Text classification on historical newspapers

How was the 1848 June Days uprising covered in the press ?

Léo Dumont – leo.dumont@univ-paris1.fr

15 novembre 2024

Université Paris 1 – PIREH, Centre d'histoire du 19e siècle

Historical issue

Historical issue

June Days were an uprising from 22 to 26 June 1848 in Paris.

- Study the event through the words used to describe it.
- Analyze the process of restoring order carried out by the triumphant republican authorities.

In search of a narrated « law and order fighter »

Towards the construction of a corpus

What sources to study the discursive construction of the « law and order fighters » ?

- 1848 as an « unprecedented media explosion [...] in the history of the press (Ambroise-Rendu, 1999) »
- 450 newspaper creations

Characteristics of a daily newspaper in 1848

A newspaper issue contains a wide variety of content :

- news,
- serialized novels,
- transcription of parliamentary debates,
- letters,
- stock prices,
- advertisements,
- ...

Not always regular rules for layout.

How to find relevant content in the press ?

- Rely on digitized historical newspapers data and distant reading (Moretti, 2013) :
  - French National Library (**BnF**)
  - **RetroNews** : digital library dedicated to historical newspapers

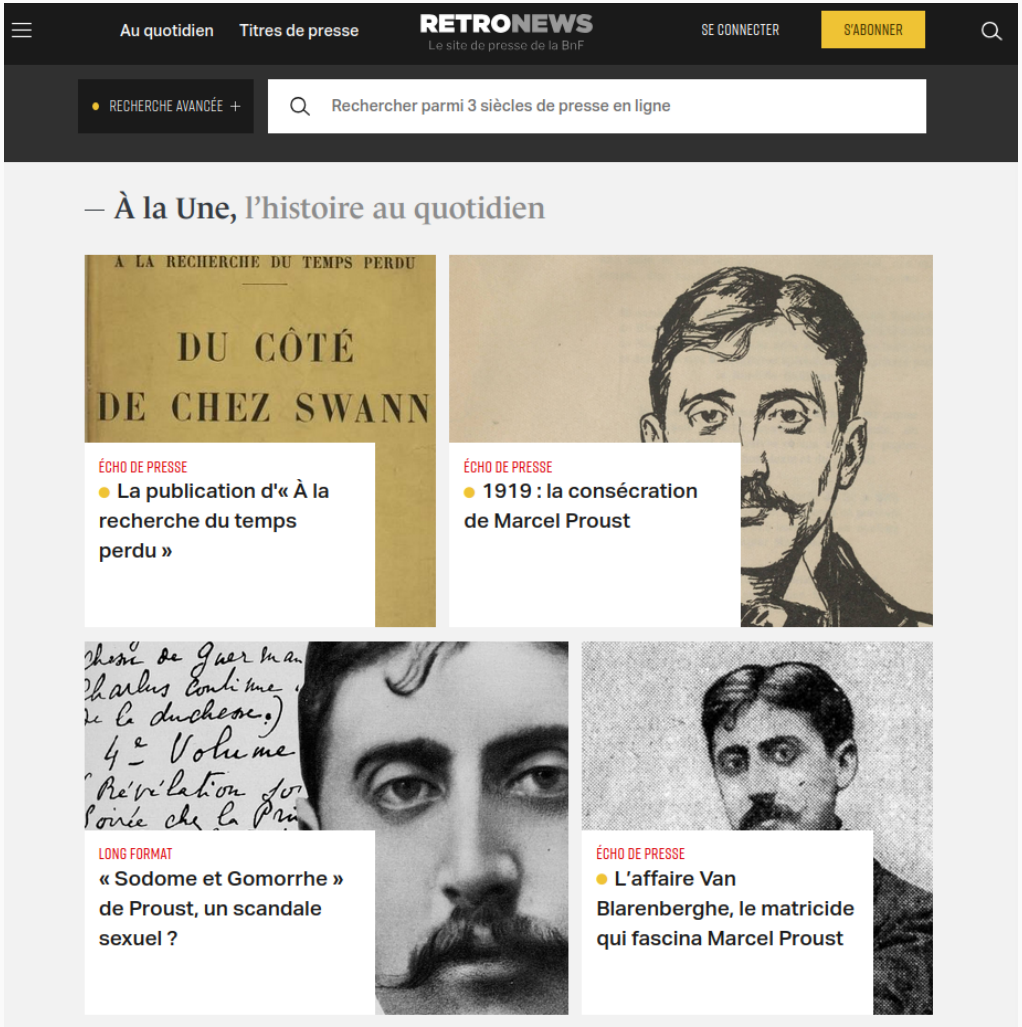


Figure 1 – RetroNews Homepage.

Digitized historical newspapers data challenges

Specific features of digitized newspapers :

OCR quality

- historical documents are still a challenge for OCR
- negative outcomes on NLP tasks

Digital materials granularity

- Two main levels of granularity (METS/ALTO XML standard) :
  - Physical Structure
  - Logicial Structure

