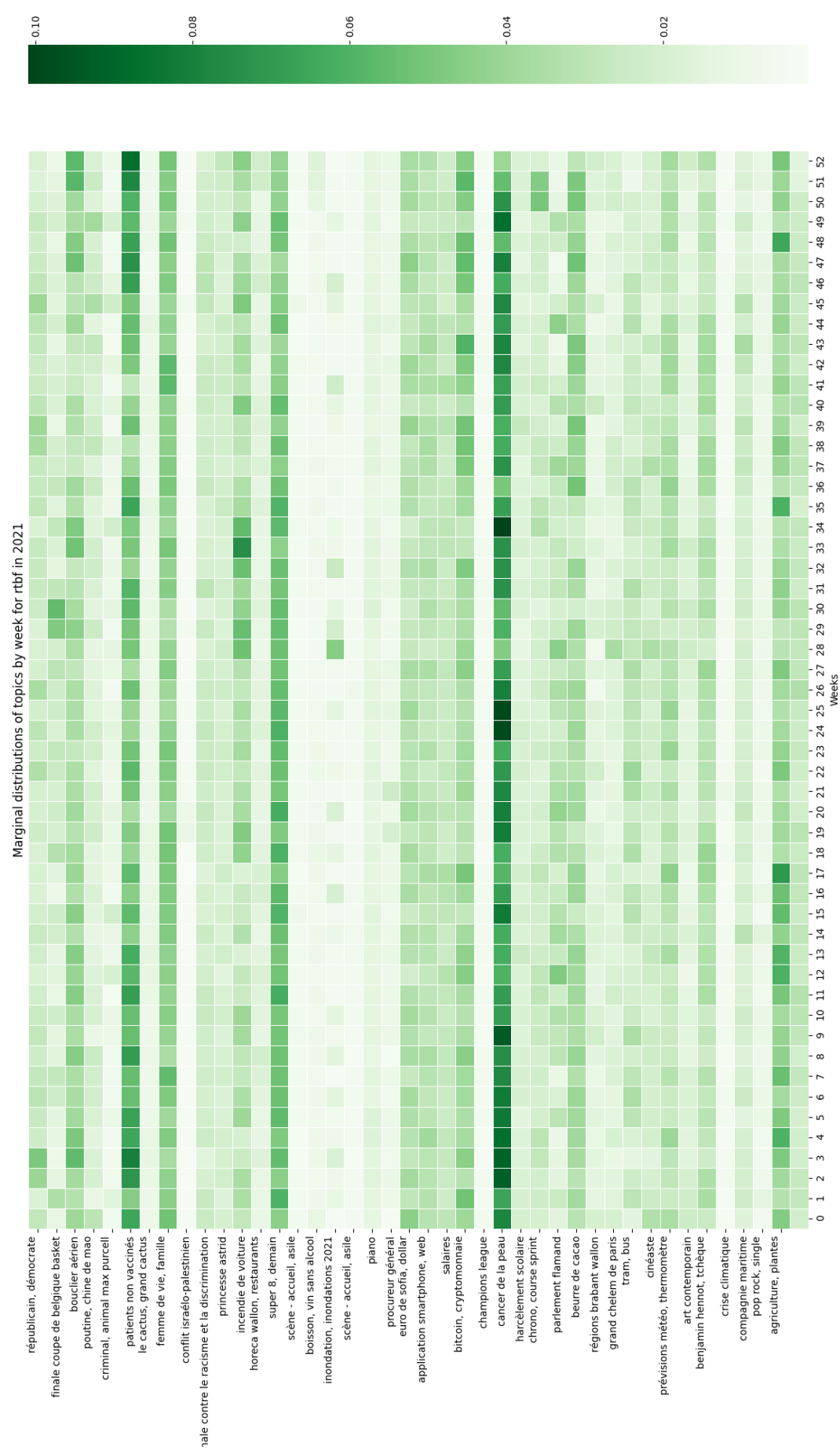


Figure B.7: Marginal distributions of topics by week for RTBF in 2021

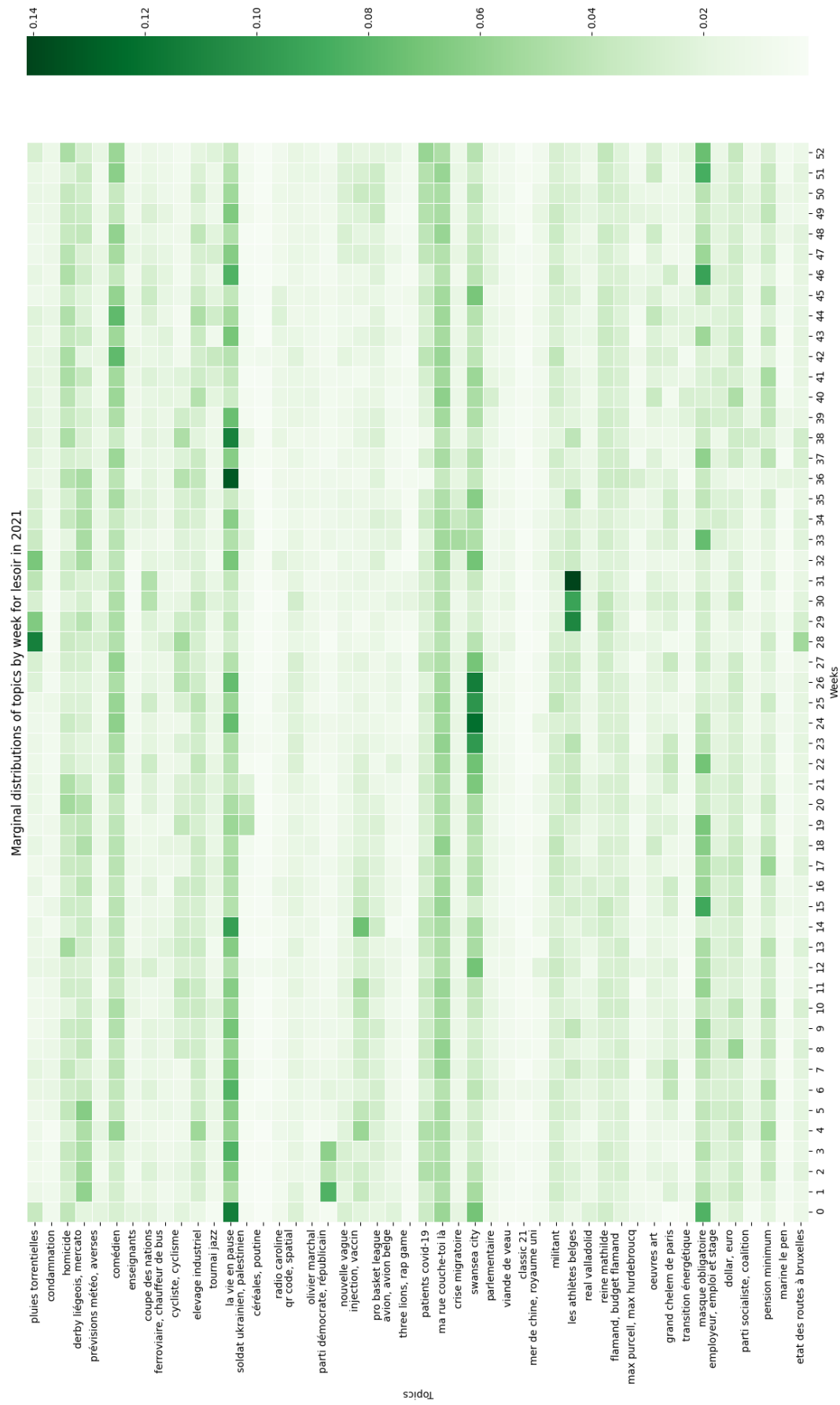


Figure B.8: Marginal distributions of topics by week for Lesoir in 2021

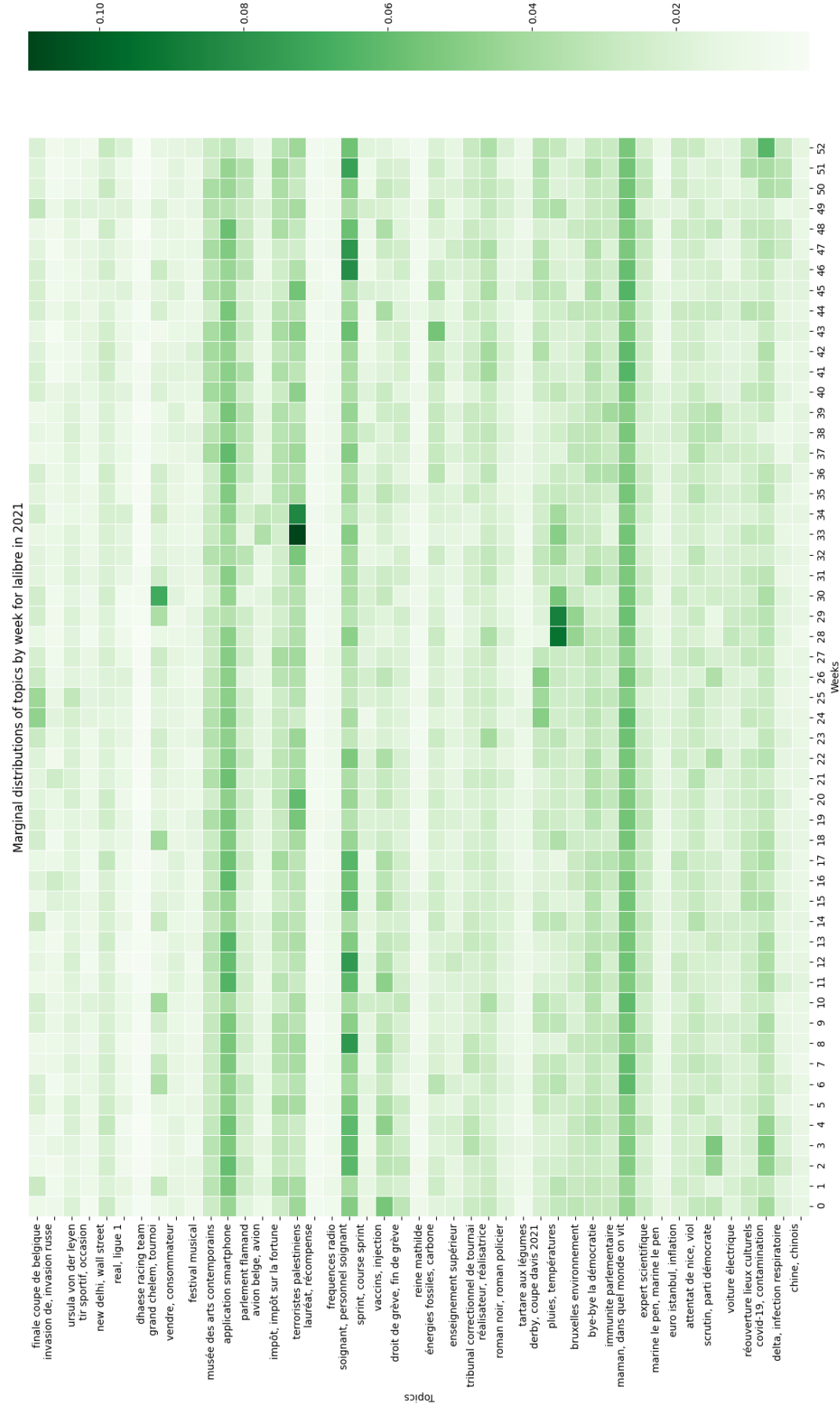


