Grand débat et big data

Gérard Weisbuch

5 août 2019

Résumé

Les réponses des participants au Grand Débat National sont analysées par un traitement numérique simple et clair. Les termes pertinents dans l'ensemble des réponses à une question sont comptabilisés et permettent ainsi l'extraction des données suivant la logique booléenne.

1 Introduction

L'exercice du Grand Débat National s'est tenu en France du 15 janvier au 15 mars 2019. Le terme est un peu ambigu puisqu'il peut recouvrir aussi bien le débat officiel par lequel des participants volontaires ont déposé sur le site officiel des réponses aux questions posées par l'administration Française, qu'aux débats alternatifs organisés en réaction, sans compter les débats oraux organisés dans les mairies par exemple. On pourrait aussi y ajouter les "cahiers de doléances". Les textes de réponses [1] et les synthèses [6] du débat officiel sont disponibles sur le site https ://grand-debat.fr/.

La taille des données, près de 2 millions de contributions en ligne, appelle un dépouillement automatique car on imagine mal une lecture exhaustive directe de toutes les réponses. Les critiques ne manquent pas sur l'utilisation des machines mais on voit mal comment mettre en oeuvre les alternatives suggérées, comme le retour aux textes.

Que recherche-t-on au juste ? L'objectif est de classer les réponses aux questions posées. Plus précisément peut-on mettre en évidence des sous-ensembles cohérents de réponses à une question donnée, proches les uns des autres mais distincts par leur sens des réponses des autres sous-ensembles ? Peut-on définir des distances entre les réponses, et en particulier appliquer cette notion à la classification des réponses ? Enfin une fois les sous-ensembles définis, on souhaite aussi évaluer le nombre des réponses de chacun des sous-ensembles.

Dans l'idéal, on souhaiterait disposer de méthodes entièrement automatiques, c'est à dire ne faire intervenir qu'un minimum de paramètres ajustables dans les algorithmes, et surtout ne pas fixer arbitrairement les classes a priori. C'est l'algorithme de classement qui serait sensé les découvrir. C'est le but des méthodes publiées depuis une vingtaine d'années en traitement des langues naturelles (NLP, natural language processing) [4, 8, 2, 3]. Ces méthodes sont basées sur un codage numérique des textes.

D'une manière générale, un texte est transformé en une liste de chiffres codant par exemple la fréquence des termes utilisés. Le terme anglo-saxon est bag of words, littéralement sac de mots.

Leurs auteurs comparent les performances des différentes méthodes sur des corpus de textes communs et affichent des résultats impressionnants. Nous sommes loin d'avoir obtenu des performances comparables sur les réponses au Grand Débat que nous avons étudiées par ces méthodes : la lecture directe de textes dont les représentations numériques sont proches ne correspond pas toujours à des réponses de même sens. Autrement dit la correspondance réponse/représentation numérique obtenue n'est pas fiable.

Peut-on au moins projeter dans un espace de dimensions réduites comme par la projection en composantes principales (PCA, principal component analysis [7]) et interpréter la nature des axes de projection ? Bien que la PCA soit d'un usage courant dans l'analyse des données nous n'avons pas non plus été capable d'en interpréter les résultats, même en poursuivant la décomposition jusqu'à l'ordre 5.

Nous avons donc été amenés à recourir à une méthode semi-automatique dans laquelle les critères à la base de la classification sont définis par l'utilisateur. La méthode est donc moins "objective" que les méthodes récentes mentionnées plus haut, mais elle a par contre le grand avantage d'être claire et compréhensible par un très large public. Le principe en est le même que celui des recherches par mots clés sur internet.

Nous décrirons d'abord la recherche Booléenne [5] dans la section suivante en illustrant notre propos sur un exemple concret de classement des réponses à une question du grand débat. Nous décrirons ensuite quelques résultats obtenus dans le classement des réponses à d'autres questions. La portée de la méthode et ses limites seront abordés dans la discussion. C'est aussi au cours de la discussion que nous comparerons la méthode Booléenne aux méthodes bag of words qui font appel à des mathématiques bien plus sophistiquées.

