A Comparative Analysis of Media Coverage of the 2017 Catalan Independence Referendum

Lambert Garceau

Melissa Knudtson

McGill University

GLIS 630

Steven H. H. Ding

April 8, 2018

Contents

Abstract

The Catalan independence referendum of 2017 was a major world event that was lengthily covered by the international press. The question that we set out to answer was whether there were any discernible patterns in terms of sentiment and vocabulary in the coverage of the event by language, particularly English and French, and region. We tried to answer to this question with the support of a multilingual dataset containing in total 67 articles, with 36 in French and 12 in English, as well as three in Catalan and 16 in Spanish. To assess the coverage in terms of sentiment, we analysed our dataset in RapidMiner with two classification models, specifically the Support Vector Machine (SVM) and the Decision Tree models, while using appropriate text processing methods for each language. To assess the coverage in terms of vocabulary as well as compare and confirm the patterns that we found using these classification models, we also produced lexical dispersion plots and word clouds by language. These methods combined appeared to reveal that the French coverage of the referendum showed more interest in the international and financial repercussions of the Catalan secession, while the English coverage focused on the local actions of separatists. However, our dataset, despite having been expanded, was not big enough to confirm the slight patterns that we discovered. We recommend that a much larger dataset be used in any further attempts to answer this question. To this end, we are pursuing the proper procedure to publish our expanded dataset.

Introduction

The 2017 Catalan independence referendum garnered significant international media attention. We felt that the tone of English- and French-language media coverage of this event was not obvious. Media coverage in English gave us the impression that many English-language sources presented the event with an unclear or inconsistent stance, perhaps for inability to

directly interpret the Catalan situation according to the source's editorial line, while others cast the event as a nationalist or populist movement. Somewhat similarly, media coverage in French seemed to usually present the referendum as a populist and unlawful procedure (probably because France faces regional separatist movements itself), while in Quebec, nationalist papers seemed to portray the referendum in a very positive light. We sought to explore our impressions by seeking hidden patterns in newspaper articles as well as certain metadata for these articles.

The major challenges we faced in carrying out our project were the great deal of time needed to enlarge a small dataset, the fact that the dataset we built was still quite small which affected the performance and results of the functions that we used, the highly subjective nature of sentiment bias and the limitations on tools to assign sentiment bias automatically, and the lack of a practical way to process and compare all languages in our multilingual dataset at once.

Despite these challenges, our project generated some insights on sentiment and vocabulary in English and French media coverage of the Catalan independence movement. It also resulted in the production of a larger and more multilingual dataset on this subject than was previously available from a public source. We intend to make this dataset publicly available pending research on relevant legal considerations.

Problem Statement

We set out to find interesting patterns in terms of sentiment and vocabulary in English- and French-language newspaper articles on the independence referendum from October 2017. Specifically, we explored correlations between, on the one hand, an article's text, language, region of origin, source publication, author and publication date, and, on the other hand, that article's sentiment (i.e. anti-independence bias, pro-independence bias or neutrality).

While we did find existing projects analyzing sentiment bias in media coverage of the independence referendum (indeed, one such project was the source of our initial dataset), we did not find any existing project comparatively analyzing French- and English-language coverage, let alone by features such as region. Therefore, we set out to answer a previously unexplored question.

Proposed Method or Solution

Our general methodology was as follows:

1. **Dataset preparation**
2. **Classification using support vector machine (SVMLib) (to explore sentiment)**
3. **Classification using decision trees (to explore sentiment)**
4. **Data visualization (to explore vocabulary)**
    a. Word clouds (to analyze across frequency)
    b. Lexical dispersion plots (to analyze across time)

Our specific methodology was as follows:

1. **Dataset preparation**
    a. We located a dataset containing as much of the necessary data as possible. The best dataset we could find was the *Bias Media CAT* dataset created by Jose Berengueres and available on Kaggle (Berengueres, 2018). This dataset consisted of 12 articles in English, 16 articles in Spanish and 3 articles in Catalan from October 2017, for a total of 31 articles. For each article, it contained the full text of the article as well as its language, region of origin and source publication. We decided to keep the Spanish and Catalan articles and include them in our analyses to be able to compare the French and English results with the Spanish and Catalan results.

b. We added two pre-selected features of interest to the above-mentioned dataset: author and publication date.

c. Since our dataset lacked French articles, we gathered 36 French-language articles from October 2017and their relevant features or metadata from a variety of regions of origin (France, Canada and Belgium), source publications and authors, for a total of 67 articles.

d. Because in our dataset each article was a single text value, and we felt that valuable distinctions in sentiment could be made sentence by sentence, we used nltk.sent_tokenize(…) from the Natural Language Toolkit (NLTK) Python package to loop through the articles and tokenize or split them into sentences, for a total of approximately 1600 sentences or records (NLTK project, 2017a).

e. To establish a means of checking the accuracy of any automated assignment of sentiment, we manually assigned a sentiment score as our target attribute to one quarter of the dataset on a 1-5 scale, where 1 represented very anti-independence sentiment, 2 anti-independence sentiment, 3 neutral sentiment, 4 pro-independence sentiment and 5 pro-independence sentiment.

f. We used the Vader sentiment analysis tool from NLTK in Python to automatically assign a sentiment score to each sentence from -1.0 to + 1.0 (NLTK project, 2017b).

g. We used raw Python to convert the floating-point Vader sentiment scores to our 1-5 scale. In doing so, we assumed that assumed that negative sentiment corresponded to anti-independence sentiment and positive sentiment corresponded to pro-independence sentiment.

h. We used raw Python to check the accuracy of the Vader scores, assuming that a Vader sentiment score corresponding to a manual score +/-1 was sufficiently accurate since sentiment in general and granular sentiment in particular is highly subjective. The accuracy of the Vader scores was 66%. Since we manually labelled one quarter of the dataset (with what we assumed is 100% accuracy) and we automatically labelled the other three quarters of the dataset, a calculation of the overall accuracy as a weighted average yielded an overall accuracy of 75%, as follows:
    - (.25 x 1.0) + (.75 x 0.66) = 75%

i. We retrieved a list of stopwords in Spanish using nltk.corpus.stopwords.words('spanish').

j. We retrieved a list of stopwords in Catalan from the website of a university in Barcelona (LATeL, n.d.).

