Analysis with 16 topics

Vestin Hategekimana

2023-06-19

## Introduction

In this papre, we delve into the fascinating world of topic modeling using the stm package in R. Our goal is to uncover latent themes and patterns within a vast collection of textual data. The journey has been both intriguing and time-consuming, requiring meticulous efforts, including a fair amount of reverse engineering. While we have made significant progress, there is still work to be done as we strive to refine our models further.

The initial stages of our analysis have predominantly focused on descriptive exploration, employing a wide array of visualization techniques to gain insights into the underlying structure of the data. Through these visualizations, we have sought to extract meaningful information and unravel the intricate relationships between topics, shedding light on the hidden dynamics within the corpus.

One crucial aspect of our investigation involved data cleaning, a meticulous process that aimed to enhance the quality and reliability of our results. Notably, after undertaking the cleaning steps, we observed significant changes in the composition of the initial 16 topics. These alterations prompted us to diligently monitor and document their transformations, ensuring a comprehensive understanding of the evolution throughout the analysis.

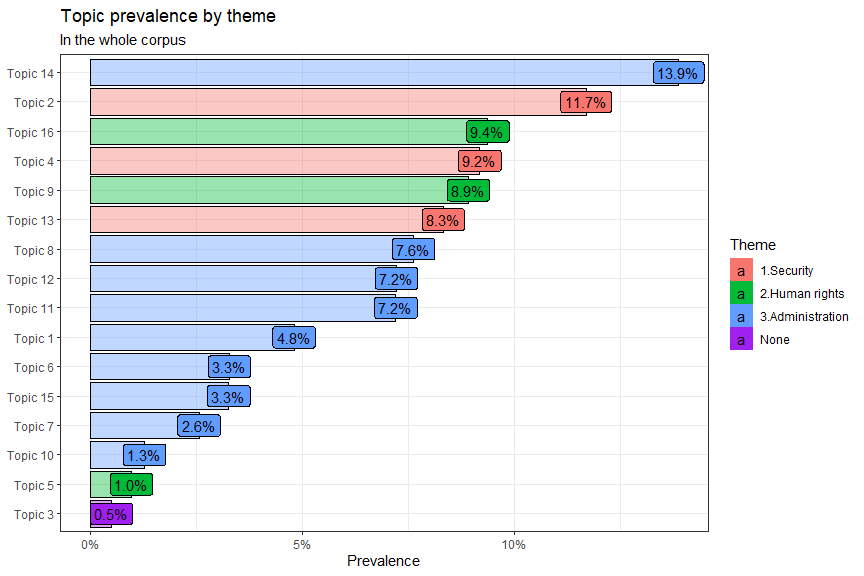
To ensure a comprehensive record, we have included a detailed description of the revised 16 topics in the annex section, allowing us to track their evolution before their eventual merging. However, we anticipate that with further data cleaning, the likelihood of requiring additional revisions to the topics will diminish, providing a more stable foundation for subsequent analyses.

## 16 topics

For the analysis we chose to use 16 topics to have a more refined distribution of topics. Then we grouped theme in 3 themes:

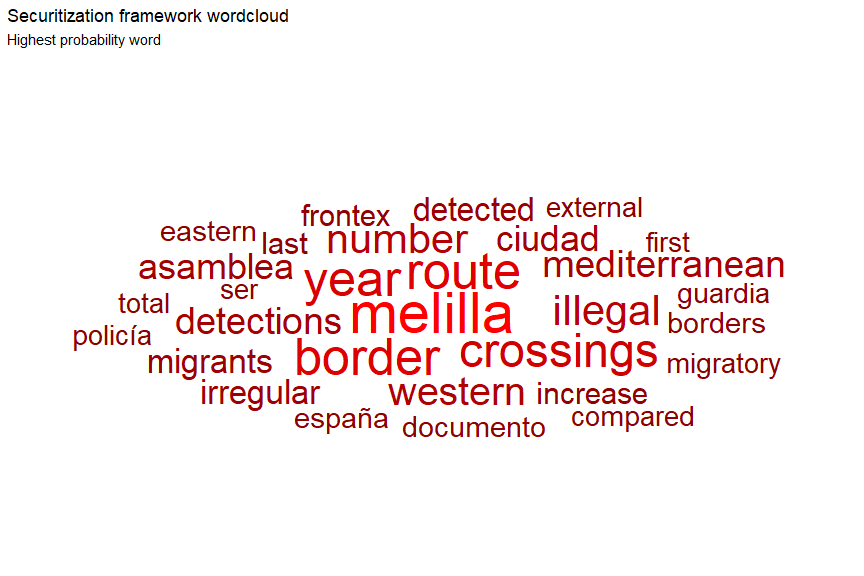
1. **(Security) Securitization framework**
   * Topics: 2, 3, 4, 13
2. **(Human right) Human rights / migrant rights framework**
   * Topics: 7, 9, 16
3. **(Administration) Administrative framework**
   * Topics: 1, 5, 6, 8, 10, 12