2 Le classement Booléen

Nous proposons ici d'utiliser une technique élémentaire de recueil d'information (information retrieval [5]) applicable aux grands recueils de données, basée sur la détection dans les textes des termes pertinents.

Nous sommes aujourd'hui familiers avec la pratique des moteurs de recherche sur internet. L'utilisateur recherche les documents contenant un ensemble de termes. C'est cette pratique que nous généralisons pour une analyse quantitative des réponses au grand débat.

Supposons que nous souhaitions analyser les réponses à une question sur l'organisation de l'Etat :

"En qui faites-vous le plus confiance pour vous faire représenter dans la société et pourquoi ? "

Après examen d'une centaine de réponses, des termes comme 'maire', 'élué, 'député', 'personne' etc. semblent pertinents. Mais en plus du décompte de chacun de ces termes exacts nous aimerions :

— Prendre en compte des réponses contenant des termes approchés : certains participants peuvent avoir écrit 'députés', d'autres 'élus' ou 'président','Président','député'. Au lieu d'un seul terme, nous testerons donc la présence de 'élu' OU bien 'élus' OU bien 'député' OU bien 'députés' ... Le OU est ici un OU logique, correspondant à la présence d'un ou plusieurs termes liés au premier. Nous nommerons lemme l'ensemble de ces termes que nous jugeons équivalents dans le contexte de la question posée.
— Il est intéressant de comptabiliser la présence d'un lemme parmi d'autres possibles, mais on aimerait aussi comptabiliser la présence simultanée de plusieurs lemmes dans la même réponse : par exemple, combien de réponses contiennent le lemme 'élu' mais aussi en plus le lemme 'maire' dans la réponse examinée. Le test fait alors intervenir un ET logique : la réponse au test n'est positive que si les deux lemmes sont présents.

Dans le cas de la question prise en exemple :

"En qui faites-vous le plus confiance pour vous faire représenter dans la société et pourquoi ? "

La lecture préliminaire d'une centaine de réponses nous conduit à choisir les trois lemmes suivants dont le représentant est le premier élément.

Le lemme ('maire','Maire','maires') représenté par le terme 'maire',

le lemme ('personne','moi','peuple') représenté par le terme 'personne',

et le lemme ('élu','élus','député','députés','président', 'Président') représenté par le terme 'élu'.

L'analyse des 89.266 réponses exprimées consiste à tester la présence dans les réponses d'un ou plusieurs lemmes.

En ne testant la présence que d'un seul lemme indépendamment des deux autres, on observe que 27.004 réponses mentionnent un des termes du lemme maire, 39.877 le lemme élu et et 18.256 le lemme personne. 21.162 réponses ne mentionnent aucun des trois.

Ce qui donne en pourcentage des lemmes reconnus (sur 68.104) : maire 39,7 % personne 26,8 % élu 58,6 % . Le test d'un seul lemme à la fois est comparable aux méthodes standard utilisées pour le dépouillement officiel par la firme Opinionway.

Les tests plus élaborés sur la présence simultanée de plusieurs lemmes donnent :

15.165 réponses ne mentionnent que le lemme maire à l'exclusion des deux autres.

11.975 réponses ne mentionnent que le lemme personne à l'exclusion des deux autres.

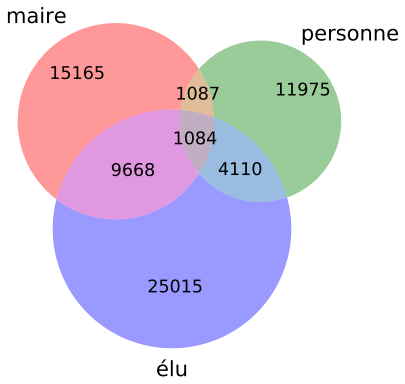
1.087 réponses mentionnent les deux lemmes maire et personne mais non le lemme élu.

