SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

PROJET INDUSTRIEL MAIN 4

Étude sur le renouveau politique en France

Samy ASMA
Achille BAUCHER
Homer DURAND
Thomas GENIN
Ibtissam LACHHAB

Encadrant Metriq:
M.Jean-Tupac QIROGA
Encadrant Académique:
M.Xavier TANNIER

23 Janvier 2020

Nous tenons particulièrement à remercier nos encadrants académiques et d'entreprise Xavier Tannier *et* Tupac Quiroga *pour leur soutien actif.*

Contents

1 Introduction

1.1 Contexte

Depuis la loi du 9 décembre 2016, la Haute Autorité pour la Transparence de la Vie Publique (HATVP) a instauré une obligation de référencement à un registre public pour les représentants d'intérêt [1]. De plus, sous chaque mandat présidentiel, l'assemblée nationnale archive les questions émises par les députés aux ministères. L'agence Proches, entreprise de communication d'influence politique, désire utiliser les nouvelles technologies d'analyse de données pour proposer une étude innovante afin de rendre accessible aux citoyens les mécanismes d'influence politique. Elle a donc fait appel à Metriq, entreprise spécialisée dans l'analyse de données, pour produire une étude en accès libre sur ce sujet.

1.2 Problématique

Notre projet est un projet exploratoire consistant à l'étude analytique de ces bases de données afin de mettre en exergue des liens entre les questions posées, les députés, leur groupes politiques, l'actualité ou encore les possibles vecteurs d'influence. La technicité de notre projet porte sur la maîtrise des outils d'analyse de bases de données et d'analyse textuelle. L'étude devant être accessible à tous, les résultats devront apparaître sous forme de graphiques explicatifs et parfois interactifs. L'interprétation des résultats est laissée aux experts en lobbying et influence de l'agence Proches, qui jugent de la pertinence de ceux-ci et nous orientent vers les pistes de recherches adéquates.

2 Préparation

2.1 Outils

Nous avons choisi le langage Python pour sa simplicité et son abondance de librairies puissantes et accessibles :

- **Gestion de données** : Pandas, Numpy
- **Analyse textuelle** : Nltk, SciKit-Learn, Spacy
- **Visualisation** : Matplotlib, Seaborn, Plotly, et Streamlit pour les interactions.

Nous travaillons avec un serveur d'environnement de développement Jupyter qui nous est fourni par l'entreprise Metriq.

En ce qui concerne le dépôt de notre code source, nous travaillons sur le service d'hébergement GitHub. Nous déployons notre application web interactive sur un serveur Heroku.

Pour la partie management du projet, c'est la plateforme Basecamp qui nous sert de support de discussion avec notre encadrant Metriq. En outre elle est aussi un outil de gestion de notre avancement, en proposant un outil de gestion et de mise à jour des tâches à effectuer.

2.2 Données

2.2.1 Présentation

Notre projet s'axe sur les bases de données de l'Assemblée Nationale disponibles sur leur site web. Les bases de données contiennent les textes des questions écrites des députés au gouvernement, et leurs réponses lorsque celles-ci sont disponibles. Nous avons eu accès aux bases de données de la XIVe et de la XVe législature, ce qui nous a permis d'effectuer un travail de comparaison entre la législature de François Hollande et celle d'Emmanuel Macron. La fonction des lettres est décrite sur le site de l'Assemblée Nationale [2] comme suit :

> Les questions écrites sont posées par un député à un ministre ; seules celles qui portent sur la politique générale du Gouvernement sont posées au Premier ministre.
>
> Les questions écrites doivent être sommairement rédigées et se limiter aux éléments strictement indispensables à la compréhension de la question. Elles ne doivent contenir aucune imputation d'ordre personnel à l'égard de tiers nommément désignés. En outre, le principe de séparation des pouvoirs et d'irresponsabilité du chef de l'État interdit à l'auteur d'une question écrite de mettre en cause les actes du Président de la République.
>
> Le texte des questions écrites est remis au Président de l'Assemblée nationale, qui le notifie au Gouvernement. Depuis 2008, les députés déposent leurs questions par voie électronique en utilisant un portail internet spécialisé. Les questions écrites sont publiées chaque semaine, durant les sessions et hors session, dans un fascicule spécial du Journal officiel qui comporte également les réponses des ministres aux questions précédemment posées. Depuis le 1er janvier 2016, ce fascicule est dématérialisé et la version authentique est consultable sur le site de l'Assemblée nationale.
>
> Les réponses aux questions n'ont en principe aucune valeur juridique et ne lient pas l'administration sauf en matière fiscale où elles sont considérées comme exprimant l'interprétation administrative des textes.
>
> En raison de sa simplicité et de son caractère illimité et facilitée par les nouvelles techniques informatiques, la procédure des questions écrites a rencontré un très large succès. Elle permet en effet aux députés d'intervenir quand ils le souhaitent (même en intersession) et autant qu'ils le souhaitent auprès des ministres pour des questions touchant souvent directement leurs électeurs. La conséquence a été une inflation du nombre de questions écrites : de 3 700 questions écrites déposées en 1959, on est passé à 12 000 en 1994 et 20 066 en 2015.
>
> Le délai moyen de réponse s'est établi à 180 jours au 30 septembre 2015. Le taux global de réponse reste constant, se situant à environ 70%.

Les lettres suivent ainsi un formalisme assez précis, comme on peut le voir sur la figure 1.

2.2.2 Extraction

Les lettres sont disponibles sur le site de l'assemblée nationale au format xml et json. Nous les avons téléchargées, puis extraites avec Python vers un DataDrame (tableau de données) de la

```
"M. Jean-Charles Colas-Roy attire l'attention de Mme la garde des sceaux, ministre de la justice, sur la rec
onnaissance de la langue des signes française dans la Constitution. Depuis le 30 mars 2007, la France a sign
é la convention relative aux droits des personnes handicapées ratifiée par décret le 1er avril 2010. Parmi c
es droits, se trouve la reconnaissance par l'État de l'ensemble des langues parlées et non parlées telles qu
e la langue des signes. De plus, le code de l'éducation dispose que la langue des signes est reconnue comme
langue à part entière. Aujourd'hui, cette reconnaissance correspond à une recommandation de l'Union européen
ne et de l'Organisation des Nations Unies, afin de permettre l'accès à la pleine citoyenneté des personnes s
ourdes, sans discrimination. Il lui demande donc de bien vouloir lui faire connaitre les intentions du Gouve
rnement à ce sujet, et plus précisément s'il entend intégrer la langue des signes française dans le futur pr
ojet de réforme constitutionnelle."
```

Figure 1: Lettre du député Colas-Roy à Mme la ministre de la justice

librairie Pandas pour faciliter les manipulations. Nous avons dû pour cela convertir les dates et les textes au bon format. Les premières lignes du DataFrame ainsi obtenues sont affichées sous Jupyter sur la figure 2.

	date_question	date_reponse	groupe_auteur	ministere_adresse	question	rubrique	titre
0	2019-04-23	NaT	LAREM	Ministère de la justice	M. Jean-Charles Colas-Roy attire l'attention d...	droits fondamentaux	Reconnaissance de la langue des signes françai...
1	2018-06-12	2018-12-25	LAREM	Ministère de l'intérieur	M. Bertrand Sorre attire l'attention de M. le ...	sécurité routière	Précision sur les 80 km/h
2	2018-06-19	2018-10-02	LAREM	Ministère de la transition écologique et solid...	Mme Françoise Dumas attire l'attention de M. l...	eau et assainissement	Aides des agences de l'eau aux projets de réha...
3	2019-07-02	NaT	FI	Ministère de l'intérieur	M. Alexis Corbière alerte M. le ministre de l'...	sécurité des biens et des personnes	Conditions de travail des pompiers professionn...
4	2018-05-01	2018-09-25	LAREM	Ministère de l'intérieur	Mme Barbara Pompili appelle l'attention de M. ...	sécurité routière	Forfait post-stationnement et loueurs courte d...

Figure 2: Affichage du DataFrame Pandas des lettres dans Jupyter

2.2.3 Attributs

- **date_question** : La date à laquelle la question a été écrite (*Année, Mois, Jour*). Il est à noter que les lettres sont répertoriées seulement une fois par semaine, le Mardi.
- **date_reponse** : Date de réception de la réponse, si réponse il y eu. Dans le cas contraire, l'attribut vaut NaT.
- **groupe_auteur** : Le groupe politique [3] de l'Assemblée auquel appartient le député auteur de la lettre. Il peut correspondre à un parti ou à un regroupement de partis politiques . Le nombre de lettres envoyé par chaque groupe est affiché en Figure 3.
- **question** : Le texte de la question.
- **rubrique** : La rubrique dans laquelle chaque lettre est placée, correspondant au thème de celle-ci. Voir la Figure 5 pour quelques exemples.
- **titre** : Le titre de la question.

2.3 Première approche

Nous avons commencé par étudier le contenu des données pour réfléchir à l'orientation de nos recherches.

Tout d'abord, les rubriques très diverses dans lesquelles sont placées les lettres nous ont paru assez représentatives du thème abordé et c'est pourquoi nous les avons étudiées plus précisément dans la partie 3.

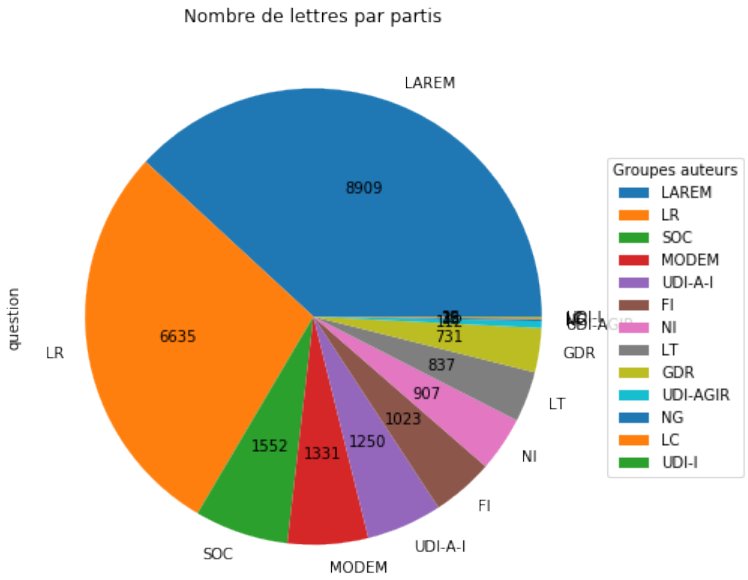


Figure 3: Les groupes auteurs de l'Assemblée Nationnale pour la législature Macron

En affichant l'évolution du nombre de lettres par semaine nous avons compris que les quantités étaient assez grandes pour analyser de manière quantitative les variations et anomalies, ce que nous avons fait dans la partie 4.

En étudiant le contenu des lettres, nous avons remarqué qu'elles suivaient un formalisme très précis (formules de politesses et manière d'aborder le sujet) comme on le voit dans la figure 1. Le ton des lettres est très neutre, ce qui rend difficile une analyse de sentiments, et les arguments sont généralement assez complexes et exprimés dans un style élaboré, une analyse sémantique paraît hors de portée. Pour l'analyse du texte, nous nous sommes donc surtout intéressés aux champs lexicaux, dans la partie 5 et par la suite.

3 Analyse des thèmes (rubriques)

. Dans un premier temps nous avons donc choisi d'analyser les thèmes des lettres écrites ainsi que leur émetteur, pour avoir une vision d'ensemble, avant de s'attarder sur le contenu des lettres en question.

3.1 Tendances générales

Nous avons d'abord voulu observer l'évolution et la fréquence des questions dans les différentes rubriques. Pour avoir une vision d'ensemble avant de s'intéresser de plus près à certaines d'entre elles, nous avons commencé par afficher les occurences moyennes des questions par rubrique, toutes rubriques confondues, au cours du mandat du gouvernement Macron (Figure 4). Cette observation devait aussi nous indiquer la présence ou non de tendances globales au cours du temps.

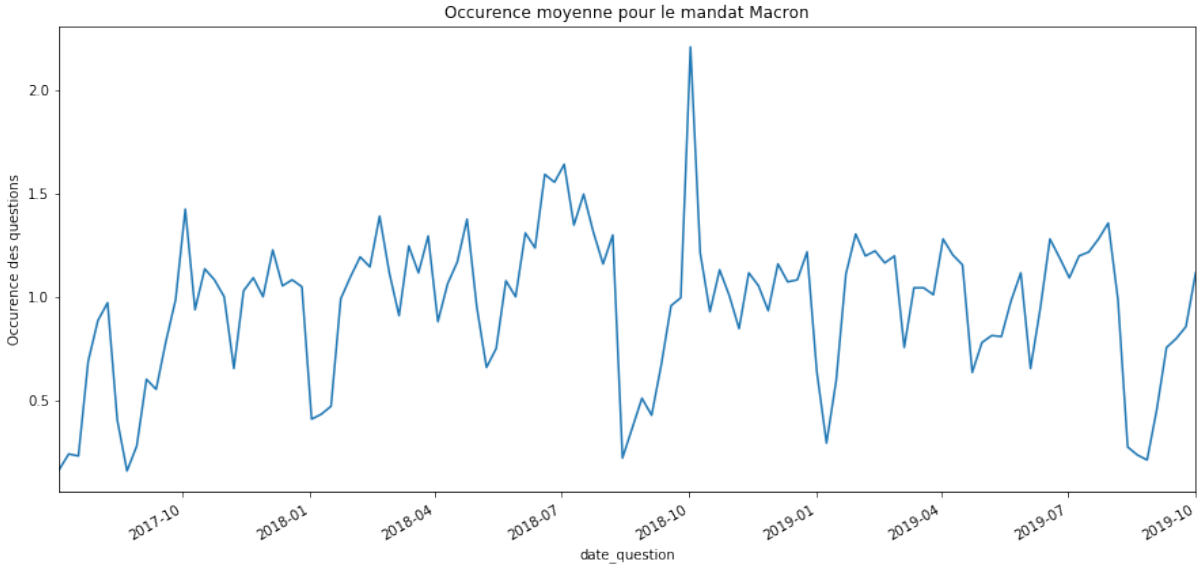


Figure 4: Occurence moyenne des lettres par rubrique

Comme nous pouvons l'observer, la fréquence des questions écrites est extrêmement variable, et présente de nombreux pics et creux très marqués. De plus, nous pouvons observer une certaine régularité, avec des périodes moins favorables aux questions, comme par exemple celle qui suit juillet. D'après le spécialiste en comunication d'influence de l'agence Proches, avec qui nous nous sommes entretenus ultérieurement, ces particularités sont facilement explicables. Les périodes de creux suivies d'un pic autour de Juillet correspondent aux périodes de fermeture de l'assemblée nationale, qui laissent la semaine suivante en brusque augmentation avec les questions préparées durant cette fermeture. La période de janvier présente des creux corespondant à la cloture de l'année, et donc à une baisse des motions présentées devant l'assemblée.

La visualisation de ces tendances nous a permis par la suite d'avoir à l'esprit que des phénomènes globaux jouaient aussi sur les occurences des lettres dans les différentes rubriques, et qu'il fallait donc les prendre en compte dans les calculs.

3.2 Répartition globale des rubriques

Nous nous sommes ensuite intéressés aux évolutions des thèmes traités sous les mandats présidentiels de 2012 et 2017. Le but est de dégager les grandes tendances des thèmes abordés sous ces deux mandats, pour pouvoir éventuellement les confronter à l'actualité politique du pays. Nous avons commencé par afficher dans la Figure 5 les 10 rubriques les plus abordées durant les deux mandats, on constate que l'on obtient des résultats sensiblement différents.

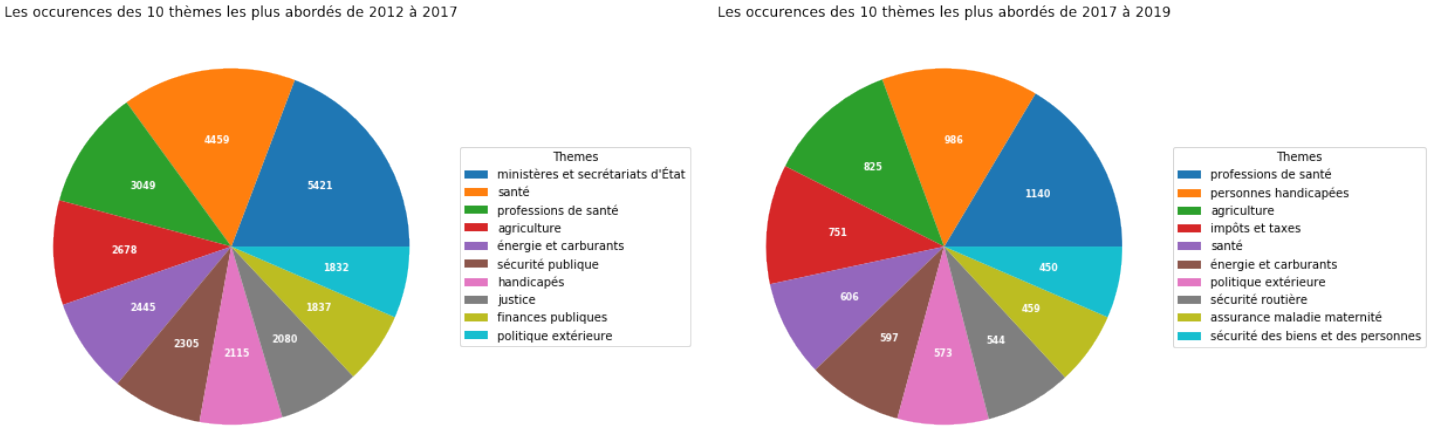


Figure 5: Thèmes les plus abordés entre 2012 et 2019

La visualisation des ces différences nous donnent des informations sur la priorité des thèmes abordés octroyée par chaque mandat. Pour analyser plus précisément l'évolution de la répartition des rubriques, nous avons affiché les occurences des lettres dans les principaux thèmes abordés chaque semaine au cours de ces deux mandats.