OCR quality

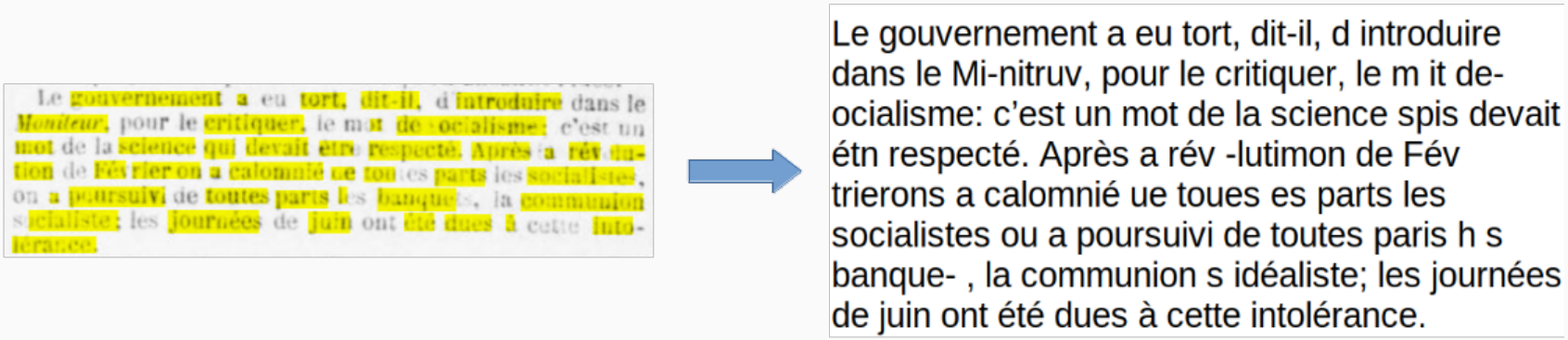


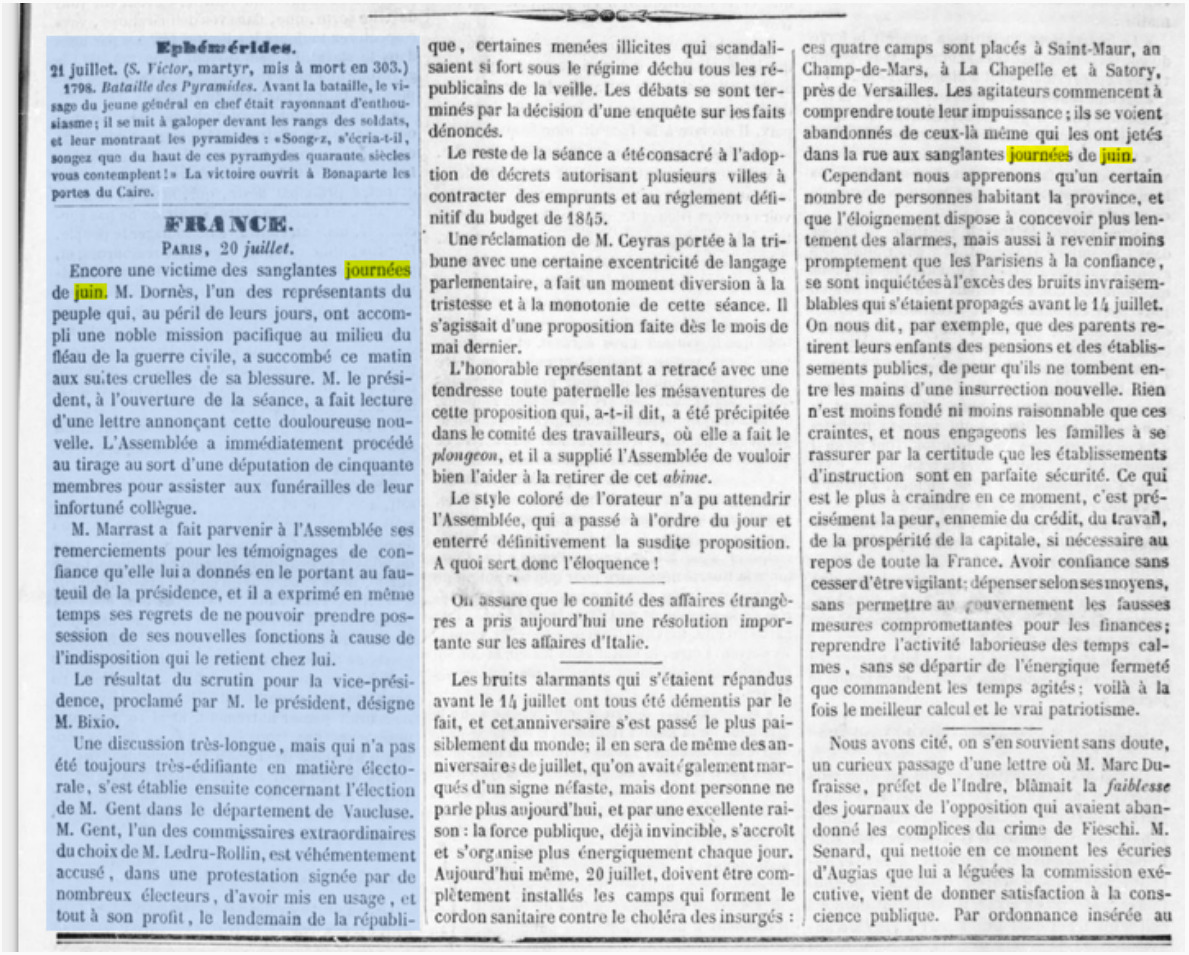
Figure 2 – Example of OCR quality.

A well known problem : (Mackovski, 2001), (Traub, 2015), (Mutuvi, 2018), (Hill, 2019), (Jiang, 2021), (Vitman, 2022).

Digital materials granularity

Two main levels of granularity with METS/ALTO XML standard :

- Physical Structure
  $\rightarrow$ pages
- Logical Structure
  $\rightarrow$ block layouts of text



Ephémérides.

21 juillet. (*S. Victor*, martyr, mis à mort en 303.)

1798. *Bataille des Pyramides.* Avant la bataille, le visage du jeune général en chef était rayonnant d'enthousiasme; il se mit à galoper devant les rangs des soldats, et leur montrant les pyramides : «Song-z, s'écria-t-il, songez que du haut de ces pyramydes quarante siècles vous contemplent!» La victoire ouvrit à Bonaparte les portes du Caire.

FRANCE.

PARIS, 20 *juillet*.