25.015 réponses ne mentionnent que le lemme élu à l'exclusion des deux autres.

9.668 réponses mentionnent les deux lemmes maire et élu mais non le lemme supprimer

4.110 réponses mentionnent les deux lemmes personne et élu mais non le lemme maire

1.084 réponses mentionnent les trois lemmes.



Le diagramme de Venn est la représentation graphique de ces résultats. La surface de chaque portion de cercle est proportionnelle au nombre de réponses mentionnant les lemmes. Les trois cercles complets de couleur brique, verte et bleue correspondent respectivement aux lemmes maire, personne et élu. Les intersections aux couleurs composites correspondent aux réponses mixtes mentionnant deux ou trois lemmes. Les chiffres sont les nombres de réponses correspondantes.

Nous voyons dans le cas présent qu'une majorité des répondants font confiance aux élus, et plus précisément aux maires

La procédure employée pour l'analyse des réponses à une question consiste donc à :

— Rechercher les lemmes utilisés dans les réponses à partir de la lecture d'une partie des réponses, disons un centaine. Il s'agit non seulement de trouver les termes pertinents, mais aussi de les regrouper en lemmes. Nous appliquons aussi la procédure de regroupement à la lemmisation au sens strict : on considère comme équivalents des termes écrits en lettres capitales ou en minuscules, les termes au masculin ou féminin et ceux au singulier ou au pluriel.

— Une fois les lemmes définis on comptabilise leur présence ou leur absence dans l'ensemble des réponses.

Plusieurs représentations graphiques des résultats sont possibles. Nous avons utilisé les diagrammes de Venn à 2 et 3 dimensions, donc pour 2 ou 3 lemmes testés, car ils conservent la proportionnalité de la surface des domaines au nombre des réponses. Mais cette propriété n'est pas conservée pour un plus grand nombre de lemmes. On peut certes tester plus de lemmes et représenter les résultats par plusieurs diagrammes. Ainsi pour

quatre lemmes, on trace 2 diagrammes pour trois lemmes suivant que le quatrième lemme est cité ou pas dans les textes (voir la section 3.1). Mais le nombre de diagramme N_d augmente alors comme une fonction puissance de nombre de lemmes testés d :

$$N_d = 2^{d-3}$$

Une autre représentation, bien adaptée à la dichotomie de la procédure, est l'arborescence hiérarchique ou dendogramme. La racine de l'arbre est l'ensemble des réponses et chaque branchement correspond à l'application d'un test sur la présence ou l'absence d'un des lemmes dans les réponses. Les dendogrammes permettent de prendre en compte plus de lemmes, mais les distances lues sur l'axe des x peuvent être ambigües : elles dépendent de l'ordre des dichotomies effectuées. Un unique diagramme de Venn à trois dimensions peut être représenté $3! = 6$ dendogrammes.

3 Autres exemples

3.1 DEMOCRATIE ET CITOYENNETE

Question 27 : Que faudrait-il faire pour consulter plus directement les citoyens sur l'utilisation de l'argent public, par l'Etat et les collectivités ?

Nos premiers essais avec trois lemmes comprenant moins de termes ne permettaient pas de 'récolter' une fraction importante des réponses supérieure à 30%. Autrement dit 70% des réponses ne contenaient aucun des termes choisis. C'est pourquoi nous avons inclus plus de termes et utilisé 4 lemmes au lieu de 3.

A la lecture d'une centaine de réponses on choisit donc les 4 lemmes suivants.