2. Classification using support vector machine (SVMLib) (to explore sentiment)

We used RapidMiner to perform this function, which we hoped would enable prediction of sentiment of articles not in our dataset based on the features included in our dataset. (See Annex A.) Some notable aspects of our process were as follows:

a. Select Attributes: We selected all potentially meaningful attributes (excluding the article ID and URL).

b. **Filter Examples**: Since text in each language required separate processing (e.g. contained different stopwords), we filtered by language and ran the process once per language.

c. **Nominal to Numerical**: Since SVMLib proved unable to process our polynomial label attribute (with the values 1, 2, 3, 4 and 5), we converted polynomial data to numerical data using this operator.

d. **Process Documents from Data**: We used this operator to tokenize text into words, filter the relevant stopwords (using external files for Spanish and Catalan), stem (in the case of English only, as stemming is not available for all languages in RapidMiner and is generally optional) and transform cases.

e. **Split Data**: We split the data into a training set (60%), a validation set to optimize parameters (10%), and a test set (30%).

f. **Performance**: The test set was used to analyze the performance of the results of the optimized parameters.

The strengths of this solution were filtering and application of language-specific processing to enhance the results and optimization of parameters to attempt to maximize the accuracy of the results.

3. Classification using decision trees (to explore sentiment)

We used RapidMiner to perform this function, which we hoped would enable prediction of sentiment of articles not in our dataset based on the features included in our dataset. (See Annex B.) The specific process was similar to the process for SVMLib, except for the following:

a. The data were split into a training set (60%) and a test set (40%) only.

b. Parameter optimization was not done.

c. Pre-pruning and post-pruning were not done in view of the small size of the dataset.

The strengths of this solution were filtering and application of language-specific processing and the fact that it required us to partition an already small dataset in two ways rather than three ways.

4. Data visualization (to explore vocabulary)

To supplement our exploration of classification efforts we used NLTK to prepare several data visualization aids. First, we explored vocabulary in terms of frequency of appearance in the English and French text by preparing comparative word clouds for the French and English text. Second, we explored vocabulary in terms of changes in appearance over time by preparing comparative lexical dispersion plots for the French and English text, using several words appearing most frequently in the dataset as well as several words that yielded interesting comparative results. (See Annex C.)

Experimental Setup and Results

We used the SVM and Decision Tree models in RapidMiner. We filtered each process by language, meaning that both processes were each repeated four times. In both the SVM and the decision tree, we found the French portion of the dataset to be the most accurate — 77.96% for the SVM and 79.31% for the decision tree — probably because the majority of the articles that we used were in that language. Overall, we were not able to find and prove clear patterns in the political tendencies of the press coverage of the referendum. However, we were able to detect interesting patterns in the frequency and choice of terms by language.

The SVM processes were our first attempt to classify our data by sentiment, according to all other available attributes (author, region, language, original text). Because of the low number of records, the results were generally not very accurate, ranging from 37.5% for Catalan to 77.96% for French. Overall, each SVM more easily classified articles with sentiments ranging

from 2 to 4, with neutral articles (3) being classified with the highest accuracy and recall. Articles classified as true 3 were also generally the most numerous according to the classifiers. Unfortunately, considering the low accuracy of our processes, we cannot seriously affirm that there might be a pattern in those results that could really confirm or deny any hypothesis regarding regional and linguistic biases in the press coverage of the referendum. The low accuracy was probably due to the fact that the charts created by the SVM contained scarce data that was further divided by language separation. In an attempt to reduce this problem of data scarcity, we also attempted to produce a SVM that encompassed all of our data, with a final accuracy of 62%, which was still too low to affirm the presence of any serious pattern.

As for the decision tree, we were able to find some very small emerging patterns in term use by language that may be of interest. The trees were also expanded by the removal of pre-pruning and pruning, as they were too small otherwise. The French decision tree seemed to reflect more interest in financial and legal considerations in French media coverage than coverage in other languages. The three top nodes in that tree were “Justice”, “Respect” and “Accusés”, which may support the idea that the French speaking journalists were concerned for the stability of the region and the respect of the rule of law. Another curious example that tends to support this hypothesis was the apparition of the term “Gas” in the tree. When we reviewed examples of this term in context, we found that they were referring to Gas Natural, a major Spanish company that moved its headquarters out of Catalonia in the aftermath of the referendum. The English perspective was slightly more removed from the event and appeared to focus more on day-to-day processes. This might have been supported by the tops terms, being “aim”, “interview” and “mayb”. The Catalan and Spanish decision trees were, perhaps unsurprisingly, antagonistic in their word use. The words “dialogar” in the Spanish decision tree

and “mediadors” in the Catalan decision trees indeed both led to a sentiment score of 1. This suggested that neither side wanted to talk, although it must be kept in mind that the dataset is too small to really support that possibility.

We also used data visualization tools to see whether a data representation strictly based on term frequency would support our patterns detected in the decision trees. We produced word clouds and lexical dispersion plots for the English and French sections of the dataset, considering they covered the majority of our data. The French word cloud supported the idea that French press gave more importance to economic and legal concerns, with words like “Union Européenne”, “Europe”, “exécutif”, “vote”, “loi” and “sous tutelle”. The English word cloud also supported the hypothesis that the English articles in our dataset were more descriptive of the process of the referendum and the actions it encompasses. Its most common words were “secession”, “independence”, “Spain”, “Spanish” and “Catalonia”. The lexical dispersion plots also proved to be a useful complementary tool, as they could show the evolution of a term usage over time. The dispersion plots we created compared certain terms of interest in French and English articles over time, from early to late October 2017. They showed that the English articles, despite being less numerous than the French ones, tended to use the words “illegal” and “separatist” far more often. This may lead to the supposition that the English coverage of the referendum was more openly hostile to the Catalan cause than the French coverage.