Figure B.9: Marginal distributions of topics by week for Lalibre in 2021

Marginal distributions of topics by week for lavenir in 2021

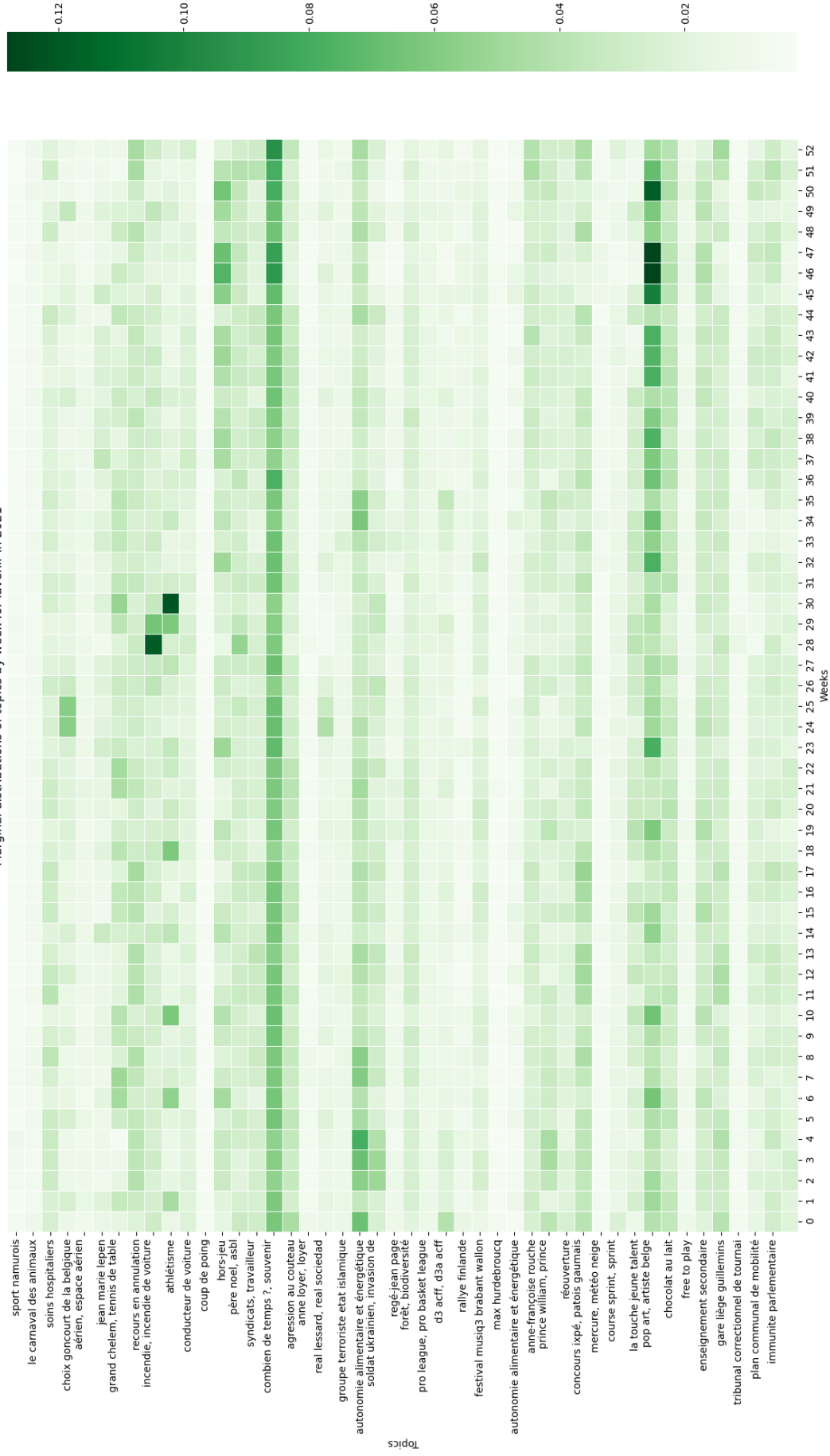


Figure B.10: Marginal distributions of topics by week for Lavenir in 2021

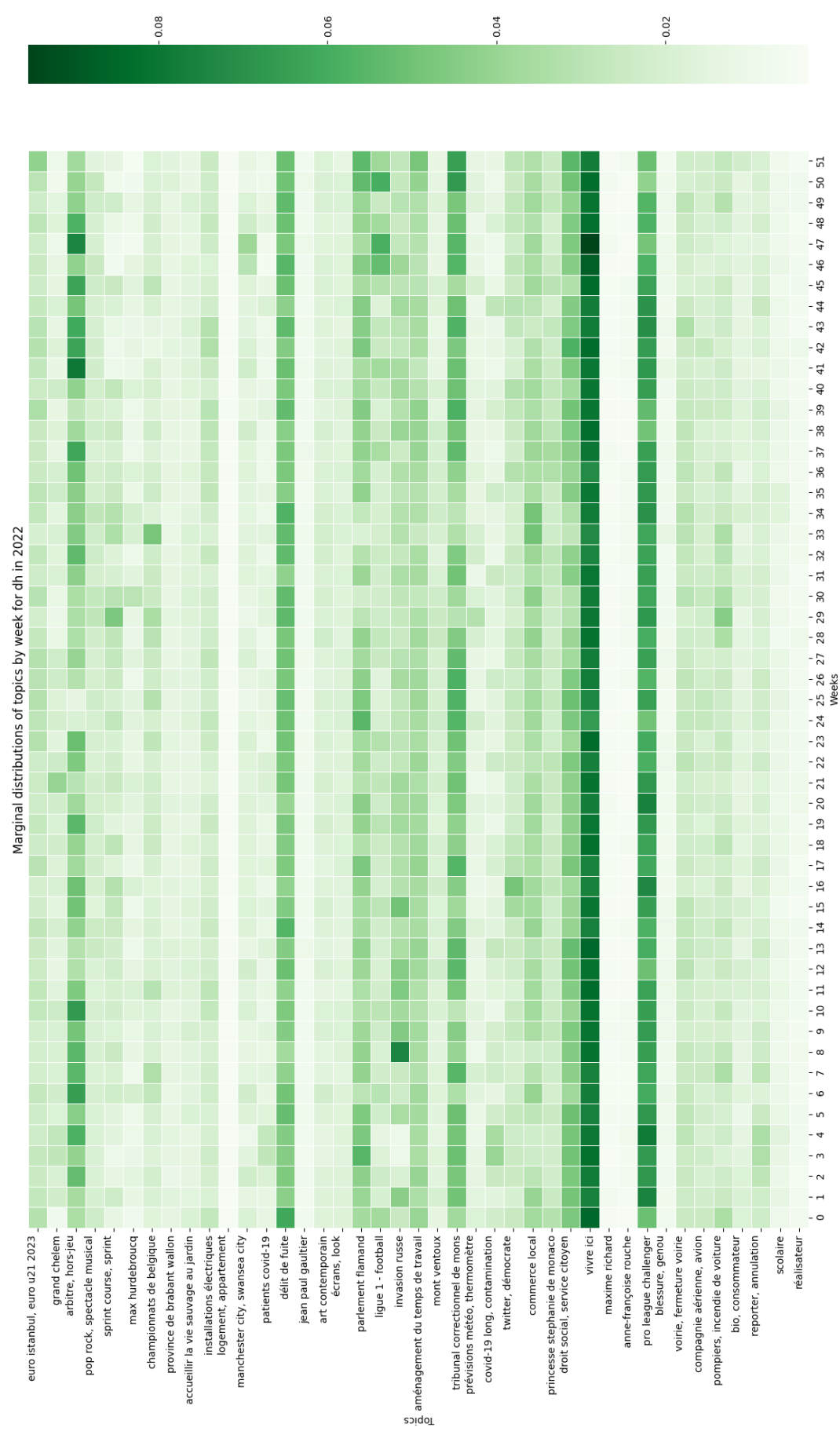


Figure B.11: Marginal distributions of topics by week for DH in 2022