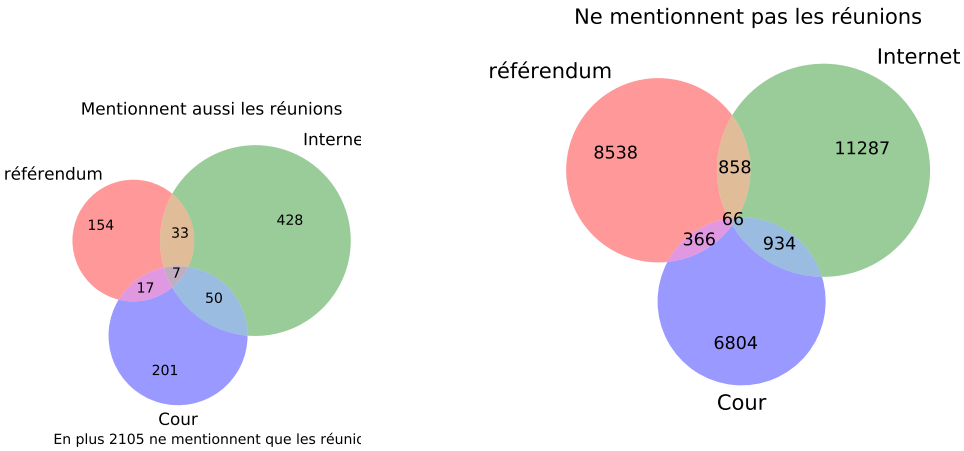
'référendum','RIC','Referendum','referendum','ric','peuple'

'Internet','internet','en ligne','site','sondages','plateforme','presse','informer'

'Cour','comptes','cour','Comptes','parlementaire','parlementaires'

'réunion','Réunion', 'réunion publique','réunions', 'assemblées','assemblés'

On teste la présence de lemmes identifiés par leur premier élément. Avec 4 lemmes deux diagrammes de Venn à trois lemmes sont nécessaires. Nous avons choisi de scinder les contributions suivant le lemme réunion. Les deux diagrammes représentent donc à gauche les réponses qui contiennent et à droite celles ne contiennent pas le lemme réunion.



Diagrammes de Venn : A gauche figurent les nombres de réponses contenant au moins l'un des termes d'un, deux ou trois des lemmes référendum, Cour des Comptes et internet, en plus du lemme réunion.

A droite, nombres de réponses ne mentionnant pas le lemme réunion mais contenant au moins l'un des termes d'un, deux ou trois des lemmes référendum, Cour des Comptes et internet. De plus le lemme réunion seul est mentionné dans 2.105 réponses.

La faible superficie des intersections entre les lemmes implique que les opinions sont divisées (clusters). Les lemmes sont donc discriminants. Par contre ils ne couvrent toujours qu'une faible fraction des réponses, environ 40 %.

8.445 répondants font confiance à la Cour des Comptes et certains souhaitent même voir ses recommandations contraignantes. La disponibilité des comptes sur l'internet est demandée dans 13.663 réponses. Le référendum, impliquant le contrôle direct reste suggéré dans 10.039 réponses.

Question 33 : Que faudrait-il faire aujourd'hui pour renforcer l'engagement citoyen dans la société ?

On teste la présence de lemmes suggérés par la lecture d'une centaine de réponses.

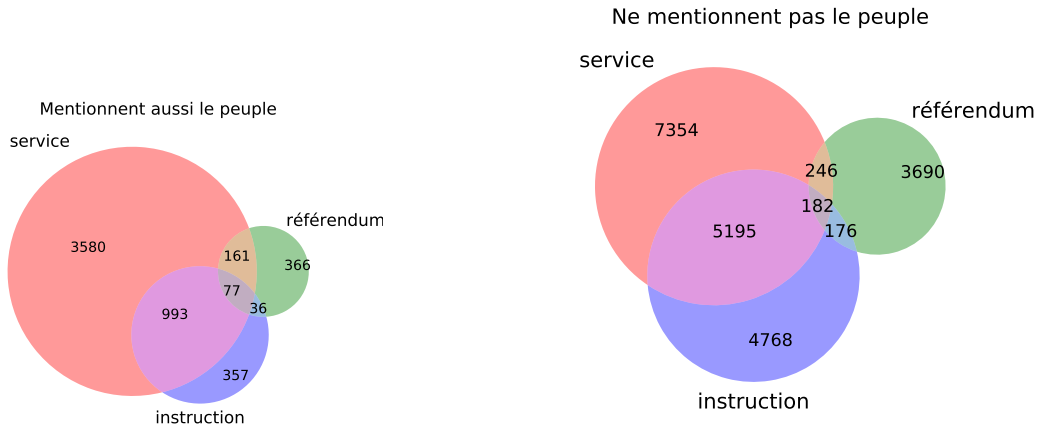
'service','engagement','écoute'

'référendum','RIC','vote','votes'

'instruction','éducation','éducations','civique','économie', 'contribution','association'

'peuple','citoyen'

72.840 réponses ont été testées, l'un des 4 lemmes au moins est présent dans 31.465 réponses.