Encore une victime des sanglantes journées de juin. M. Dornès, l'un des représentants du peuple qui, au péril de leurs jours, ont accompli une noble mission pacifique au milieu du fléau de la guerre civile, a succombé ce matin aux suites cruelles de sa blessure. M. le président, à l'ouverture de la séance, a fait lecture d'une lettre annonçant cette douloureuse nouvelle. L'Assemblée a immédiatement procédé au tirage au sort d'une députation de cinquante membres pour assister aux funérailles de leur infortuné collègue.

M. Marrast a fait parvenir à l'Assemblée ses remerciements pour les témoignages de confiance qu'elle lui a donnés en le portant au fauteuil de la présidence, et il a exprimé en même temps ses regrets de ne pouvoir prendre possession de ses nouvelles fonctions à cause de l'indisposition qui le retient chez lui.

Le résultat du scrutin pour la vice-présidence, proclamé par M. le président, désigne M. Bixio.

Une discussion très-longue, mais qui n'a pas été toujours très-édifiante en matière électorale, s'est établie ensuite concernant l'élection de M. Gent dans le département de Vaucluse. M. Gent, l'un des commissaires extraordinaires du choix de M. Ledru-Rollin, est véhémentement accusé, dans une protestation signée par de nombreux électeurs, d'avoir mis en usage, et tout à son profit, le lendemain de la république, certaines menées illicites qui scandalisaient si fort sous le régime déchu tous les républicains de la veille. Les débats se sont terminés par la décision d'une enquête sur les faits dénoncés.

Le reste de la séance a été consacré à l'adoption de décrets autorisant plusieurs villes à contracter des emprunts et au réglement définitif du budget de 1845.

Une réclamation de M. Ceyras portée à la tribune avec une certaine excentricité de langage parlementaire, a fait un moment diversion à la tristesse et à la monotonie de cette séance. Il s'agissait d'une proposition faite dès le mois de mai dernier.

L'honorable représentant a retracé avec une tendresse toute paternelle les mésaventures de cette proposition qui, a-t-il dit, a été précipitée dans le comité des travailleurs, où elle a fait le *plongeon*, et il a supplié l'Assemblée de vouloir bien l'aider à la retirer de cet *abime*.

Le style coloré de l'orateur n'a pu attendrir l'Assemblée, qui a passé à l'ordre du jour et enterré définitivement la susdite proposition. A quoi sert donc l'éloquence !

On assure que le comité des affaires étrangères a pris aujourd'hui une résolution importante sur les affaires d'Italie.

Les bruits alarmants qui s'étaient répandus avant le 14 juillet ont tous été démentis par le fait, et cet anniversaire s'est passé le plus paisiblement du monde; il en sera de même des anniversaires de juillet, qu'on avait également marqués d'un signe néfaste, mais dont personne ne parle plus aujourd'hui, et par une excellente raison : la force publique, déjà invincible, s'accroît et s'organise plus énergiquement chaque jour. Aujourd'hui même, 20 juillet, doivent être complètement installés les camps qui forment le cordon sanitaire contre le choléra des insurgés : ces quatre camps sont placés à Saint-Maur, au Champ-de-Mars, à La Chapelle et à Satory, près de Versailles. Les agitateurs commencent à comprendre toute leur impuissance; ils se voient abandonnés de ceux-là même qui les ont jetés dans la rue aux sanglantes journées de juin.

Cependant nous apprenons qu'un certain nombre de personnes habitant la province, et que l'éloignement dispose à concevoir plus lentement des alarmes, mais aussi à revenir moins promptement que les Parisiens à la confiance, se sont inquiétées à l'excès des bruits invraisemblables qui s'étaient propagés avant le 14 juillet. On nous dit, par exemple, que des parents retirent leurs enfants des pensions et des établissements publics, de peur qu'ils ne tombent entre les mains d'une insurrection nouvelle. Rien n'est moins fondé ni moins raisonnable que ces craintes, et nous engageons les familles à se rassurer par la certitude que les établissements d'instruction sont en parfaite sécurité. Ce qui est le plus à craindre en ce moment, c'est précisément la peur, ennemie du crédit, du travail, de la prospérité de la capitale, si nécessaire au repos de toute la France. Avoir confiance sans cesser d'être vigilant; dépenser selon ses moyens, sans permettre au gouvernement les fausses mesures compromettantes pour les finances; reprendre l'activité laborieuse des temps calmes, sans se départir de l'énergique fermeté que commandent les temps agités; voilà à la fois le meilleur calcul et le vrai patriotisme.