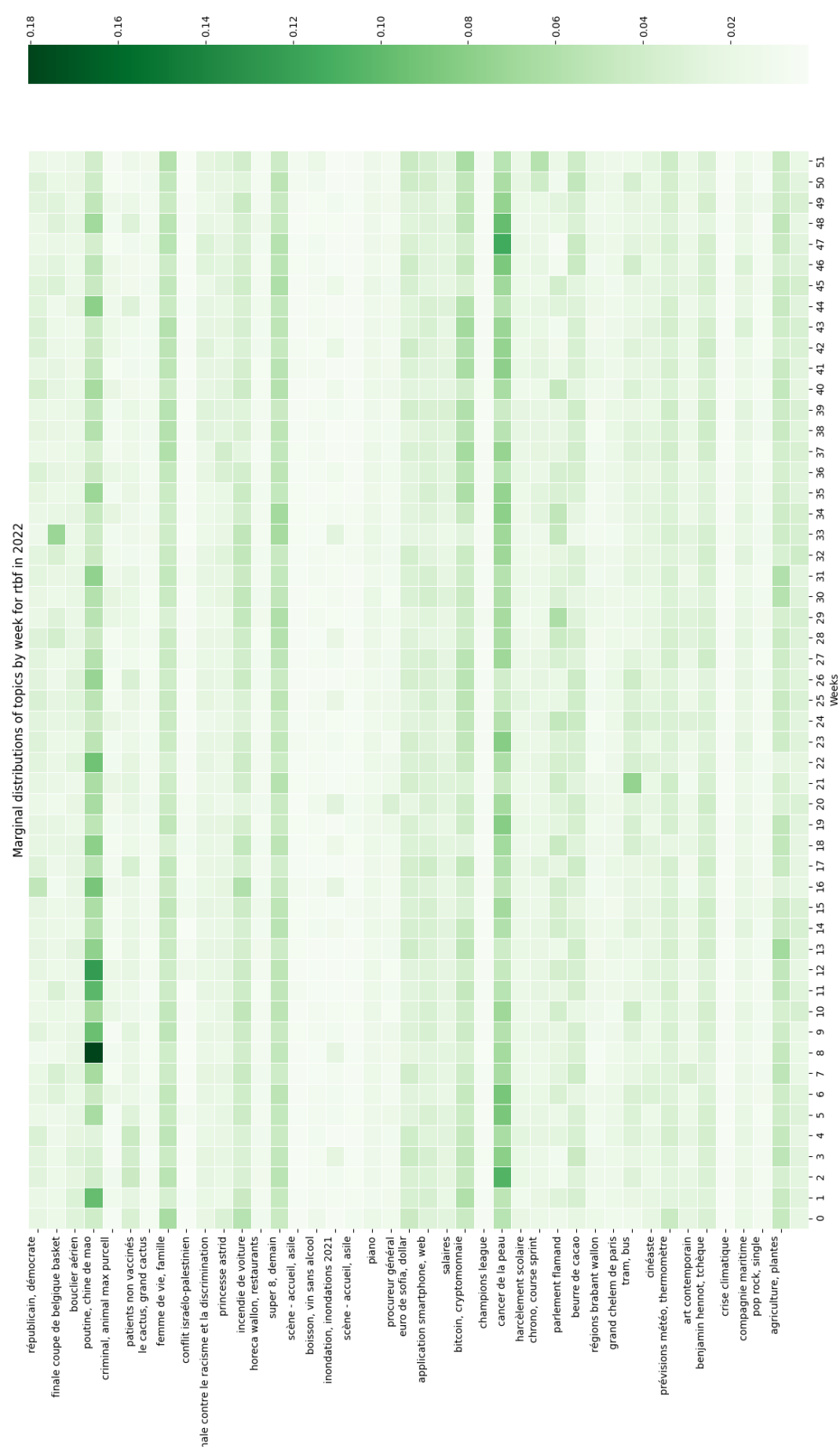


Figure B.12: Marginal distributions of topics by week for RTBF in 2022

Marginal distributions of topics by week for lesoir in 2022

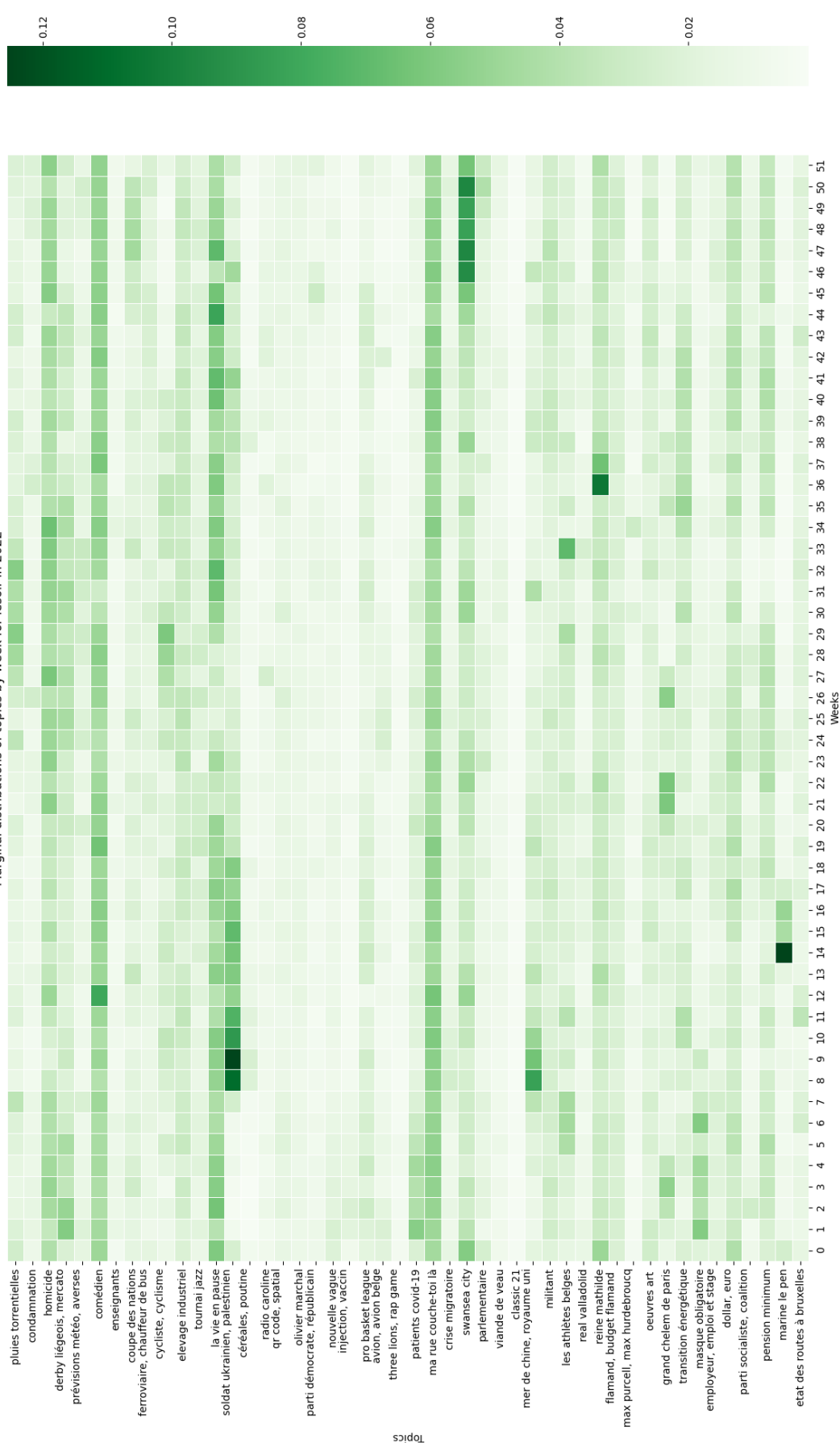


Figure B.13: Marginal distributions of topics by week for lesoir in 2022