En plus 4284 ne mentionnent que le peuple

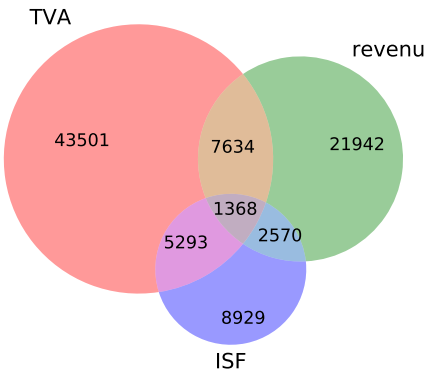
Prévalence du lemme service (17.788 réponses positives). Référendum relativement peu fréquent (4.934). Faible intersection de référendum avec le deux autres lemmes.

3.2 LA FISCALITE ET LES DEPENSES PUBLIQUES

Question 14 : Quels sont selon vous les impôts qu'il faut baisser en priorité ?

On teste la présence des lemmes dans 140.589 réponses. Un des trois lemmes est présent dans 91.237 réponses. Les lemmes sont :

'TVA','tva','CSG','csg'
'revenu','travail','production','salaires','salaire'
'ISF','isf','habitation'



On obtient un grand nombre de réponses(140.589), les citoyens semblent très motivés sur le sujet des impôts. De plus, les réponses mettent en cause une petite fraction de l'ensemble de la fiscalité et les opinions sont tranchées. 57.796 réponses parlent de baisser la TVA, 33.514 l'impôt sur le revenu et 18.160 l'ISF.

Question 17 : Quels sont les domaines prioritaires où notre protection sociale doit être renforcée ?

Résultats pour 131.066 réponses non vides.

On teste la présence de lemmes représentés par le premier élément.

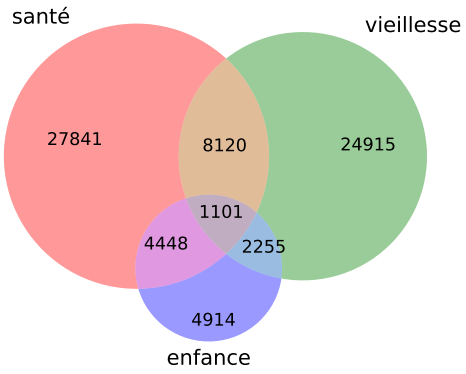
'santé','maladie','hôpital','hopitaux'

'vieillesse','retraite','autonomie','handicap','personnes'

'enfance','éducation','enseignement'

Un des trois lemmes est présent dans 73.594 réponses, c'est à dire environ 56%.

Les préoccupations majeures sont la santé et la vieillesse, dont la dépendance. Ce serait intéressant d'avoir les âges des répondants. On a vérifié par ailleurs que le lemme chômage n'est mentionné que dans 5% des réponses.

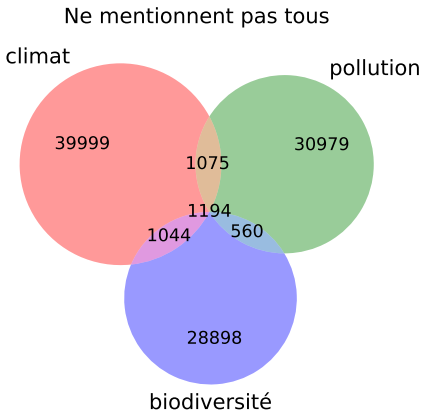


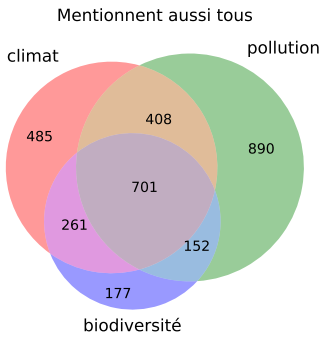
3.3 LA TRANSITION ECOLOGIQUE

Question 12 : Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ?