Here is an illustration on how the topics are distributed:

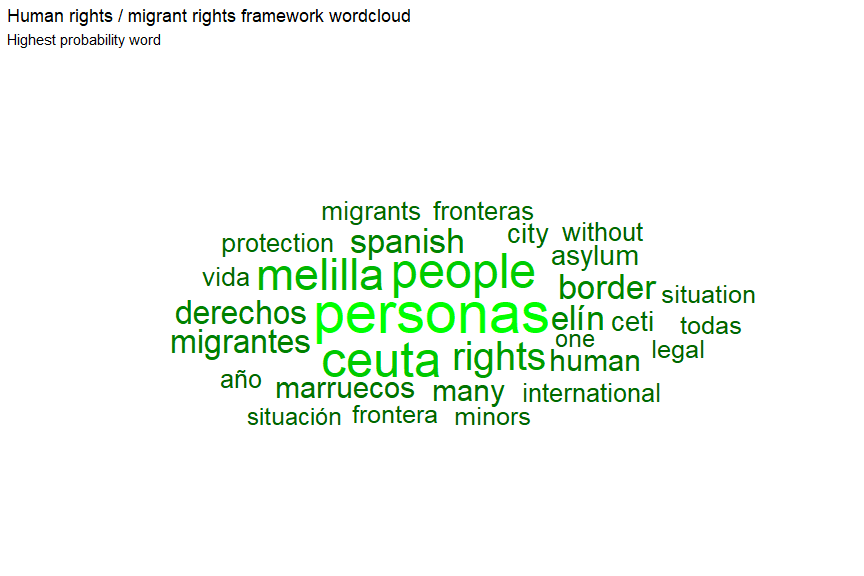


Each theme will be described.