Marginal distributions of topics by week for lalibre in 2022

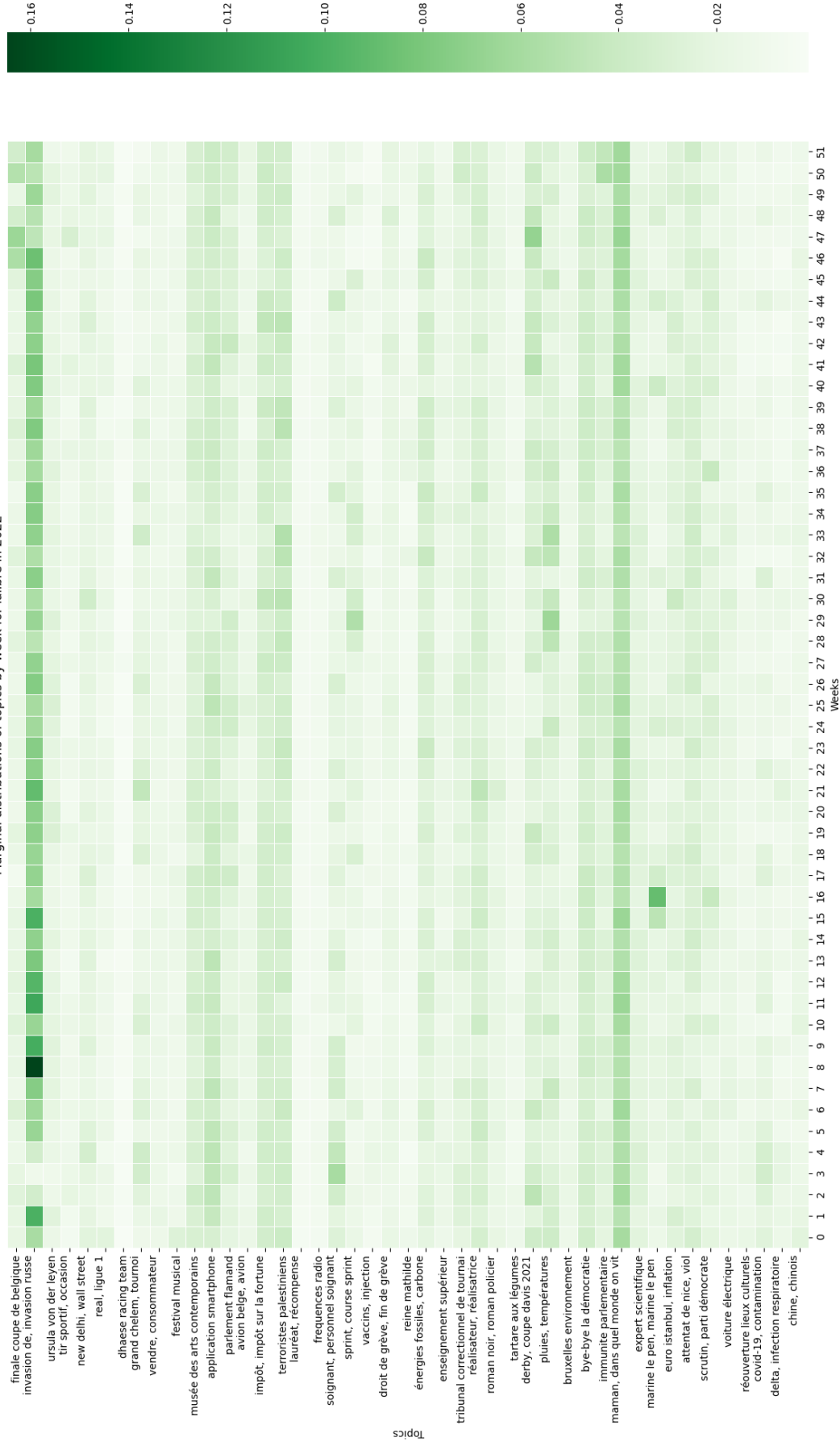


Figure B.14: Marginal distributions of topics by week for Lalibre in 2022

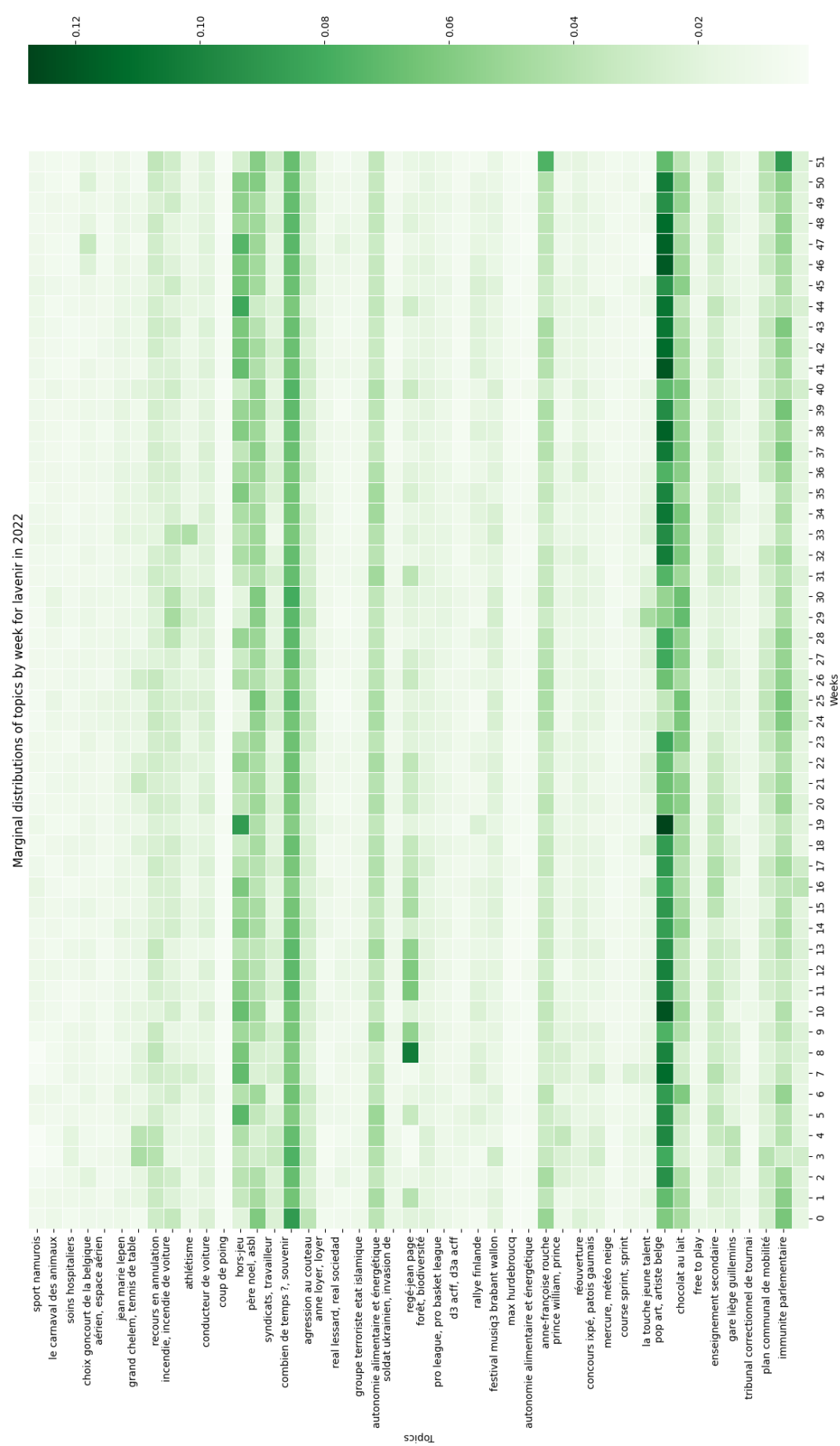


Figure B.15: Marginal distributions of topics by week for Lavenir in 2022