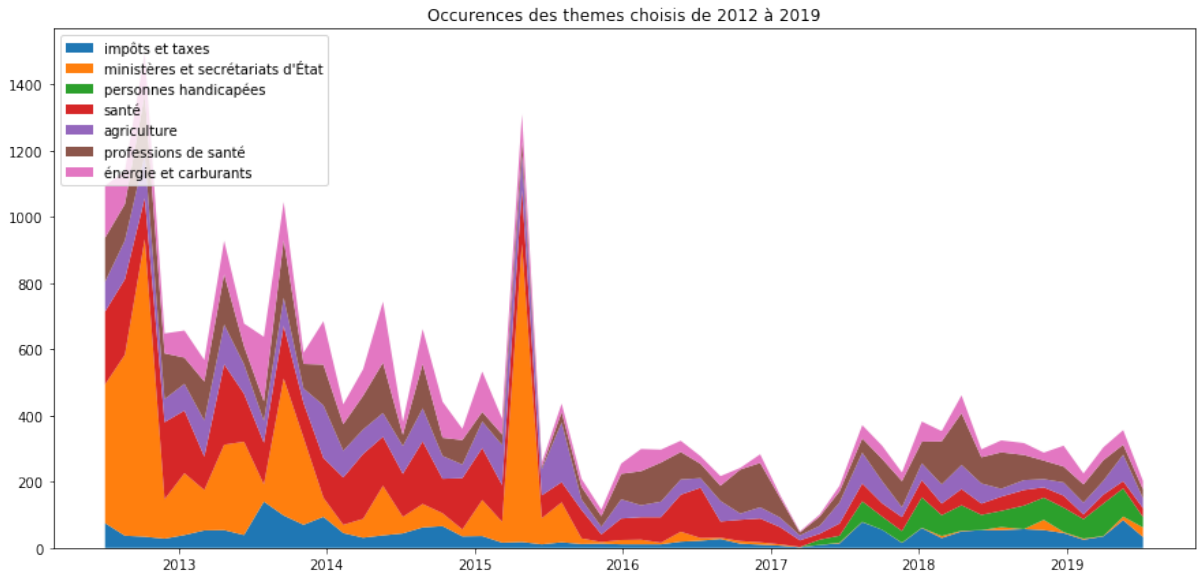


Figure 6: Occurence des thèmes majeurs abordés entre 2012 et 2019

On peut y remarquer, entre autres, l'émergence de thèmes nouveaux à partir du mandat du gouvernement Macron, comme la rubrique *personnes handicapées*. Cependant, les résultats obtenus sont à prendre avec précaution, car d'après notre discussion avec l'entreprise proches,

la répartition des lettres dans les différentes rubriques dépend de l'arbitraire de l'administration. Etant donné que les changements de noms de rubriques peuvent biaiser la comparaison des deux mandats, nous nous sommes par la suite intéressés à un mandat en particulier, celui du gouvernement Macron.

3.3 Focalisation sur une rubrique

Maintenant que nous avons une idée des principaux thèmes des questions écrites et de leurs variations sous les différents mandats présidentiels, nous pouvons nous intéresser plus précisément à certaines d'entre elles. L'outil de visualisation que nous souhaitons proposer ici concerne l'analyse d'un thème en particulier, nous prendrons dans l'exemple celui de la sécurité routière. On aimerait savoir, d'abord comment cette rubrique a été traitée au cours du temps, mais aussi quels sont les partis politiques français qui abordent le plus ce thème dans les questions.

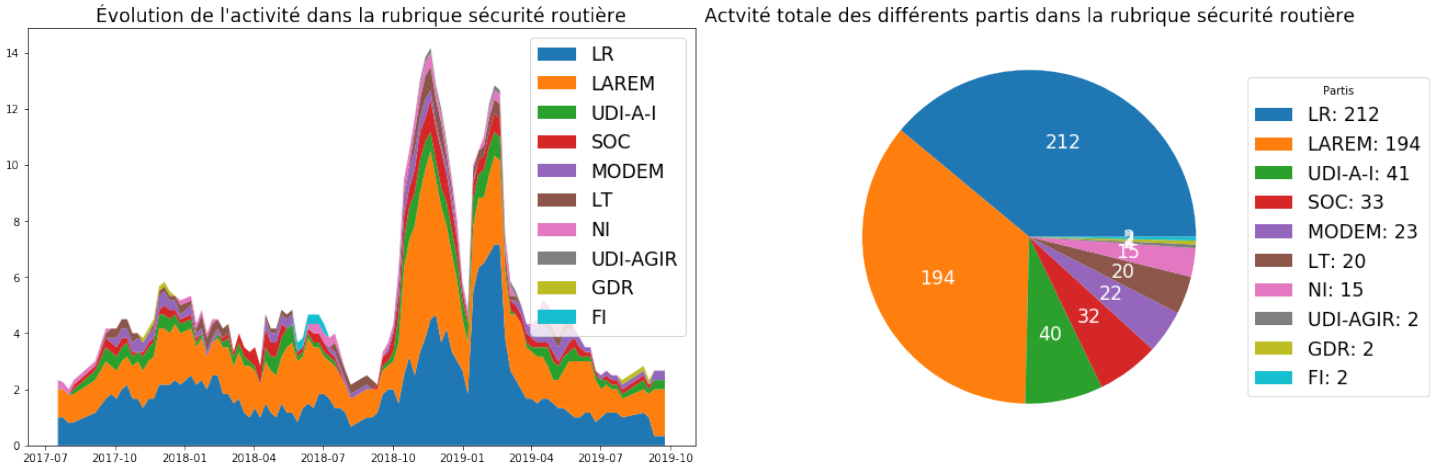


Figure 7: Activité des partis sur le thème de la sécurité routière

On remarque alors que les partis les plus actifs sur le thème de la sécurité routière sont majoritairement Les Républicains puis la République En Marche tout au long des deux premières années du mandat présidentiel de 2017. Cette représentation est à nuancer par le fait que ce sont les deux partis les plus actifs dans l'assemblée nationale, qui comptent le plus grand nombre de députés. Des corrections pour prendre en compte ce biais sont proposées dans la partie 7.

De plus, si on regarde l'évolution des questions abordant ce thème, on peut y interpréter les pics comme des réactions à des actualités politiques, ici le passage aux 80 km/h. La détection du thème abordé au cours est plus amplement traitée dans la partie 5. Ce graphique nous aide aussi à repérer si des partis ont été particulièrement actifs à certaines périodes.

3.4 Distances entre partis

Nous avons remarqué au cours de notre étude que les partis n'abordaient pas chaque thème avec la même fréquence. Nous avons affiché les 10 thèmes les plus abordés par quelques groupes auteurs dans la figure 8, ce qui confirme notre intuition.

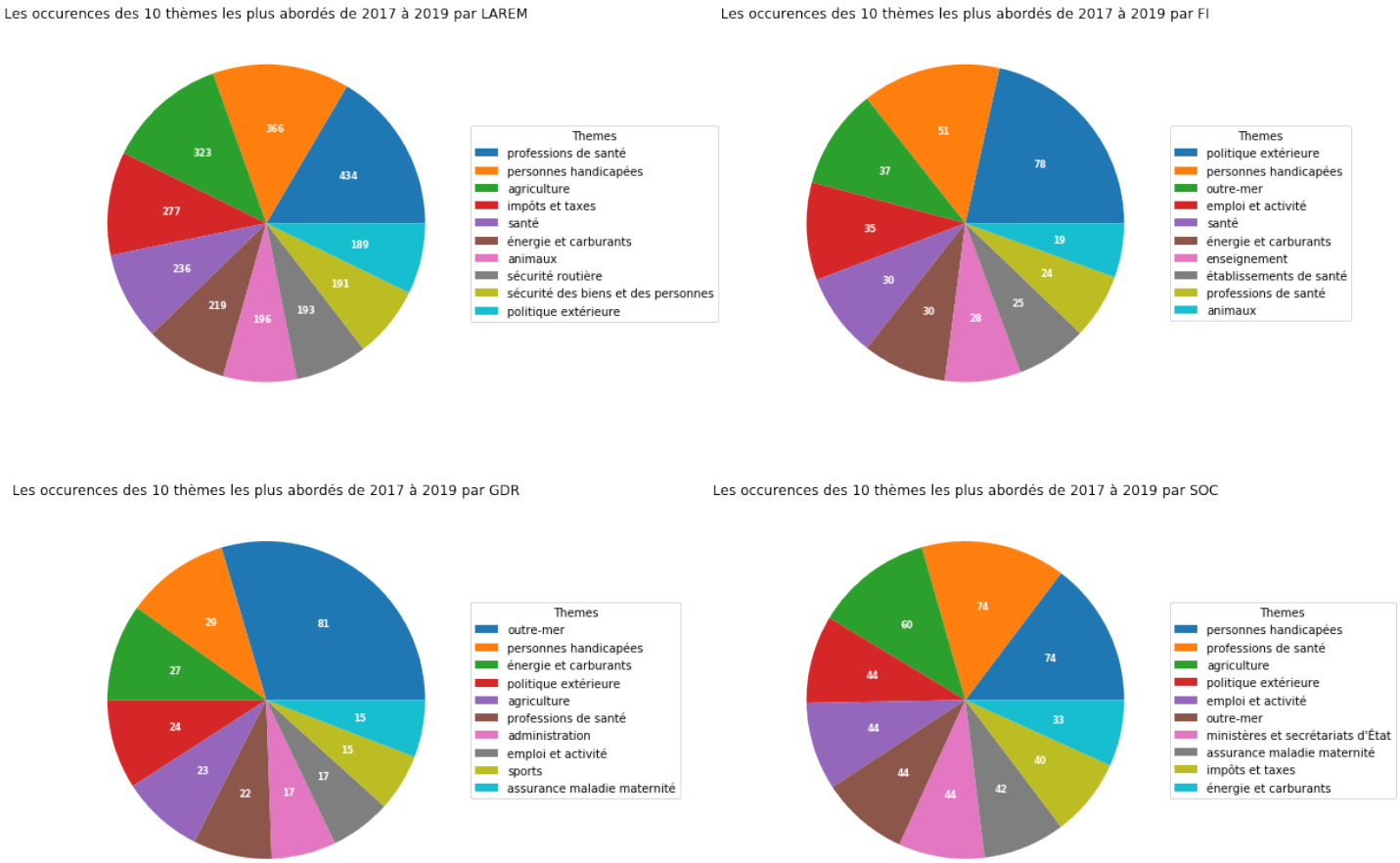


Figure 8: Occurrence des thèmes par partis

On remarque en effet que les députés LAREM posent principalement leurs questions dans les thèmes *profession de santé*, *personnes handicapées* et *agriculture* quand les députés FI abordent majoritairement les thèmes *politique extérieure*, *personnes handicapées* et *outre-mer*.

Nous avons donc cherché un outil qui nous permettrait de se représenter plus précisément les différences et similarités qu'il existait entre les partis. Il nous a semblé intéressant de calculer une distance séparant les partis selon leur répartition dans les rubriques. Pour cela on se place dans un espace à n dimensions (pour les n thèmes de la base de données) et on y place chaque parti en un point X en fonction de la proportion avec laquelle ils abordent chaque thème :

$$X = \{\frac{v_i}{\sum_{i=1}^{n} v_i}, 1 \leq i \leq n\}$$

Avec v_i le nombre de lettres envoyées par le parti dans la rubrique i. On peut alors calculer une distance entre ces deux vecteurs dans cet espace. Nous avons choisi d'étudier

deux formules différentes. La première est la distance **Euclidienne**, qui est la plus intuitive.

$$Dist\_eucl(X1, X2) = \sqrt{\sum_{i=1}^{n}(X1_i - X2_i)^2}$$

La seconde est la distance de **Manhattan**. Cette dernière permet de réduire l'impact des grandes valeurs de différences, car les termes ne sont pas mis au carré. La distance usuelle nous donnant forcément un résultat compris entre 0 et 2, nous la divisons par deux afin d'obtenir un pourcentage de différence, ce qui est plus parlant.

$$Dist\_manh(X1, X2) = \frac{1}{2}\sum_{i=1}^{n}|X1_i - X2_i|$$

Pour notre analyse, la distance de **Manhattan** répond mieux au problème posé. En effet, les proportions des thèmes pour les petits partis peuvent n'être pas pertinentes, dans le sens où ils peuvent n'avoir pas assez de lettres pour que la proportion mesurée soit représentative. Ainsi, un parti ayant écrit très peu de lettres aura de très grandes valeurs de proportion dans certaines rubriques, et 0 dans beaucoup d'autres, ce qui impliquera de grandes différences avec les proportions des autres partis. Nous souhaitons réduire l'influence de ces grandes valeurs de différences provoquées par les petits partis.

En calculant la *Dist_manh* de chacun des partis par rapport aux autres et en plaçant les résultats dans une matrice, on peut afficher le résultat sous forme de Heatmap (où les coefficients sont remplacés par des couleurs pour plus de visibllité) dans la Figure 9) .

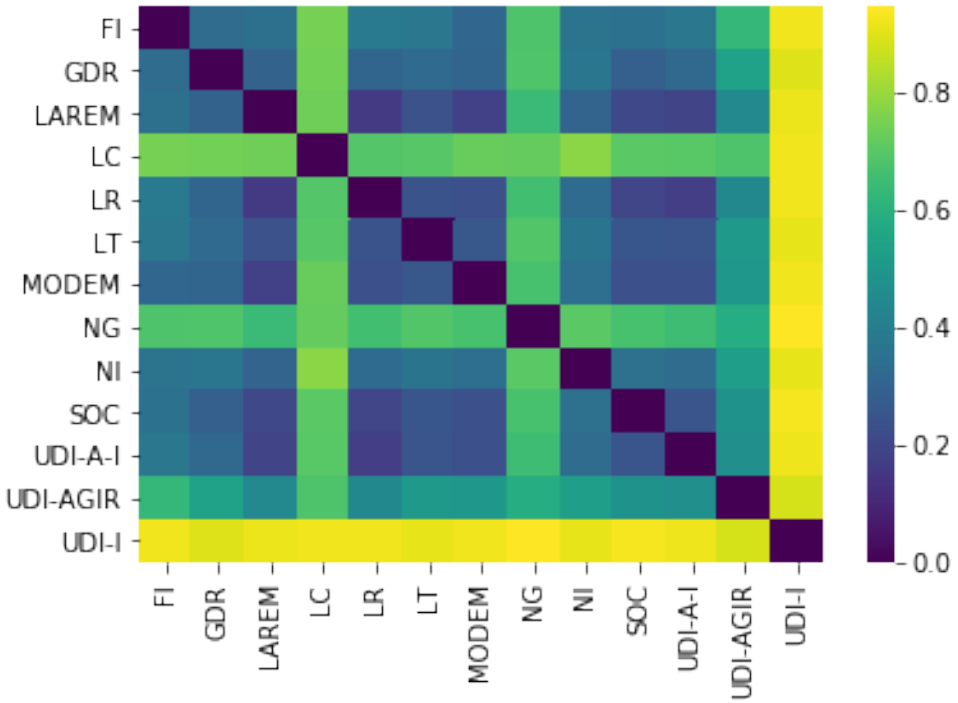


Figure 9: Heatmap de la distance de Manhattan entre les partis

On remarque ainsi que certains partis se distinguent des autres dans leur façon d'aborder les thèmes. C'est par exemple le cas de LC, NG ou UDI-I dont les distances avec

les autres groupes sont respectivement supérieures à 0.6, 0.6 et 0.8. En revanche certains groupes tels que LR et LAREM sont proches avec une distance d'environ 0.4.

Cependant, cette mesure de distance est biaisée par le fait que les petits partis, comme LC et NG, écrivent très peu de lettres et n'abordent donc pas assez de thèmes pour pouvoir comparer leurs proportions aux autres. Lors de notre entretien avec l'agence Proche, Armand nous a donc conseillé de pallier à ce problème en regroupant certains petits partis, qui avaient des opinions et des affiliations proches. Il nous a donné la liste des groupement concernés, c'est à dire des 3 UDI et de LC (UDI-AGIR, UDI-A-i, UDI-i, LC regroupés en UDI-LC), ainsi que de NG et SOC (NG-SOC). En plus d'éliminier en partie le biais des petites valeurs, ce regroupement permet aussi sur la heatmap de comparer plus précisément les groupes, les nuances de couleur étant plus remarquables puisqu'on a éliminé les valeurs trop extrêmes (Figure 10).

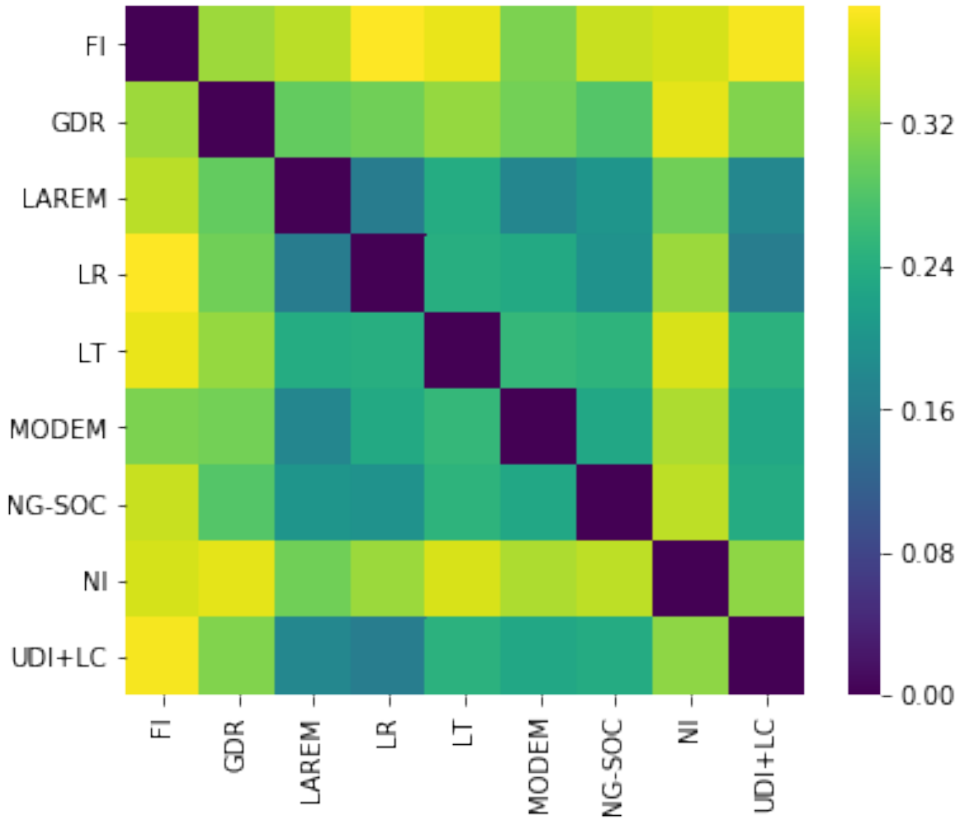


Figure 10: Heatmap de la distance de Mahattan entre les partis regroupés

Nous avons continué dans la suite de notre étude la suite avec ces groupements, qui facilitent beaucoup tout ce qui concerne les comparaisons entre partis.

4 Anomalies

Nous avons pu remarquer une forte variabilité dans les graphiques de l'étude temporelle des occurrences des questions (Figure 4). Particulièrement, on peut observer sur certains thèmes des périodes de pic du nombre de questions posées (Figure 11). En observant de plus près le contenu de ces lettres au cours de ces périodes anormales, nous avons parfois pu comprendre que ce pic était du à un événement particulier, comme un nouveau projet de loi (**Exemple :** Passage aux 80 Km/h). Il nous a donc paru intéressant de réfléchir à une méthode de détection automatique des pics de question. Ces résultats pourraient ensuite servir de base à une analyse textuelle des causes de cette anomalie pour l'identifier à un sujet précis de l'actualité, ou bien à une forte activité spontanée de lobbying .