The processes could be improved by adding certain operators that would increase the accuracy of the sentiment analysis. The SVM could benefit from data discretization, which would hierarchize our attributes according to intervals. It would also benefit from the inclusion of fewer classes in the sentiment analysis attributes, which would divide an already limited dataset less. We could use three instead of five levels of sentiment concerning independence,

which could be positive, neutral and negative. The decision tree could also be improved by ensuring the highest maximum depth to include every possible concept in the tree.

Finally, the results would have been more accurate had we had access to or built a much bigger dataset. Our results were consistently limited by the low number of articles that we assembled, as data preparation was a lengthy and time-consuming process. The dataset itself would also benefit from more equal representation of our four languages.

Conclusion

Ultimately, our original impressions that French and English media coverage of the 2017 Catalan independence referendum were neither confirmed nor definitively denied. Language as an attribute was not shown to decisively impact sentiment bias.

The primary limitation of our project was the small volume of records and number of attributes comprising our original dataset which rendered our findings generally inconclusive, despite the fact that we enlarged the original dataset from 30 to approximately 1,600 records and added several attributes for each record. We did manage to offset this limitation in several ways, such as eliminating pre-pruning and post-pruning in our decision tree process. To witness the effect if any of attributes other than text on the results of our classification efforts, a much larger dataset would have to be used. Another limitation was the fact there was an unequal representation of languages in the dataset, ranging from three to 36 articles, which restricted direct comparisons.

Despite these limitations, our project did produce some interesting results. From our decision trees and data visualization tools, we were able to notice a slight difference between the French and English coverage of the events in Catalonia. The French press was more focused on financial, international and legal concerns. Its English counterpart tended to analyze the

referendum in its local context and tended to consider the Catalan actions as illegal. We were able to build processes that could be easily improved and that would have the potential to show more interesting patterns, if we were to add more articles to analyze. With more training and more data, our processes could help to predict the sentiment bias of unread articles.

Multilingual text mining is a challenging operation, and the scientific community would benefit from trying to improve text mining in other languages than English, perhaps by developing more tools to work with non-English material.

Summary

We used support vector machine and classification models to compare sentiment bias or tone and vocabulary choices across media coverage of the 2017 Catalan independence referendum by language (particularly French and English) and region. To do this, we used the Natural Language Toolkit (NLTK) package to prepare a novel dataset from an existing partial dataset and RapidMiner to apply the chosen classification models. Our results regarding sentiment were inconclusive, but our results regarding vocabulary choices did suggest that English coverage tended to be expository whereas French coverage did tend to emphasize the financial and legal repercussions of the event. We recommend that a much larger dataset be prepared to explore this topic further and come to more decisive conclusions, and to this end we intend to make this dataset publicly available pending research on relevant legal considerations.

References

Berengueres, Jose. (2018). *Bias Media CAT* [Data file]. Retrieved from: https://www.kaggle.com/harriken/bias-media-cat

LATeL. (n.d.). *Llista de mots buits del català*. Retrieved from: http://latel.upf.edu/morgana/altres/pub/ca_stop.htm

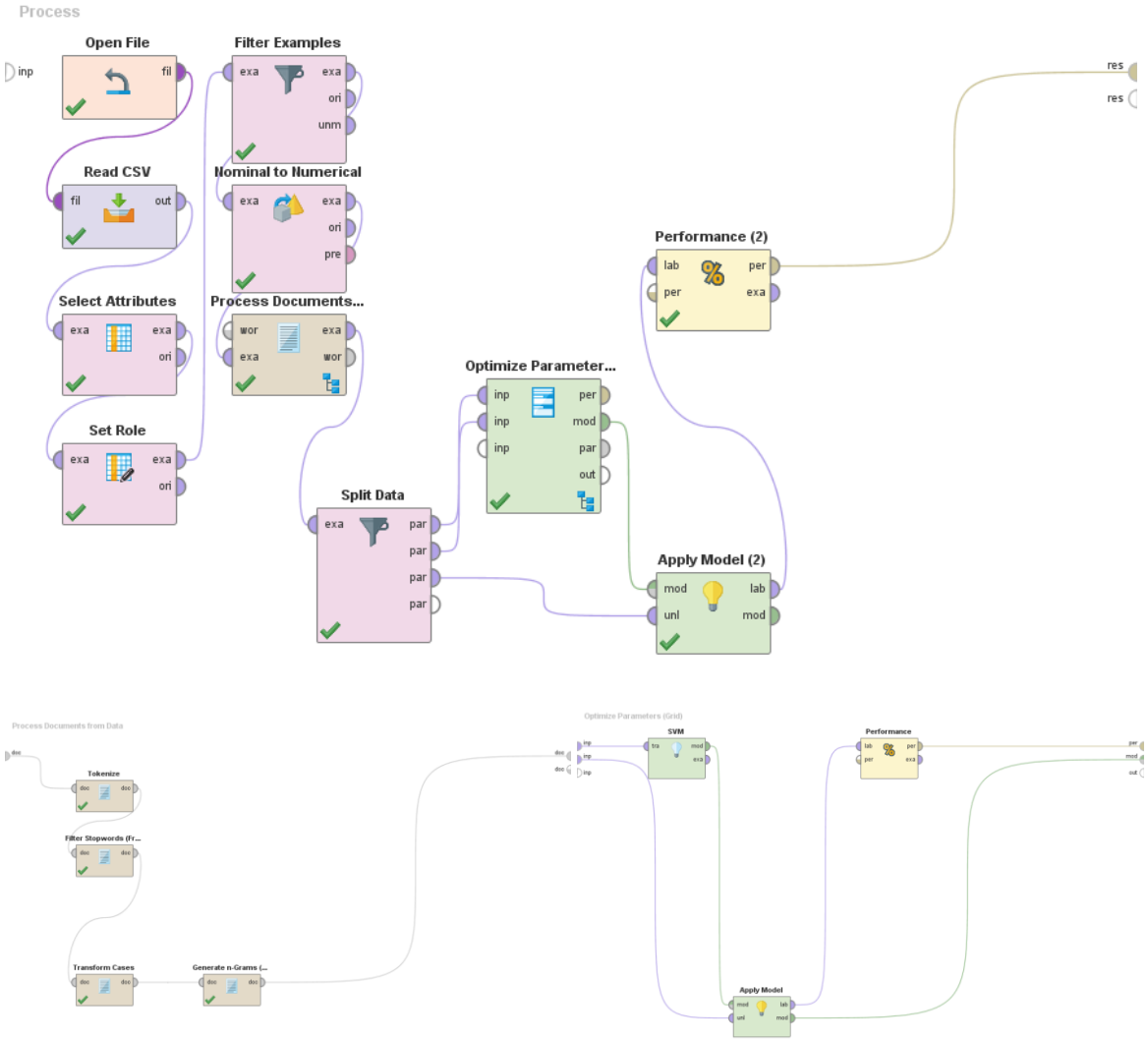
Mueller, Andreas. (2018). *Word Cloud.* Retrieved from: https://github.com/amueller/word_cloud

NLTK Project. (2017a). *Natural Language Toolkit*. Retrieved from: https://www.nltk.org/

NLTK Project. (2017b). *Source code for nltk.sentiment.vader*. Retrieved from: http://www.nltk.org/_modules/nltk/sentiment/vader.html

Annex A: Support Vector Machine (SVMLib) Process and Results

Process — FRENCH (as a representative example)



Results — ENGLISH

accuracy: 48.21%

	true 3	true 4	true 2	true 1	true 5	class precision
pred. 3	26	6	11	10	2	47.27%
pred. 4	0	7	0	1	3	63.64%
pred. 2	3	10	13	6	1	39.39%
pred. 1	2	1	2	6	0	54.55%
pred. 5	0	0	0	0	2	100.00%
class recall	83.87%	29.17%	50.00%	26.09%	25.00%	

Results — FRENCH

accuracy: 77.96%

	true 4	true 3	true 2	true 1	true 5	class precision
pred. 4	0	0	0	0	0	0.00%
pred. 3	16	237	29	18	4	77.96%
pred. 2	0	0	0	0	0	0.00%
pred. 1	0	0	0	0	0	0.00%
pred. 5	0	0	0	0	0	0.00%
class recall	0.00%	100.00%	0.00%	0.00%	0.00%	

Results — SPANISH

accuracy: 50.59%

	true 3	true 2	true 1	true 4	true 5	class precision
pred. 3	37	10	6	11	2	56.06%
pred. 2	2	1	0	2	0	20.00%
pred. 1	0	0	0	0	0	0.00%
pred. 4	3	4	0	3	0	30.00%
pred. 5	1	0	0	1	2	50.00%
class recall	86.05%	6.67%	0.00%	17.65%	50.00%	

Results — CATALAN

accuracy: 37.50%

	true 3	true 4	true 5	true 1	true 2	class precision
pred. 3	8	4	2	4	4	36.36%
pred. 4	0	0	0	0	0	0.00%
pred. 5	0	0	0	0	0	0.00%
pred. 1	0	0	0	0	0	0.00%
pred. 2	0	1	0	0	1	50.00%
class recall	100.00%	0.00%	0.00%	0.00%	20.00%	

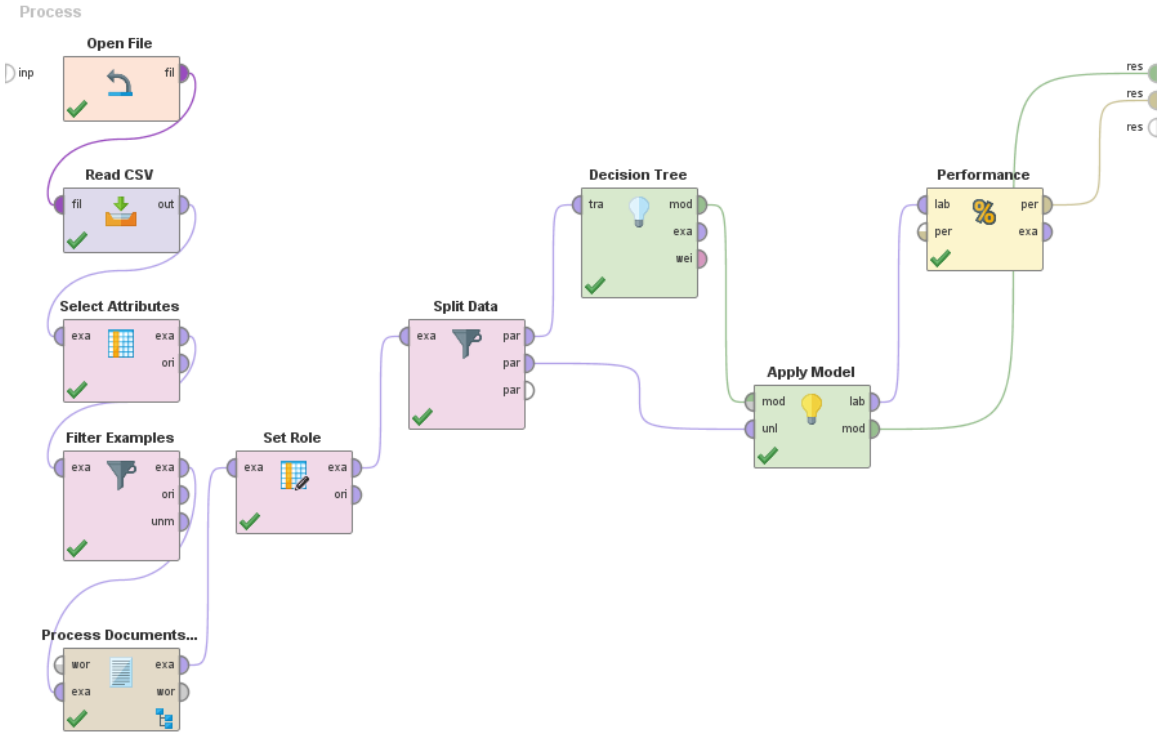
Results — ALL

accuracy: 62.08%

	true 3	true 2	true 1	true 4	true 5	class precision
pred. 3	312	66	44	50	12	64.46%
pred. 2	7	10	7	7	7	26.32%
pred. 1	0	0	1	0	0	100.00%
pred. 4	0	0	1	6	0	85.71%
pred. 5	0	0	0	0	0	0.00%
class recall	97.81%	13.16%	1.89%	9.52%	0.00%	

Annex B: Decision Tree Process and Results

Process — FRENCH (as a representative example)



Results — ENGLISH

accuracy: 34.00%

	true 3	true 4	true 2	true 1	true 5	class precision
pred. 3	40	22	31	29	5	31.50%
pred. 4	0	7	1	0	3	63.64%
pred. 2	0	0	0	0	0	0.00%
pred. 1	0	1	0	1	0	50.00%
pred. 5	1	2	3	1	3	30.00%
class recall	97.56%	21.88%	0.00%	3.23%	27.27%	

Tree

```
aim > 0.179: 5 {3=0, 4=0, 2=0, 1=0, 5=2}
aim ≤ 0.179
|   interview > 0.219: 5 {3=0, 4=0, 2=0, 1=0, 5=2}
|   interview ≤ 0.219
|   |   mayb > 0.125: 5 {3=0, 4=0, 2=0, 1=0, 5=2}
|   |   mayb ≤ 0.125
|   |   |   absolut > 0.158: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
```

```
|   |   |   absolut ≤ 0.158
|   |   |   |   angela > 0.124: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   angela ≤ 0.124
|   |   |   |   |   appeal > 0.145: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   appeal ≤ 0.145
|   |   |   |   |   |   banner > 0.107: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   banner ≤ 0.107
|   |   |   |   |   |   |   break > 0.251: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   break ≤ 0.251
|   |   |   |   |   |   |   |   campaign > 0.161: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   |   campaign ≤ 0.161
|   |   |   |   |   |   |   |   |   centr > 0.182: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   |   |   centr ≤ 0.182
|   |   |   |   |   |   |   |   |   |   coercion > 0.188: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   |   |   |   coercion ≤ 0.188
|   |   |   |   |   |   |   |   |   |   |   commit > 0.331: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   commit ≤ 0.331
|   |   |   |   |   |   |   |   |   |   |   |   matter > 0.334: 5 {3=0, 4=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   matter ≤ 0.334
|   |   |   |   |   |   |   |   |   |   |   |   |   rule > 0.201: 1 {3=0, 4=0, 2=0, 1=5, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   rule ≤ 0.201
|   |   |   |   |   |   |   |   |   |   |   |   |   |   violat > 0.092: 1 {3=0, 4=0, 2=0, 1=3, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   violat ≤ 0.092
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   author_0 > 0.267: 4 {3=0, 4=4, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   author_0 ≤ 0.267
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   support > 0.202: 4 {3=0, 4=4, 2=0, 1=0,
5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   support ≤ 0.202
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   back > 0.229: 4 {3=0, 4=3, 2=0, 1=0,
5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   back ≤ 0.229
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   immedi > 0.244: 4 {3=0, 4=3, 2=0,
1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   immedi ≤ 0.244: 3 {3=62, 4=34,
2=52, 1=39, 5=0}
```

Results — FRENCH

accuracy: 79.31%

	true 4	true 3	true 2	true 1	true 5	class precision
pred. 4	5	0	0	0	0	100.00%
pred. 3	14	316	39	24	5	79.40%
pred. 2	0	0	0	0	0	0.00%
pred. 1	0	0	0	0	0	0.00%
pred. 5	2	0	0	0	1	33.33%
class recall	23.81%	100.00%	0.00%	0.00%	16.67%	

Tree

```
justice > 0.336: 5 {4=0, 3=0, 2=0, 1=0, 5=2}
justice ≤ 0.336
|   respect > 0.220: 5 {4=0, 3=0, 2=0, 1=0, 5=2}
|   respect ≤ 0.220
|   |   accusés > 0.168: 5 {4=0, 3=0, 2=0, 1=0, 5=1}
|   |   accusés ≤ 0.168
|   |   |   adéquatement > 0.137: 5 {4=0, 3=0, 2=0, 1=0, 5=1}
|   |   |   adéquatement ≤ 0.137
|   |   |   |   cadeau > 0.188: 5 {4=0, 3=0, 2=0, 1=0, 5=1}
|   |   |   |   cadeau ≤ 0.188
|   |   |   |   |   engrenage > 0.133: 5 {4=0, 3=0, 2=0, 1=0, 5=1}
|   |   |   |   |   engrenage ≤ 0.133
|   |   |   |   |   |   fondement > 0.150: 5 {4=0, 3=0, 2=0, 1=0, 5=1}
|   |   |   |   |   |   fondement ≤ 0.150
|   |   |   |   |   |   |   respecter > 0.070: 4 {4=5, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   respecter ≤ 0.070
|   |   |   |   |   |   |   |   solution > 0.096: 4 {4=5, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   solution ≤ 0.096
|   |   |   |   |   |   |   |   |   parties > 0.082: 4 {4=4, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   parties ≤ 0.082
|   |   |   |   |   |   |   |   |   |   attendait > 0.118: 4 {4=2, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   attendait ≤ 0.118
```

```
|   |   |   |   |   |   |   |   |   |   |   certain > 0.123: 4 {4=2, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   certain ≤ 0.123
|   |   |   |   |   |   |   |   |   |   |   |   gas > 0.207: 4 {4=2, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   gas ≤ 0.207
|   |   |   |   |   |   |   |   |   |   |   |   |   importance > 0.129: 4 {4=2, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   importance ≤ 0.129
|   |   |   |   |   |   |   |   |   |   |   |   |   |   accroît > 0.127: 4 {4=1, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   accroît ≤ 0.127
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   accuse > 0.327: 4 {4=1, 3=0, 2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   accuse ≤ 0.327
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   arrestations > 0.141: 4 {4=1, 3=0, 2=0,
1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   arrestations ≤ 0.141
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   barrage > 0.140: 4 {4=1, 3=0, 2=0, 1=0,
5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   barrage ≤ 0.140
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   champagne > 0.229: 4 {4=1, 3=0,
2=0, 1=0, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   champagne ≤ 0.229: 3 {4=4, 3=474,
2=59, 1=35, 5=0}
```

Results — SPANISH

accuracy: 49.56%

	true 3	true 2	true 1	true 4	true 5	class precision
pred. 3	49	16	5	15	3	55.68%
pred. 2	0	0	0	0	0	0.00%
pred. 1	2	1	2	1	1	28.57%
pred. 4	3	3	1	4	1	33.33%
pred. 5	3	0	0	2	1	16.67%
class recall	85.96%	0.00%	25.00%	18.18%	16.67%	

Tree

```
confirma > 0.125: 1 {3=0, 2=0, 1=4, 4=0, 5=0}
confirma ≤ 0.125
|   activado > 0.162: 5 {3=0, 2=0, 1=0, 4=0, 5=2}
|   activado ≤ 0.162
|   |   dialogar > 0.109: 1 {3=0, 2=0, 1=2, 4=0, 5=0}
|   |   dialogar ≤ 0.109
|   |   |   gravemente > 0.074: 1 {3=0, 2=0, 1=2, 4=0, 5=0}
|   |   |   gravemente ≤ 0.074
|   |   |   |   actualizado > 0.147: 1 {3=0, 2=0, 1=1, 4=0, 5=0}
|   |   |   |   actualizado ≤ 0.147
|   |   |   |   |   atender > 0.298: 1 {3=0, 2=0, 1=1, 4=0, 5=0}
|   |   |   |   |   atender ≤ 0.298
|   |   |   |   |   |   contumaz > 0.202: 1 {3=0, 2=0, 1=1, 4=0, 5=0}
|   |   |   |   |   |   contumaz ≤ 0.202
|   |   |   |   |   |   |   cosas > 0.172: 1 {3=0, 2=0, 1=1, 4=0, 5=0}
|   |   |   |   |   |   |   cosas ≤ 0.172
|   |   |   |   |   |   |   |   desobediencia > 0.375: 1 {3=0, 2=0, 1=1, 4=0, 5=0}
|   |   |   |   |   |   |   |   desobediencia ≤ 0.375
|   |   |   |   |   |   |   |   |   acordado > 0.283: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   acordado ≤ 0.283
|   |   |   |   |   |   |   |   |   |   ambigua > 0.128: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   ambigua ≤ 0.128
|   |   |   |   |   |   |   |   |   |   |   anunciar > 0.251: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   anunciar ≤ 0.251
|   |   |   |   |   |   |   |   |   |   |   |   apostado > 0.165: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   apostado ≤ 0.165
|   |   |   |   |   |   |   |   |   |   |   |   |   catalana > 0.203: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   catalana ≤ 0.203
|   |   |   |   |   |   |   |   |   |   |   |   |   |   comparecido > 0.218: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   comparecido ≤ 0.218
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dado > 0.235: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dado ≤ 0.235
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   espa > 0.123: 4 {3=0, 2=0, 1=0, 4=5, 5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   espa ≤ 0.123
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tras > 0.179: 4 {3=0, 2=0, 1=0, 4=5,
5=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tras ≤ 0.179
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   puigdemont > 0.141: 4 {3=0, 2=0,
1=0, 4=4, 5=0}
```

```
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   puigdemont ≤ 0.141: 3 {3=86, 2=31,
1=0, 4=20, 5=0}
```

Results — CATALAN

accuracy: 24.24%

	true 3	true 4	true 5	true 1	true 2	class precision
pred. 3	7	6	2	3	5	30.43%
pred. 4	0	1	0	1	0	50.00%
pred. 5	0	0	0	2	1	0.00%
pred. 1	2	0	0	0	0	0.00%
pred. 2	2	0	1	0	0	0.00%
class recall	63.64%	14.29%	0.00%	0.00%	0.00%	

Tree

```
perqu > 0.166: 1 {3=0, 4=0, 5=0, 1=3, 2=0}
perqu ≤ 0.166
|   mediadors > 0.100: 1 {3=0, 4=0, 5=0, 1=2, 2=0}
|   mediadors ≤ 0.100
|   |   actuat > 0.128: 1 {3=0, 4=0, 5=0, 1=1, 2=0}
|   |   actuat ≤ 0.128
|   |   |   agrair > 0.154: 1 {3=0, 4=0, 5=0, 1=1, 2=0}
|   |   |   agrair ≤ 0.154
|   |   |   |   allà > 0.108: 1 {3=0, 4=0, 5=0, 1=1, 2=0}
|   |   |   |   allà ≤ 0.108
|   |   |   |   |   abast > 0.092: 5 {3=0, 4=0, 5=1, 1=0, 2=0}
|   |   |   |   |   abast ≤ 0.092
|   |   |   |   |   |   acabar > 0.155: 5 {3=0, 4=0, 5=1, 1=0, 2=0}
|   |   |   |   |   |   acabar ≤ 0.155
|   |   |   |   |   |   |   als > 0.143: 5 {3=0, 4=0, 5=1, 1=0, 2=0}
|   |   |   |   |   |   |   als ≤ 0.143
|   |   |   |   |   |   |   |   anys > 0.225: 5 {3=0, 4=0, 5=1, 1=0, 2=0}
|   |   |   |   |   |   |   |   anys ≤ 0.225
|   |   |   |   |   |   |   |   |   civils > 0.157: 5 {3=0, 4=0, 5=1, 1=0, 2=0}
```

```
|   |   |   |   |   |   |   |   |   civils ≤ 0.157
|   |   |   |   |   |   |   |   |   |   pot > 0.067: 4 {3=0, 4=3, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   pot ≤ 0.067
|   |   |   |   |   |   |   |   |   |   |   aixä > 0.063: 4 {3=0, 4=2, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   |   aixä ≤ 0.063
|   |   |   |   |   |   |   |   |   |   |   |   autonomisme > 0.092: 4 {3=0, 4=2, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   |   |   autonomisme ≤ 0.092
|   |   |   |   |   |   |   |   |   |   |   |   |   espanyola > 0.096: 4 {3=0, 4=2, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   espanyola ≤ 0.096
|   |   |   |   |   |   |   |   |   |   |   |   |   |   aprovat > 0.244: 4 {3=0, 4=1, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   aprovat ≤ 0.244
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   banda > 0.284: 4 {3=0, 4=1, 5=0, 1=0, 2=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   banda ≤ 0.284
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dit > 0.110: 2 {3=0, 4=0, 5=0, 1=0, 2=3}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dit ≤ 0.110
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   catalunya > 0.157: 2 {3=0, 4=0, 5=0,
1=0, 2=2}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   catalunya ≤ 0.157
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   espanyol > 0.112: 2 {3=0, 4=0, 5=0,
1=0, 2=2}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   espanyol ≤ 0.112: 3 {3=16, 4=0,
5=0, 1=0, 2=3}
```

Results — ALL

accuracy: 60.54%

	true 3	true 2	true 1	true 4	true 5	class precision
pred. 3	420	100	68	77	18	61.49%
pred. 2	0	0	0	0	0	0.00%
pred. 1	0	0	0	0	0	0.00%
pred. 4	0	0	0	0	0	0.00%
pred. 5	6	1	2	7	8	33.33%
class recall	98.59%	0.00%	0.00%	0.00%	30.77%	

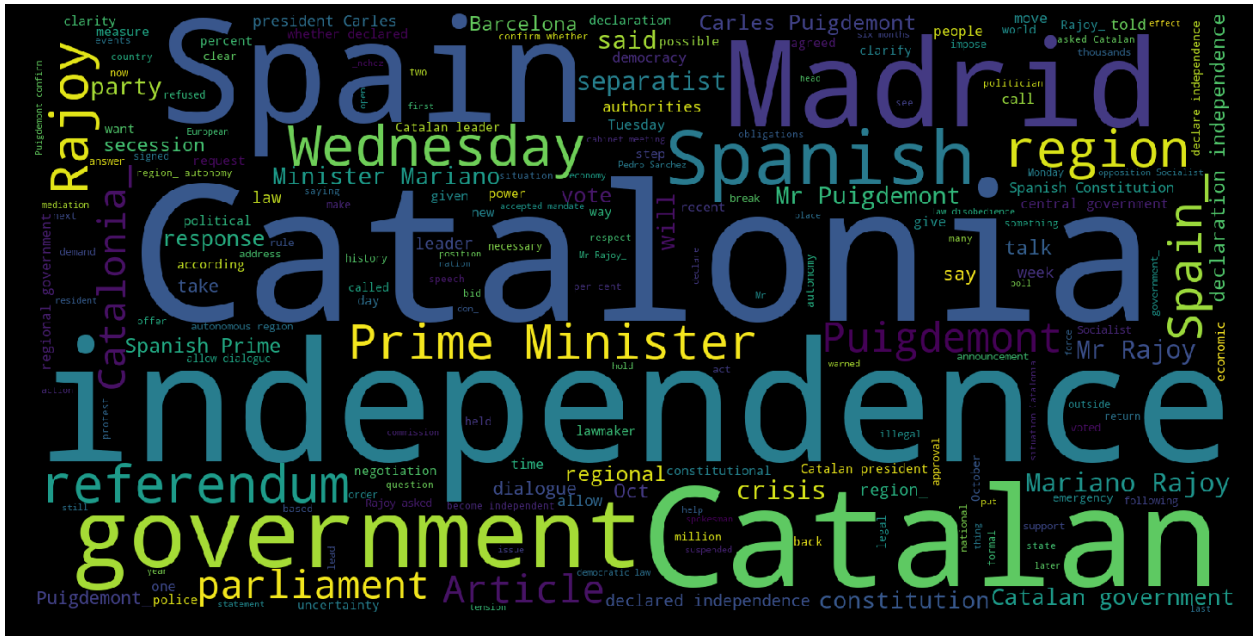
Tree

```
absolute > 0.043: 5 {3=0, 2=0, 1=0, 4=0, 5=3}
absolute ≤ 0.043
|   justice > 0.193: 5 {3=0, 2=0, 1=0, 4=0, 5=3}
|   justice ≤ 0.193
|   |   acordado > 0.089: 5 {3=0, 2=0, 1=0, 4=0, 5=2}
|   |   acordado ≤ 0.089
|   |   |   aimed > 0.134: 5 {3=0, 2=0, 1=0, 4=0, 5=2}
|   |   |   aimed ≤ 0.134
|   |   |   |   catalana > 0.139: 5 {3=0, 2=0, 1=0, 4=0, 5=2}
|   |   |   |   catalana ≤ 0.139
|   |   |   |   |   decidir > 0.064: 5 {3=0, 2=0, 1=0, 4=0, 5=2}
|   |   |   |   |   decidir ≤ 0.064
|   |   |   |   |   |   activado > 0.148: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   activado ≤ 0.148
|   |   |   |   |   |   |   actualizado > 0.148: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   actualizado ≤ 0.148
|   |   |   |   |   |   |   |   adéquatement > 0.122: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   adéquatement ≤ 0.122
|   |   |   |   |   |   |   |   |   agradecer > 0.134: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   agradecer ≤ 0.134
|   |   |   |   |   |   |   |   |   |   agree > 0.182: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   agree ≤ 0.182
|   |   |   |   |   |   |   |   |   |   |   albeit > 0.114: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   albeit ≤ 0.114
|   |   |   |   |   |   |   |   |   |   |   |   algunas > 0.130: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   algunas ≤ 0.130
|   |   |   |   |   |   |   |   |   |   |   |   |   arch > 0.159: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   arch ≤ 0.159
|   |   |   |   |   |   |   |   |   |   |   |   |   |   banner > 0.094: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   banner ≤ 0.094
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   break > 0.220: 5 {3=0, 2=0, 1=0, 4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   break ≤ 0.220
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   calmness > 0.074: 5 {3=0, 2=0, 1=0, 4=0,
5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   calmness ≤ 0.074
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   campaign > 0.129: 5 {3=0, 2=0, 1=0,
4=0, 5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   campaign ≤ 0.129
```

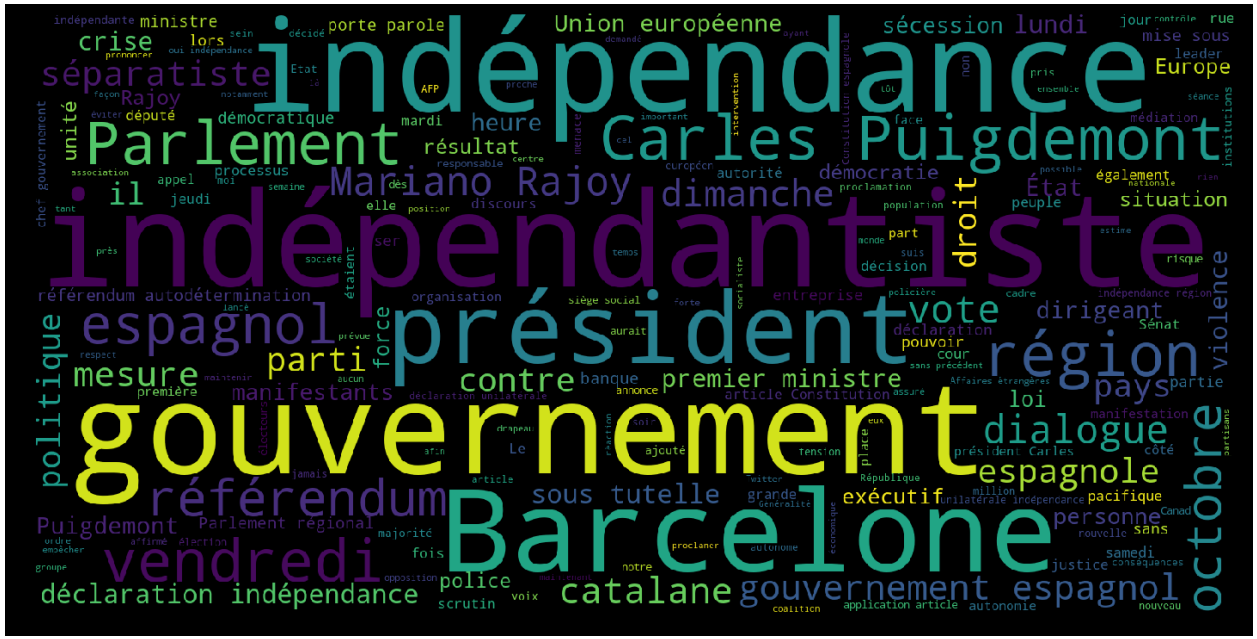
```
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cap > 0.379: 5 {3=0, 2=0, 1=0, 4=0,
5=1}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cap ≤ 0.379: 3 {3=638, 2=151,
1=105, 4=125, 5=12}
```

Annex C: Word Clouds and Lexical Dispersion Plots

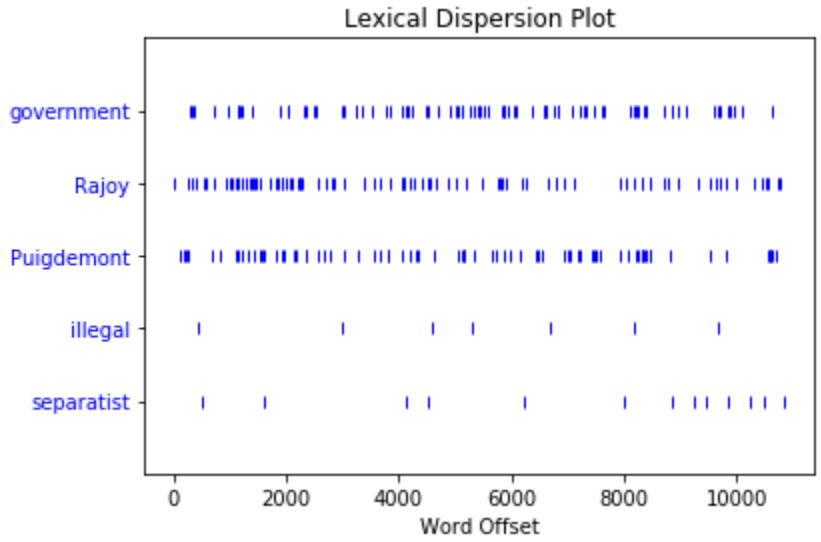
Word Cloud — ENGLISH



Word Cloud — FRENCH



Lexical Dispersion Plot — ENGLISH



Lexical Dispersion Plot — FRENCH