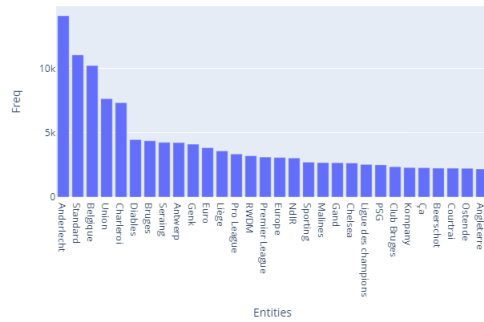
Appendix C

Named entity recognition plots

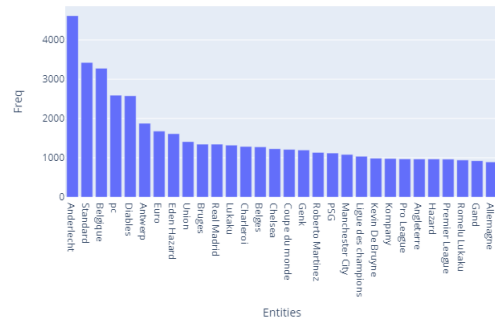
This chapter displays the plots of the most cited entities by topic in the different newspapers.

Football topic

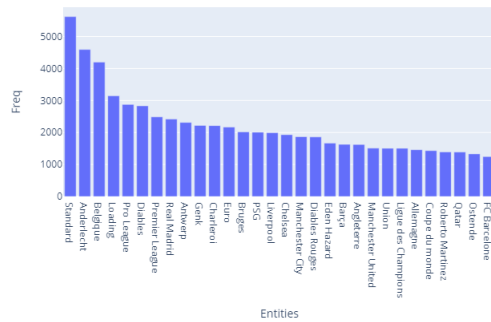