Nous avons cité, on s'en souvient sans doute, un curieux passage d'une lettre où M. Marc Dufraisse, préfet de l'Indre, blâmait la *faiblesse* des journaux de l'opposition qui avaient abandonné les complices du crime de Fieschi. M. Senard, qui nettoie en ce moment les écuries d'Augias que lui a léguées la commission exécutive, vient de donner satisfaction à la conscience publique. Par ordonnance insérée au

Figure 3 – Logicial Structure Detection.

Block layout is rarely a proper content.

Corpus construction

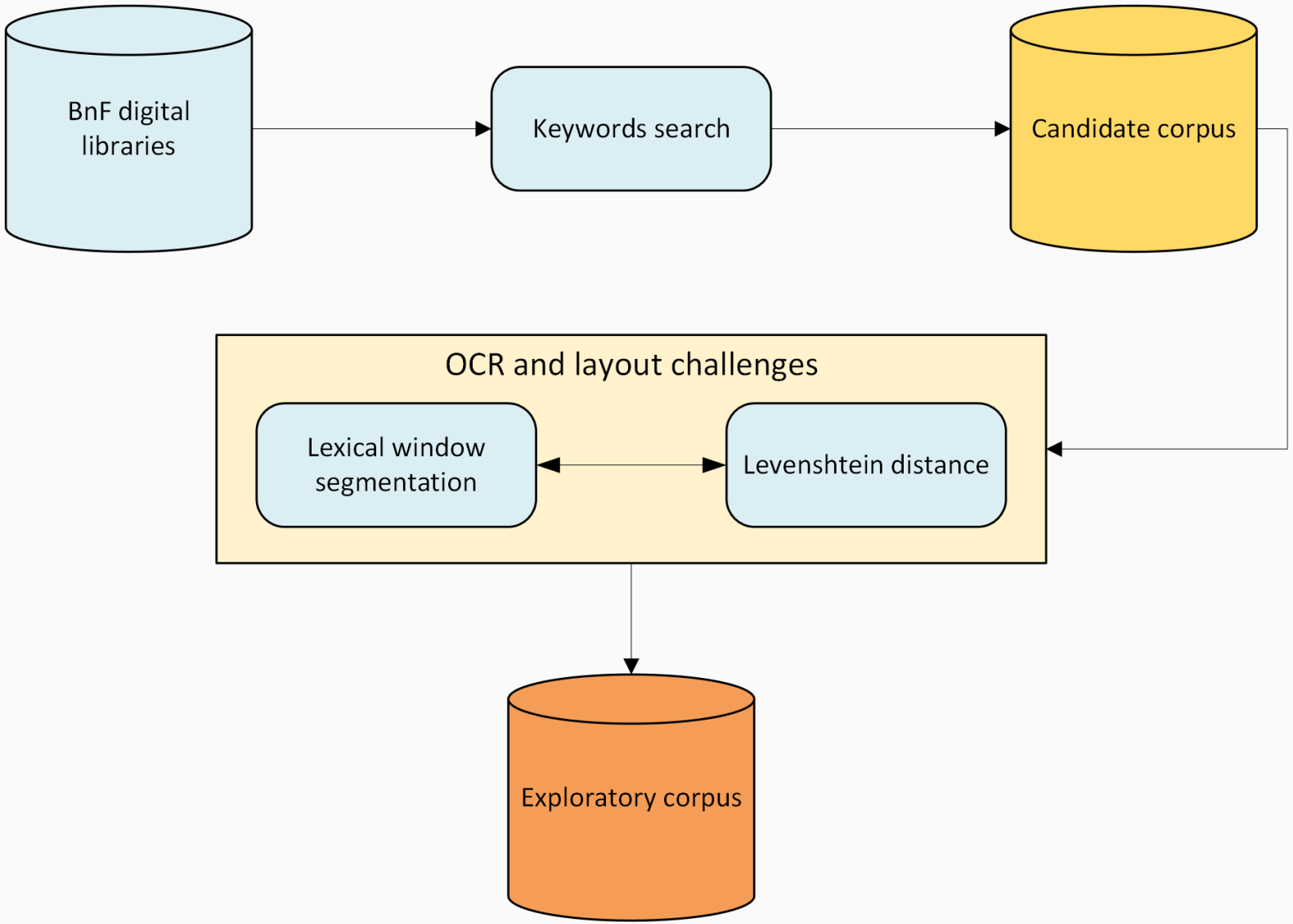


Figure 4 – Corpus construction flow chart.

Corpus construction

Corpus construction steps :

1. Candidate corpus construction :
   - 1.1 all occurrences of the phrase « journées de juin » in the national daily press (June 1848 – December 1852)
   - 1.2 candidate corpus (6k results, 1.7 Gb json files)
2. Exploratory corpus construction :
   - 2.1 Lexical window slicing of approximately 500 words around the searched keywords
   - 2.2 1.7 million words to explore !

Previous research : (Wilson Black , 2022), (Ehrmann et al., 2022), (Beach & Walker Hanlon, 2023), (Owens & Padilla, 2021), (Sternfeld, 2011).

Corpus Exploration with Distant Reading

Which Distant Reading methodology ?

Two approaches have been chosen :

1. Supervised classification from a training dataset
2. Unsupervised classification from the labeled dataset

Each utilizing BERT based language models :

> *« fine-tuning pre-trained BERT on OCR'd texts significantly improves its resilience to OCR noise in classification tasks » (Jiang et al., 2021)*

Corpus Exploration with Distant Reading

Supervised classification

Generating A Training dataset

Random Sampling of a training dataset (5 % of the candidate corpus) :

- 373 documents
- 90k words

Iterative labeling heuristic trials :

- From 5 to 3 labels

Final tradeoff :

- Political debates (n = 185)
- Law and order fighters (n = 72)
- Insurgents (n = 116)

Few-Shot Classification

SetFit framework (Tunstall et al. 2022) :

- based on *Sentence Transformers*
- fast to fine-tune with few-shot
- prompt-free
- well-documented

Based-model used for sentence embeddings : DistilCamemBERT (Delestre & Amar, 2022)

- distillation version of CamemBERT (Martin et al., 2020)

Model evaluation

- training dataset : 20 labeled examples per class (60 documents)
- test dataset : 313 documents

Class	Precision	Recall	F1-Score	Support
Political debates	0.92	0.80	0.86	96
Law and order fighters	0.86	0.96	0.91	165
Insurgents	0.85	0.75	0.80	52
Accuracy			0.88	313
Macro Avg	0.88	0.84	0.85	313
Weighted Avg	0.88	0.88	0.87	313

Table 1 – Classification Report

Corpus Exploration with Distant Reading

Unsupervised classification

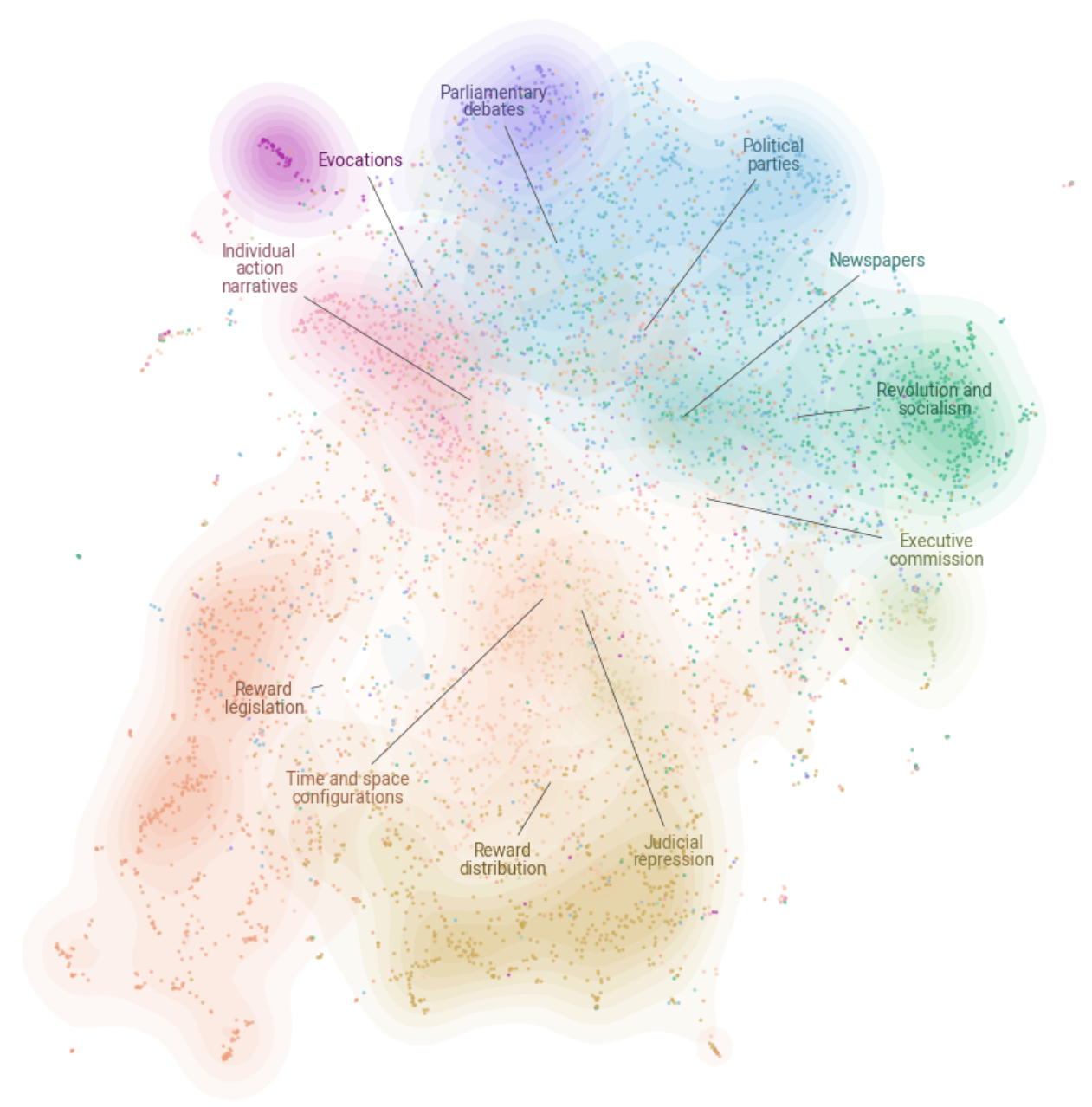
Topic Modeling

BERTopic (Grootendorst, 2022)

Topic modeling approach based on 5 main steps :

1. Sentence embeddings
2. Dimensionality Reduction
3. Clustering
4. Tokenizer
5. Weighting scheme
6. Fine-tune topics representations

Documents and topics map



Topics and labels

Figure 5 – Topics and predicted labels heatmap
(Pearson's Chi-squared test :X-squared = 3491.2, df = 20, p-value < 2.2e-16).

Conclusions

- Usable results with sentence-transformers based models and noisy textual data.
- A combination of supervised and unsupervised classification is a way to get a better sense of the discursive phenomena at work.
- Main future research directions :
  - expand the corpus,
  - OCR post-correction based on LLMs before the candidate corpus stage.