On teste la présence des lemmes dans 143.329 réponses . Un des quatre lemmes est présent dans 119.171 réponses.

Les lemmes sont :
'climat','climatique','climatiques'
'pollution'
'biodiversité'
'tous','toute','tout'





En plus 12 346 ne mentionnent que tous

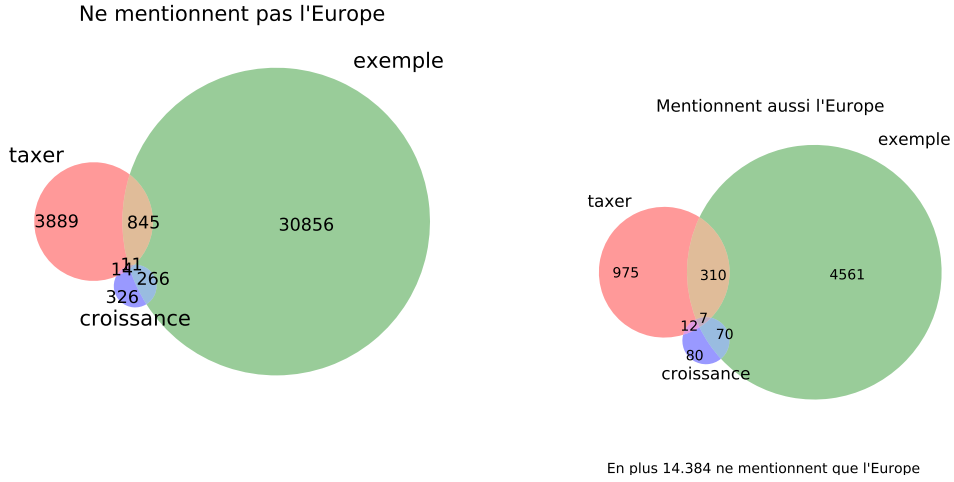
Grand nombre de réponses : 143.329 , dont 24.158 seulement échappent à la classification parce qu'ils ne mentionnent aucun des lemmes choisis. Les 13% qui répondent par le lemme "tous" refusent d'établir des priorités. Assez logiquement, les autres réponses n'ont qu'un faible recouvrement entre les lemmes.

Que pourrait faire la France pour faire partager ses choix en matière d'environnement au niveau européen et international ?

On teste la présence de lemmes représentés par le premier élément.
'taxer','Taxer','Renforcer','renforcer','Taxation'
'exemple','exemplaire','exemplarité','pilote','leader','national','modèle'
'croissance','décroissance'
'Europe','europe','européen','européenne'
Résultats pour 112.917 réponses.

Un des 4 lemmes est présent dans 57.056 réponses mais 55.861 réponses n'en contiennent aucun.

Les 4 lemmes conduisent à deux diagrammes, à gauche sans le lemme Europe et à droite avec.



Conclusions

Grand nombre de réponses : 112.917 dont les lemmes captent la moitié. L'autre moitié sont surtout des sceptiques sur les possibilités d'action de la France.

4 Discussion

Notre approche comparée au dépouillement officiel de réponses par la firme opinionway

La méthodologie d'exploitation des données par opinionway n'est pas publique mais au vu des résultats on peut imaginer qu'elle est du même type que la nôtre. Opinionway donne des résultats correspondant à la recherche d'un seul terme indépendamment des autres termes utilisés, alors que nous testons aussi l'usage simultané de trois ou quatre lemmes.

Comptabiliser les termes individuellement est une première approche. Rechercher comme nous le faisons les co-expressions de lemmes prend en compte la pluralité des réponses possibles aux défis socio-politiques et même leur possible complémentarité . Il peut être important par exemple de voir qu'à la question "Que faudrait-il faire aujourd'hui pour renforcer l'engagement citoyen dans la société ?" une fraction importante des réponses mentionnent à la fois les deux lemmes "service, engagement, écoute" et "instruction...".

Par contre opinionway recherche en général un plus grand nombre de termes que nous et leurs statistiques plus complètes détaillent les résultats en fonction de la région, du type d'acteurs (individus ou collectivité) et des modalités de la consultation.

Notre approche comparée aux méthodes basées sur le bag of words [4, 8, 2, 3]

La méthode Booléenne que nous utilisons est basée sur le choix par l'examinateur des lemmes pertinents ; elle possède donc un caractère plus "arbitraire" que les techniques bag of words basées sur l'extraction "objective" d'un grand nombre de termes, de l'ordre d'une centaine en fonction de leur fréquence. A partir des représentations vectorielles de grande dimension des textes la suite de l'analyse mathématique consiste à projeter ces vecteurs sur des espaces de dimension plus réduite dans le but d'observer des amas (clusters) d'opinions semblables. Même dans les cas où cette classification est possible, ce qui n'est pas le nôtre, l'interprétation des axes de projection n'est pas directe ; il faut en revenir à la lecture des réponses. Au contraire notre interprétation est directe.

Les réponses "manquantes"

D'une manière générale, la proportion de réponses contenant au moins l'un des lemmes n'est que d'environ 50%. C'est aussi la proportion mentionnée dans l'analyse officielle d'opinionway[6]. Une analyse plus approfondie serait nécessaire comme par exemple tester le pourcentage de réponses reconnues en fonction du nombre des lemmes. Ce pourcentage ne peut qu'augmenter à chaque fois que l'on rajoute un lemme aux précédents mais cette augmentation devrait ralentir rapidement.

Par ailleurs l'examen direct des réponses montre que souvent les réponses ne contenant pas les lemmes choisis pour leur fréquence sont des réponses plus originales ou mieux développées. On retrouve ici un problème classique de la classification : doit-on avant tout compter les échantillons classables ou au contraire s'intéresser à ce qui est original, donc inclassable. Si l'objectif du grand débat était de rechercher de nouvelles idées, ce sont les réponses inclassables qui pourraient être les plus intéressantes.

Conclusions

Le choix des lemmes par l'opérateur ne permet pas de prétendre à une parfaite objectivité. Mais d'un autre côté, l'une des raisons de l'opposition aux applications de l'intelligence artificielle dans les politiques publiques est que ses algorithmes sont opaques. Ce n'est pas le cas de la méthode que

nous proposons ici. Tous les citoyens peuvent en comprendre les tenants et les aboutissants : le choix des lemmes et les résultats obtenus sont clairs. Les algorithmes sont très simples et disponibles sur le site github[9]. Chacun peut donc faire tourner l'algorithme avec un choix différent des questions et des lemmes utilisés pour le dépouillement. La méthode est donc bien adaptée au dépouillement d'une consultation politique comme le grand débat national. Bien entendu elle a vocation à être améliorée, par exemple à partir des méthodes classiques de recueil d'information [5].

Références

[1] Réponses des participants au grand débat national. `https://granddebat.fr/pages/donnees-ouvertes`, 2019.

[2] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[3] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov) :2579–2605, 2008.

[5] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1) :100–103, 2010.

[6] Opinionway. Synthèses du grand débat national. `https://granddebat.fr/pages/syntheses-du-grand-debat`, juin 2019.

[7] Jake VanderPlas. *Python data science handbook : essential tools for working with data.* " O'Reilly Media, Inc.", 2016.

[8] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[9] Gérard Weisbuch. Dépot logiciel github. `https://github.com/weisbuch/Grand-debat/`, 2019.