dh - Most cited entities for football topic



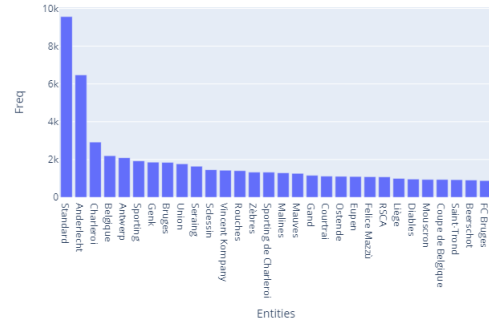
lalibre - Most cited entities for football topic



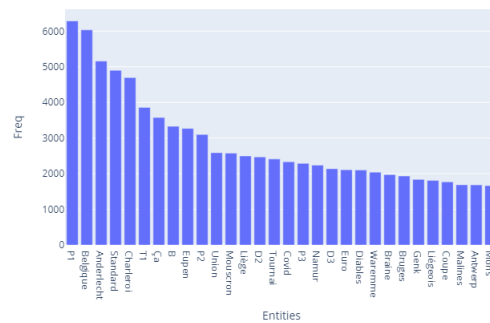
rtbf - Most cited entities for football topic



lesoir - Most cited entities for football topic

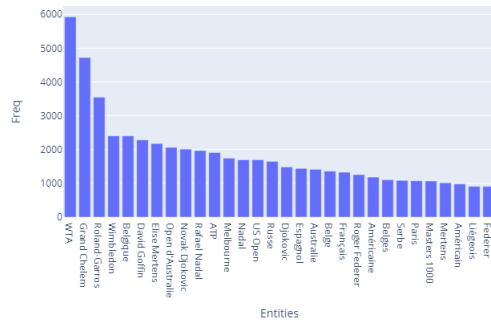