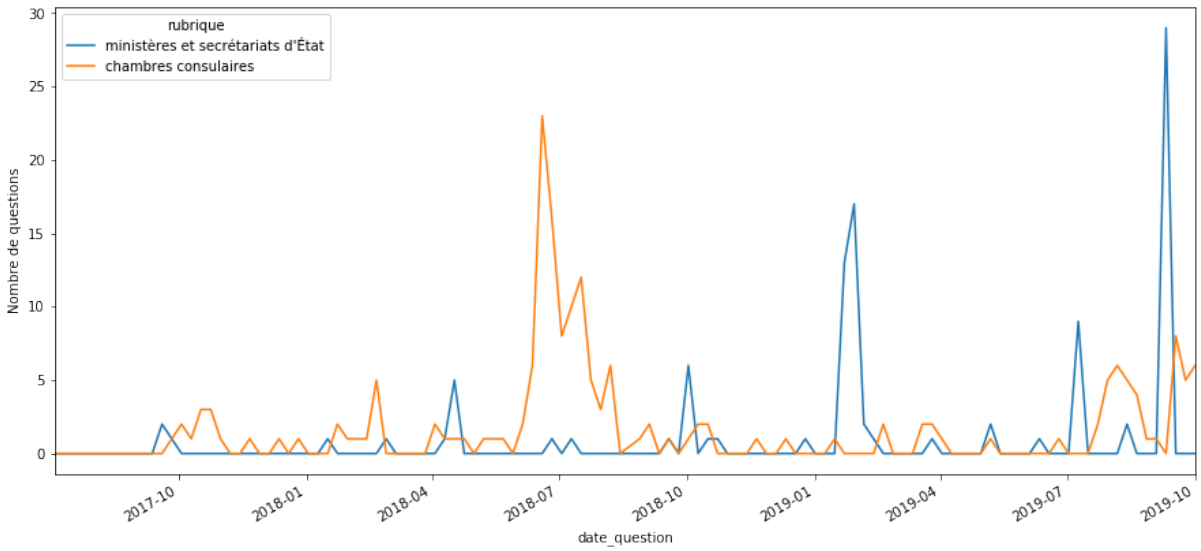


Figure 11: Thèmes comportant des pics de questions

4.1 Repérer les anomalies

La première étape consiste à détecter les rubriques dans lesquelles il y a effectivement un ou plusieurs pics significatifs de questions posées. Sur les conseils de notre professeur de statistiques Fanny Villers, nous nous sommes dirigés vers des méthodes simples et intuitives de calculs avec des moyennes et des variances, qui sont généralement suffisament efficaces.

L'idée est de détecter lorsqu'il y a une (ou plusieurs) grande variation ponctuelle des occurrences des questions. Cette caractéristique peut se manifester dans la somme des écarts absolus à la moyenne (variance empirique):

$$EcartAbs(V) = \sum_{i=1}^{n} |V_i - \bar{V}| \tag{1}$$

avec V le vecteur du nombre de question de chaque semaine, $\bar{V}$ sa moyenne et n le nombre de semaines étudiées.

Nous avons donc ordonné les thèmes par leur plus grande variance empirique. Cependant, ce sont simplement les thèmes disposant du plus grand nombre de questions qui ont occuppé les premières places du classement, alors qu'il ne comportent pas nécessairement

de pics significatifs. Nous avons donc voulu pondérer cette valeur par la quantité des questions posées, pour qu'on puisse détecter une variation relative qui correspond à notre définition d'un pic.

$$EcartRel(V) = \frac{\sum_{i=1}^{n} |V_i - \bar{V}|}{\bar{V}} \tag{2}$$

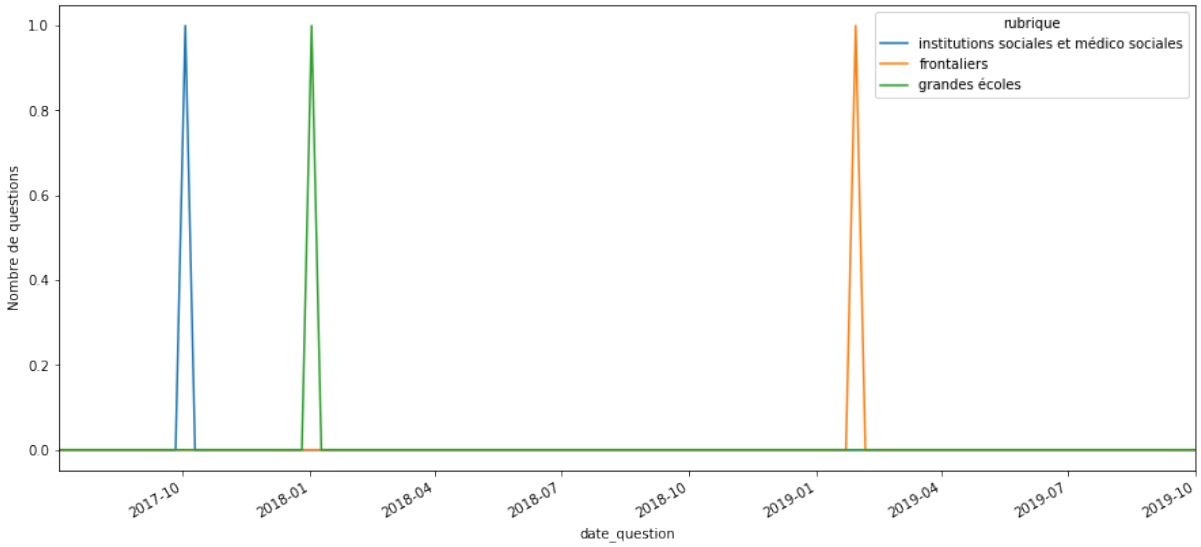


Figure 12: Les 3 thèmes les mieux classés avec la méthode EcartRel

Cette fois-ci, on peut voir dans la Figure 12 que ce sont des thèmes extrêmements peu abordés qui ont fait leur apparition dans les premiers. En effet, un thème qui n'a été abordé que 1 ou 2 fois aura une valeur de EcartRel très élevée à cause de sa moyenne très faible. Cela correspond mieux à notre définition d'un pic, cependant un pic d'une seule question n'est pas du tout intéressant à observer en tant que réaction spontanée sur un sujet précis. Nous avons pensé à fixer une limite minimale du nombre de questions, mais celle-ci étant délicate à déterminer nous nous sommes tournés vers un autre type de pondération accordant plus d'importance au nombre total d'occurrences, en ajoutant une racine carrée:

$$RelSqrt(V) = \frac{\sum_{i=1}^{n} |V_i - \bar{V}|}{\sqrt{\bar{V}}} \tag{3}$$

Cette méthode de classement s'est révélée la plus efficace en mettant en avant des thèmes dans lesquels on obtient des pics assez satisfaisants pour l'analyse, comme le montre la Figure 13.

Nous pouvons donc à présent nous placer sur des thèmes qui comportent une variabilité suffisante et qui correspondent sûrement à un ou plusieurs pics de questions, les creux étant pratiquement inexistants.

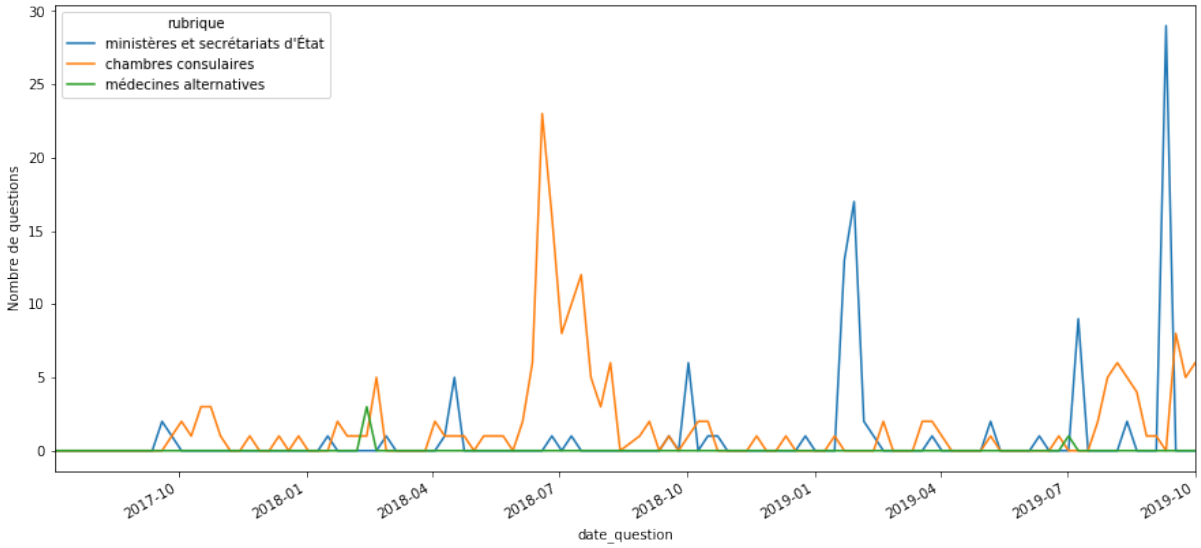


Figure 13: Les 3 thèmes les mieux classés avec la méthode RelSqrt

4.2 Repérer la période

En se plaçant donc sur un thème à pics, il faut maintenant les détecter et déterminer leur durée. Cependant, il est assez délicat de déterminer un coefficient à partir duquel une valeur est suffisament éloignée des autres pour qu'on puisse la considérer comme appartenant à un pic. C'est pourquoi nous avons commencé par ne prendre qu'un seul pic, correspondant à la valeur maximale atteinte sur toute la période.

Pour déterminer la période de ce pic, c'est à dire la période autour de ce pic durant laquelle on a vu un nombre de questions particulièrement élevée, nous avons commencé avec une méthode naïve. On y ajoute simplement toutes les semaines précédant et suivant le pic jusqu'à ce qu'on tombe en dessous de la moyenne, comme en orange sur la Figure 14.

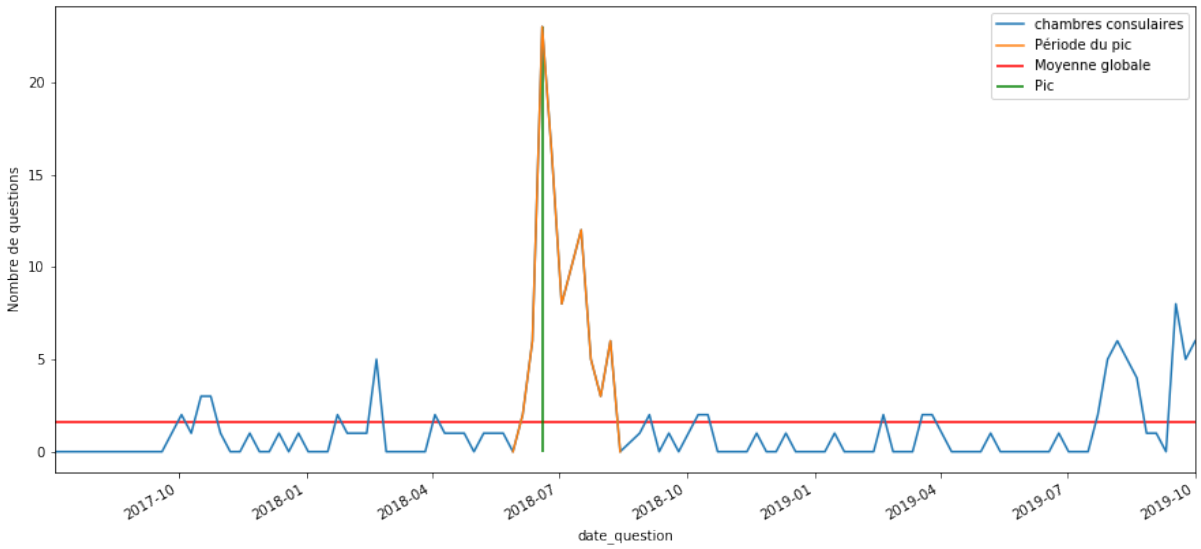


Figure 14: Méthode de détection de la période d'un pic pour un thème exemplaire

Cette méthode s'est révélée assez efficace en pratique, même si nous avons cherché à l'améliorer.

4.3 Résultats

Nous avons industrialisé le processus décrit précédemment pour que le programme affiche automatiquement les graphes des n meilleurs (au sens de la contenance d'anomalies selon le classement de la troisième méthode) thèmes en surlignant la période de pic maximal (voir Figure 15).

Nous avons aussi écrit une panoplie de fonctions permettant d'obtenir les titres des lettres, la répartition des des ministres destinataires et des partis concernés par une période de pic choisie, afin de mieux comprendre la cause et les caractéristiques de ce pic (voir Figure 16).

Nous verrons dans la partie 5.2.2 comment nous avons utilisé la détection de ces périodes pour tenter de connaître le sujet qui a provoqué un pic.

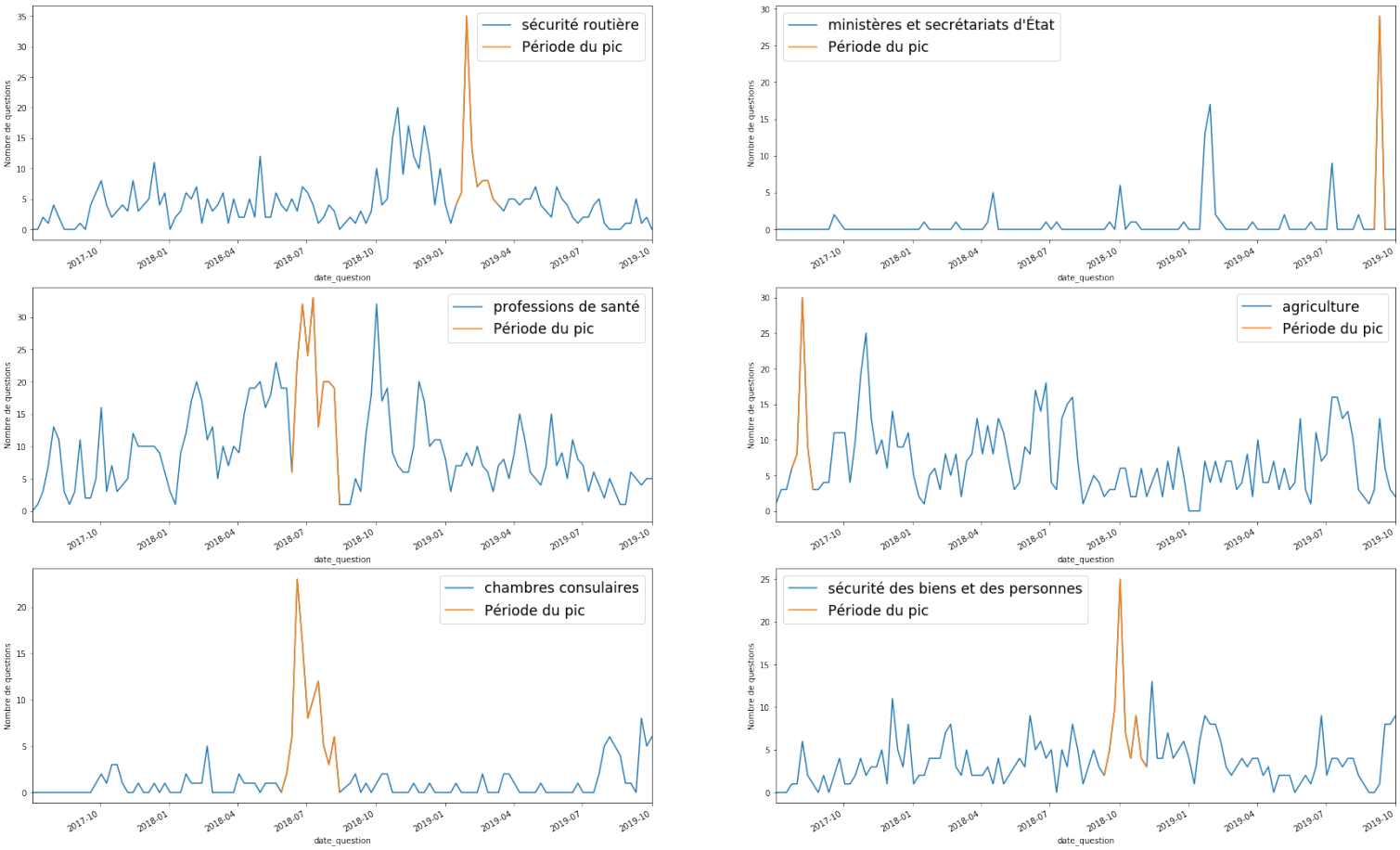


Figure 15: Repérage des meilleurs thèmes selon le classement Abs

date_question	titre
2018-06-19	Ressources des CCI - Engagement du Gouvernemen...
2018-06-19	Ressources des CCI - Évolutions prévues
2018-06-19	Engagements gouvernementaux concernant le budg...
2018-06-19	Évolution des missions et financements des ch...
2018-06-19	Stabilisation des taxes affectées aux chambres...
2018-06-19	Chambres des métiers et de l'artisanat - aveni...
2018-06-19	Ressources des chambres de commerce et d'indus...
2018-06-19	Avenir des chambres de métiers et de l'artisanat
2018-06-19	Ressources des chambres de commerce et'industrie
2018-06-19	Baisse de la taxe pour frais des chambres de c...

groupe_auteur	question
LR	16
LAREM	5
SOC	5
MODEM	4
UDI-A-I	4
FI	2
LT	2
NI	1

ministere_adresse	question
Ministère de l'économie et des finances	19
Ministère de l'action et des comptes publics	14
Ministère du travail	4
Secrétariat d'État auprès du ministre de l'action et des comptes publics	2

Figure 16: Analyse plus précise d'un pic dans le thème *chambres consulaires*

5 Analyse du texte

5.1 Normalisation

Avant de commencer l'analyse des termes du textes, nous l'avons normalisé à l'aide de la bibliothèque Spacy. Notre fonction de normalisation se charge de nettoyer tous les mots inutiles dans l'analyse (que, le, est ...) grâce à une liste de *stopwords* que Monsieur Tannier nous a fourni. Grâce à la fonction lemma.strip de Spacy, tous les verbes sont mis à l'infinitif et les noms au singulier. Tous les caractères sont aussi mis en minuscule. Cette normalisation permet de mieux analyser l'utilsation des différents termes.

5.2 Sur-représentations

Nous essaierons dans cette partie de faire ressortir les mots-clés les sur-représentés dans certains textes, c'est à dire des termes qui apparaissent particulièrement dans un document par rapport aux autres. Cette analyse pourrra être utile pour comprendre le sujet abordé dans un corpus, ainsi que les apétences des différents groupes et auteurs pour certains termes.

5.2.1 Le TF-IDF

Le TF-IDF [4] (de l'anglais Term Frequency-Inverse document frequency) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Cet algorithme est très pertinent pour notre projet car il permet de faire resortir du corpus (ensemble de texte), les mots les plus "uniques".On peut décomposer cet algorithme en 2 partie

TF (*term-frequency*) : Comme son nom l'indique, le but est de calculer la fréquence des termes dans le corpus. Pour cela, il existe differente variante pour calculer la fréquence des termes (Nombres d'occurences, calcul binaire, normalisation, ect). Pour notre projet, nous avons choisi d'utiliser la méthode de normalisation.
Nous calculons le nombre d'occurences que l'on pondère par le nombre de mots dans les corpus. Nous avons choisi cette méthode car les corpus sont de différentes longueurs et certains thèmes sont plus récurrents que d'autres, nous voulions donc minimiser au maximum le risque d'avoir de trop grand écart entre les différents TF calculés.

$$TF[mot] = \frac{OC}{MC}$$

OC : Nombre d'occurence du mot dans le corpus
MC : Nombre de mots dans le corpus

IDF (*inverse document frequency*) : L'IDF est la fréquence inverse de la présence du mot dans le corpus (ensemble de textes). Son but est de donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

$$IDF[mot] = Log(\frac{|D|}{T})$$

|D| :Nombre total de documents dans le corpus
T :Nombre de textes où le mot apparait

date_question	agriculture	baisse	cci	chambre	finance	ministre	public	ressource
2017-09-26	0.0	0.000000	0.000000	0.126452	0.049407	0.041417	0.000000	0.118631
2017-10-03	0.0	0.028963	0.086888	0.144070	0.120624	0.040447	0.000000	0.057926
2017-10-10	0.0	0.000000	0.129217	0.068868	0.080724	0.022557	0.028914	0.096913
2017-10-17	0.0	0.287049	0.208763	0.092719	0.108681	0.054664	0.023357	0.104382
2017-10-24	0.0	0.044339	0.266033	0.115529	0.086175	0.030960	0.039686	0.103457

Figure 17: Score du TF-IDF dans le thème *chambres consulaires*

Les scores de TF-IDF 17 de tous les mots sont donnés avec une granularité de 1 semaine. Ils peuvent prendre des valeurs comprises entre 0 et 1 (ceci est dû à la pondération du TF). Si le score d'un mot est proches de 1, cela signifie que ce mot est "unique" dans l'ensemble des textes. Pour notre analyse, nous ne nous attardons pas tant sur la valeur du score mais plutôt sur la valeur par rapport aux autres. C'est pour cela que nous classons les 5 mots ayant le plus grands scores de TF-IDF (voir Figure 18).

date_question	0	1	2	3	4
2017-09-26	amputer	industrie	participe	152	487
2017-10-03	péréquation	rural	situer	classer	commune
2017-10-10	réseau	entreprise	chef	évoquer	côté
2017-10-17	baisse	diminution	cci	tfc	pourcent
2017-10-24	taux	cci	région	projet	commerce

Figure 18: Mots ayant les plus grands scores de TF-IDF dans le thème *chambres consulaires*

Pour la semaine du 26-09-2019, le mot "amputer" est le mot qui ressort le plus, on peut donc avoir des informations sur le sujet abordé lors de cette période. Un des objectifs avec le TF-IDF est de pouvoir faire resortir des mots afin de pouvoir utiliser ces mots dans la matrice de co-occurrence que l'on expliquera dans la partie suivante.

5.2.2 Application aux anomalies

L'objectif est à présent de se servir des scores du TF-IDF pour tenter de mieux repérer le sujet associé à un pic de questions dans une certaine rubrique, comme on peut voir dans la Figure 14. Une fois la période du pic obtenue avec les méthodes présentées dans la partie 4.2, nous nous attardons sur les scores de TF-IDF des mots des lettres envoyées au cours de cette période. Afin de visualiser plus facilement la diversité et l'importance de ces mots, nous avons opté pour une représentation en nuage de mots. Plus un mot a un score de TF-IDF élevé, plus il est affiché en grand. Dans la figure 19, nous avons affiché le nuage des mots sur-représentés dans les textes des questions envoyées au plus gros pic repéré. On peut y remarquer principalement les mots **80** et **kilomètre**, en lien avec le passage aux 80 km/h.

Figure 19: Mots les plus sur-représentés dans un pic de questions dans le thème *sécurité routière*

Un autre exemple sur un thème qui a représenté un pic significatif dans l'étude 4.2, les *chambres consulaires*. Nous avons pu appliquer le tf-idf à la fois aux textes des questions et aux titres de questions, ce qui nous a donné, avec la méthode précédemment décrite, les nuages de mots suivants de la figure 21. On remarque que le sujet de la "baisse des ressources des chambres de commerce et d'industrie (CCI) (que nous avons confirmé en lisant plus en détail les lettres concernées) est clairement plus lisible dans l'analyse par les titres.

Afin de tenter de comprendre les sujets abordés lors de chaque pic d'une rubrique choisie, nous avons proposé de réunir un grand nombre d'informations dans un graphique comme le 20:

- Évolution du nombre de lettres (en bleu)
- Moyenne du nombre de lettres par semaine (en rouge)
- Principaux pics repérés (en vert) : tous les pics qui dépassent la moyenne d'une valeur seuil fixée à l'écart type multipliée par une valeur arbitraire (ici 0.03)
- Période des pics (en noir, en bas) : la période associée à chaque pic selon le mode de calcul présenté dans la partie 4

Dans le tableau, on a, pour chaque période de pic, les mots sur-représentés par rapport à d'habitude dans les questions et dans le titres, ainsi que la somme des lettres envoyées. On peut ainsi suivre l'évolution des thèmes abordés au cours du mandat.

	date du pic	Sur-représentés dans les questions	Sur-représentés dans les titres	Somme des lettres durant la période du pic
0	2017-08-08	ichn, aide, retard, paiement, versement	retard, ichn, pac, état, aide	47.0
1	2017-10-31	miel, pays, label, acheter, origine	miel, traçabilité, pays, origine, label	85.0
2	2018-06-26	cuivre, palme, huile, programme, agriculture	européen, cuivre, commander, toxicité, conclusion	66.0
3	2019-07-16	royaume, uni, produit, 44, tartrique	44, article, egalim, loi, application	77.0
4	2018-07-31	apiculteur, sinistré, agricole, hiver, aquitaine	surmortalité, abeille, exonération, occasionne...	44.0
5	2019-05-28	fongique, maladie, em, ceranae, nosema	fongique, apicole, maladie, filière, abeille	13.0
6	2019-09-10	chambre, agriculture, taxe, réseau, proximité	chambre, agriculture, diminution, épandage, fi...	13.0

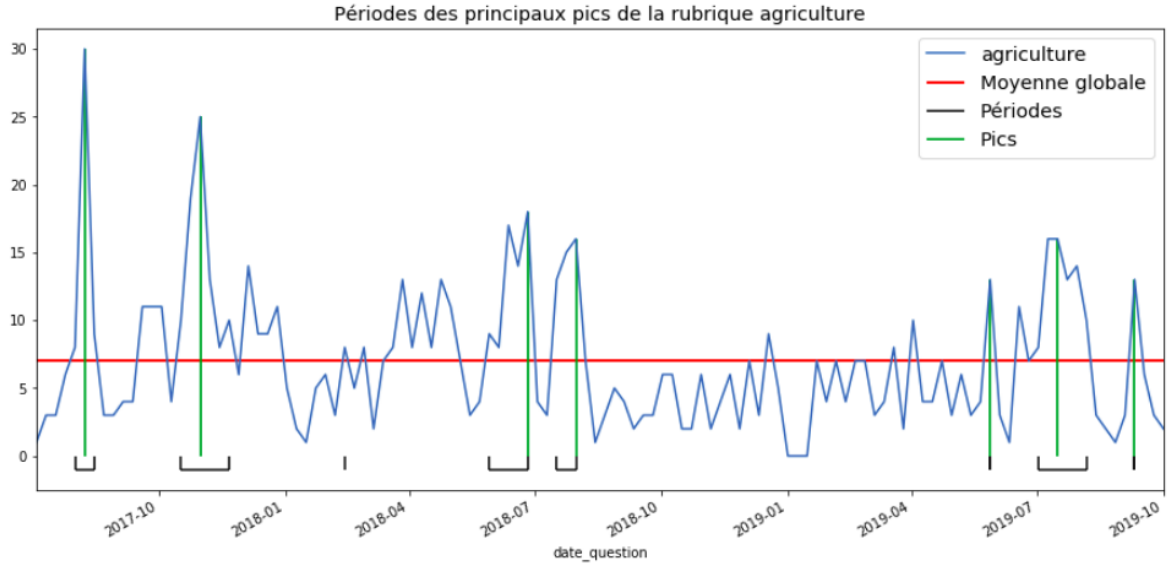


Figure 20: Période des pics dans la rubrique *santé*

Nous avons aussi observé quelle était l'occurrence de ces mots sur une période plus longue, afin de vérifier si il y avait vraiment une sur-représentation à la période en question. Afin de nous représenter de la manière la plus complète possible cette évolution, nous avons affiché 4 types de graphiques juxtaposés, montrés sur la Figure 22.

- L'occurrence brute des mots.
- L'occurrence des mots pondérés par le nombre total de mots écrits cette semaine là, afin de corriger les trop grandes apparitions en cas d'un envoi massif de lettres, ou l'inverse.
- Le score attribué par le tf-idf à chacun des mots pour chaque semaine, afin de vérifier sa cohérence avec les deux graphiques du haut.
- Le nombre total de questions envoyées par semaine dans le thème que nous analysons.

La période du pic repérée est délimitée par des traits noirs.

Nous remarquons sur le graphique une certaine cohérence entre les occurrences, leur proportions et les scores du tf-idf. Cependant, on peut observer étrangement que certains mots, comme ceci, ont des scores très élevés pendant de longues périodes, ce qui ne devrait pas être le cas. Un approfondissement pour écrire notre propre tf-idf et choisir ses caractéristiques sera sûrement nécessaire pour mieux comprendre et peut-être corriger cette observation.

Figure 21: Dans les textes (gauche) et les titres (droite) des questions d'un pic de *chambres consulaires*

Un autre format pour observer l'évolution des occurrences des mots clés est un graphique empilé (Figure 23 et 24), qui permet de se rendre compte facilement de l'importance ponctuelle des différents mots. On peut le représenter en relatif et en absolu, deux façons complémentaires de se représenter la situation.

5.2.3 Comparaison par partis

Nous cherchons ici à comparer les groupes politiques à travers le lexique qu'ils utilisent dans leurs lettres.

Dans un premier temps nous avons recherché sur une période données (en général sur une anomalie) quels étaient les mots singuliers utilisés par les différents groupes politiques. Nous avons pour cela concaténé les lettres en texte regroupant une semaine entière de lettre aux ministres et ceux pour chaque groupe. En appliquant le tf-idf à l'ensemble de ces textes pour un parti on peut voir quelle sont les mots intéressants sur l'anomalie observée. On peut ensuite les repésenter sous forme de word cloud (Figure 25) afin de faire ressortir les plus importants.

On peut ici voir que sur une période donnée (Figure 26), LAREM utilise particulièrement les mots "santé" ou "gynécologue" quand les Socialistes emploient les mots "pratique" et "infirmier".

On peut, dans un second temps, concaténer l'ensemble des lettres d'un parti et d'appliquer un tf-idf aux textes ainsi créés. Cela permet de faire ressortir les mots singuliers les plus utilisés par chacun des partis.

En affichant le résultat sous forme de word cloud on peut par exemple voir que les mots "ne", "se" ou "ministre" sont fortement représentés dans la plupart des groupes politiques ce qui semble incohérent avec le principe du tf-idf. En effet cet outil est supposé relever les mots singuliers à chaque partis, un mot ne peut donc pas avoir un tf-idf élevé dans chaque groupe. Cela nous a amener à remettre en question l'utilisation de la librairie pour le calcul du tf-idf car nous n'avions pas de contrôle sur la formule exacte utilisée. Il s'est finalement avéré que l'algorithme était très peu efficace s'il y avait peu de documents.

Figure 22: Évolution détaillée de la représentation certains mots dans la rubrique *chambres consulaires*

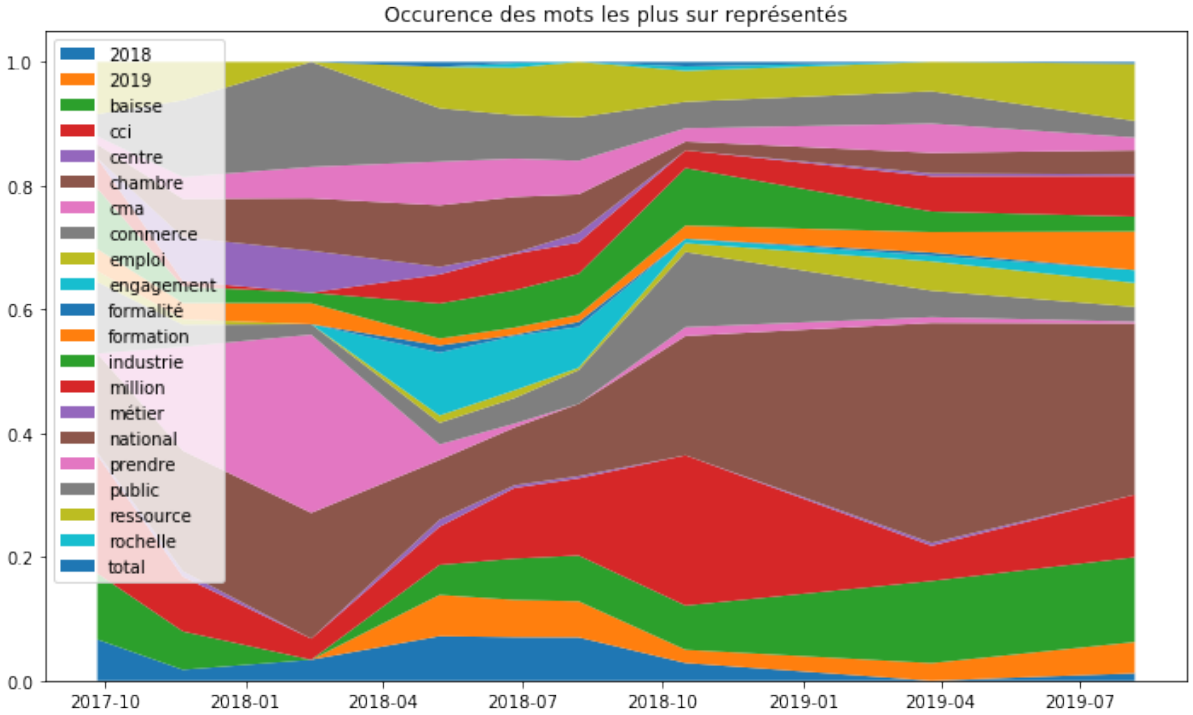


Figure 23: Représentation normalisée de certains mots dans la rubrique *chambres consulaires*

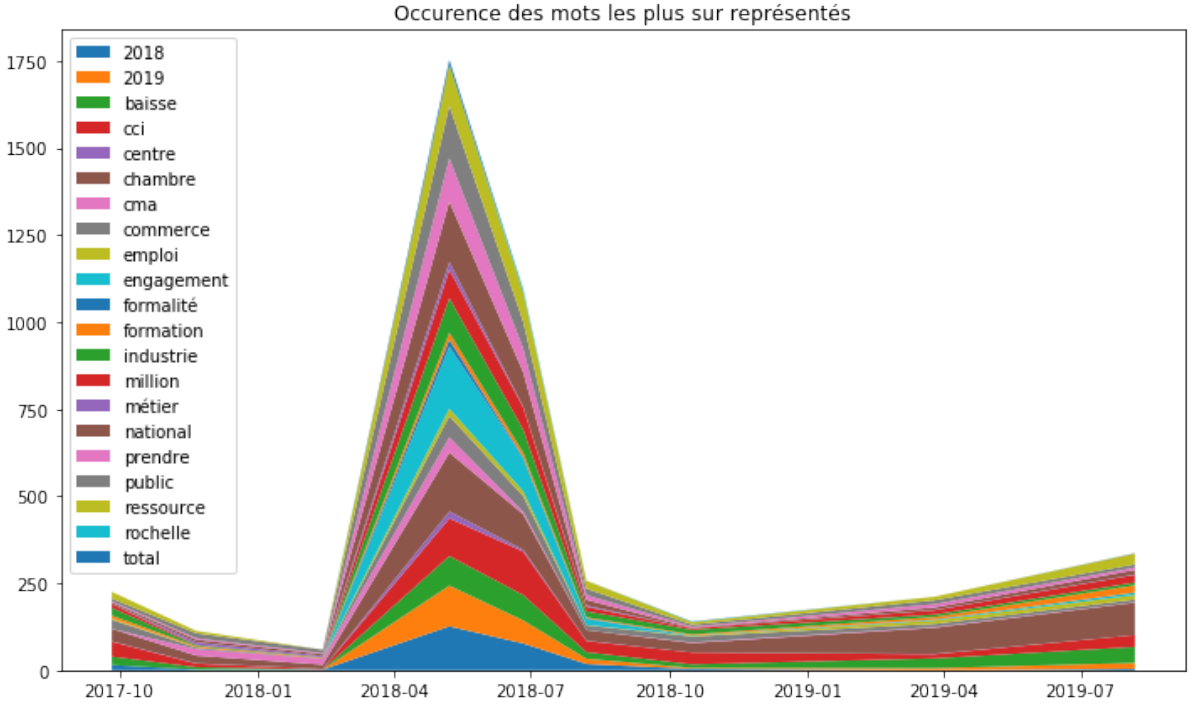


Figure 24: Représentation de certains mots dans la rubrique *chambres consulaires*

Figure 25: Mots les plus sur-représentés dans la rubrique Santé pour une période selon les partis

Figure 26: Mots les plus sur-représentés dans chaque parti pour la rubrique Santé

5.3 Co-occurrences

Nous nous appuyons sur l'idée que l'intensité avec laquelle certains mots sont associés dans un texte, peut être révélateur de la vision et des arguments de celui-ci.

Exemple de vision: Une vision sécuritaire d'une manifestation associerait plus facilement ce terme avec danger, ou dégradation, etc., tandis qu'un soutien l'associerait plus avec revendication, justice, etc.

Exemple d'argument: En regardant les associations des mots nucléaire et sécurité, les présences des mots norme et respect seraient ici révélatrices.

Nous verrons si cette manière d'analyser un corpus pourra permettre de distinguer différentes façon de penser, de faire des regroupements, et surtout de repérer des influences par la présence de liens similaires dans le vocabulaire.

Une matrice de co-occurrence de mots, associe chaque lien entre deux mots avec la façon dont ils sont liés.

Les exemples des différentes matrices de cooccurrence ci-dessous mettent en lien des mots arbitrairement choisis dans l'ensemble des lettres écrites dans la rubrique ordre public dans le mandat de 2017.

5.3.1 Co-occurrence binaire

Dans une matrice de co-occurrence binaire, on compte le nombre de fois où les deux mots sont apparus dans le même texte, au sein d'un ensemble de textes (Figure 27).

	blessé	gilet jaune	justice	manifestation	ordre	police	social	syndicat	sécurité
blessé	0	6	5	8	9	4	1	0	3
gilet jaune	6	0	6	17	16	8	5	1	6
justice	5	6	0	5	9	5	4	2	6
manifestation	8	17	5	0	37	23	9	1	18
ordre	9	16	9	37	0	29	12	2	33
police	4	8	5	23	29	0	5	1	24
social	1	5	4	9	12	5	0	2	14
syndicat	0	1	2	1	2	1	2	0	4
sécurité	3	6	6	18	33	24	14	4	0

Figure 27: Matrice de co-occurence binaire pour une liste de mot donnée

5.3.2 Co-occurrence multiple

A chaque texte rencontré, on multiplie les occurrences des deux mots. Permet de mettre en valeur les mots qui sont répétés plusieurs fois ensemble dans un même texte (Figure 28).

5.3.3 Co-occurrence en fourchette

Pour chaque texte, on le fait parcourir par un intervalle d'une certaine taille , et à chaque fois qu'on y retrouve deux mots ensemble on ajoute 1 à leur lien. Permet de donner de

	blessé	gilet jaune	justice	manifestation	ordre	police	social	syndicat	sécurité
blessé	0	21	8	57	52	15	4	0	4
gilet jaune	21	0	19	63	83	18	12	2	16
justice	8	19	0	8	19	8	5	2	9
manifestation	57	63	8	0	289	82	21	2	82
ordre	52	83	19	289	0	206	37	5	208
police	15	18	8	82	206	0	6	1	129
social	4	12	5	21	37	6	0	3	30
syndicat	0	2	2	2	5	1	3	0	10
sécurité	4	16	9	82	208	129	30	10	0

Figure 28: Matrice de co-occurence multiple pour une liste de mot donnée

l'importance à la proximité des mots dans un texte. En effet, plus 2 mots sont proches, plus ils se retrouveront ensemble lors du parcours de l'intervalle (Figure 29).

	blessé	gilet jaune	justice	manifestation	ordre	police	social	syndicat	sécurité
blessé	0	10	0	48	23	7	0	0	0
gilet jaune	10	0	14	87	18	5	0	0	7
justice	0	14	0	7	9	17	4	0	7
manifestation	48	87	7	0	166	36	19	0	60
ordre	23	18	9	166	0	60	19	0	79
police	7	5	17	36	60	0	4	8	69
social	0	0	4	19	19	4	0	0	11
syndicat	0	0	0	0	0	8	0	0	9
sécurité	0	7	7	60	79	69	11	9	0

Figure 29: Matrice de co-occurence en fourchette pour une liste de mot donnée

5.3.4 Co-occurrence par phrase

On compatbilise cette fois les points pour le lien entre deux mots à chaque fois qu'ils sont rencontrés ensemble dans une phrase. Cette méthode est censé permettre de rapprocher les mots qui ont un lien sémantique, Figure 30.

5.3.5 Utilisation

La principale difficulté de l'utilisation de ces matrices réside dans le choix des mots pertinents à montrer ensemble. Il y a la possibilité qu'un expert ou on curieux les choisisse arbitrairement. Nous avons tenté par la suite de détecter automatiquement les mots qui semblent les plus pertinents de considérer les liens grâce à leur sur-représentation calculée par le TF-IDF. Cependant, cette approche nous oblige à un certain nombre de choix arbitraires,

	blessé	gilet jaune	justice	manifestation	ordre	police	social	syndicat	sécurité
blessé	0	0	2	7	8	3	0	0	5
gilet jaune	0	0	0	0	0	0	0	0	0
justice	2	0	0	3	18	31	118	5	53
manifestation	7	0	3	0	40	9	7	2	32
ordre	8	0	18	40	0	62	60	11	106
police	3	0	31	9	62	0	18	6	177
social	0	0	118	7	60	18	0	43	1517
syndicat	0	0	5	2	11	6	43	0	10
sécurité	5	0	53	32	106	177	1517	10	0

Figure 30: Matrice de co-occurence par phrase pour une liste de mot donnée

sur les documents à considérer ou les méthodes de TF-IDF, et nous avons eu du mal à nous décider sur un algorithme en particulier, même après de nombreux essais. Nous avons donc décidé de ne pas développer cet aspect en priorité, mais de le laisser de côté pour le réutiliser au cas où il semble pertinent au cours de notre étude.

6 Rencontre avec Proches

Notre démarche a jusque là été très exploratoire, et nous avons essayé de multiples approches pour mettre en valeur les données des lettres écrites. Cependant, nous arrivons à un point où les résultats que nous obtenons sont difficiles à évaluer pour des non-initiés aux rouages de l'assemblée nationnale, et il serait imprudent de continuer nos recherches sans savoir quelles approches sont les plus pertinentes dans le cadre de l'étude. C'est pourquoi la réunion de Janvier avec l'entreprise Proches, au cours de laquelle nous avons exposé nos méthodes et résultats, nous a permis de mieux comprendre le contexte et de nous orienter selon les conseils de l'expert en communication d'inflence.

Globalement, tous les types de résultats que nous avons présenté ont beaucoup intéressé l'agence et particulièrement la possibilité de jouer avec les données avec les algorithmes que nous avons développé. Mettre à disposition cet outil à tout type de public est donc un des principaux aspects que nous avons convenu de développer, dans la partie 8.3. Notre travail sur le repérage des anomalies et des singluarités leur a paru une approche intéressante nous l'avons donc approfondit dans la partie 7.1. Armand nous a aussi fait part du besoin de comprendre la façon dont un sujet particulier a été traité par les différents groupes politiques. Nous avons développé cette approche dans la partie 7.5.

7 Observer les singularités

Pour plusieurs aspects de notre étude, nous avons eu besoin de nous confronter au problème suivant : **comment quantifier, repérer et visualiser des singularités qui pourraient nous intéresser ?** Par exemple, dans le cas de la répartition des lettres envoyées par les partis dans les différentes rubriques, nous aimerions pouvoir savoir si un parti s'est particulièrement investit, ou l'inverse, dans une rubrique particulière. Nous sommes cofrontés à plusieurs difficultés :

- **Quantification** Comment mesurer que la valeur est particulièrement élevée ? *Ex: qu'un parti s'est particulièrement investit dans une rubrique ?*
- **Seuil de singularité :** À partir de quelle seuil peut-on dire qu'une valeur est anormale ? *Ex: qu'un parti s'est anormalement investit dans une rubrique ?*
- **Visualisation :** Comment représenter et comparer les différences de singluarité, dans des rubriques souvent très nombreuses ? *Ex: comparer les rubriques abordées par un parti*
- **Variation de taille :** Comment un score de singularité pourrait prendre en compte les variations de tailles dans le calcul ? *Ex: Les gros partis ont plus de chance d'avoir beaucoup de lettres dans une rubrique*

Nous avons expérimenté plusieurs méthodes pour répondre à ces problématiques. Nous avons commencé par tenter de pondérer les valeurs par les tailles pour résoudre le problème 1 et 4, mais subsistait l'absence d'un seuil de singularité compréhensible et le fait que les valeurs proches de 0 ne pouvaient pas être prises en compte. La visualisation des résultats de sous représentation était de plus très délicate.

Puis, l'approche des tests statistiques s'est révélée extrêmement efficace, et nous avons choisi de l'approfondir.

7.1 Tests statistiques

L'idée qui nous a amené aux tests statistiques [5] est de voir la singularité comme une probabilité d'obtenir un tel résultat. Nous voulons qu'une probabilité faible indique qu'une valeur est particulièrement faible ou élevée. *Ex: Un parti a très peu de chances d'envoyer ce nombre de lettres, par rapport à ce ce qu'on pourrait attendre de lui a priori.* L'idée est donc de considérer une hypothèse a priori H_0 sur les données et d'en déduire les probabilités selon cette hypothèse des différentes valeurs disponibles, et ainsi de pouvoir repérer facilement les valeurs qui s'en écartent le plus. Comme nous considérons dans cette étude des lois normales, le score de singularité sera en fait la p-value de la valeur observée selon l'hypothèse observée. Cette méthode résout toutes les difficultés rencontrées précédemment :

- **Quantification** Le score est une probabilité facilement interprétable.
- **Seuil de singularité :** Il peut être choisi arbitrairement à un niveau α : les quantités singulières sont celles qui ont une p-value inférieure à α.
- **Visualisation :** On peut comparer les score sur un même graphique.

- **Variation de taille :** La taille de l'échantillon est prise en compte dans l'hypothèse nulle et donc dans le score de singularité.

Nous avons détaillé les calculs de cette approche dans la partie suivante, qui en est une application au problème de la répartition des questions d'un parti dans les différentes rubriques.

7.2 Répartition des rubriques pour un parti

L'objectif est d'analyser la manière dont les différentes rubriques sont utilisées par un parti. L'idée est d'observer en quelle quantité un parti envoie envoie ses lettres dans les différentes rubriques, et d'en déduire si cette quantité est anormalement élevée ou faible pour chaque rubrique. Pour procéder à cette mesure, nous avons appliqué la démarche d'un test statistique basé sur une hypothèse nulle d'homogénéité, pour mieux repérer les valeurs qui s'en écartent.

7.2.1 Hypothèse nulle

Nous nous plaçons dans l'hypothèse H_0 : *Chaque rubrique est utilisée de la même manière par tous les partis.* Autrement dit, chaque parti est supposé avoir la même probabilité que les autres d'envoyer une lettre dans une rubrique, et celle-ci correspond donc à la moyenne d'utilisation de la rubrique. Si on a n_r lettres envoyées dans la rubrique r sur un total de n lettre, alors chaque groupe g est censé avoir $p_{g,r} = p_r = \frac{n_r}{n}$ chances d'envoyer une lettre dans la rubrique r .

7.2.2 Loi binomiale

Plus précisemment, si on considère l'évènement : *Le parti r envoie sa lettre dans la rubrique g*, elle suit selon notre hyptothèse H_0 une loi de Bernouilli avec une probabilité p_r. On répète cette expérience à chaque tirage d'une lettre du groupe g, soit n_g fois en tout, et le nombre de lettres $v_{g,r}$ envoyées par le groupe g dans la rubrique r suit ainsi une loi binomiale $\mathcal{B}(p_r, n_g)$:

$$\mathbb{P}(v_{g,r} = k) = \binom{n_g}{k} p_r^k (1 - p_r)^{n_g - k} \tag{4}$$

7.2.3 P-value

La p-value de $v_{g,r}$ est la probabilité selon H_0 d'obtenir une valeur plus extrême que $v_{g,r}$. Pour l'assimiler à un score de singularité, on calcule donc une p-value unilatérale selon les deux extrêmes :

- **Cas Supérieur :** Si on a une valeur supérieure à l'espérance, ce sera la probabilité d'envoyer $v_{g,r}$ ou plus de lettres dans cettre rubrique.
- **Cas Inférieur :** Sinon, ce sera la probabilité d'envoyer $v_{g,r}$ ou moins de lettres dans cettre rubrique.

. Pour calculer ces scores, il suffit de sommer les probabilités $\mathbb{P}(v = k)$ pour tous les v plus extrêmes que $v_{g,r}$.

7.2.4 Exemple pour la loi binomiale

Intéressons-nous par exemple à la rubrique *environnement*. Nous allons comparer le score de singularité de deux groupes, *FI* et *LR*. Il y a $n_r = 518$ lettres envoyées dans cette rubrique sur un total de $n = 23365$, c'est à dire une probabilité moyenne $p_r = \frac{518}{23365} = 0.0044$ d'y envoyer une lettre.

Pour FI : A envoyé en tout $n_{FI} = 1023$ lettres en tout, la loi selon H_0 suit donc une $\mathcal{B}(0.0044, 1023)$, dont l'espérance est ≈ 4. *FI* a envoyé 8 lettres dans cette rubrique. Son score de singularité est donc calculé en sommant des probabilités des valeurs supérieures ou égales à 8 (en rouge sur le graphique 31), $= 0.05$.

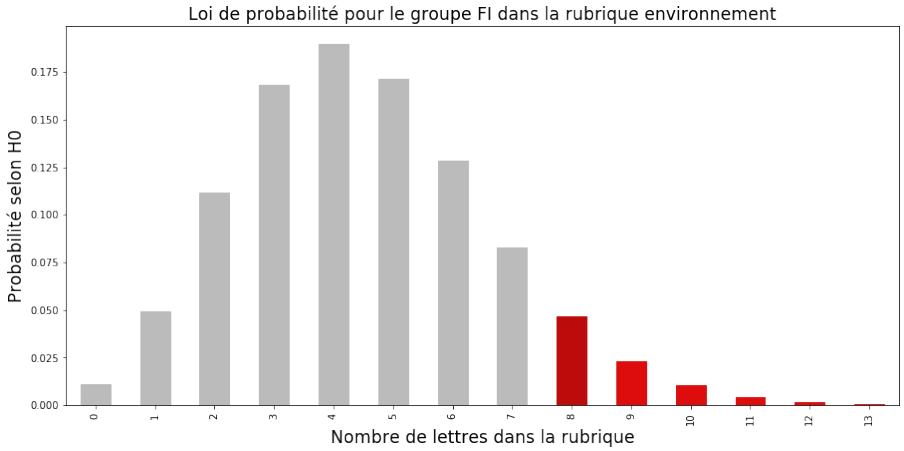


Figure 31: Calcul du score de singularité pour FI dans la rubrique environement

Pour LR : A envoyé en tout $n_{LR} = 6635$ lettres en tout, la loi selon H_0 suit donc une $\mathcal{B}(0.0044, 6635)$, dont l'espérance est ≈ 30. *LR* a envoyé 25 lettres dans cette rubrique. Son score de singularité est donc calculé en sommant des probabilités des valeurs inférieures ou égales à 25 (en bleu sur le graphique 32), $= 0.22$.

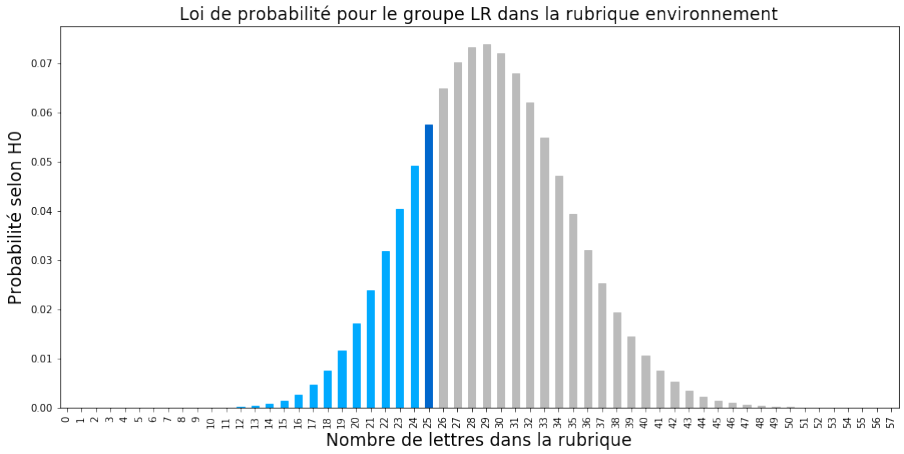


Figure 32: Calcul du score de singularité pour LR dans la rubrique environement

Remarque : Les probabilités obtenues sont dans ces exemples assez élevées, c'est à dire que leur singularité est faible, surtout pour *LR*. On trouve habituellement des valeurs qui peuvent aller jusqu'à 10^{-5}, car l'hypothèse H_0 est complétement fausse, elle ne nous sert qu'à repérer les valeurs qui s'en écartent le plus.

Ce score de singularité trouve donc une solution au problème des différences entre le nombre de lettres envoyées par les différents partis, car ce nombre est complètement pris en compte dans le calcul de la probabilité. Le score obtenu a de plus un sens très précis et est facile à intérpréter.

7.2.5 Loi normale

Sachant que dans notre étude, le nombre de lettres est assez important, on peut aussi approcher la loi binomiale avec une loi normale d'espérance np_r et de variance , $np_r(1-p_r)$ lorsque $np_r > 9$:

$$\mathcal{B}(p_r, n) \sim \mathcal{N}(n_g p_r, n_g p_r(1-p_r)) \tag{5}$$

Cette approche facilite le calcul de la pvalue pour des nombre de lettres élevés. On se ramène à une loi normale centrée réduite :

$$\frac{v_{g,r} - n_g p_r}{\sqrt{n_g p_r(1-p_r)}} \sim \mathcal{N}(0,1) \tag{6}$$

On peut en déduire, avec les tables de correspondance ou avec les fonctions déjà implementées d'intégration de la loi normale, la p-value de la variable $v_{g,r}$ sous l'hypoyhèse H_0.

Exemple pour LAREM : A envoyé en tout $n_{LAREM} = 8910$ lettres, la loi selon H_0 suit donc une $\mathcal{N}(E, \sigma^2)$ avec l'espérance $E = n_{LAREM} p_r \approx 139$ et une variance $\sigma^2 = n_{LAREM} p_r(1-p_r) = 139$. *LAREM* a envoyé 145 lettres dans cette rubrique. Son score de singularité est donc calculé en intégrant la loi de 145 jusqu'à l'infini (en jaune sur le graphique 33), $= 0.25$.

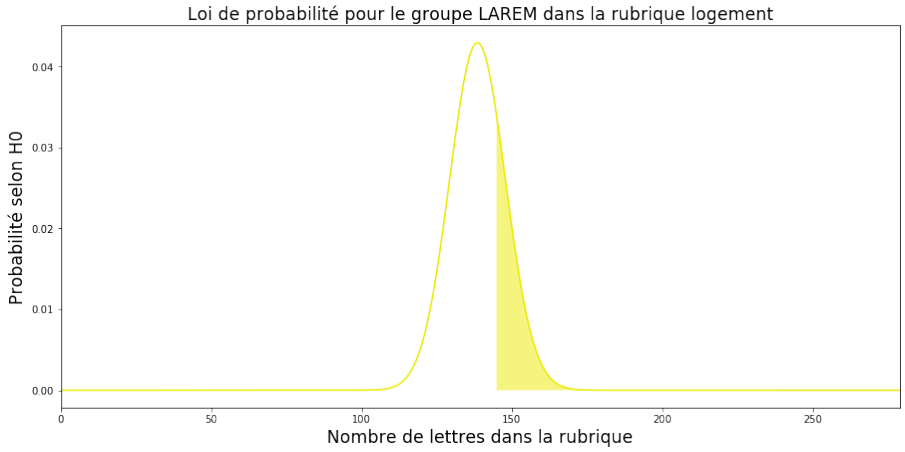


Figure 33: Calcul du score de singularité pour LAREM dans la rubrique logement

7.2.6 Repérage et visualisation

La mesure de la p-value est très utile pour repérer les valeurs extrêmes à un niveau α, c'est à dire lorsque la p-value descend en dessous d'un seuil $\alpha << 1$. En revanche, comme dit plus haut, les p-values les plus intéressantes ont une valeur extrêmement faible, et il est assez difficile de se représenter ce qu'elle signifie. Elles sont parfois si faibles qu'un objet float 64 la confond avec 0, et leur visualisation n'apporte que peu interprétation intéressante (figure 34).

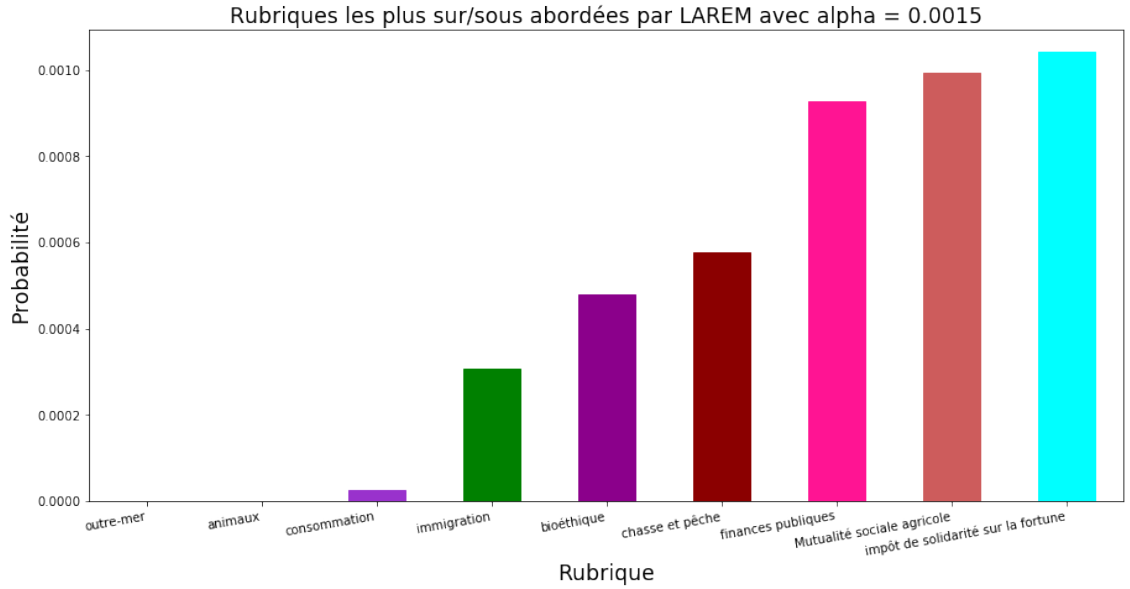


Figure 34: P-value des rubriques ayant la singularité la plus élevée chez LAREM

Une autre méthode de visualisation consiste à comparer la quantité obtenue à la quantité attendue, qui est l'espérance selon l'hypoyhèse H_0. La comparaison ainsi obtenue est beaucoup plus facilement interprétable par un non-initié (figure 35).

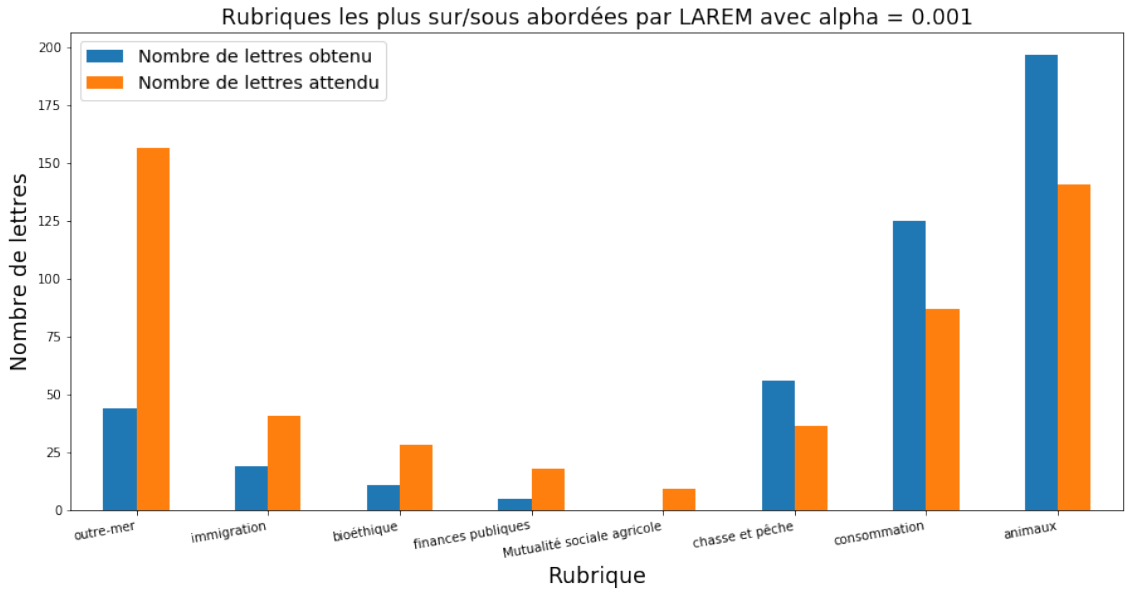


Figure 35: Comparaison des rubriques ayant la singularité la plus élevée chez LAREM

Nous pouvons aussi remarquer que les comparaisons entre p-values sont similaires aux comparaisons entre les variables centrées réduites des valeurs obtenues. La visualisation de ces variables offrent un score qui a l'avantage de présenter la sur-représentation en positif et la sous-représentation en négatif (figure 36).

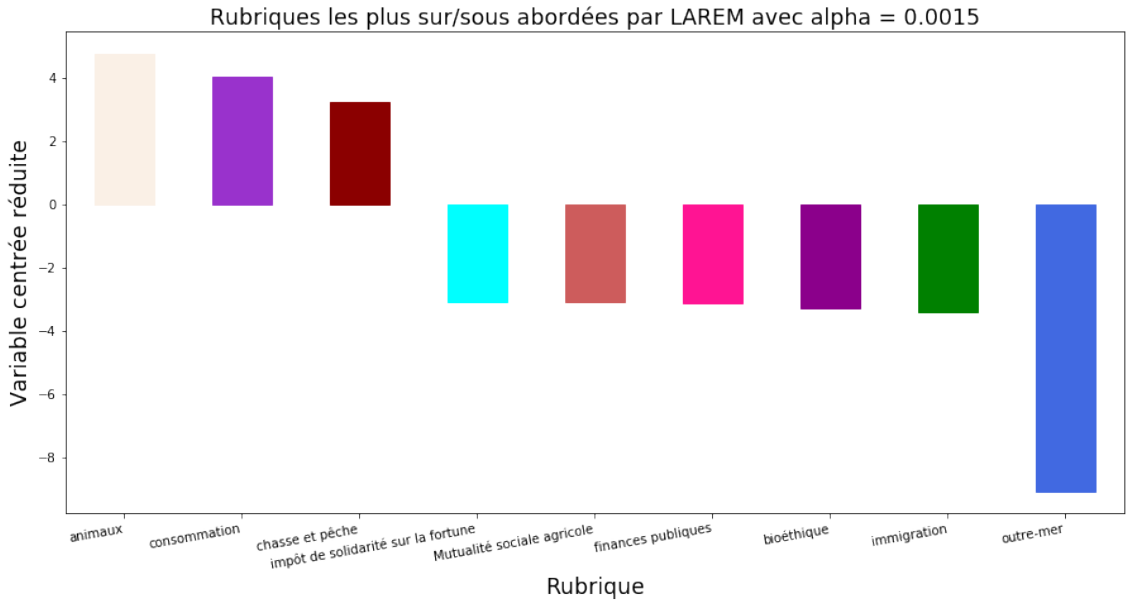


Figure 36: Score des rubriques ayant la singularité la plus élevée chez LAREM

7.3 Remarques

Sur le principe : Cette approche de partir d'une hypothèse nulle qui considère que les variables sont réparties de manière homogène, pour ensuite détecter et visualiser celles qui s'écartent le plus de cette hypothèse, répond ainsi parfaitement à notre problème. Nous avons donc utilisé ce principe à plusieurs reprises.

P-value ou Variable centrée réduite ? Utiliser la variable centrée réduite est peut-être finalement plus pertinent que la p-value. En effet, cette dernière va avoir tendance à prendre des valeurs extrêmement faibles lorsqu'on la mesure pour un groupe qui envoie un très grand nombre de lettres. Pour cause, selon l'hypothèse H_0, représentée par une loi normale, plus on écrit de lettres, plus il est improbable de s'éloigner de l'espérance. Si c'est très pratique pour prendre en compte les petites valeurs, la détection de valeurs extrêmes de p-value aura forcément plus de chance de repérer des groupes qui envoient beaucoup de lettres de manière générale. La variable centrée réduite, en revanche, s'adapte parfaitement pusiqu'elle pondère la différence à l'espérance par l'écart type, sans s'occuper des probabilités.

7.4 Problème dual : répartition des partis dans une rubrique

Si nous voulons à présent nous intéresser à une rubrique en particuier, procédons comme précédemment mais avec le problème dual: quels partis sont anormalement représentés dans une rubrique choisie ?

7.4.1 Hypothèse nulle et lois

Nous faisons donc l'hypothèse H_0 : *Dans une rubrique, la répartition des partis est la même que la moyenne.* Chaque lettre de la rubrique a donc une probabilité $p_{g,r} = p_g = \frac{n_g}{n}$ d'être envoyée par le parti g, avec n_g le nombre de lettres envoyées par g et n le nombre de lettres total. Le nombre de lettre d'un parti dans une rubrique suit donc une loi binomiale $\mathcal{B}(n_r, p_g) \sim \mathcal{N}(n_r p_g, n_r p_g(1 - p_g)$.

On peut remarquer qu'en fait c'est exactement la même hypothèse que dans le problème précédent, et donc exactement la même loi. Simplement, nous nous plaçons à présent dans une rubrique pour comparer les partis plutôt que l'inverse.

7.4.2 Repérage et visualisation

En calculant la p-value avec la loi binomiale ou normale de la même manière que précédemment, nous obtenons le graphique 37. On y interprète que les partis les plus extrêmes dans la rubrique environnement sont FI et NI, mais on ne peut déduire sur ce graphique dans quel sens le sont-ils.

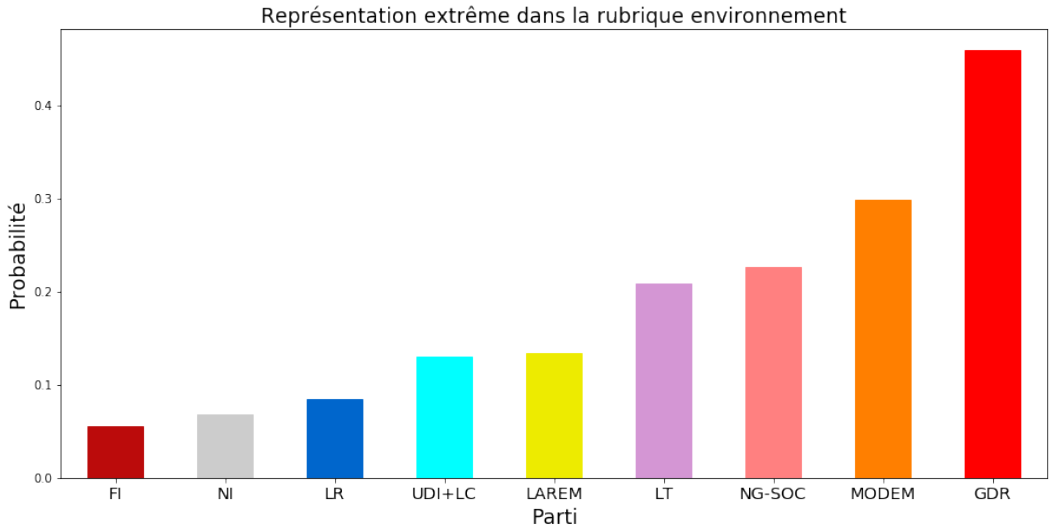


Figure 37: Singularité des partis dans la rubrique environnement

Cependant, la fonctionnalité de détection à un niveau α est rendue inituile par la possibilité de comparer tous les partis sur un même graphique, puisqu'il n'y en a que 9. On peut donc observer et confronter les différents partis sur les graphiques 38 et 39 sans se soucier de la p-value.

Noter aussi que les valeurs de la variable centrée réduite et de la p-value indiquent à quel point la rubrique s'éloigne de l'hypothèse H_0, à quel point elle est inégalement traitée par les partis. Ainsi, on observe sur la rubrique environnement une bien plus gande régularité que

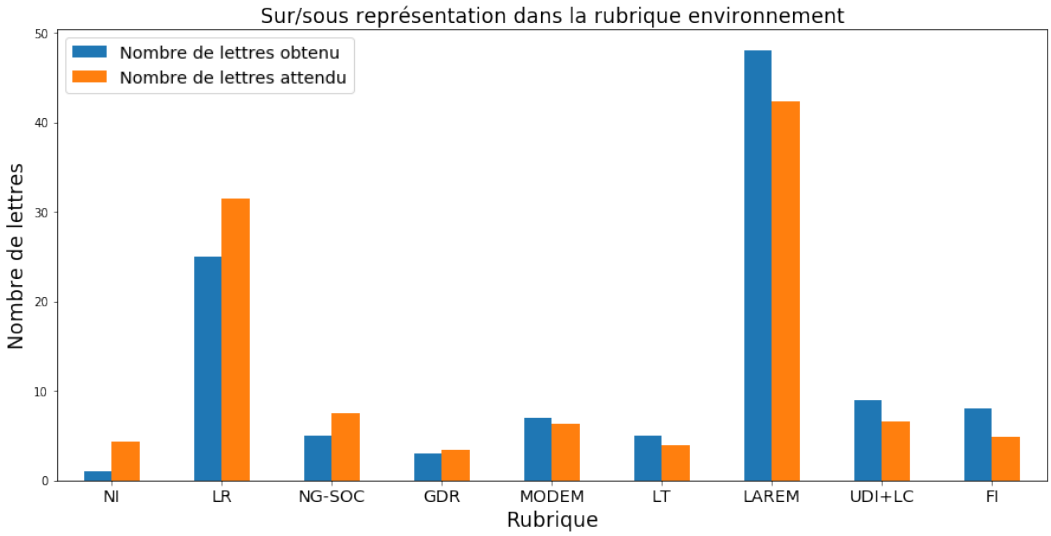


Figure 38: Utilisation de la rubrique environnement par les différents partis

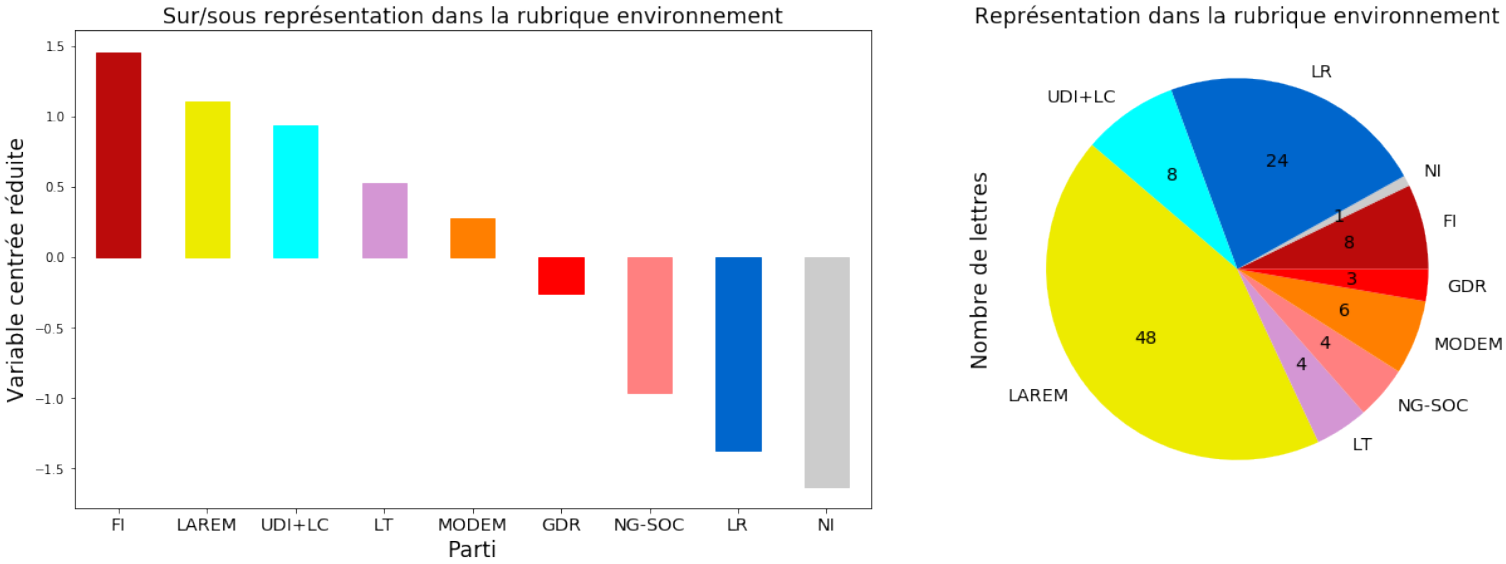
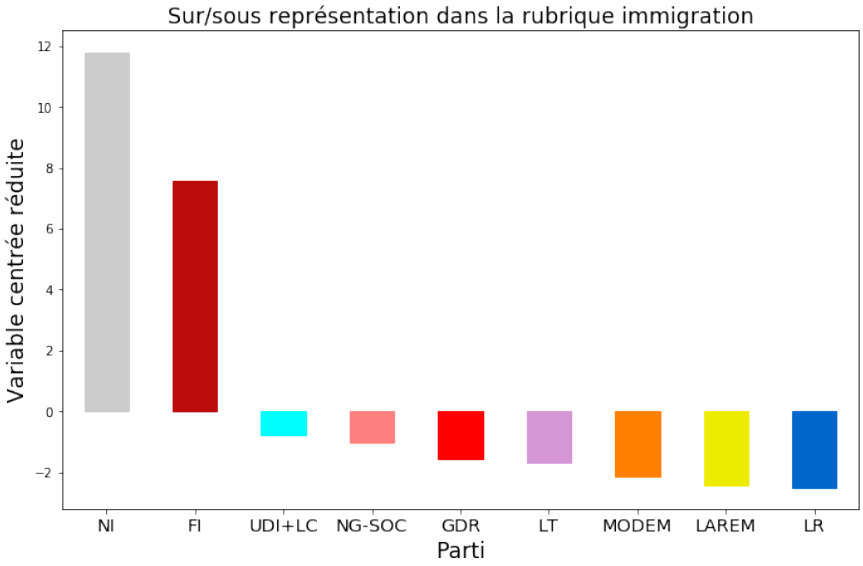


Figure 39: Utilisation de la rubrique environnement par les différents partis

sur la rubrique immigartion, qui est bien plus inégale comme on le voit à ses grandes valeurs du score de singularité des variables centrées réduites (Figure 40).



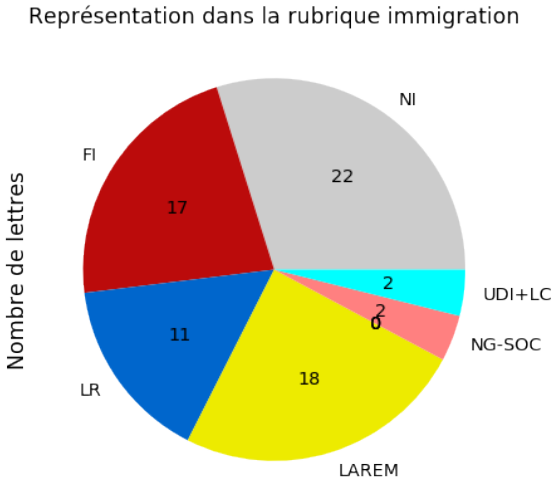


Figure 40: Comparaison de l'utilisation du terme révolution entre les différents groupes

7.5 Utilisation d'une expression par les partis

L'objectif est d'analyser la manière dont les différents partis abrodent un sujet. Une idée est d'observer en quelle quantité un terme choisi est utilisé dans les lettres, et d'en déduire si cette quantité est anormalement élevée ou faible pour chaque parti. Pour procéder à cette mesure, nous avons repris exactement l'aproche expliquée en détail dans la partie précédente (7.2), en redifénissant les objets étudiés.

7.5.1 Hypothèse nulle

L'hypothèse nulle est ici que tous les partis utilisent l'expression choisie *Exp* de la même manière, c'est à dire que H_0 : *Chaque parti a la même probabilité d'utiliser* Exp *dans une lettre*. Ainsi, si n_e lettres contiennent le terme *Exp* sur n au total, chaque groupe g aura une probabilité $p_{g,e} = p_e = \frac{n_e}{n}$ d'écrire une lettre contenant *Exp*.

7.5.2 Loi, Variable et p-value

Ainsi, la probabilité de l'évènement : *Le groupe g utilise le terme Exp dans sa lettre* suit une loi de bernouilli avec un probabilité p_e. Si le groupe a envoyé au total n_g lettres, le nombre de lettres du groupe contenant cette expression doit donc suivre une loi binomiale $\mathcal{B}(p_e, n_g) \sim \mathcal{N}(n_g p_e, n_g p_e(1-p_e)$. On calcule la variable centrée réduite du nombre de lettre $v_{g,e}$ du groupe g utilisant le terme *Exp*:

$$\frac{v_{g,e} - n_g p_e}{\sqrt{n_g p_e (1-p_e)}} \sim \mathcal{N}(0,1) \tag{7}$$

On en déduit la p-value associée à chacune des valeurs $v_{g,e}, g \in G$.

7.5.3 Repérage et visualisation

On applique donc ce calcul à chacun des partis. Le nombre de groupes étant assez faible, une bonne visualisation pour comparer les sur/sous utilisations consiste à tous les confronter sur un même graphique, en affichant la variable centrée réduite de leur valeur en tant que score. Cependant, afin d'éviter de confondre la sur-utilisation d'un terme et son utilisation absolue, nous ajoutons un diagramme circulaire des utilisations brutes de l'expression.

La nuance entre anomalie d'utilisation et utilisation absolue est illustrée sur le graphique 41: même si LAREM et LR sont les partis qui ont le plus utilisé le terme révolution (à droite), c'est en réalité très peu rapporté au nombre total de lettres qu'ils ont envoyé. En revanche FI a beaucoup plus utilisé ce terme que ce qu'on pourrait attendre de son nombre de lettre total (à gauche).

Une autre façon de visualiser cette information est d'afficher, comme dans la section précédente, le nombre d'occurences attendu devant celui obtenu, sur la figure 42. Ce sont des notions plus concrêtes et plus sûres. Cependant, même si on affiche ces deux valeurs dans l'ordre du plus au moins extrême, la comparaison est moins directement visible, et on ne profite pas des couleurs.

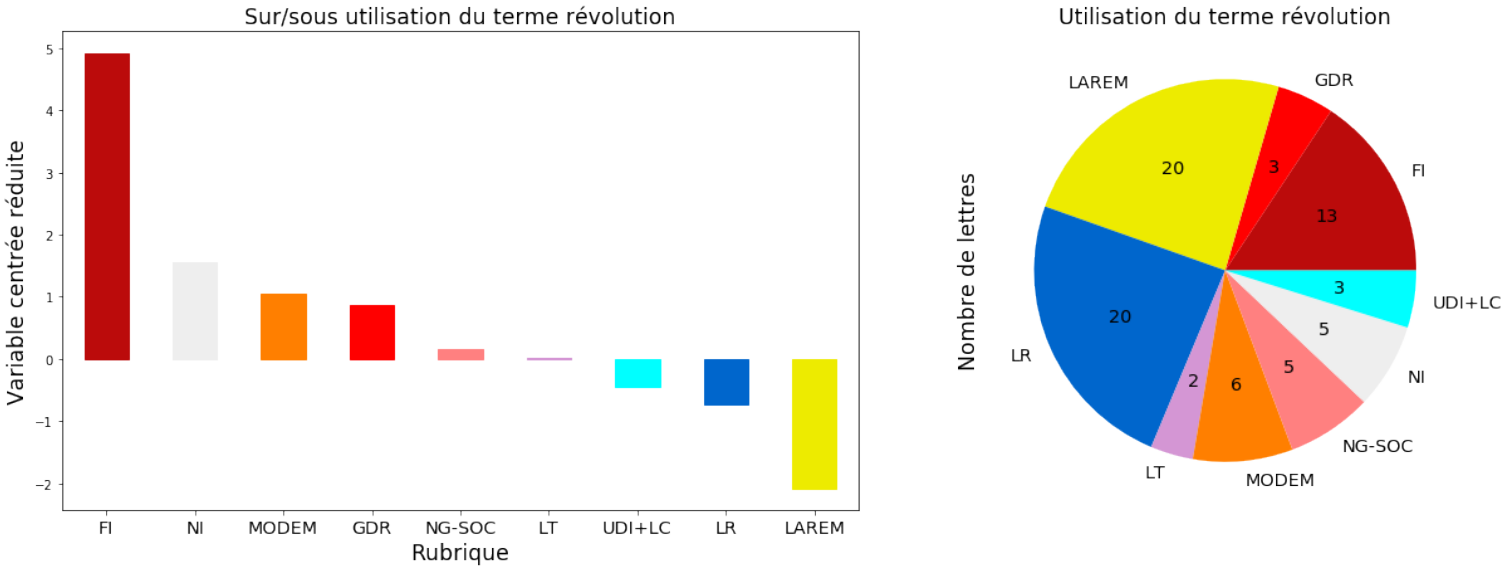


Figure 41: Comparaison de l'utilisation du terme révolution entre les différents groupes

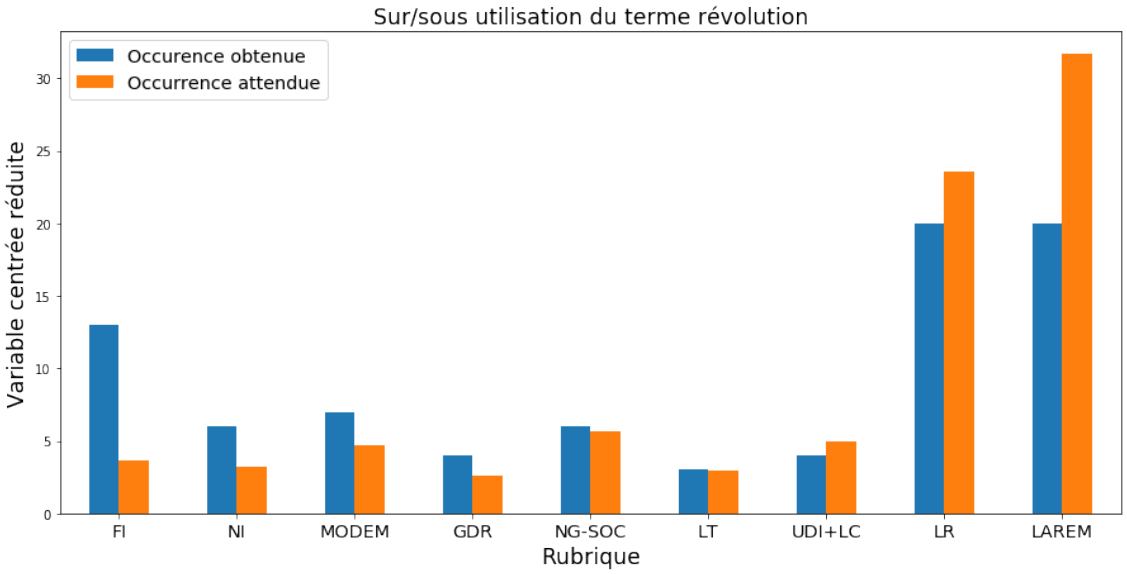


Figure 42: Comparaison de l'utilisation du terme révolution entre les différents groupes

8 Visualisation interactive en ligne

8.1 Les challenges de Proches

L'agence Proches a pour ambition de proposer une analyse des données à différents niveaux accessible à tous. Les challenges consistent dans un premier temps à apporter de la visibilité aux résultats. Ainsi, nous avons décidé lors de notre réunion avec l'agence de rendre disponible les résultats obtenus via un site internet. Ce choix permet cibler une plus grande diversité de personnes qu'avec le dépôt git, et donc de rendre l'étude beaucoup plus accessible.

La visée de ce site est la mise à disposition du grand public comme de l'expert, d'un outil permettant de fournir les résultats de l'analyse des données, initiée par l'utilisateur. Ainsi l'utilisateur doit pouvoir renseigner ses critères de sélection (l'outil de visualisation, la période de temps, etc..), afin d'obtenir un rendu personnalisé et visuel des résultats. L'objectif est que chacun puisse y trouver son compte. Car rappellons qu'un des principes fondateurs de la société est de démocratiser et démystifier l'activité de lobbying. Il est donc primordial dans un premier temps que les gens s'y intéressent, et pour cela chacun doit pouvoir y trouver des résultats qui "lui parlent" personnelement, et qu'il comprend. D'où l'intérêt d'une expérience utilisateur personnalisée, ainsi que d'un rendu visuel explicite, au travers de ce site.

Enfin l'utilisation d'un site internet contribue à l'image de marque de la société. En effet, ce site fera partie de la devanture de l'entreprise. Dans un premier temps il pourra servir aux futurs clients de la société d'avoir un apperçu de l'entreprise et de ses activités. Et dans un second temps, le site permettre d'accroitre la présence de Proches sur le marché de l'analyse lobbying.

Pour résumer, l'utilisation d'un site permet à l'entreprise de démocratiser son activité en rendant les fruits de son activité accessibles à tous. Pour cela il propose une expérience visuelle et personnalisée, qui contribue à son image de marque, et renforce sa présence.

8.2 Outils utilisés

Nous avons utilisé différents outils afin d'implémenter et de déployer le site internet.

- **Streamlit** : Une librairie Python qui permet de créer des applications graphiques déployables assez facilement sur un réseau local avec des scripts Python. Chaque Widget (Bouton, Slider, Barre de recherche, etc.) est considéré comme une variable ce qui facilite la lecture et l'écriture du code. Un des autres avantages est la réutilisation des résultats déjà caclulés, de manière sécurisée. Streamlit introduit une primitive de cache qui se comporte comme une mémoire de données persistante, immuable par défaut, qui permet aux applications Streamlit de réutiliser les informations en toute sécurité et sans effort.

- **Plotly** : Une librairie Python qui fournit des outils de visualisation interactives (zoom, indications au passage de la souris) de données en ligne. Elle fournit des outils graphiques, analytiques et statistiques.

- **Heroku** : Heroku est une plate-forme cloud en tant que service (PaaS) qui permettent le déploiement d'applications web (comme streamlit) en utilisant la technologie git. La plateforme est capable de supporter plusieurs langages tels que Python, Java, Scala, ect.

8.3 Le site Etude-assemblée

Nous avons implémenté l'application internet grâce à Streamlit en utilisant les différentes fonctionnalités de cette librairie, puis nous avons déployer notre application en utilisant Heroku et git. L'application est disponible sur ce lien : `https://etude-assemblee.herokuapp.com/`.

Le site internet se décompose en 4 pages traitant chacune d'une analyse différentes des donnée.

8.3.1 Homepage

Cette page (Figure 43), est une introduction au projet avec présentation des données utilisés, ainsi que quelques visualisations basiques sur la répartition des groupes auteurs, la quantité de lettres envoyées au cours du temps et la répartition des rubriques.

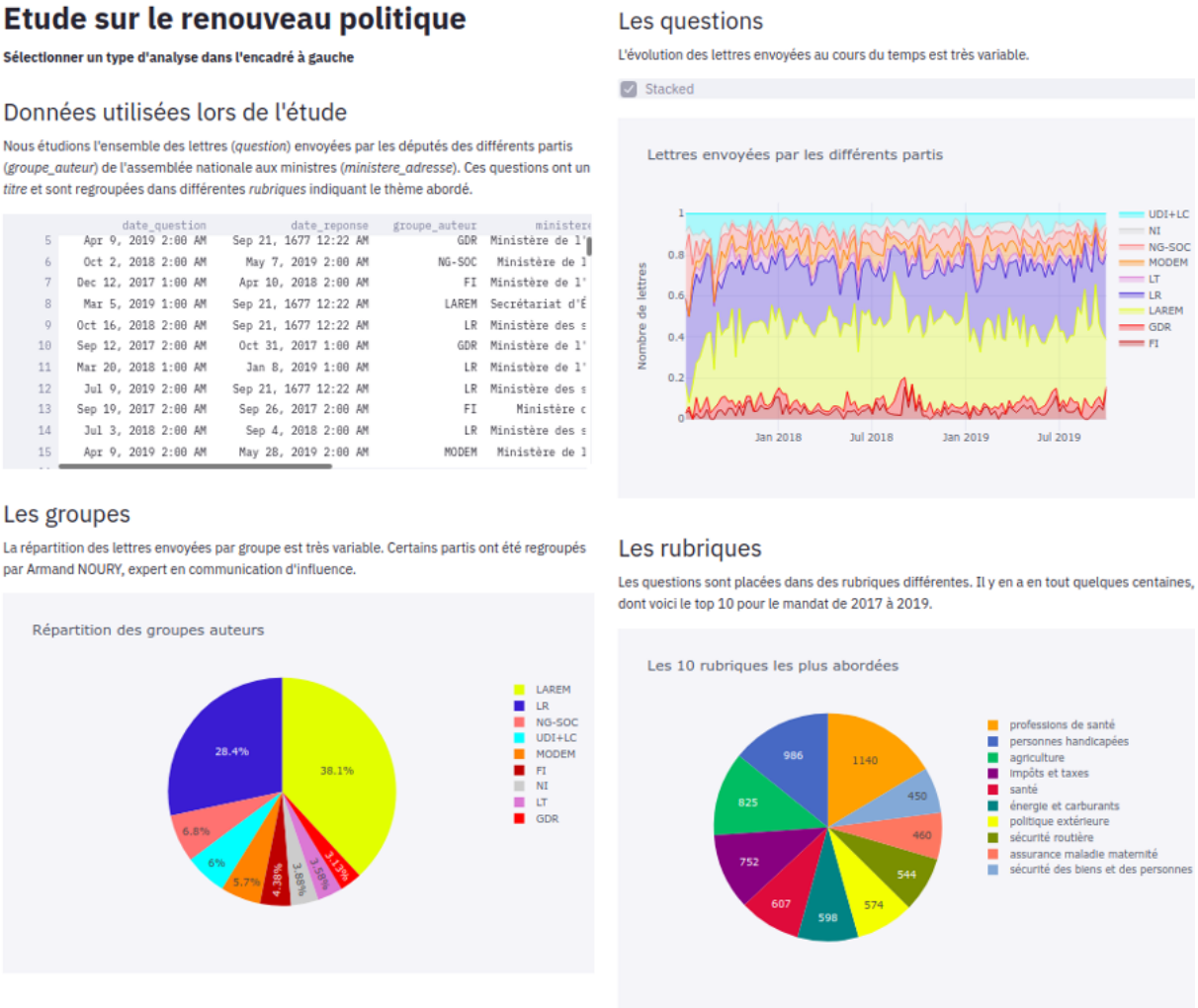


Figure 43: **Homepage** du site etude-assemblée

8.3.2 Anomalies

Cette page (Figure 44) recherche les anomalies présent dans les lettres envoyés pour une thématique choisie. Pour cela, elle va analyser les occurrences d'envoi de lettre et repérer les pics d'envoi de lettre et faire ressortir grâce à la technique du TF-IDF les mots les plus sur-representer durant cette période sous forme d'un nuage de mot. Nous pouvons aussi voir la répartion par groupes des auteurs des lettres et à quel ministère ceux-ci se sont adressé.Enfin nous pouvons lire les lettres envoyées durant la période analysée afin de connaitre le contenu exact des lettres. Cette page répond à la demande de Proches de repérer rapidement les anomalies dans l'envoi des lettres par les députés afin de déceler des informations politiques (Réponse à l'actualité, pression des lobbys, etc).

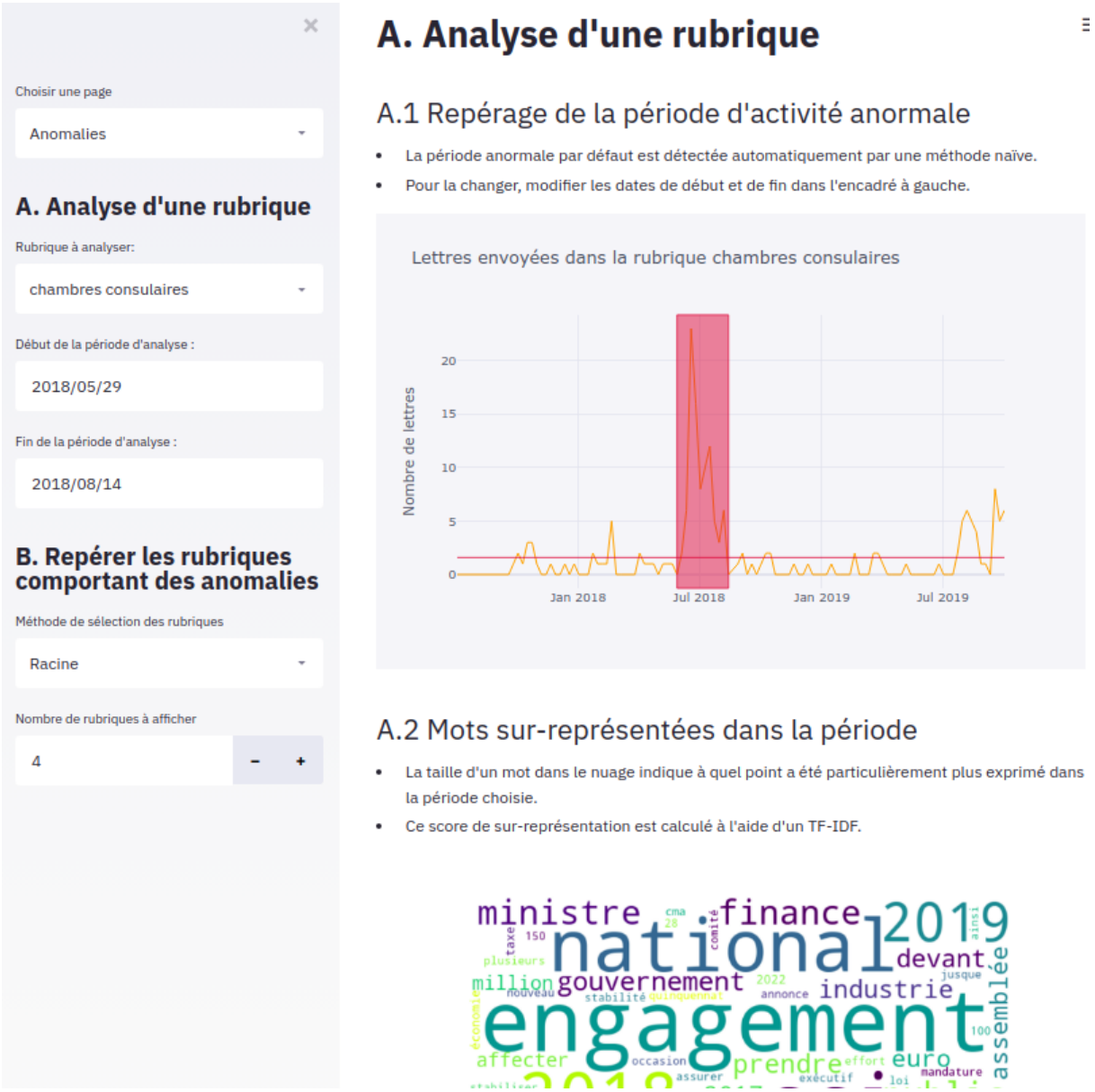


Figure 44: Page **Anomalie** du site etude-assemblée

8.3.3 Expression

Cette page (Figure 45) répond au besoin de comprendre la manière dont un certain sujet a été abordé par les différents groupes et au cours du temps. On peut y analyser l'utilisation d'une expression entrée par l'utilisateur dans la barre de recherche, qui peut être un mot ou une combinaison de mots. Les options permettent de choisir si on veut appliquer la recherche au texte normalisé (comme expliqué dans la section 5.1), si on veut ne repérer que les termes isolés (*plat* ne sera pas compté dans *omoplate*), et enfin si on veut rentrer nous même une expression REGEX (REGular EXpression, outil de représentation d'une chaîne de caractère [6] et [7]).

Les graphiques affichés représentent d'abord l'utilisation et la sur-utilisation par les différents groupes de l'expression choisie, comme expliqué dans la section 7.5. L'évolution des occurences de cette expression est aussi affichée en dessous.

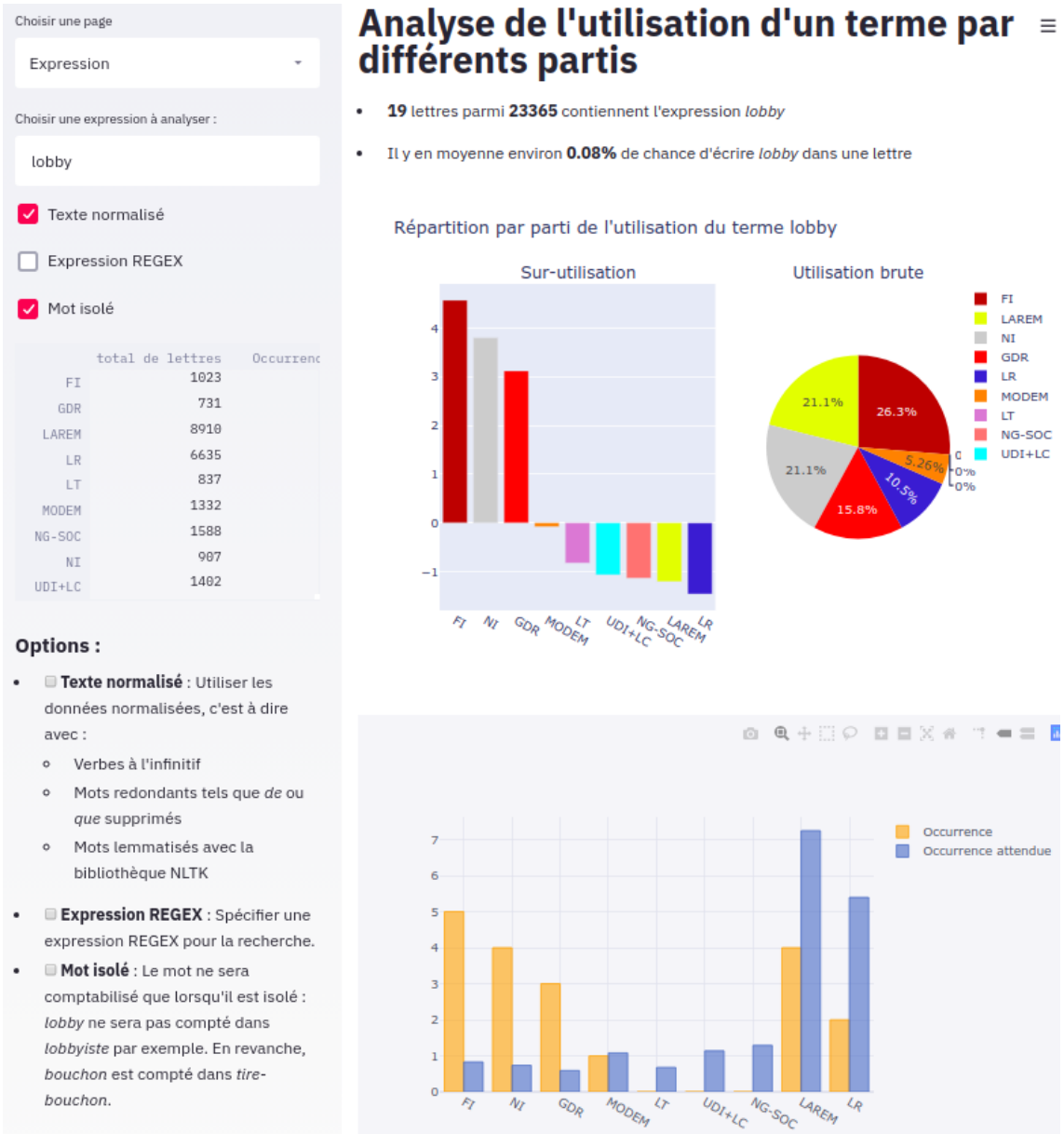


Figure 45: Page **Expression** du site etude-assemblée, pour l'expression *lobby*

8.3.4 Répartition

Cette page, (Figure 46) s'intéresse à la façon dont les lettres d'un parti choisi par l'utilisateur sont dispersées dans les différentes rubriques. Elle reprend les graphiques de sur-utilisation d'une rubrique expliqués dans la section 7.2, en permettant de choisir entre deux modes de sélection. Dans le mode par défaut, on indique avec la slidebar combien de rubriques veut-on voir afficher parmi les plus anormales du parti. Dans le mode de sélection avec alpha, ce sont toutes les rubriques anormales à un seuils supérieur à alpha (sur la Figure 47, dont la précision est choisie sur la slidebar, qui sont affichées. Enfin, on peut choisir dans la boîte de sélection au dessous le mode de tri : selon les rubriques plus extrêmes (les p-values les plus faibles), les plus sur ou sous représentées.

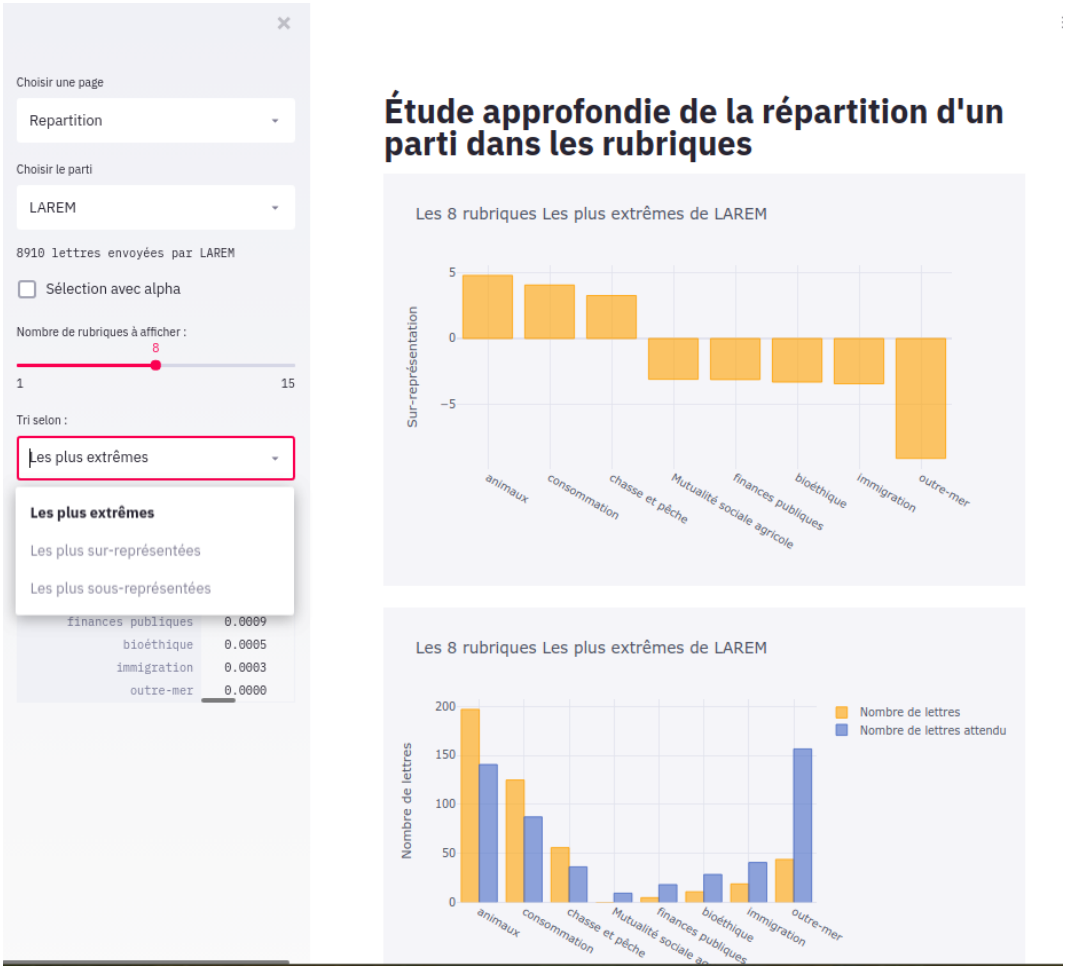


Figure 46: Page **Répartition** du site etude-assemblée, pour le parti *LAREM*

Figure 47: Option de sélection avec alpha pour la page **Répartition**

9 Impacts sur la société

«L'espace global de l'information et de la communication est un bien commun de l'humanité qui doit être protégé comme tel. Son organisation relève de la responsabilité de l'humanité tout entière, par l'intermédiaire d'institutions démocratiques, dans le but de faciliter la communication entre les individus, les cultures, les peuples et les nations, au service des droits humains, de la concorde civile, la paix, la vie et l'environnement» – Déclaration internationale sur l'information et la démocratie, novembre 2018 [9].

En 2019, Transparency International classe la France 23$^{\text{ème}}$ (avec un score de 69%) – à égalité avec les États-Unis d'Amérique mais derrière le Danemark (1$^{\text{er}}$), l'Allemagne (9$^{\text{ème}}$) ou l'Uruguay (21$^{\text{ème}}$) – sur son indice de perception de la corruption (IPC) [8]. Même si certaines critiques légitimes ont été formulées sur cet indicateur [10], la régression de la France dans le classement témoigne de l'enrayement de la dynamique de transparence dans laquelle la France s'était inscrite avec la *loi relative à la transparence de la vie publique* (octobre 2013) et la *loi relative à la transparence, à la lutte contre la corruption et à la modernisation de la vie économique* dite loi Sapin II (décembre 2016). Si une nouvelle impulsion politique semble nécessaire pour développer la transparence et lutter contre la corruption, les actions privées et/ou citoyennes ont également le pouvoir de renforcer l'accès à l'information de toutes et de tous afin qu'ils puissent exercer pleinement leur droit à la liberté d'expression et d'opinion. Notre projet se fixe pour objectif de fournir aux citoyens un outil leur permettant d'extraire un maximum d'informations d'une partie des données mise en libre service par l'assemblée nationnale.

L'article 19 de la déclaration universelle des droits de l'homme et du citoyen garantit à tout individu le « droit à la liberté d'opinion et d'expression… » et ces droits supposent que soient garantis les moyens de les exercer [11]. La liberté d'opinion nécessite donc l'accès libre aux informations fiables et factuelles. C'est dans ce cadre que la France abolit le secret administratif en garantissant l'accés aux documents qu'elle détient (qu'elle les ait produit ou non) avec la loi du 17 juillet 1978, permettant ainsi aux citoyens de consulter librement tout document ne relevant pas du secret d'état – *article 6 de la loi de juillet 1978* [12]. Cette volonté de transparence est renforcé en 2000 avec la création de Légifrance – *plateforme rassemblant la majorité des textes législatifs du droit français* – ou avec Data.gouv.fr – *plateforme de diffusion des données publiques de l'état français* – mais également avec la mise en accés libre du registre des représentant d'intéret créé dans le cadre de la loi Sapin II ou des questions au gouvernement sur le site de l'assemblée nationale. Si la libre diffusion des données relatives à la vie publique est un pas en avant dans la recherche de la transparence administrative et peut contribuer à redonner confiance aux citoyens dans leurs élus, elle nécessite des outils d'analyse afin d'en extraire un maximum d'information fiable et pertinente.

Les sciences des données, qui peuvent se définir comme l'ensemble des sciences – s'appuyant sur des outils mathématiques, statistiques ou informatique – permettant d'extraire de la connaissance d'un ensemble de données, interviennent ici : lorsque la quantité de données est trop importante pour pouvoir être traitée en un temps raisonnable par l'homme ou lorsque ces dernières sont trop complexes et nécessitent d'être simplifiées ou résumées [13].

Notre projet se présente, dans ce contexte, comme un outil de recherche, de condensation et plus largement d'analyse des données contenues dans les lettres des députés au gouvernement et de leurs réponses. Nous avons créé un ensemble de fonctionnalités – basé sur des outils statistiques d'analyse textuelle (TF-IDF, matrice de co-occurence, etc..) ou de visualisation numérique (Matplotlib, Plotly, etc..) – qui permettent à tous, profanes ou initiés

des outils utilisés, d'extraire des connaissances de ce jeu de données à travers l'application web hébergé sur Heroku ou avec les ressources du git.

Le site etude-assemblee permet d'extraire différents types d'anomalies, de repérer le champs lexical utilisé par un groupe politique ou voir la façon dont celui ci aborde préférentiellement certains thèmes. On pourra par exemple repérer rapidement des périodes d'activité anormalement élevé dans certaines rubriques et voir le champs lexical des lettres qui lui sont associées. On pourra également repérer les thèmes les plus anormalement abordés ou les mots les plus anormalement employés par chacun des partis.

Nous sommes conscient que les outils que nous avons créé ne permettent pas d'extraire toute l'information pertinente de ce jeu de données. Chaque fonctionnalité permet d'extraire des renseignements ciblés dans un cadre précis, avec des hypothèse faîtes au préalable (Les lois étudiées suivent des lois normales, etc...) et certains paramètres choisient arbitrairement (seuil de singularité, etc...). Le site *etude-assemblee* donne une première approche global de notre travail et permet d'obtenir plusieurs résultats visuels clairs et compréhensibles par tous mais n'est pas exhaustif des fonctionnalités mise à disposition dans le git. C'est pourquoi nous invitons les utilisateurs, initiés ou non aux méthodes statistiques utilisées et à la programmation en python, à s'approprier notre travail, à le modifier ou le compléter si nécessaire et à l'utiliser dans un cadre qui lui est propre et qui lui semble pertinent.

10 Conclusion

Ce projet a été une occasion de nous épanouir dans différents domaines. D'abord, nous avons découvert et apppris à utiliser une grande diversité d'outils informatiques. Nous maîtrisons à présent une panoplie de bibliothèques Python efficaces dans la gestion et la visualisation de données. Nous sommes aussi beaucoup plus à l'aise pour travailler dans un cadre d'exploration (Jupyter), de collaboration et de partage (Git, Wiki, LateX). Ensuite, le caractère exploratoire du projet nous a confié une grande liberté et nous a permis d'apprendre à proposer des idées en s'appuyant sur notre intuition comme sur nos connaissances académiques. Nous avons passé beaucoup de temps à réfléchir ensemble pour orienter notre travail à la lumière des derniers résultats obtenus. Nous nous sommes aussi beaucoup servis de nos connaissances scolaires, particulièrement en base de données et en statistiques. Enfin, l'application à un problème concret était très enrichissant pour comprendre comment mettre l'innovation et les outils techniques au service d'un projet utile.

Les résultats ainsi obtenus sont dans l'ensemble assez satisfaisants. Parmi les nombreuses pistes explorées, nous avons trouvé que les résultats d'une grande partie d'entre elles étaient assez intéressants. Ces derniers ont même parfois dépassé nos attentes comme celles de nos encadrants, et les deux entreprises nous ont félicité pour nos travaux. Si toutes nos études n'ont pas forcément abouti à un rendu concret, et qu'il reste de nombreux aspects à approfondir, nous sommes contents de voir qu'une grande partie de notre travail a été appréciée et sera utilisée.

Si nous avions plus de temps pour appronfondir le sujet, nous aurions aimé améliorer le site internet avec les conseils de l'agence pour le rendre plus agréable et puissant. Il aurait été aussi très intéressant de se pencher sur des algorithmes d'analyse textuelle plus poussés.

Références

[1] Loi HATVP : https://www.hatvp.fr/textes-de-reference/

[2] Description des questions : Fiche de synthèse 51, http://www2.assemblee-nationale.fr

[3] Groupes : http://www2.assemblee-nationale.fr/15/les-groupes-politiques/

[4] Informations sur le TF-IDF: https://fr.wikipedia.org/wiki/TF-IDF

[5] Ressources en statistiques : Cours de Statistiques Inférentielles de Mme Fanny Villers

[6] Informations à propos des expressions régulières :
https://fr.wikipedia.org/wiki/Expression_régulière

[7] Utilisation de REGEX : https://www.rexegg.com/regex-quickstart.html

[8] Transparency Internationnal :
https://www.transparency.org/cpi2019?/news/feature/cpi-2019

[9] Déclaration Internationale sur l'information et la démocratie : https://rsf.org/fr/
lespace-global-de-linformation-et-de-la-communication-un-bien-commun-de-lhumanite

[10] Wikipédia – Indice de perception de la corruption
:https://fr.wikipedia.org/wiki/Indice_de_perception_de_la_corruption

[11] Déclaration universelle des droits de l'homme et du citoyen :
http://www.textes.justice.gouv.fr/textes-fondamentaux-10086/
droits-de-lhomme-et-libertes-fondamentales-10087/
declaration-universelle-des-droits-de-lhomme-de-1948-11038.html

[12] Loi du 17 juillet 1978 :
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000339241

[13] Wikipédia – Sciences des Données :
https://fr.wikipedia.org/wiki/Science_des_données

Analyse de données

[1] Documentation Pandas : https://pandas.pydata.org/docs/

[2] Documentation Numpy : https://numpy.org/

[3] Documentation Spacy : https://spacy.io/

[4] Documentation Scikit-Learn: https://scikit-learn.org/stable/

[5] Documentation Nltk : https://www.nltk.org/

Visualisation

[1] Documentation Matplotlib : https://matplotlib.org/

[2] Documentation Plotly : https://plotly.com/

[3] Documentation Streamlit : https://www.streamlit.io/

[4] Documentation Seaborn : https://seaborn.pydata.org/

Autres

[1] Jupyter : https://jupyter.org/

[2] Heroku : https://www.heroku.com/