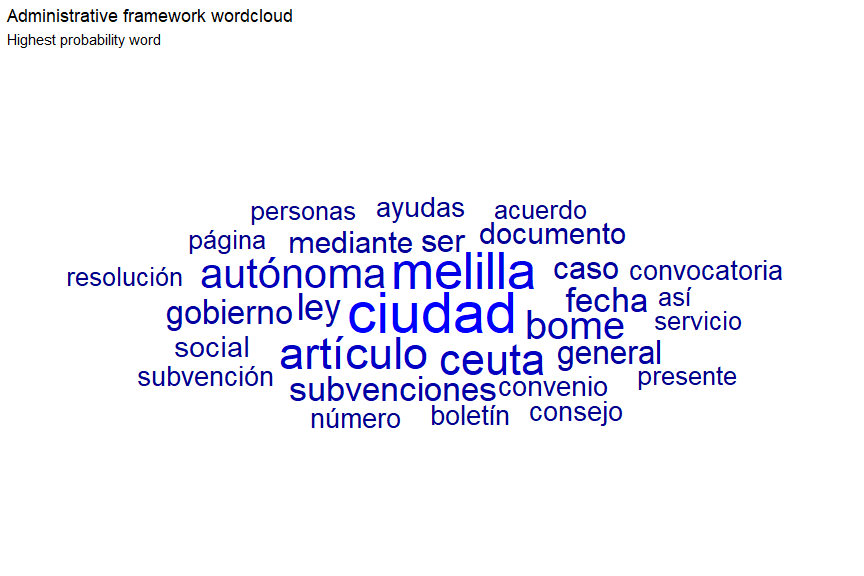
### (Security) Securitization framework



### (Human right) Human rights / migrant rights framework



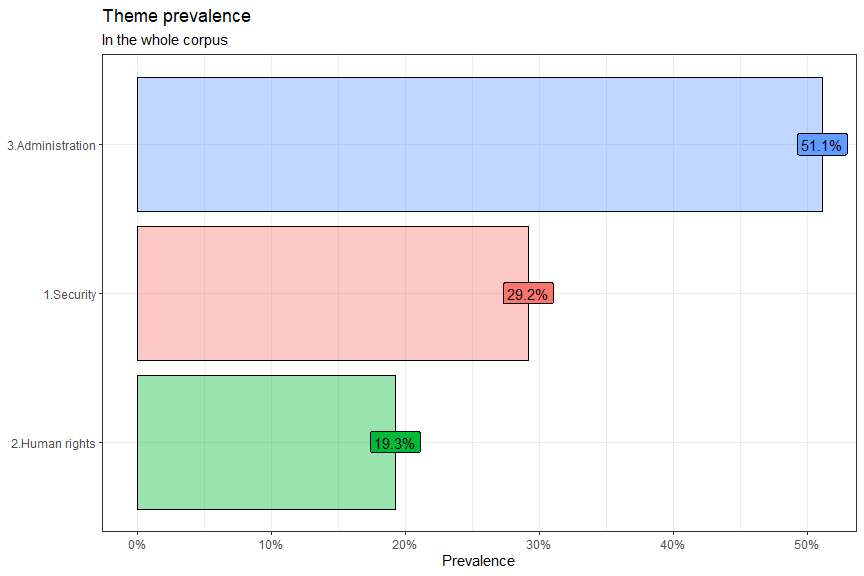
### (Administration) Administrative framework



## Topic prevalence

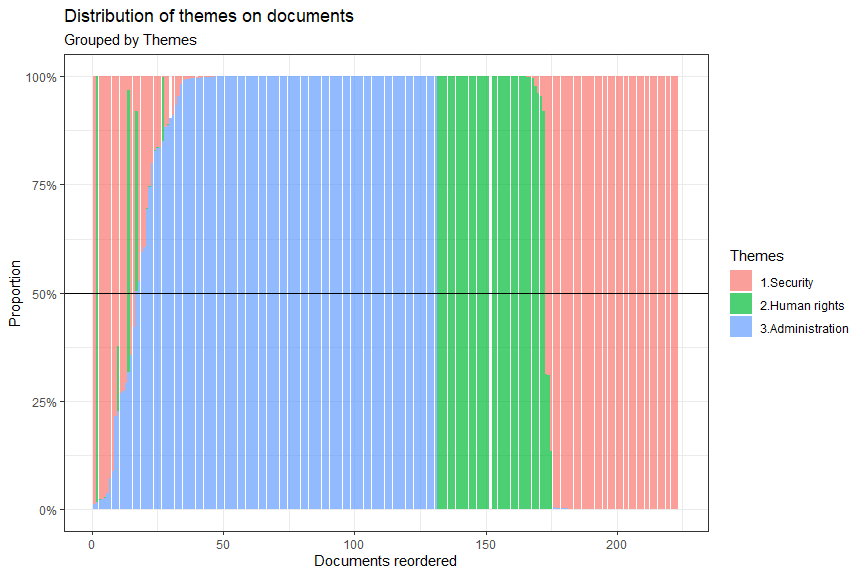
### Overall

The Administrative framework represents a half of the theme, wheras the Securitization framework represent ~30% and the human/migrant rights framework represent ~20%.



### Documents aligned

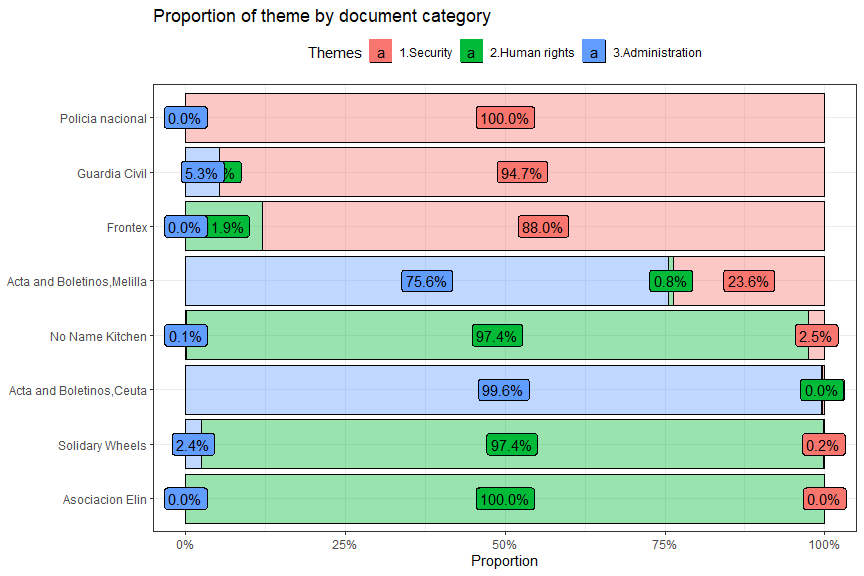
When we align all the document to see the proportion of each topic on them, we can clearly see that almost all the documents have a dominant topic (more thant 50% of their content). Indeed only one document doesn’t have any dominant topics because it was the topic 3 that we removed for inconsistance. it meens that the topic 3 was highly influenced by this document (BOME-AX-2022-6.pdf).



### Prevalence by author/category

When we plot the themes prevalence by author/category, we clearly see that some topic are more relate to certain institutions. Whithout surprise, human right oriented institution tend to have a bigger proportion of human/migrant right framework wheras institution from Mellila and Ceuta have more administrative framework. The “Policia Nacional”, the “Guarida Civil” and Frontex tend to have more Securitization framework. We also highlight that the institution of Meililla have a good share of Securitization framework (23.6%), which is more than the administration of Ceuta.

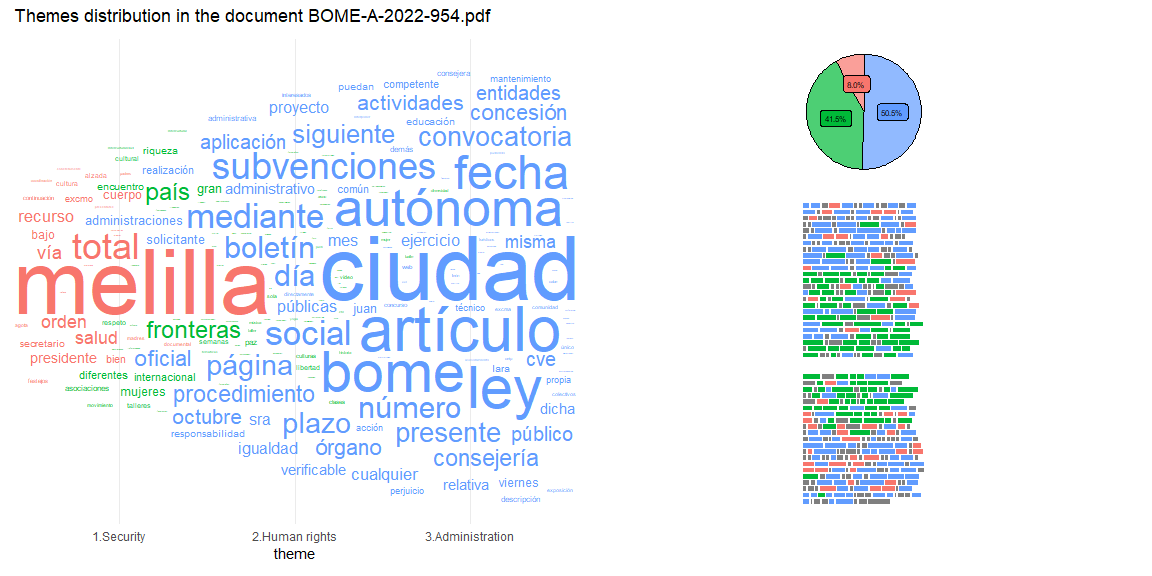
*Note: The graph is awful, I will think of a better way to draw it.*



## In depth document analysis

Has we have seen before, (almost) all the documents have a dominant theme. But here we are interested by documents that have less predominant themes, mining more balanced distribution of themes. To assess that we compute the gini coefficient for each document to see which one are the less inequal (more balanced). A gini coefficient of 0 inidicate no inequality (all the themes in a document have the same proportion) wheras a value of 1 indicate that only one topic si dominant with 100% of the share in a given document. Generally we except more values in between, knowing that a Gini of 0.5 already indicate a high level of ditribution inequality. In our case, the document with the lowest gini score (the most equal) already scores 0.42 (quite high inequality), it was expected since all the document have a topic dominance. It is interesting to see what document it is and which topic tend to be distributed more evenly.

The most even document is BOME-A-2022-954.pdf, with a gini score of 0.4249721. By ploting various way of visualizing the distribution of the topic, we can see that indeed the document share a big proportion of administrative framework and human/migrant rights. Even though this is an administrative text, we can see that other terms related to other themes are also part of the document.



## Regression models

Finally, we check if metadata such as year of publication and author/category are more related to specific topics.

*Note: In our case, we grouped the topics mannually, meaning that our construct doesn’t fit the stm package ecosystem. Therefore we didn’t have access to the regression method recommanded by the package. We used instead a simple linear regression that give results relly close to the model used in the stm pacakge when uncertainty is not considered. Knowing that we computed a regression model using bootstraping to have better mesure of uncertainty. But ultimately, it is not enough and we must do further reverse engeenering to rebuild the regression model we are looking for latter.*

In the model our reference level for the categories (so the group we will compare all the other group against) is “Acta and Boletinos Ciudad autonoma de Ceuta”. For the year our reference level is 2021. For all the model, year have no relationship with the themes, meaning that they don’t tend to appear more during one of the year (2021, 2022 and 2023), so the themes are constant in the amount of time.

**Securitization Model**

As we have seen with previously the only categories that have more chance on having the Securitization theme are oubvious ones (“Frontex”, “Guardia Civil”, “Policia nacional”). All af theses categories have positive coefficient and p-values lower than 0.05. We can also see as previously noticed that Melilla has more topic on Securitization framework than ceuta.

**Human rights model**

Compared to the administration of Ceuta, “Asociacion Elin”, (to a lesser extent) “Frontex”, “No Name Kitchen”, “Solidary Wheels” show higher chance of having human/migrant rights framework in their documents.

**Administration model**

Finally all the categories have lesser chance than Ceuta to have the Administrative framework in their document. But the difference is lower for Melilla than the other categories.

|  |  |  |  |
| --- | --- | --- | --- |
| **Relationship between themes and metadata** | | | |
|  | **1.Security** | **2.Human rights** | **3.Administration** |
| *Predictors* | *Estimates* | *Estimates* | *Estimates* |
| (Intercept) | 0.01  (0.655) | 0.01  (0.191) | 0.98 \*\*\* (**<0.001**) |
| fyear [2022] | -0.01  (0.812) | -0.01  (0.314) | 0.02  (0.596) |
| fyear [2023] | -0.03  (0.549) | -0.05  (0.068) | 0.08  (0.082) |
| category [Acta and Boletinos Ciudad autonoma de Melilla] | 0.23 \*\*\* (**<0.001**) | 0.01  (0.171) | -0.25 \*\*\* (**<0.001**) |
| category [Asociacion Elin] | -0.00  (0.663) | 1.00 \*\*\* (**<0.001**) | -1.00 \*\*\* (**<0.001**) |
| category [Frontex] | 0.88 \*\*\* (**<0.001**) | 0.12 \*\* (**0.002**) | -0.99 \*\*\* (**<0.001**) |
| category [Guardia Civil] | 0.95 \*\*\* (**<0.001**) | -0.00  (0.456) | -0.95 \*\*\* (**<0.001**) |
| category [No Name Kitchen] | 0.02  (0.123) | 0.98 \*\*\* (**<0.001**) | -1.00 \*\*\* (**<0.001**) |
| category [Policia nacional] | 0.99 \*\*\* (**<0.001**) | -0.00  (0.271) | -0.99 \*\*\* (**<0.001**) |
| category [Solidary Wheels] | -0.00  (0.933) | 0.98 \*\*\* (**<0.001**) | -0.98 \*\*\* (**<0.001**) |
| Observations | 223 | 223 | 223 |
| R2 / R2 adjusted | 0.638 / 0.622 | 0.936 / 0.933 | 0.731 / 0.720 |
| * *p<0.05   \*\* p<0.01   \*\*\* p<0.001* | | | |

## Annex

### New topics number

Tracking changes in topics

| Previous topic number | New topic number | Theme | Theme based on… | Language |
| --- | --- | --- | --- | --- |
| 8 | 1 | 1. Administration | Previous topic | Spanish |
| 1 | 2 | 1. Security | Previous topic | Spanish |
| 3 | 3 | Removed | Removed | Spanish |
| 4 | 4 | 1. Security | Previous topic | English |
| 6 | 5 | 1. Administration | Previous topic | English |
| 15 | 6 | 1. Administration | Previous topic | Spanish |
| 5 | 7 | 1. Human rigth | Previous topic | Spanish |
| 1 | 8 | 1. Administration | Previous topic | Spanish |
| 9 | 9 | 1. Human rigth | Previous topic | English |
| 10 | 10 | 1. Administration | Previous topic | Spanish |
| 11 | 11 | 1. Administration | Previous topic | Spanish |
| 12 | 12 | 1. Administration | Previous topic | Spanish |
| No previous topic | 13 | 1. Security | Approximation | Spanish |
| 7 | 14 | 1. Administration | Previous topic | Spanish |
| 14 | 15 | 1. Administration | Previous topic | Spanish |
| 16 | 16 | 1. Human right | Previous topic | Spanish |

## New topics related words

## Topic 1 Top Words:  
## Highest Prob: ceuta, ciudad, mohamed, autónoma   
## FREX: cánovas, sres, cifuentes, abstenciones   
## Lift: abdelkarim, airsoft, arnaiz, bacha   
## Score: mohamed, carpeta, contrastada, sres   
## Topic 2 Top Words:  
## Highest Prob: melilla, guardia, ciudad, policía   
## FREX: rabia, droga, detenido, alerta   
## Lift: arrestado, arrestados, asfáltico, caninos   
## Score: rabia, minerva, guardia, criminal   
## Topic 3 Top Words:  
## Highest Prob: demás, incluso, excepto, similares   
## FREX: manufacturas, preparaciones, hilados, fibras   
## Lift: ácidos, cacao, cueros, lana   
## Score: manufacturas, preparaciones, aparatos, hilados   
## Topic 4 Top Words:  
## Highest Prob: route, border, year, crossings   
## FREX: detections, route, western, accounted   
## Lift: abusing, accounted, acquired, affiliated   
## Score: route, border, crossings, year   
## Topic 5 Top Words:  
## Highest Prob: rights, fundamental, frontex, consultative   
## FREX: consultative, forum, fro, froms   
## Lift: forum, standards, aangeklaagd, abilities   
## Score: rights, consultative, forum, frontex   
## Topic 6 Top Words:  
## Highest Prob: ciudad, ley, convenio, artículo   
## FREX: jerez, salle, familia, residencial   
## Lift: aalali, abalada, abolladuras, acabados   
## Score: convenio, menores, artículo, familia   
## Topic 7 Top Words:  
## Highest Prob: ciudad, gobierno, ceuta, consejo   
## FREX: expd, suministrados, procesa, resiliencia   
## Lift: alineamiento, bentolila, cca, cece   
## Score: procesa, ciudad, consejo, autónoma   
## Topic 8 Top Words:  
## Highest Prob: ciudad, ceuta, autónoma, servicio   
## FREX: interpelación, viajero, presentada, explicaciones   
## Lift: confirmó, deseamos, desnarigado, elige   
## Score: contrastada, carpeta, interpelación, ciudad   
## Topic 9 Top Words:  
## Highest Prob: people, melilla, spanish, rights   
## FREX: sudán, wheels, solidary, streets   
## Lift: abandon, abandonment, abazar, abdulaziz   
## Score: people, rights, border, human   
## Topic 10 Top Words:  
## Highest Prob: ceuta, ciudad, gastos, subvención   
## FREX: retrib, ayto, inc, pdt   
## Lift: encrucijada, hiperbárica, pap, servcios   
## Score: bocce, ayto, gastos, ciudad   
## Topic 11 Top Words:  
## Highest Prob: melilla, documento, gobierno, consejo   
## FREX: idic, tsd, daños, identificativo   
## Lift: abselam, abunda, académicamente, aceptados   
## Score: autenticidad, documento, electrónicamente, comprobada   
## Topic 12 Top Words:  
## Highest Prob: trabajo, artículo, empresa, bome   
## FREX: negociadora, segundos, acoso, plus   
## Lift: abderrahaman, acordase, alejandra, alicia   
## Score: bome, empresa, convenio, trabajadores   
## Topic 13 Top Words:  
## Highest Prob: melilla, asamblea, documento, gobierno   
## FREX: ordinal, pregunta, expropiación, contesta   
## Lift: abadía, abanderadas, abatibles, abatidos   
## Score: autenticidad, documento, electrónicamente, comprobada   
## Topic 14 Top Words:  
## Highest Prob: melilla, bome, artículo, subvenciones   
## FREX: iae, epígrafes, concedida, concedidas   
## Lift: a:aplicar, aakioui, aautonoma, abdou   
## Score: bome, subvenciones, artículo, subvención   
## Topic 15 Top Words:  
## Highest Prob: ceuta, ciudad, ley, oficial   
## FREX: aspirantes, calificador, selectivas, oposición   
## Lift: sonal, epig, abastecen, abdeluaheb   
## Score: aspirantes, ciudad, convocatoria, calificador   
## Topic 16 Top Words:  
## Highest Prob: personas, ceuta, elín, derechos   
## FREX: racismo, elin, elín, marchas   
## Lift: acampada, agradable, ahogadas, aplicaron   
## Score: elín, migrantes, personas, migración

## New topics related documents

##   
## Topic 1:   
## ActaPlenaria\_21-09-09.pdf  
## ActaPlenaria\_21-07-28.pdf  
## ActaPlenaria\_22-06-29.pdf   
## Topic 2:   
## 1 December 2021.pdf  
## 21 June 2021.pdf  
## 21 May 2021.pdf   
## Topic 3:   
## BOME-AX-2022-6.pdf  
## BOME-AX-2021-54.pdf  
## BOME-A-2022-377.pdf   
## Topic 4:   
## Data Frontex 2023.pdf  
## Data migratory routes November 2021.pdf  
## Data migratory routes September 2021.pdf   
## Topic 5:   
## cf-annual-report-2021-final.pdf  
## The\_Fundamental\_Rights\_Officer\_Annual\_Report\_2021.pdf  
## Support to Spain.pdf   
## Topic 6:   
## Acta\_C.Gob\_21-09-17.pdf  
## Acta\_C.Gob\_21-07-21.pdf  
## BOME-A-2022-377.pdf   
## Topic 7:   
## BOME-A-2021-918.pdf  
## Acta\_C.Gob\_22-12-30A.pdf  
## BOME-AX-2023-4.pdf   
## Topic 8:   
## ActaPlenaria\_23-02-28C.pdf  
## ActaPlenaria\_23-01-30C.pdf  
## ActaPlenaria\_22-07-22.pdf   
## Topic 9:   
## August 2022.pdf  
## May 2022.pdf  
## October 2022.pdf   
## Topic 10:   
## BOCCE\_Extra87\_27-12-2022.pdf  
## BOCCE\_Extra03\_21-01-2022.pdf  
## ActaPlenaria\_22-05-30.pdf   
## Topic 11:   
## 0\_27425\_1.pdf  
## 1\_27356\_1.pdf  
## 0\_26758\_1.pdf   
## Topic 12:   
## BOME-A-2022-516.pdf  
## BOME-A-2021-570.pdf  
## BOME-A-2022-6.pdf   
## Topic 13:   
## 0\_25209\_1.pdf  
## PTSASAM\_CI-Acta\_formalizada-1-25112022.pdf  
## 0\_26384\_1.pdf   
## Topic 14:   
## BOME-AX-2021-45.pdf  
## BOME-AX-2022-370.pdf  
## BOME-A-2022-446.pdf   
## Topic 15:   
## BOCCE\_Extra52\_09-07-2021.pdf  
## BOCCE\_6251\_11-11-2022.pdf  
## BOCCE\_6167\_21-01-2022.pdf   
## Topic 16:   
## Memoria Experiencial - Elín\_ un oasis en la frontera.pdf  
## May 2021.pdf  
## December 2021.pdf