lavenir - Most cited entities for football topic

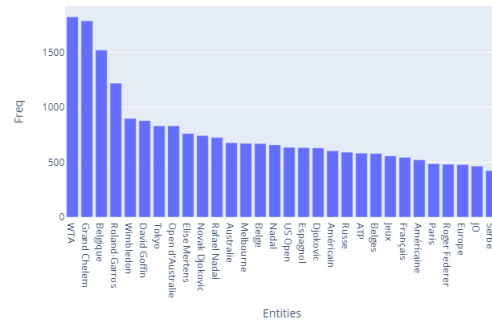


Tennis topic

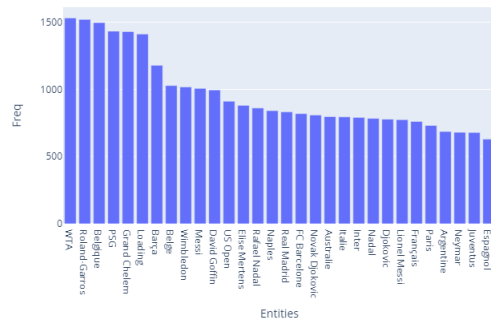
DH - Most cited entities for tennis topic



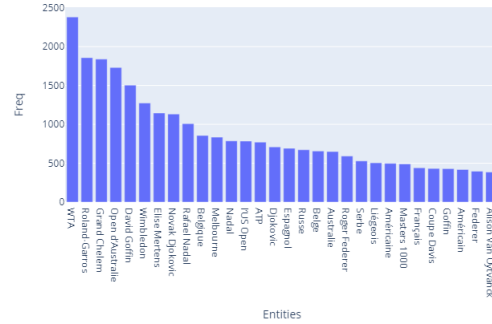
lalibre - Most cited entities for tennis topic



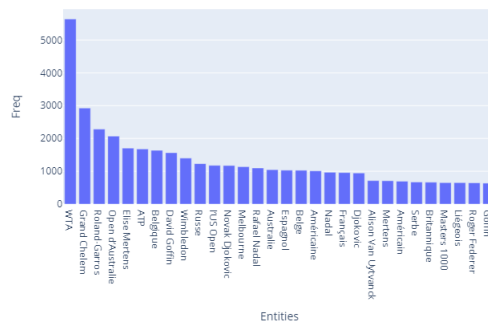
rtbf - Most cited entities for tennis topic



lesoir - Most cited entities for tennis topic

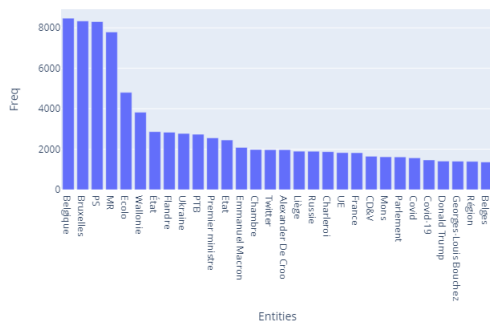


lavenir - Most cited entities for tennis topic

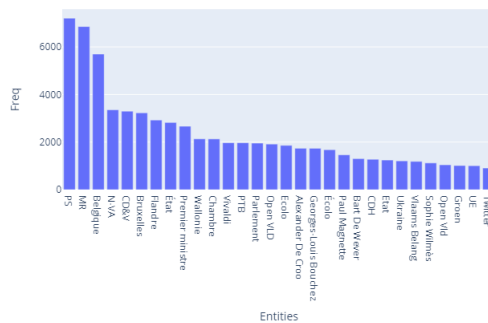


Political topic

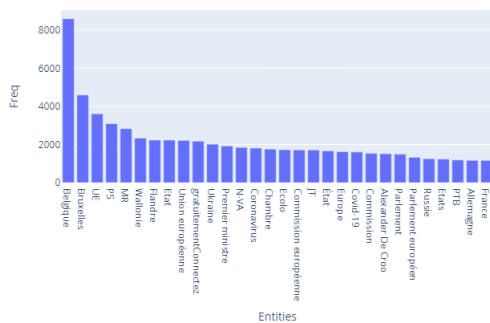
DH - Most cited entities for political topic



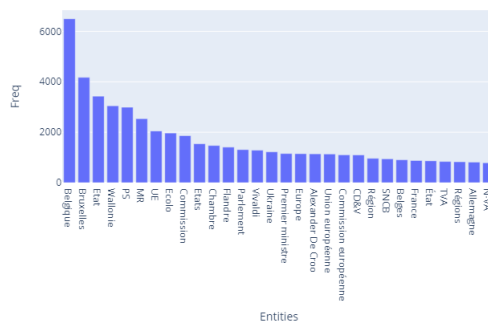
Lalibre - Most cited entities for political topic



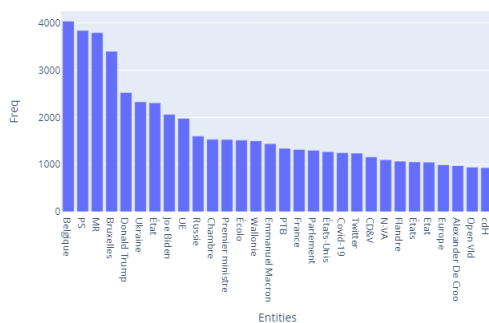
RTBF - Most cited entities for political topic



Lesoir - Most cited entities for political topic

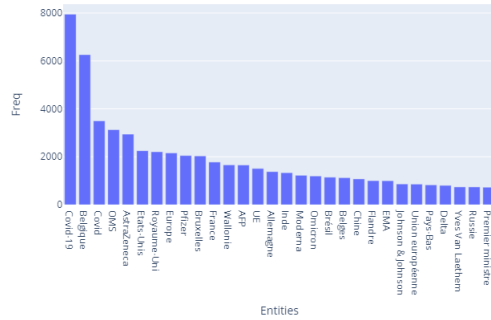


Lavenir - Most cited entities for political topic

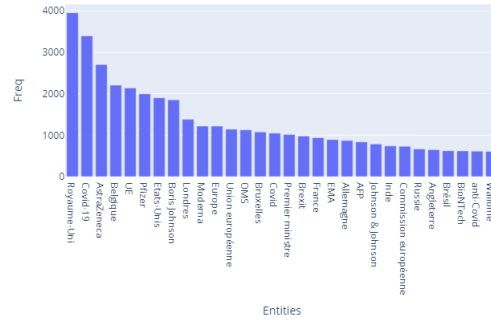


Covid vaccine topic

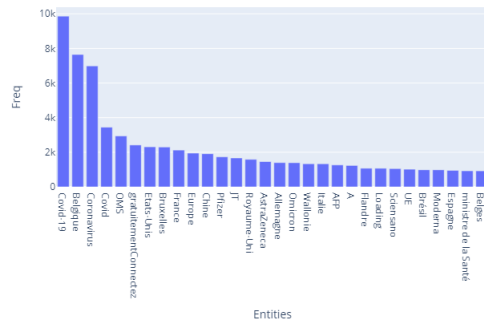
DH - Most cited entities for Covid vaccine



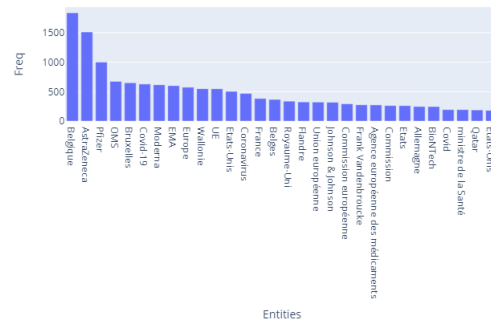
Lalibre - Most cited entities for Covid vaccine



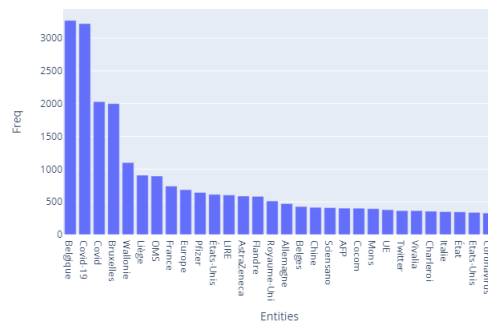
RTBF - Most cited entities for Covid vaccine



Lesoir - Most cited entities for Covid vaccine



Lavenir - Most cited entities for Covid vaccine



UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl