Projet Statistique

Rapport Final

Thème

Analyse d'une consultation citoyenne par NLP "Crise Covid-19 : comment inventer tous ensemble le monde d'après ?"

ÉTUDIANTS :

Bouchiar Wahib

Ibrahim Kassoum Habibou

Kaindje Fondjo Veronique Carelle

Sano Batourou

TUTRICE :

Gay Anne-Cécile

COACH :

Arnaud LE HESRAN

Avril 2021

Résumé

Le monde fait perpétuellement face à des problématiques nouvelles dans des sphères diverses et variées. Le constat est frappant autant dans le domaine de la politique, que celui de la santé ou encore celui de l'environnement. Il va sans dire que nous sommes tous concernés par ces problématiques en tant qu'humain en général, et en tant que citoyen en particulier.

Make.org offre l'opportunité à chacun de s'engager en proposant des solutions à des problématiques d'ordre national, voire européenne, et c'est à travers sa plateforme où l'avis de chaque acteur de la société compte et le tout dans le but de satisfaire l'intérêt général.

Les propositions citoyennes auxquelles nous nous intéressons dans le cadre de ce travail ont pour but de répondre à la question : « Crise Covid-19 : comment inventer tous ensemble le monde d'après ? ». Notre mission étant de trouver le moyen d'attribuer de manière automatique un thème (ou tag) à chacune des propositions effectuées.

Jusque là, chaque proposition entrée par les citoyens sur la plateforme est taguée de manière manuelle et, au vu du nombre de propositions effectuées, la tâche peut rapidement s'avérer lourde. L'objectif au final est donc de proposer un modèle permettant de taguer automatiquement les propositions entrantes. Pour ce faire, nous avons passé en revue plusieurs méthodes statistiques susceptibles d'aider à y parvenir.

Nous verrons que parmi ces méthodes, les unes seront uniquement basées sur les propositions dont nous disposons déjà (modélisation non supervisée), tandis que les autres utiliseront en plus l'information supplémentaire disponible liée aux tags ayant été affectés manuellement aux propositions (modélisation non supervisée).

Table des matières

Table des figures

Introduction

L'explosion des données survenue il y a quelques décennies a permis de remettre en question non seulement l'usage que l'on en fait, mais aussi et surtout les techniques d'analyse en vigueur à l'époque. Parmi les problématiques qui se posent parallèlement à cet événement, il y a : comment traiter des données de ce calibre ? Comment les stocker ? etc.... On assiste ainsi à l'avènement de nouveaux outils informatiques et à la mise en place de nouvelles méthodologies permettant de gérer des données massives.
De nos jours, de plus en plus d'organismes privés ou publics tirent profit des données qu'ils collectent dans le but de réaliser leurs objectifs pouvant aller de l'amélioration de leurs produits à l'évaluation de politiques publiques en passant par l'aide à la décision. Par ailleurs, l'usage de méthodes statistiques telles que l'apprentissage supervisé ou encore la régression linéaire pour y parvenir est devenu monnaie courante.

Ce projet a été sollicité par la plateforme de consultation massive Make.org, qui a pour objectif de déterminer des solutions concrètes en se servant des propositions citoyennes ayant abouties à un consensus. Notre mission consiste à modéliser les thématiques issues des propositions faites par des citoyens au sujet de la consultation suivante : "Crise Covid-19 : comment inventer tous ensemble le monde d'après ?". L'objectif in fine sera d'affecter une ou plusieurs thématiques (aussi appelées tags) de manière automatique à chaque proposition.
Pour ce faire, deux approches s'offrent à nous : la modélisation supervisée en se servant des thématiques ayant préalablement été affectées manuellement ; et la modélisation non supervisée par **Word embedding** ; tout cela nécessitant la mise en oeuvre des méthodes de **Natural Language Processing** (NLP) afin de rendre exploitable les données textuelles à disposition. La comparaison de ces deux approches permettra de statuer sur la méthode la plus adéquate et d'atteindre l'objectif fixé.

Le présent rapport constitue un compte rendu détaillé du travail que nous avons eu à faire dans le cadre de ce projet. Dans un soucis d'organisation, ce rapport sera répartis en quatre axes principaux partant de l'analyse descriptive des données à disposition à la mise en oeuvre des modélisation supervisée et non supervisée, en passant évidemment par le pré-traitement des données. Dans la première section intitulée "Analyse descriptive", on procédera à une analyse des variables d'intérêt afin de résumer l'information qu'elles contiennent. Dans la section suivante, nous tâcherons de donner une "feuille de route" décrivant la démarche suivie pour mener à bien ce projet ; le reste de la section sera dédiée à la description des méthodes et des résultats obtenus lors du pré-traitement des données. Les deux dernières sections seront consacrées à la description ainsi qu'à la mise en oeuvre des méthodes de modélisation supervisée et non supervisée. Nous finirons bien évidemment par apporter une conclusion à notre travail.

1 Analyses descriptives

Comme pour tout projet à vocation statistique, effectuer une étude descriptive des données est une étape incontournable. S'agissant dans notre cas d'une étude se basant sur des données volumineuses, il est difficile d'appréhender l'information fournie par la base de donnée en partant d'une simple observation de cette dernière. Ainsi, cette étape consiste à réaliser de simples études permettant de mieux connaître le contenu de la base de donnée, tout en répondant à des questions liées aux effectifs, à la moyenne ou encore à des proportions.
Dans le cadre de notre étude, l'analyse descriptive consistera à résumer les informations concernant les auteurs des propositions (ici les citoyens) ainsi que les propositions elles-mêmes. Partant des données disponibles, les informations les plus utiles dans cette phase sont celles liées d'une part à l'âge des citoyens ainsi que leurs codes postaux ; d'autre part aux propositions et aux votes.
Ainsi, dans les lignes qui suivent, nous tenterons de donner autant d'informations pertinentes que possible concernant les données en répondant à un certain nombre de questions.

1.1 Les valeurs manquantes et autres anomalies

Avant toute chose, il est important de savoir si la base de données comporte des anomalies afin de voir s'il est nécessaire, mais surtout s'il est possible d'apporter des modifications en vue de l'améliorer. Dans notre cas, il s'agit de valeurs manquantes.
En effet, il manque un grand nombre d'informations concernant notamment l'âge et les codes postaux. Il manque en effet environ 64% des âges et 61% des codes postaux. Cela représente des proportions énormes de la base de données, ce qui rend impossible la suppression de ces données manquantes. Par ailleurs, bien qu'on pourrait par exemple remplacer les âges manquants par l'âge médian, cela s'avère être contre-productif car n'apporterait pas d'information pertinente et biaiserais les résultats.
Par conséquent, nous avons choisi de conserver les données telles qu'elles car cela n'aura non seulement pas d'impact négatif sur la suite de l'analyse, mais permettra aussi d'aboutir à des conclusions fidèles à la réalité. Par ailleurs, en observant le tableau récapitulatif ci-dessous, on note facilement quelques anomalies dans la base de données. Il s'agit en fait de valeurs négatives apparaissant dans des colonnes représentant des nombres de votes.

	agree_count	agreeLikeit_count	agreeDoable_count	agreePlatitudeAgree_count	disagree_count	disagreeNoWay_count	disagreeImpossible_count
count	18681.000000	18681.000000	18681.000000	18681.000000	18681.000000	18681.000000	18681.000000
mean	50.959852	13.972592	15.627643	3.317488	5.279268	1.510144	1.348750
std	89.614630	34.338029	29.970790	4.062941	9.214126	4.156239	2.580325
min	-1.000000	-1.000000	-2.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	12.000000	2.000000	3.000000	1.000000	1.000000	0.000000	0.000000
50%	28.000000	7.000000	7.000000	2.000000	3.000000	0.000000	1.000000
75%	60.000000	16.000000	18.000000	5.000000	6.000000	2.000000	2.000000
max	3490.000000	1443.000000	1251.000000	75.000000	332.000000	177.000000	86.000000

FIGURE 1 – Extrait du tableau récapitulatif des données

Pour pallier à ce problème, deux options s'offrent à nous. Supprimer les lignes correspondantes, ou alors remplacer ces valeurs négatives par 0. Nous avons opté pour la seconde

option afin de ne pas perdre de l'information liées à des propositions, et ce, sachant que l'étude n'en sera pas significativement impactée.

1.2 Signalétique citoyenne

1.2.1 Caractéristiques démographiques

Dans un premier temps, nous allons aborder les questions liées aux caractéristiques démographiques des citoyens ayant effectué des propositions dans le cadre de cette consultation.

Notons tout d'abord que les auteurs des propositions ont un âge compris entre 13 et 86 ans. Ceci s'explique par le fait qu'il faut être âgé d'au moins 13 ans pour pouvoir effectuer des propositions. De plus, les auteurs des propositions ont en moyenne 45 ans, et la moité des propositions ont été faites par des citoyens ayant moins de 44 ans.

Caractéristique	Age
Moyenne	44,67
Ecart-type	13,98
Minimum	13
Q1	34
Médiane	44
Q3	56
Maximum	86

Cependant, la distribution de l'âge des citoyens peut être assimilée à une distribution normale, avec des effectifs faibles au niveau des bornes. Ce qui fait que parmi toutes les personnes ayant effectué des propositions, peu d'entre elles ont autour de 13 ans (respectivement de 86 ans). Tandis que les citoyens qui ont autour de 35 ans sont ceux qui font le plus de propositions.

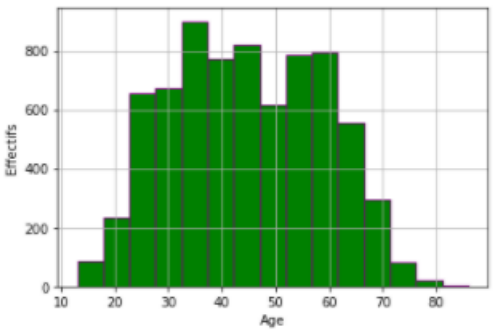


FIGURE 2 – Répartition des âges des citoyens

Pour pousser plus loin l'analyse de la répartition des âges des individus, nous nous sommes intéressés à la comparaison de cette dernière avec la pyramide des âges en France.
Nous remarquons que globalement, la distribution des âges des individus ayant effectué des propositions est similaire de la pyramide des âges en France. Cependant, on remarque une différence significative au niveau de l'écart d'effectif entre les classes d'âge successives (les 34-54 ans et les 55-75 ans). En dépit de tout cela, on pourrait a priori affirmer que les propositions sont faites par des français, ou du moins en majorité. Nous verrons par la suite si ce résultat se confirme en se basant sur la zone géographique des citoyens.

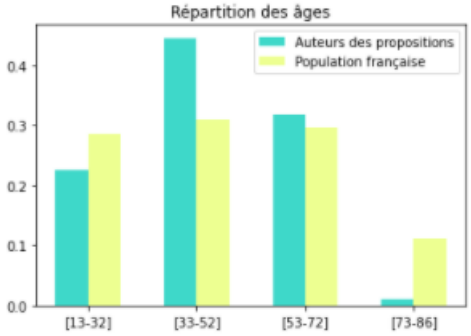


FIGURE 3 – Comparaison de la répartition des âges

La répartition des âges de la population française a été construite en se basant sur les données du 19 Janvier 2021 de l'INSEE. Évidemment, dans l'objectif de rendre la comparaison fidèle à la réalité, les individus de moins de 13 ans et ceux de plus de 86 ans n'ont pas été pris en compte dans la représentation de la population française.

1.2.2 Les zones géographiques

On a vu que les auteurs des propositions ont entre 13 et 86 ans et qu'ils ont en moyenne 45 ans. On va maintenant voir comment l'âge et les propositions sont répartis dans le territoire.

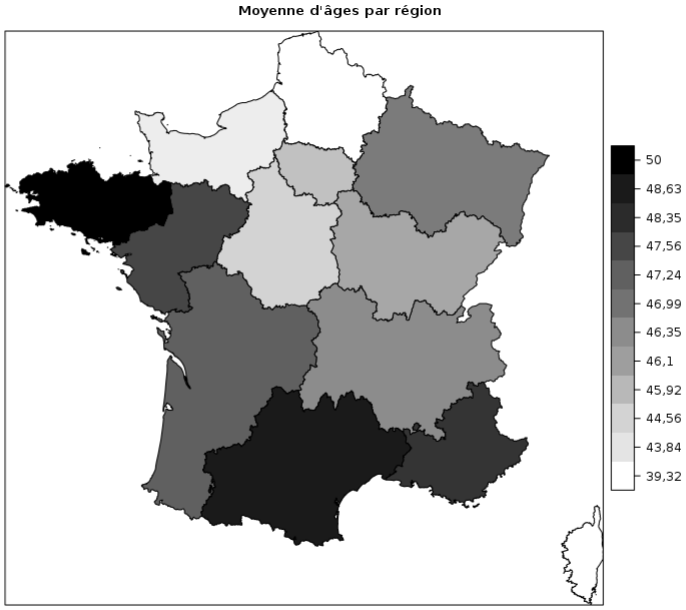


FIGURE 4 – Moyenne d'âges par région

En analysant cette carte, on peut affirmer que le Sud et l'Ouest comptent des participants en moyenne plus âgés que dans le Nord et le Nord-Est de la France. Il faudrait maintenant analyser comment est réparti l'ensemble des propositions dans la métropole.

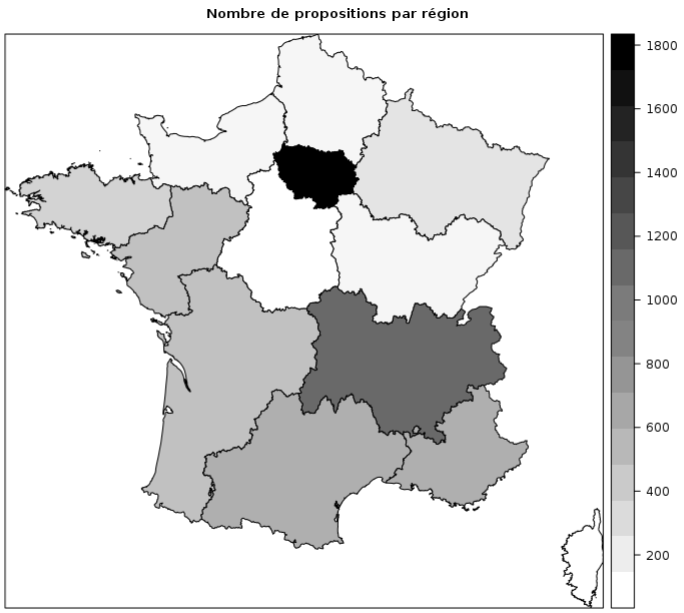


FIGURE 5 – Nombre de propositions par région

On remarque que la grande majorité des votants se situent en Ile-de-France et dans la région Auvergne-Rhône-Alpes. Cela semble coller aux nombres d'habitants par région en France actuellement que l'on peut observer sur la carte suivante.

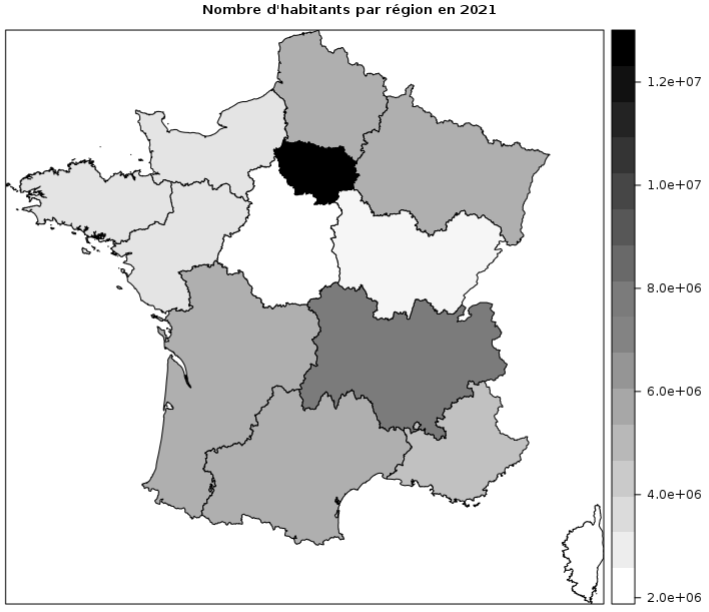


FIGURE 6 – Nombre d'habitants par région en 2021

Source : Données INSEE sur les populations légales de 2021

Pour mieux comparer, on étudie le rapport de la proportion du nombre de votes par région sur la proportion du nombre d'habitants de cette région. Ainsi, on peut savoir si le nombre

de votes pour une région donnée sera représentatif de la population globale en France.

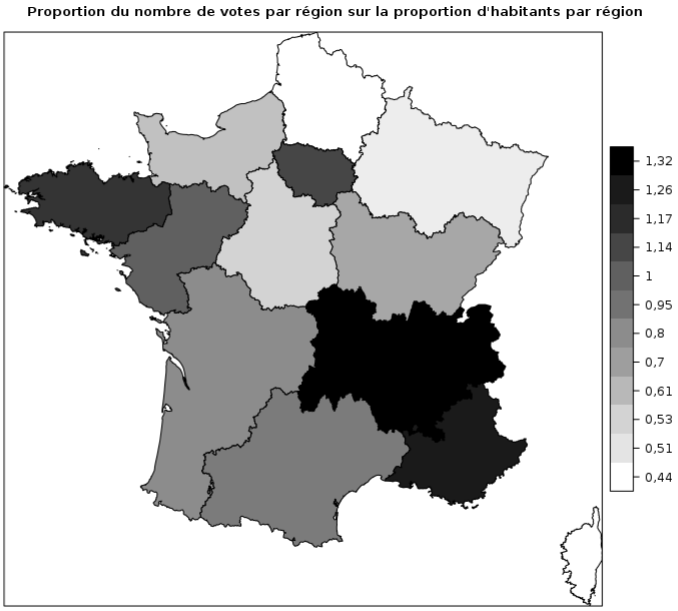


FIGURE 7 – Proportion du nombre de votes par région sur la proportion d'habitants par région

On observe que les régions avec le plus de propositions sont sur-représentées par rapport à la population légale alors que les régions au Nord et au Nord-Ouest sont celles qui sont le moins représentatives de la population française. Ces régions étaient les régions les plus jeunes en moyenne d'âges. De plus, les régions qui sont sur-représentées comme la Bretagne ou l'Occitanie sont celles qui ont les moyenne d'âges les plus élevées. Cela est donc cohérent avec l'analyse de la partie précédente où l'on observe que la classe d'âge des 53 à 72 ans est celle qui représente le mieux la population française tandis que les autres le sont beaucoup moins.

Sachant où se trouvent les commentaires sur le territoire et comment ils sont réparties en fonction de leurs âges, on peut maintenant analyser plus en profondeur les propositions.

1.3 Analyse des propositions

L'objectif ici est de relever autant d'informations que possible au sujet des propositions. Tout d'abord, notre base de données comporte 18 681 propositions, avec un total de 1 245 731 votes.

- Nombre total de votes par proposition Le tableau ci-dessous contient les statistiques pertinentes concernant le nombre total de votes concernant le sujet posé. Comme on peut le lire, il y a en moyenne 67 votes par proposition, et la moitié des propositions décomptent moins de 40 votes. De plus, le quart des propositions ont reçu moins de 20 votes. Et bien que certaines propositions ont eu un franc succès avec plus de 3000 votes, seulement 75% d'entre

elles ont reçu pus de 84 votes.

Caractéristique	Nombre de votes
Total	18681
Moyenne	66,68
Ecart-type	97,09
Minimum	0
Q1	20
Médiane	40
Q3	84
Maximum	3539

TABLE 1 – Statistiques concernant le nombre de votes

Ci-dessous, on peut voir les propositions ayant reçu le plus grand nombre de votes :

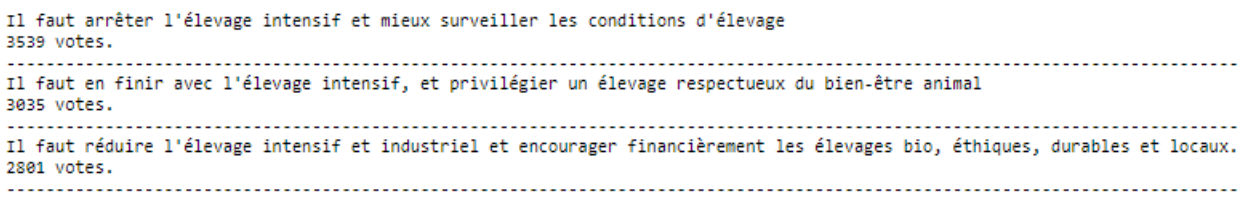


FIGURE 8 – Les trois propositions ayant reçu le plus de votes

Le nombre total de votes par proposition révèle que 239 d'entre elles, soit 1,28%, n'ont fait l'objet d'aucun vote. Cela pourrait s'expliquer par le fait que la base de données a été extraite à un instant t.

- Propositions apportant le plus/moins de divergence L'un des intérêts de l'analyse des propositions est de déterminer quelles sont celles qui apportent le plus de divergence, et celles qui en apportent le moins. Pour cela, on s'aidera du rapport suivant :

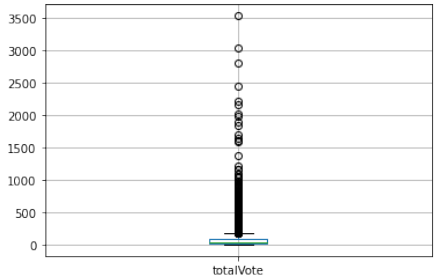
$$unanimite = poids * |\frac{nombre\ de\ votes\ d'accord - nombre\ de\ votes\ pas\ d'accord}{nombre\ de\ votes\ d'accord + nombre\ de\ votes\ pas\ d'accord}|$$

Plus le rapport se rapprochera de 0, plus la proposition concernée sera considérée comme étant source de divergence. A l'opposé, plus le rapport se rapprochera de 1, plus la proposition correspondante sera considérée comme apportant de l'unanimité. Il est à noter que les propositions ne présentant aucun vote sont omises de ce classement car on ne peut objectivement pas juger du caractère unanime d'une proposition n'ayant reçu aucun vote. Par ailleurs, le poids de chaque proposition est établi sur la base suivante :

	Nombre de votes	Poids
0	Moins de 20	0.15
1	]20-40]	0.20
2	]40-84]	0.25
3	Plus de 84	0.40

FIGURE 9 – Affectation des poids en fonction du nombre de vote

Les intervalles de votes utilisés pour établir les poids ont été répartis sur la base de la médiane et des quartiles 1 et 3. En effet, 25% des votes ont moins de 20 votes, et 75% en ont plus de 84 et ce, tels que répartis sur la boîte à moustache ci-contre :



Pour les raisons qui précèdent, nous accordons ainsi plus de poids aux propositions ayant plus de 84 votes, et moins de poids à celles qui en admettent moins de 20. Partant de cela, on remarque de prime abord que le rapport le plus élevé est d'environ 0,40 . Ce qui signifie que selon la pondération établie, les propositions ont tendance à ne pas apporter une totale unanimité.
Par ailleurs, il arrive que pour plusieurs propositions ayant le même poids, la valeur de l'indicateur calculé précédemment soit la même, ce qui complique la comparaison. Pour palier à ce problème, nous ajoutons un second indicateur qui permettra de comparer dans un second temps les propositions ayant été classées ex æquo selon le premier indicateur :

$$tri = unanimite * |nombre\ de\ votes\ d'accord - nombre\ de\ votes\ pas\ d'accord|$$

A cet effet, parmi les propositions qui apportent le moins de divergence, nous avons :

```
Il faut valoriser la végétalisation de l'alimentation, meilleure pour la santé, la planète et les autres êtres vivants.
2184 votes pour.        4 votes contre.        22 votes neutres.
-------------------------------------------------------------------------------------------------------------------------
Il faut abandonner l'élevage industriel afin de respecter une éthique concernant la souffrance animale.
2435 votes pour.        8 votes contre.        10 votes neutres.
-------------------------------------------------------------------------------------------------------------------------
Il faut arrêter l'élevage intensif et mieux surveiller les conditions d'élevage
3490 votes pour.        18 votes contre.       31 votes neutres.
-------------------------------------------------------------------------------------------------------------------------
```

FIGURE 10 – Propositions de plus de 2000 votes apportant peu de divergence

A l'opposé il existe un grand nombre de propositions pour lesquelles le rapport *unanimite* est nul ou proche de 0. Ces propositions correspondent à celles qui apportent le plus de divergence. Parmi elles, nous avons :

```
Il faut libérer le travail notamment hors limite de 35h et primer les résultats collectifs des salariés
238 votes pour.  225 votes contre.       128 votes neutres.
--------------------------------------------------------------------------------------------------------------------------------
Il faut mettre en place un revenu universel pour que le travail soit un choix et non plus une obligation
390 votes pour.  308 votes contre.       117 votes neutres.
--------------------------------------------------------------------------------------------------------------------------------
Il faut créer un mouvement de désobéissance civile, force citoyenne décisive pour redéfinir les priorités de nos sociétés.
453 votes pour.  332 votes contre.       183 votes neutres.
--------------------------------------------------------------------------------------------------------------------------------
```

FIGURE 11 – Propositions de plus de 500 votes apportant de la divergence

- Analyse des tags manuels en fonction des classes d'âges Nous allons maintenant nous intéresser aux tags liés aux propositions. Les propositions ayant été taguées de manière manuelle, nous allons donc partir de cet état de fait pour savoir quels tags sont les plus utilisés. Nous pourrons par la suite voir si le classement varie en fonction d'une certaine catégorisation des citoyens.

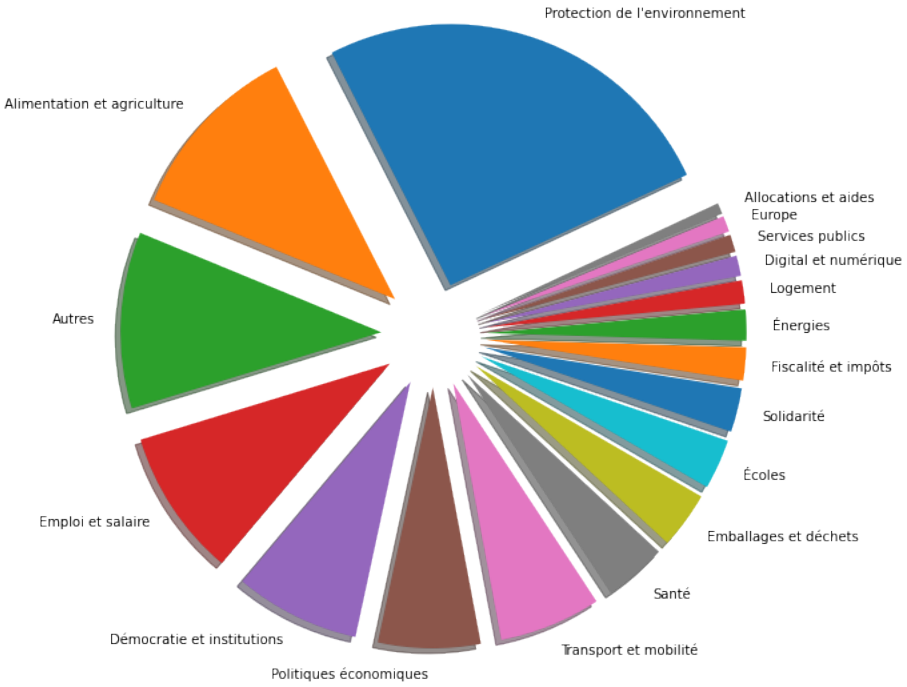


FIGURE 12 – Répartition des tags/thématiques

Les tags en question sont au nombre de 18 et tel qu'on peut le lire sur le diagramme précédent, la plupart des propositions concernent la protection de l'environnement, puis l'alimentation et l'agriculture. Il y a cependant très peu de propositions concernant par exemple les aides et allocations. Par ailleurs, plus de 10% des propositions n'ont pas de catégories spécifiques.

Ensuite, en s'interrogeant sur le fait de savoir quelle catégorie d'âge effectue le plus de propositions, nous avons défini cinq classes d'âge tel qu'on peut le voir sur le tableau ci-dessous :

	Classes âge	Intervalles	Effectifs	Fréquence
0	Jeunes	[13-25]	653	0.09
1	Jeunes actifs	[26-39]	2225	0.30
2	Actifs	[40-61]	3458	0.47
3	Jeunes retraités	[62-71]	802	0.11
4	Retraités	[72-86]	159	0.02

FIGURE 13 – Répartition des classes d'âge

Les actifs (entre 40 et 61 ans) sont ceux qui effectuent le plus de propositions. Il serait par ailleurs intéressant de savoir quels sont les domaines dans lesquelles les actifs effectuent le plus de propositions :

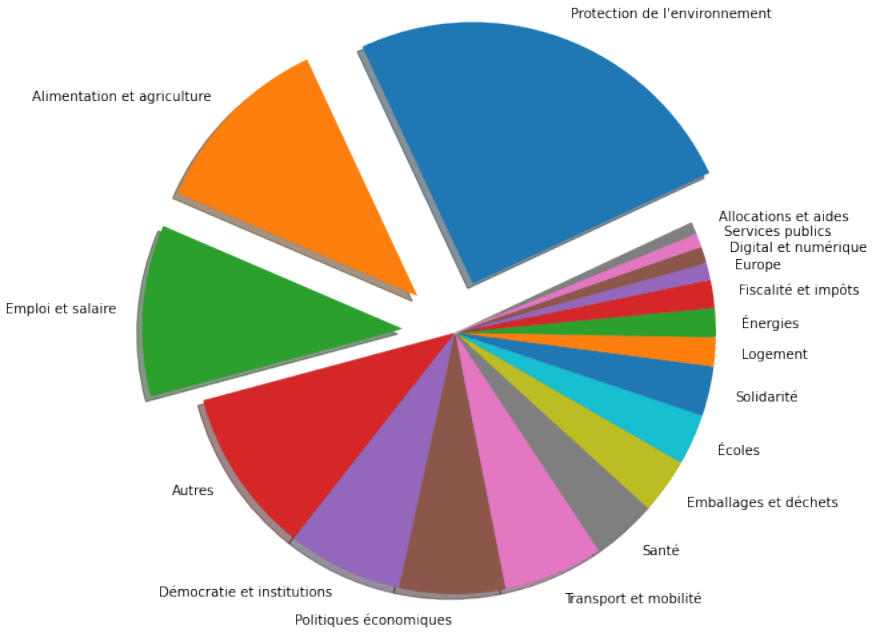


FIGURE 14 – Thématiques auxquelles s'intéressent les actifs

Les propositions faites par les actifs concernent en majorité la protection de l'environnement. Après, viennent les thématiques liées à l'alimentation et l'agriculture, puis celles liées à l'emploi et au salaire. Ce qui est en adéquation avec les problématiques qui concernent les citoyens appartenant à cette tranche d'âge. Cependant, les thématiques les moins abordées sont celles liées aux allocations et aides, puis aux services publics. Pour voir la répartition des tags les plus utilisés selon les tranches d'âge, se référer aux figures 36 à 39 en annexes.

2 Méthodologie du travail et préparation des données

Comme nous l'avons précisé depuis l'introduction l'objectif de ce travail est de modéliser les idées citoyennes (propositions). Dans un premier temps, cela se fera sur la base des tags manuels à partir de la modélisation supervisée puis par le bais de la modélisation non supervisée par word embedding.
De manière détaillée, le travail sera effectué selon la méthodologie ci-dessous :

2.1 Méthodologie du travail

Afin de venir à bout de notre objectif, nous adopterons les étapes suivantes :

— Préparation des données : Avant de se lancer dans une quelconque forme de modélisation, compte tenu des données textuelles à disposition, il s'agira d'utiliser le Natural Langugage Processing afin de préparer les propositions à leur exploitation future ;

— Modélisation supervisée : Il s'agira, une fois les données prêtes, d'utiliser un classifieur afin de catégoriser les propositions. En effet, l'on considéra les données comme étiquetées, X représentant les propositions et Y les tags manuels ;

— Modélisation non supervisée : L'objectif sera toujours de classer les propositions mais cette fois, en considérant les données non étiquetées. C'est à dire que les tags existants ne seront pas pris en compte dans le processus de classification des propositions.

2.2 Préparation des données

L'ensemble des propositions constituent un vaste dictionnaire de 1 787 102 mots. L'image ci-dessous permet d'avoir une représentation visuelle de la fréquence des mots : plus le terme revient fréquemment parmi les propositions, plus sa taille est importante. On peut voir que les propositions sont fortes de verbes d'action telles que développer, créer, arrêter mais également font mention des entreprises, de la santé etc...

FIGURE 15 – Nuage des mots

La préparation de cet ensemble de mots est une phase très importante lors de la classification textuelle. Elle consiste en différentes étapes conduisant à la représentation vectorielle des mots. En effet, dans un premier temps, des techniques découlant du Natural Language Processing vont permettre de scinder les phrases en mots bien spécifiques , de réduire au maximum le dictionnaire de mots en éliminant les "superflus" et également de mettre ces mots sous des formes permettant d'effectuer autant que possible les rapprochements entre les mêmes mots utilisés sous différentes formes. Par la suite, différents indicateurs seront utilisés afin de transformer ces mots sous formes numériques, exploitable par les classifieurs.

2.2.1 Préparation par NLP

Le Natural Language Processing désigne un champ de l'intelligence artificiel permettant de lire, comprendre et interpréter le langage humain. Il s'agit là donc de tout un processus qui fait intervenir différentes techniques telles que la tokénisation, la stemmatisation et la lemmatisation, la suppression des stopwords, pour ne citer que ceux-là car nous y auront recours dans le cadre de notre travail.

1. **La Tokénisation** désigne un processus de séparation d'une phrase ou d'un document en une séquence d'unité mots tout en écartant la ponctuation. A titre illustratif :

```
input :
Il faut fournir des masques gratuitement pour protéger la vie
output :
['Il', 'faut', 'fournir', 'des', 'masques', 'gratuitement', 'pour', 'protéger', 'la', 'vie']
```

2. **La suppression des stopwords** : Les stopwords désignent les mots du vocabulaire qui en soit n'apportent aucune information telles quel les prépositions (la, le, ce,...), les verbes d'état sous différentes formes, etc. Après la tokénisation, ces stopwords sont donc supprimés ;

```
input :
['Il', 'faut', 'fournir', 'des', 'masques', 'gratuitement', 'pour', 'protéger', 'la', 'vie']
output :
['fournir', 'masques', 'gratuitement', 'protéger', 'vie']
```

3. **La stemmatisation** quant à elle permet d'obtenir la racine des mots. Exemple : continua, continue, continuer deviennent après stemmatisation continu.

```
input :
['Il', 'faut', 'fournir', 'des', 'masques', 'gratuitement', 'pour', 'protéger', 'la', 'vie']
output :
['fourn', 'masqu', 'gratuit', 'proteg', 'vi']
```

4. **La lemmatisation** est un processus par lequel le mot est transformé en sa forme la plus simple. Un mot au pluriel ramené au singulier, un mot conjugué ramené à sa forme infinitive etc...

Aucune règle ne précise le recours à la lemmatisation et/ou à la stemmatisation. En effet, le recours à ces méthodes dépend de leur incidence sur la précision. Dans le cas de notre travail, nous avons eu de meilleurs résultats en appliquant les 2 méthodes.

2.2.2 Représentation vectorielle des mots

Il existe plusieurs méthodes permettant de représenter les mots sous forme vectorielle. Dans la suite, nous présenterons 4 techniques auxquelles, nous avons eu recours selon la situation en présence :le Bag of Word (BoW), le TF-IDF (Term Frequency — Inverse Document Frequency),le Word2vec et le Doc2vec.

1. **Bag of Words** Le bag of words est une simple méthode de vectorisation consistant à compter le nombre d'apparition d'un mot dans chaque proposition. L'application du bag of words sur les données traitées selon les méthodes préalablement présentées nous permet d'obtenir le graphe ci-dessous :

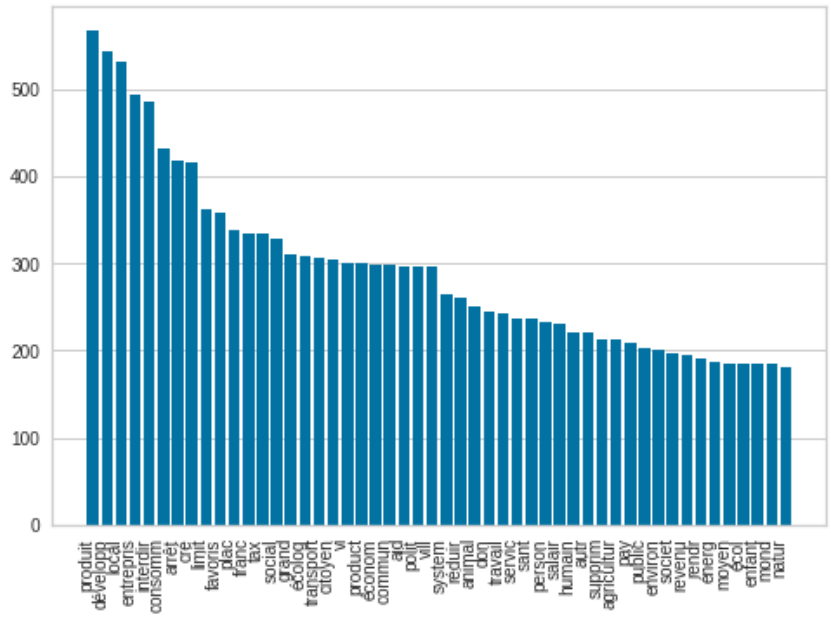


FIGURE 16 – Fréquence de distribution des 50 premiers tokens - BoW

A noter qu'il s'agit là du cumul pour l'ensemble des propositions. Contrairement au nuage de mots précédents, l'on peut voir qu'une fois débarrassés des formes conjugués, pluriel etc. des mots, les tokens tels que produit, developp, local etc. ont un poids plus important.

2. **TF-IDF** C'est une technique permettant de quantifier l'importance d'un mot dans un document à partir du poids qui lui est attribué. Spécifiquement il s'agit d'une combinaison de 2 indicateurs : le Term Frequency (TF) qui donne la fréquence d'un mot dans chaque proposition et l'IDF qui est le logarithme du rapport entre le nombre de propositions et le nombre de proposition contenant le mot. Le TF-IDF permet de surpondérer les mots qui apparaissent fréquemment dans certains documents mais peu dans d'autres. Formellement, on a :

$$tf(t,d) = \frac{\text{Nombre de répétitions du mot t dans le document d}}{\text{Nombre de mots du document d}}$$

$$idf(t, D) = log(\frac{\text{Nombre total de documents D}}{\text{Nombre de documents contenant le mot t}})$$

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

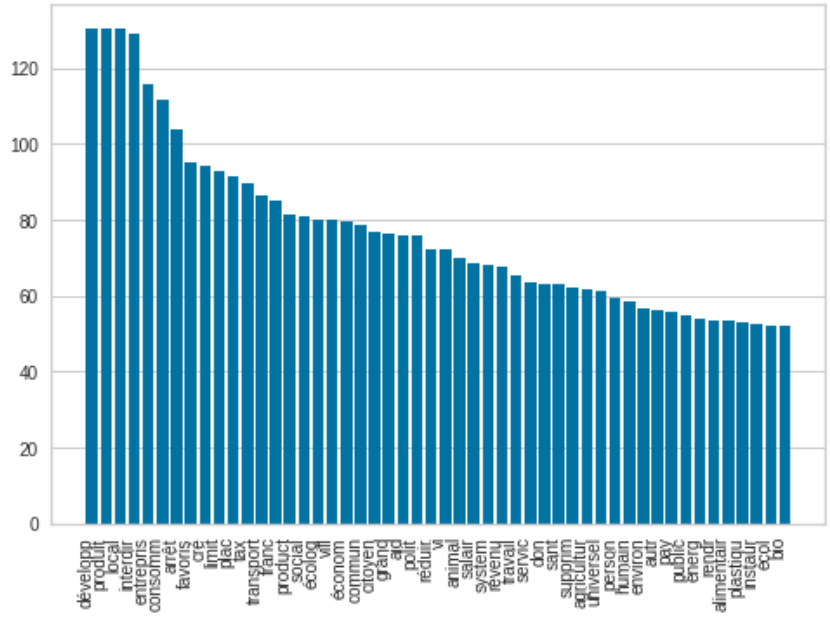


FIGURE 17 – Fréquence de distribution des 50 premiers tokens - TfIdf

Si l'on prend en compte le classement en termes des mots ayant le plus de poids, les résultats de la vectorisation par Tf-Idf sont assez proches de ceux par bag of words.

3. **Word2vec** Le Word2vec est un algorithme basé sur les réseaux de neurones, dont le but est de fournir une représentation vectorielle des mots contenus dans un texte. L'algorithme du Word2vec présente deux architectures selon l'objectif que l'on souhaite atteindre : le CBOW et le Skip-Gram.
   - Le CBOW (Continuous Bag Of Words) : sert à prédire un mot en se basant sur son contexte.
   - Le Skim-Gram : sert à prédire le contexte d'un mot à partir du mot en question.

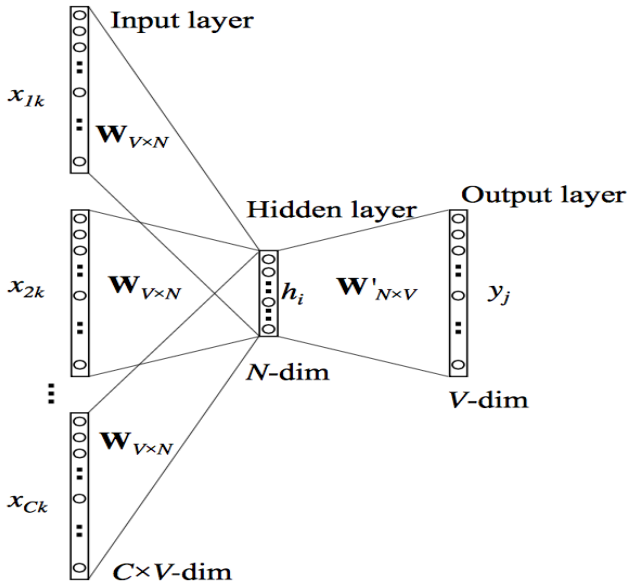


FIGURE 18 – Architecture CBOW [2]

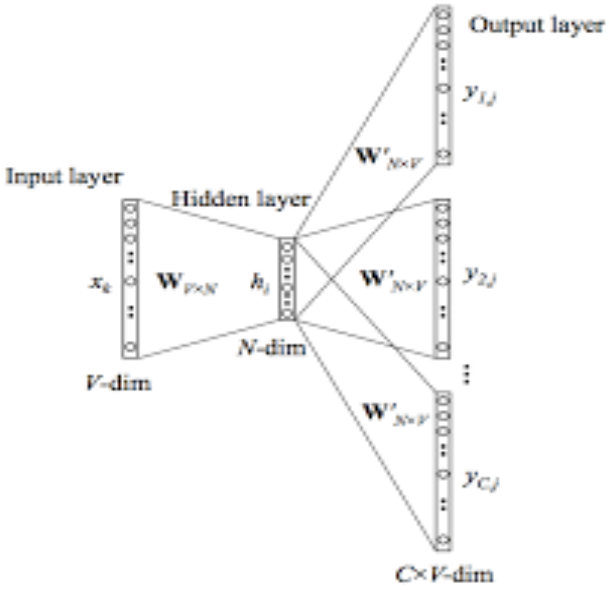


FIGURE 19 – Architecture Skip-Gram [4]

Pour créer un contexte de mots propre à nos données, l'entraînement de l'algorithme sur nos données nécessite un traitement préalable selon les méthodes décrites précédemment (au moins la tokenisation).
De manière concrète, lors de l'application du word2vec sur nos données, le paramètre le plus important auquel nous nous sommes intéressés est la taille (ou dimension) des vecteurs-mots car il s'agit d'un algorithme sensible à la taille des vecteurs de mots.

```
Vectorisation du mot 'consommer' en utilisant le contexte fournit par nos propositions pour une taille de vecteur fixée à 25:

array([-0.8212638 ,  0.0322031 ,  0.28801504, -0.33461335, -0.01937155,
       -0.16372244,  0.03719271,  0.48569146, -0.6687895 , -0.52594715,
        0.15809698,  0.14464705, -0.06718724,  0.16539797, -0.20546836,
        1.8276292 ,  0.5870471 , -0.13685633, -0.6675726 ,  0.2646178 ,
        0.7620475 ,  1.1521982 ,  0.88241893, -0.69257337,  0.8822716 ],
      dtype=float32)
```

S'agissant d'un algorithme qui prend en compte le contexte, deux mots ayant des contextes similaires seront représentés par des vecteurs proches en terme de distance.

```
La distance entre les mots 'consommer' et 'manger' est : 0.03858309984207153
La distance entre les mots 'consommer' et 'finir' est : 0.33345818519592285
Dans notre contexte, le mot 'consommer' est donc plus proche du mot 'manger' que du mot 'finir'.
```

Ainsi, plus la distance est grande, plus les mots sont sémentiquement éloignés. A l'opposé, plus l'indice de similarité est élevé, plus les mots sont similaires.

```
En termes de similarité, les mots les plus proches du mot 'consommer' sont les suivants:

[('manger', 0.9614169001579285),
 ('produire', 0.9492008090019226),
 ('poisson', 0.9464389681816101),
 ('supporter', 0.9198810458183289),
 ('maniere', 0.9006558060646057),
 ('facon', 0.8983810544013977),
 ('sol', 0.8961219191551208),
 ('acheter', 0.8945857882499695),
 ('detruire', 0.8924869298934937),
 ('considerer', 0.8898935317993164)]
```

Enfin, il est important de préciser que dans le cadre de notre travail, l'objectif étant de passer d'une proposition à un vecteur (et non d'un simple mot à un vecteur), la démarche a consisté à représenter une proposition par la moyenne de ses mots ayant préalablement été vectorisés.

4. **Dord2vec** Le Doc2vec quant à lui est un algorithme similaire au Word2vec. Si le but final du Word2vec est de représenter un mot par un vecteur, le Doc2vec servira quant à lui, à représenter un document par un vecteur. La particularité du Doc2vec est qu'il prend en considération un vecteur supplémentaire : le vecteur de document (ID du document). Ainsi, l'ID du document est pris en compte lors de l'apprentissage des vecteurs de mots. Par ailleurs, tout comme le Word2vec, le Doc2vec présente deux architectures :
   - Le DM (Distributed Memory) qui est similaire au CBOW, sert à prédire un mot en se basant sur son contexte ainsi que l'ID du document.
   - Le DBOW (Distributed Bag-Of-Words) qui est similaire au Skip-gram.

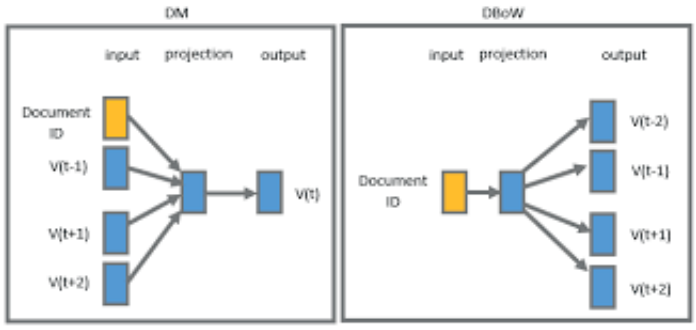


FIGURE 20 – Illustration Doc2vec [3]

Dans notre cas, une proposition correspond à un document. Dans l'exemple qui suit, l'apprentissage du modèle a été effectué sur nos données :

```
Dimension fixée à 25
input: 'arrêter la pollution'
output:  [-0.01308893 -0.00406206 -0.01801518  0.005209   -0.00362839 -0.00793862
 -0.00659705  0.01394398  0.0074224   0.0126859  -0.00986351 -0.01743358
 -0.00566566 -0.0076683   0.00968755 -0.0172959   0.01850751 -0.01982349
  0.00363616  0.00563082  0.01404888  0.00065341 -0.00615105 -0.01265833
 -0.00024807]
```

3 Modélisation supervisée

La modélisation supervisée a fait l'objet de multiples travaux exploratoires. En effet, la $1^{ère}$ difficulté qui nous a été donné d'appréhender est le fait que les propositions puissent être associées à plus d'un tag. En effet, 17,6% des propositions sont associées à au moins 2 tags. Cette contrainte nous a fait considérer 2 formes de classification textuelle : la classification multi label et la classification multi classe. Bien qu'une seule des 2 n'a finalement été concluante, nous présenterons sommairement ce qui a été fait afin de parvenir aux résultats. Par ailleurs, un arbitrage a également été nécessaire en fonction du type de vectorisation apportant les meilleurs résultats.

3.1 Présentation des méthodes

3.1.1 formes Classification

1. **La classification multi label** : Une classification est dite multi label lorsqu'un document (texte, proposition etc.) peut être rangé dans plus d'une catégorie. Il s'agit d'une configuration particulière car la plupart des classifieurs ne sont pas originellement conçus pour cadrer avec. Afin donc d'effectuer la classification, chacune des 17 modalités est transformée en dummies (1 lors que la proposition est associée à ce tag et 0 sinon - voir figure 50 en annexe) le problème est transformé selon les 3 formes ci dessous :
   - Binary Relevance : Pour ce cas, la classification est réalisée avec chacune des variables créée. Cela revient donc à faire 17 classifications binaires indépendantes ;
   - Classifier Chain : Cette transformation est similaire à la précédente à la seule différence que la dépendance entre les modalités (nouvelle variable binaire créée) est prise en compte. En effet, une fois le classifieur choisi, les propositions sont entraînées avec le 1^{ieme} label, la combinaison des 2 permet d'entraîner le second, ainsi de suite ;
   - Label Powerset : Cette méthode transforme le problème en modélisation multi class. Chaque combinaison de label (vecteur de taille de 17 pour chacune des propositions) est considéré comme une classe distincte et unique. Sous cette configuration, l'on obtient 391 classes. Plus tard nous allons écarter les tags qui reviennent le moins(moins de 1%), on sera ainsi rendu à 40 classes dont la fréquence de distribution est représentée dans le graphique ci-dessous :

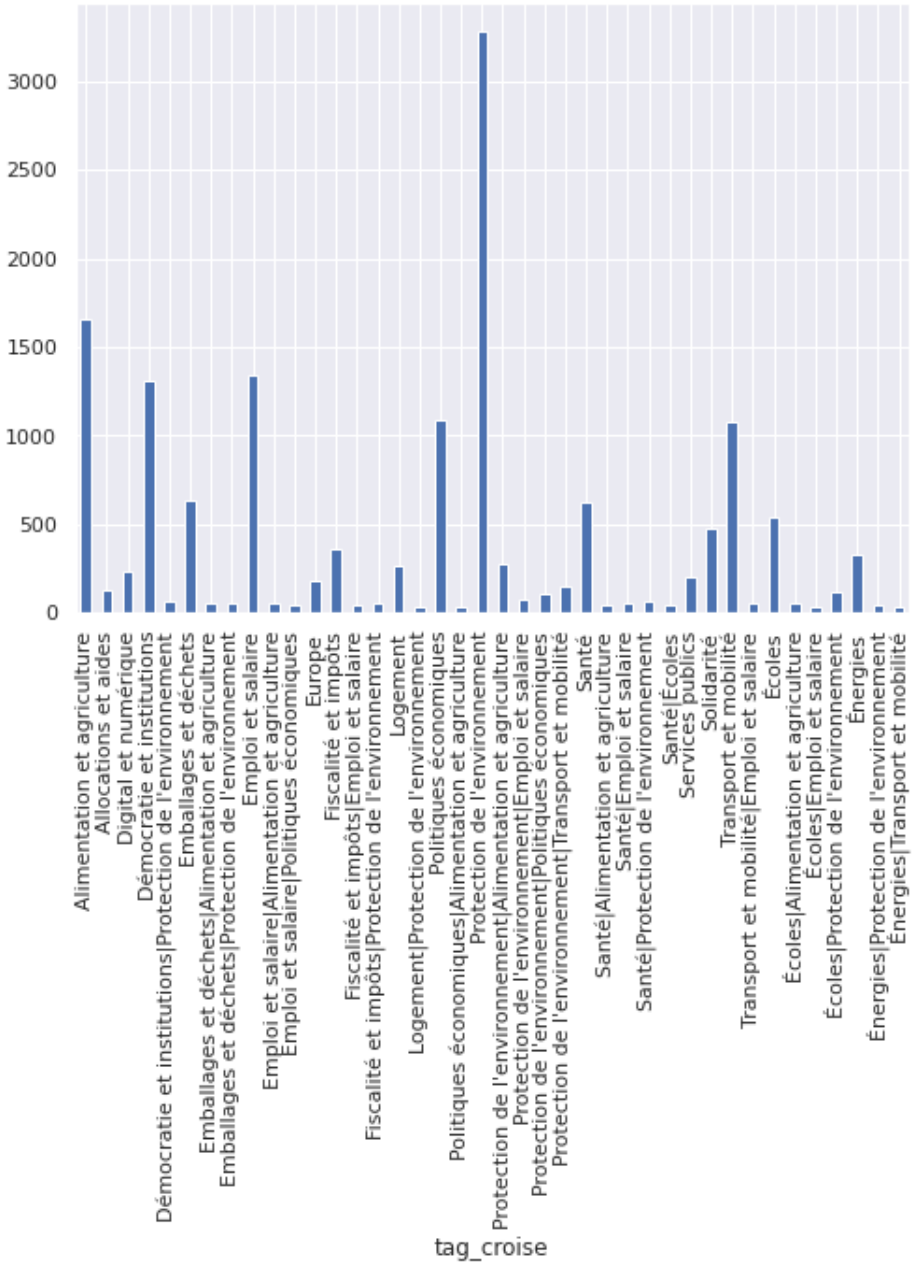


FIGURE 21 – Distribution des propositions associées à au moins 2 tags

2. **La classification multi classe** : Nous avons également considéré la classification multi classe en éliminant l'ensemble des propositions étant associés à au moins 2 tags ; L'objectif étant d'explorer l'ensemble des possibilités afin d'avoir les meilleurs résultats possibles. Dans ce cas, la variable cible reste unique et constitué des 17 tags dont nous avons fait la description en 1 ère partie du travail.

3.1.2 Classifieur : Bayésien naïf multiomial

De façon générale, le terme bayésien naïf renvoie à une hypothèse d'indépendance, très peu réaliste, mais qui conduit généralement à de bons résultats. Si l'on considère un problème de classification textuelle avec d un document appartenant à l'espace X des documents et $c = c_1, c_2, ..., c_j$ un ensemble fixe de classes. On a :

$$P(d|c) = P(t_1, t_2, ..., t_n|c)$$

où $t_1, t_2, ..., t_n$ désigne la séquence des mots (vectorisés) tels qu'il apparaît dans le document. Ainsi, le bayésien naïf suppose que les valeurs (vectorisées) des mots sont indépendantes les une des autres conditionnellement aux classes. Formellement cela se traduit par :

$$P(t_1, t_2, ..., t_n|c) = \prod_{k=1}^{n} P(t_k|c)$$

Cette hypothèse étant soutenue, la probabilité qu'un document d appartienne à la classe est donné par :

$$P(c|d) \propto P(c) \prod_{k=1}^{n} P(t_k|c)$$

$P(c)$ désigne la probabilité à priori que le document soit catégorisé dans la classe c tandis que $P(c|d)$ désigne la probabilité à posteriori. Jusque là, nous ne savons rien de la loi de distribution associée aux mots donc la probabilité à posteriori est inconnue. C'est à partir de là que le terme "multinomial" prend tout son sens car l'on va considérer que chaque $P(t_k|c)$ est une distribution multinomiale et on décide ainsi de l'appartenance d'une classe à un document, en l'assignant à la classe qui maximise la probabilité à postériori estimée (connaissant la loi) :

$$\hat{c} = argmax\hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c)$$

3.2 Résultats de la modélisation supervisée

3.2.1 Résultats de la classification multi label

1. **Classifieur : Bayésien naïf multinomial**

   Dans cette partie, le modèle a été compilé avec 2 jeux de données différents. En effet, le premier jeu de données inclut toutes les combinaisons de tags associée aux propositions : une proposition pour 2 tags, 3 tags, etc. (le maximum étant 6) ; On est ainsi rendu à 391 combinaisons distinctes. Pour ce qui est du deuxième jeu de données, nous avons éliminé les associations les moins fréquentes (moins de 1%) car vu leur faible fréquence, le modèle peut, vraisemblablement, difficilement bien les classer.
   Avec les premières données, nous pouvons constater que la précision est particulièrement faible pour les 2 premières transformations, contrairement au label powerset, qui bien que pas idéal se démarque avec une précision de 46,28%. Avec le deuxième jeu de donnée, dont la distribution des propositions par classe correspond à la figure 21 précédente,

les classes restent assez mal balancées ce qui découle sur une performance très peu variante. On lit néanmoins dans le tableau ci-dessous que la transformation par label powerset demeure la meilleure.

TABLE 2 – Taux de bonne prévision selon la transformation (%)

Transformation	Binary relevance	Classifier Chain	Label Powerset
Jeu de données 1	23,25	28,95	46,28
Jeu de données 2	21,97	27,64	48,27

2. **Classifieur : KNN**

   Les 2 premières transformations se sont avérées tout aussi peu pertinentes avec le KNN. Cependant, pour ce qui est du label powerset, qui considère chaque combinaison de tags comme un tag distinct, la précision en est meilleure. Nous obtenons avec 20 plus proches voisins, une précision 61%.

Prendre en compte le fait que les propositions soient associées à plusieurs tags ne semble être pertinent que lorsque chacune des combinaisons de tags associée à chaque proposition est considérée comme une classe à part entière. Malgré une précision de 61% obtenu avec le KNN, il demeure qu'il s'agit là de modéliser 40 tags pour lesquels les propositions sont assez mal balancée (non équilibrée, figure 19). Nous décidons donc d'éliminer les 2947 propositions associées à au moins 2 tags et donc d'explorer la classification multi classe.

3.2.2 Résultats de la classification multi classe

1. **Classifieur : classifieur bayésien naif**
   Il a été compilé sur les données vectorisées d'une part à partir du bag of words et d'autre part à partir du TfIdf. Nous avons obtenu de meilleurs résultats, par validation croisée 10 blocs avec les valeurs numériques issues du BoW. Les résultats sont condensés dans le tableau ci-dessous :

TABLE 3 – Taux de bonne prévision selon la vectorisation

Vectorisation	Précision(%)
BoW	71
TfIdf	67

2. **Classifieur : K plus proche voisin (KNN)**
   Nous avons appliqué une validation croisée 10 blocs pour le choix du nombre de voisins ; Il en découle les graphiques ci-dessous en fonction de la vectorisation appliquée :

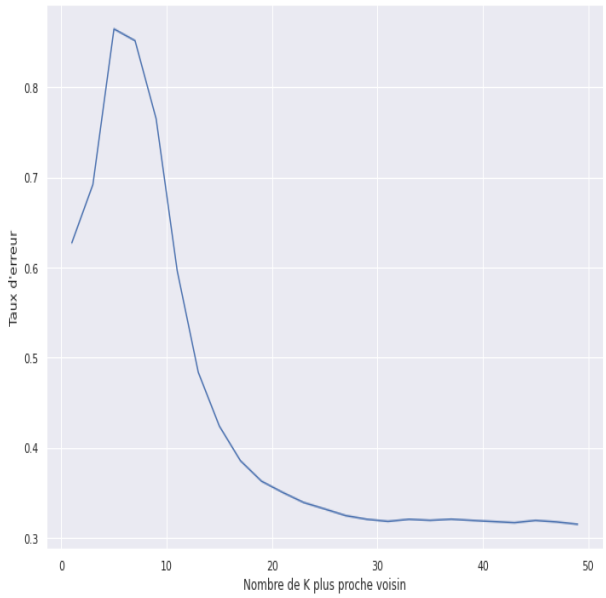


FIGURE 22 – Nombre de voisins optimal - TfIdf

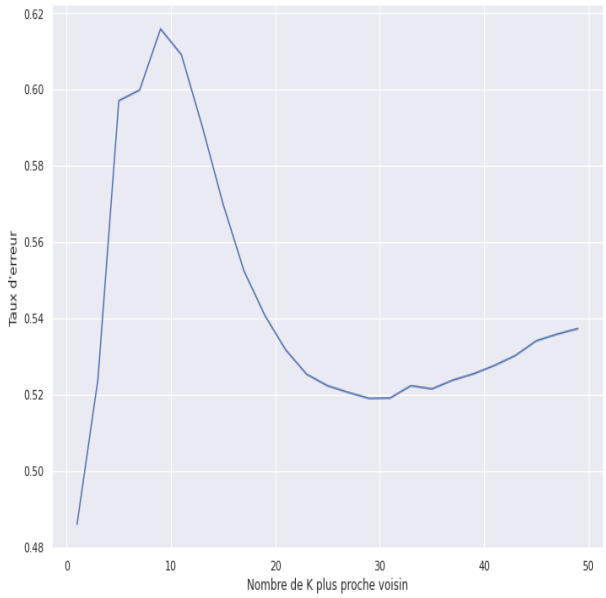


FIGURE 23 – Nombre de voisin optimal BoW

La figure ci-dessus indique k = 49 comme nombre de plus proches voisins optimal, pour ce qui est des données issues de la vectorisation à l'aide du TfIdf et k = 1 pour le BoW. Avec le tfidf (figure 22), le gain en précision n'étant que très léger, nous avons retenu 25 plus proches voisins et le modèle KNN a abouti à une précision de 67%. Avec la vectorisation par BoW par contre, les résultats sont peu satisfaisants : une précision de 48% avec 1 plus proche voisin.

En prenant en compte les modèles ayant offerts la meilleure précision, NB multinomial avec BoW et KNN avec Tfidf, nous avons représenté les matrices de confusion afin d'explorer les résultats. On peut voir sur les matrices ci-dessous, que peu importe le modèle, l'algorithme est particulièrement confus lorsqu'il s'agit de distinguer entre les autres tags (alimentation et agriculture, énergie, etc.) du tag Protection de l'environnement. Cela émane sûrement du fait que la fréquence d'association des propositions à ce tag est largement supérieure aux autres.

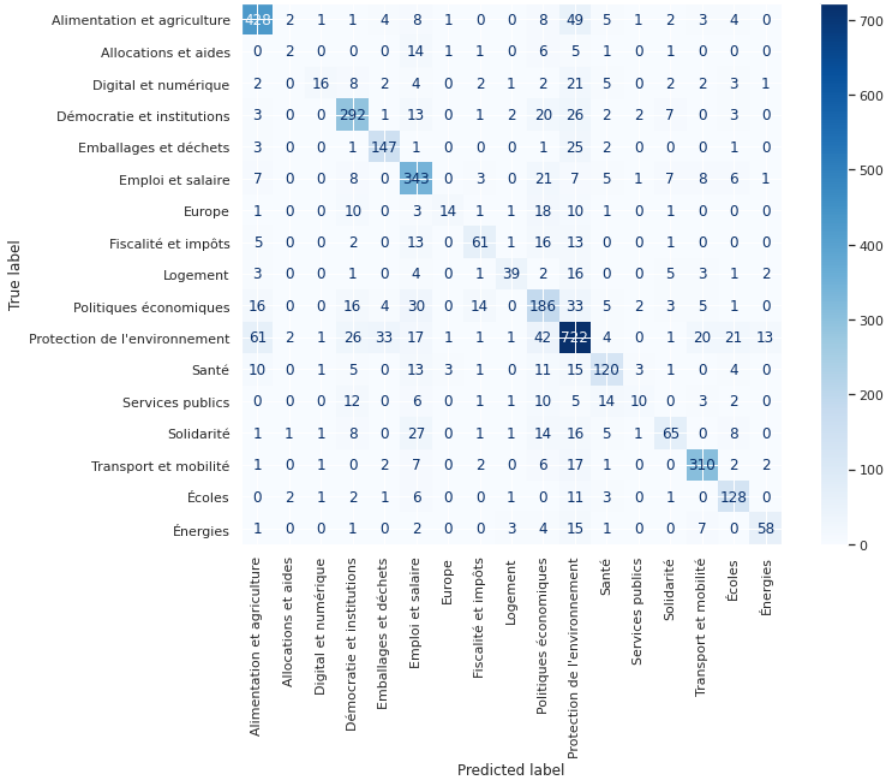


FIGURE 24 – Matrice de confusion NB multinomial + BoW

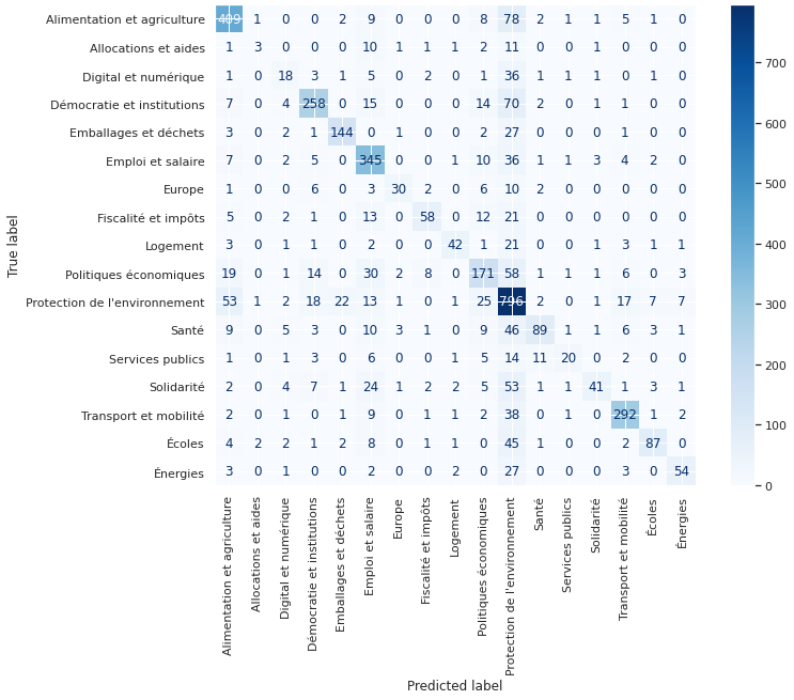


FIGURE 25 – Matrice de confusion KNN + TfIdf

4 Modélisation non supervisée

La modélisation non supervisée regroupe l'ensemble des techniques permettant de créer des classes "homogène" de proposition. Plus précisément dans cette partie, nous disposons des données d'entrée (X) mais pas de donnée de sortie (Y), l'objectif étant de modéliser la structure ou la distribution sous-jacente dans les données afin d'en apprendre davantage sur les données.
Pour ce faire, la plupart des algorithmes calculent des distances entre les observations (propositions) et regroupent ainsi les groupes de propositions les plus proches. Il faut donc être en mesure de quantifier la similarité ou la distance entre deux observations. Cette première étape peut parfois être la plus difficile de tout le processus de classification, mais elle est essentielle et demeure le premier pas de toute analyse de partitionnement. Un bon partitionnement devra permettre d'avoir une distance intra-groupe faible et une distance inter-groupe élevée.

Dans le cadre de notre travail, trois méthodes d'apprentissage non supervisée ont été implémentées : le Kmeans (ou K-moyennes), Latent Dirichlet Allocation (LDA), le Nearest Neighbors Search et la Nonnegative Matrix Factorization (NMF).

4.1 Présentation des méthodes

4.1.1 Kmeans (K-moyennes)

La méthode des Kmeans s'applique dans des cas où les p variables d'entrée sont numériques (et habituellement standardisées). L'objectif de la méthode est de partitionner les données en K groupes homogènes, elle nécessite la connaissance de la valeur de K à l'avance. De manière spécifique, cet algorithme se présente comme suit :

- choix du nombre de groupes K ;
- répartition aléatoire des observations dans les K groupes ;
- calcul des centroïdes (centre de classe ou vecteur-moyenne) pour chacun des K groupes.

$$\mu_k^i = \frac{1}{N_k} \sum_{j \in G_k} x_i^j$$

  k = 1,...K et i = 1,...p avec $\mu_k = (\mu_k^1, \mu_k^2, ...\mu_k^p)$
- calcul de la distance entre chaque observations et chacun des K centroïdes.
- on assigne chacune des n observations au groupe dont le centroïde est le plus près.
- on répètes les trois dernières étapes jusqu'à ce qu'aucune observation ne puisse être affectée à un nouveau groupe (on dit que l'algorithme converge).

Dans le cadre de notre étude, la vectorisation utilisée est le TF-IDF et le Word2Vec. La distance euclidienne fait office de mesure de proximité.

Choix du nombre de groupe K

Pour déterminer le nombre optimal de groupes K à utiliser dans la modélisation, nous allons appliquer la méthode dite du "coude" (The elbow method) qui examine le pourcentage de variance expliqué en fonction du nombre de groupes K. Elle consiste à choisir un certain nombre de groupes K afin que l'ajout d'un nouveau groupe ne donne pas une meilleure modélisation des données. Plus précisément, si l'on trace le pourcentage de variance expliqué par les groupes par rapport au nombre de groupe, les premiers groupes ajouteront beaucoup d'informations (expliquent beaucoup de variance totale), mais à un moment donné le gain marginal diminuera, donnant un angle dans le graphique. Le nombre de groupes est choisi à ce niveau, d'où le "critère du coude". Toutefois ce coude ne peut pas toujours être identifié sans ambiguïté.

Qualité du modèle

L'analyse de silhouette est utilisée pour étudier la séparation entre les groupes résultants. Elle indique la proximité entre un point d'un groupe et un groupe voisin : une valeur de 1 indique une séparation nette des groupes, une valeur de 0 indique que les groupes sont proches, et une valeur négative indique que les points ont été potentiellement affectés au mauvais groupe.

Le coefficient de silhouette est calculé en utilisant la distance moyenne intra-cluster (a) et la distance moyenne au groupe le plus proche (b) pour chaque échantillon, autrement dit, b est la distance entre un point et le groupe le plus proche auquel il n'est pas assigné. Le coefficient de silhouette pour un échantillon donnée est calculé comme suit :

$$\frac{(b-a)}{max(a,b)}$$

4.1.2 Latent Dirichlet Allocation (LDA)

Le modèle Latent Dirichlet Allocation (LDA) est un modèle probabiliste qui permet de décrire des collections de documents de texte ou d'autres types de données discrètes. La LDA fait partie d'une catégorie de modèles appelés “thème models”, qui cherchent à regrouper par thème des vastes archives de documents. Ceci permet d'obtenir des méthodes efficaces pour le traitement et l'organisation des documents : organisation automatique des documents par sujet, recherche, compréhension et analyse du texte ...etc. Le modèle LDA se base sur un calcul bayesien hiérarchique qui suppose que :

- chaque document peut être défini par une distribution (masqués) de thème ;
- chaque thème est défini comme une distribution de mots ;
- la probabilité a posteriori de ces variables latentes étant donnée, la collection de documents détermine une décomposition cachée de la collection des thèmes.

Si on suppose que :

— M est le nombre de documents,

— N_i est le nombre de mots dans le document i (avec N la taille du corpus) ;

— α est le paramètre de Dirichlet de la distribution a priori des thème par document ;

— β est le paramètre de Dirichlet de la distribution a priori des mots par thème ;

— θ_i est la distribution des thèmes pour le document i ;

— ψ_k est la distribution du thème k ;

— $z_{d,n}$ est le thème du n^{ieme} mot dans le document d $w_{d,n}$;

L'algorithme utilisé pour un document w spécifique est le suivant :

1. choisir $\theta_i \hookrightarrow Dirichlet(\alpha)$; avec $i \in 1, ...M$
2. choisir $\psi_k \hookrightarrow Dirichlet(\beta)$, avec $k \in 1, ...K$;
3. Pour chaque mot $w_{i,j}$:
   - choisir un thème $z_{i,j} \hookrightarrow Multinomial(\theta_i)$
   - choisir un mot $w_{i,j} \hookrightarrow Multinomial(\psi_{z_{i,j}})$

La loi de Dirichlet permettant de tirer une variable θ, telle que $\forall \theta_i \geq 0$ et $\sum_{i=1}^{k} \theta_i$. Sa densité s'écrit :

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_1 - 1}$$

Cette distribution permet donc d'obtenir une distribution multinomiale de paramètre , correspondant pour LDA au mélange de topics d'un document w. Ainsi, la probabilité jointe d'un mélange de thème θ, des N thèmes z et de N mots w s'écrit :

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n})$$

Connaissant les paramètres α et β, le rôle de l'inférence consiste à estimer les variables cachées θ et z_n d'un document w, connaissant la liste des mots w_n du document. Les principales méthodes d'inférence (approchée) pour LDA sont les méthodes d'échantillonage (notamment le collapsed Gibbs sampling voir) et les méthodes variationnelles (principalement les méthodes mean-field (voir ici) qui peuvent se faire en batch ou en ligne). On peut ensuite avoir recours à ce procédé d'inférence pour estimer les paramètres α, β et ψ du modèle grâce à l'algorithme d'Espérance-Maximisation (EM).

Choix du nombre de Thème

Pour que la LDA puisse être applicable il faut lui indiquer le nombre de thèmes au préalable avant de lancer la procédure. Pour ce faire, nous allons avoir recours au score dit de "cohérence" de la LDA. Les sujets sont considérés comme cohérents si la totalité ou la plupart des les mots, par exemple les N premiers mots du sujet, sont liés. Il s'agit ici de calculer le score pour différentes valeurs du nombre de thèmes et de retenir le nombre de thèmes qui correspond au score le plus élevé avant que la valeur du score commence à décroître.
Le score se calcule comme suit :

$$Score = \frac{1}{n} \sum_{i<j} \hat{m}_{cos(nlr,1)}(wi, wj)$$

Avec,

$$\hat{m}_{cos(nlr,1)}(w_i, w_j) = s_{cos}(\overrightarrow{v}_{nlr,1}(wi), \overrightarrow{v}_{nlr,1}(wj)$$

et

$$s_{cos}(\overrightarrow{v}, \overrightarrow{u}) = \frac{\sum_{i=1}^{|W|} u_i \times w_i}{||\overrightarrow{u}||_2 \times ||\overrightarrow{w}||_2}$$

et

$$\overrightarrow{v}_{nlr,1}(w_i) = (m_{nlr}(w_i, w_j))_{j=1,...|W|}$$

en prenant

$$m_{nlr}(w_i, w_j) = \frac{log\frac{P(w_i,w_j)+\epsilon}{P(w_i)*P(w_j)}}{-log(P(w_i, w_j)) + \epsilon}$$

ϵ est de l'ordre de 10^{-12} pour ne pas avoir des valeurs indéfinies, et $P(w_i)$ la probabilité associée au mot w_i et $P(w_i, w_j)$ la probabilité associée aux deux mots w_i et w_j.

Pour facilité la mise en place de la méthode, les bigrammes et les trigrammes ont étés utilises. Les bigrammes sont deux mots qui apparaissent fréquemment ensemble dans le document. Pour les trigrammes ils sont au nombre de 3. Pour regrouper les mots en bigrammes la formule suivante est utilisée :

$$Score(w_i, w_j) = \frac{Nombre(w_i, w_j) - \delta}{Nombre(w_i) \times Nombre(w_j)}$$

δ est un coefficient permettant de d'éliminer les croisements qui sont non significatifs. On regroupe en bigramme les mots pour lesquels le score est supérieur à un seuil fixé. Le même principe est utilisé pour les trigrammes.

4.1.3 Nonnegative Matrix Factorization (NMF)

La NMF est donc une technique de réduction de dimension adaptée aux matrices creuses contenant des données positives, par exemple des occurrences ou dénombrements de mots. Elle très utilisée dans le cas de l'analyse textuelle. Elle consiste à décompose une matrice de données en un produit de deux matrices ne contenant que des valeurs positives ou nulles et dont le produit rapproche la matrice des données. Soit une matrice document-termes A $\mathbb{R}^{m \times n}$ représentant m termes uniques présents dans un corpus de n documents. La méthode de la NMF génère une matrice réduite de rang k approximant la matrice A comme produit de deux matrices ayant des termes non négatifs $A \approx W \times H$ en minimisant l'erreur de reconstruction entre A et la matrice de faible dimension approximant A. Les colonnes de $W \in \mathbb{R}^{m \times k}$ peuvent être interprétées comme des sujets définis avec des poids non négatifs relatifs aux m termes. La matrice $H \in \mathbb{R}^{k \times n}$ est celle permettant de croiser les documents aux sujets.

La méthode du NMF s'applique parfaitement aux matrices résultant d'une vectorisation de type TF-IDF.

Le choix du nombre de thèmes est basé sur la recherche d'un optimum local au problème de minimisation suivant :

$$min[L(X, WH) + P(W, H)]$$

Ou L est une fonction perte mesurant la qualité d'approximation et P une fonction de pénalisation optionnelle. L est généralement un critère de moindres carrés et P est une pénalisation optionnelle de régularisation utilisée pour forcer les propriétés recherchées des matrices W et H, par exemple, la parcimonie des matrices ou la régularité des solutions.

Plusieurs algorithmes dans la littérature permettent d'obtenir les matrice W et H optimales mais chacun d'entre eux possèdent des critères de convergences différents. Toutefois, c'est la famille des moindres carrés alternés (ALS) qui est plus utilisée dans la littérature. Cette famille exploite le fait que si le problème n'est pas convexe en à la fois W et H, il l'est soit en W soit en H. Il se présente comme suit et possède de bonnes propriétés (convergence, complexité).

- Tirer W = random(n, r)
- Pour i = 1 à Maxi_ter
  - Résoudre en H : $W^tWH = W^tX$
  - Mettre à 0 les termes négatifs de H
  - Résoudre en W : $HH^tW^t = HX^t$
  - Mettre à 0 les termes négatifs de W

4.1.4 Nearest Neighbors Search

La recherche du plus proche voisin est une méthode simple qui s'applique généralement en classification et qui permet, pour un ensemble donné, de trouver le point le (ou les) plus proche(s) d'un point donné. L'évaluation de cette proximité se base sur un distance entre les points.
La recherche du plus proche voisin d'un vecteur x donné dans un ensemble E peut être perçu comme suit : un problème d'optimisation dont l'objectif est de minimiser une fonction de coût qui correspond à la distance entre le vecteur x et chacun des vecteurs y de E. Ainsi,

$$y* = \arg\min_{y \in E}(d(x, y))$$

Le calcul de la distance nécessite le choix d'une métrique qui peut être la distance euclidienne ou encore la distance Manhattan, la distance cosinus, etc... Ci-après les formules correspondant à quelques métriques [1]. Pour n points et deux vecteurs x et y donnés, on a :

- Distance euclidienne : $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distance de Manhattan : $\sum_{i=1}^{n}|x_i - y_i|$
- Distance cosinus : $Cos(\theta)$ où θ est l'angle formé par les deux vecteurs. Cette métrique se base sur un produit scalaire.

L'application de la recherche du plus proche voisin nécessitera ainsi un traitement préalable des propositions consistant à les transformer en vecteur. D'où l'utilisation des résultats obtenus par le TF-IDF, le Word2vec, et enfin, le Doc2vec.

4.2 Résultats de la modélisation non supervisée

4.2.1 Résultats du Kmeans

La première étape avant de réaliser la méthode des Kmeans consiste à déterminer le nombre de groupes (ou thèmes) optimal pour effectuer la modélisation. Les graphiques 26 et 27 représentent le taux de variance expliqué par rapport au nombre de groupes K ("règle du coude") avec les vectorisations TF-IDF et Word2Vec respectivement.

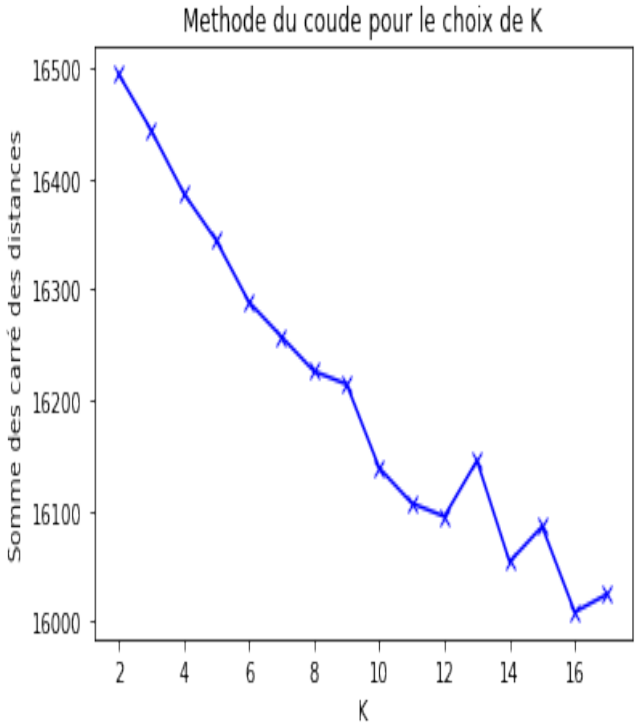


FIGURE 26 – Méthode du coude pour le choix de K avec le TF-IDF

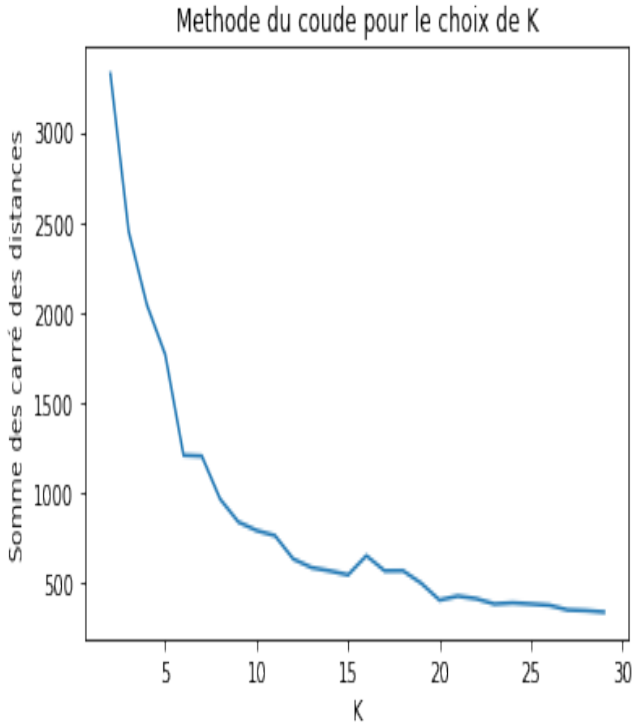


FIGURE 27 – Méthode du coude pour le choix de K avec le Word2vec

La perte d'inertie semble s'être stabilisé autour de la K = 11 pour le TF-IDF et K = 18 pour le Word2vec, c'est donc ces valeurs qui seront retenues dans la suite.
Le tableau 4 (voir annexe) représente le nombre de propositions affectées dans chacun des groupes crées avec la vectorisation TF-IDF. Il est ainsi immédiat de constaté qu'un nombre important de propositions sont affectées aux groupes 2 et 0 qui regroupent à eux seuls environ 70% du nombre total de propositions. D'autre part, la vectorisation Word2vec 5 (voir annexe) indique une répartition entre les groupes plus homogènes avec des nombres de propositions par groupes autour de 1000.

Les résultats du test de la silhouette renvoie la valeur de 0,008 avec la vectorisation TF-IDF, ce qui indique que les groupes créés ne sont pas bien distinguables. Autrement dit, les distances entre les centres de groupes ne sont pas très grandes et donc le modèle ne discrimine pas bien les propositions. Pour ce qui est du Word2vec, la valeur du test de

silhouette correspondant à cette modélisation est de 0,264. Cette valeur indique une meilleur séparation des groupes comparativement à la modélisation effectuée avec le TF-IDF.
Tout ceci, est confirmé graphiquement grâce à l'ACP effectué dans les figures 28, 29. En effet, l'on peut remarquer que les groupes formés grâce à la vectorisation Word2vec semblent être plus distinguables par rapport à ceux associés au TF-IDF.

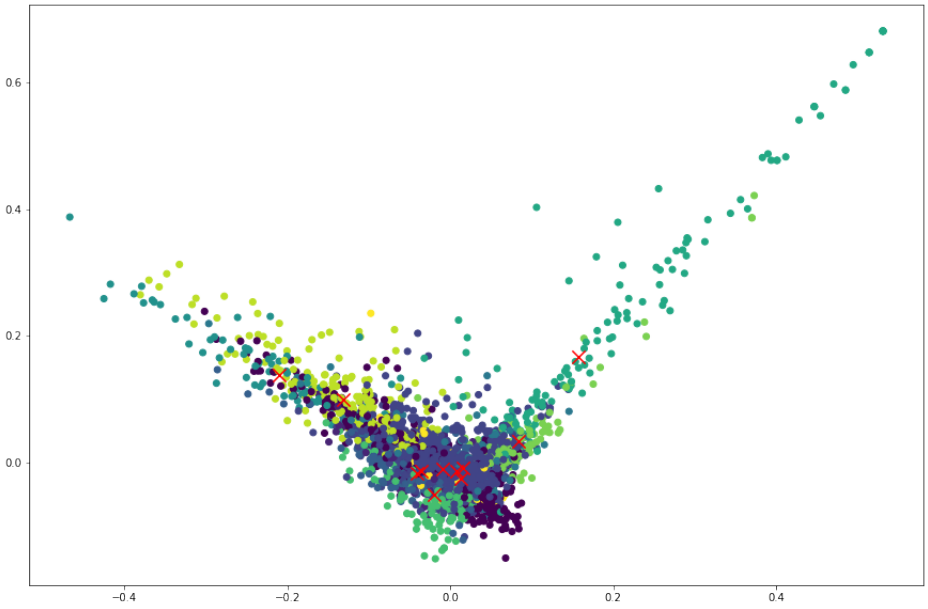


FIGURE 28 – ACP sur les groupes résultants du Kmeans avec le TF-IDF

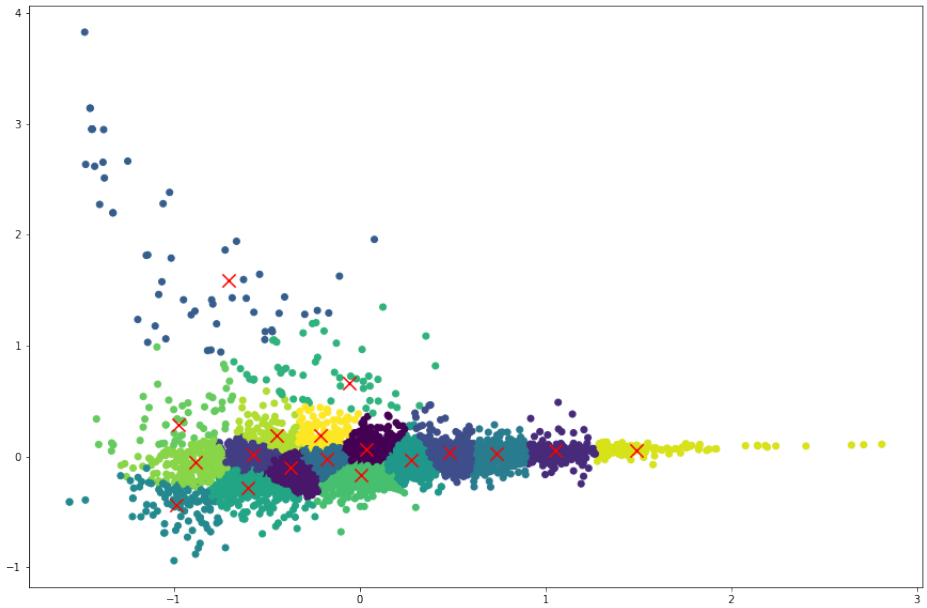


FIGURE 29 – ACP sur les groupes résultants du Kmeans avec le Word2vec

Compte tenue de ce qui précède, l'on peut conclure que les résultats de la vectorisation

Word2vec sont bien meilleurs que ceux du TF-IDF.

Le regroupement effectué grâce au Word2vec permet de remarquer que les groupes 0, 1, 9 et 13 regroupent les propositions relatives au recyclage, à la pollution, la protection de l'environnement et l'écologie, la promotion l'agriculture local pour le groupe 2, l'éducation des enfants pour le groupe 3, le bien être social et personnes âgées pour le groupe 4, énergie pour le groupe 5, la politique pour le thème 6, la rémunération salariale pour les groupe 7 et 15, transport pour le groupe 14 etc... C'est résultats sont représenter dans les figures 41, 42, 43, 44 (voir annexes).
Bien qu'associer le Word2vec à la méthode des Kmeans par rapport au Tf-Idf a amélioré les résultats, il en demeure que plusieurs propositions sont assez souvent hors contexte par rapport aux autres présentes dans le même groupes. D'ailleurs la valeur du score de silhouette reste faible.

4.2.2 Résultats du LDA

Le modèle LDA a fait l'objet de multiples allers et retours. Nous avons par exemple supprimer la stemmatisation de l'étape de pré-traitement car les résultats issus de cela étaient assez pauvres. Par ailleurs, la liste des stopwords a été renfloués afin d'éliminer les mots qui sont très souvent utilisés mais qui n'apportent aucune information (favoriser, France, francais, etc.).
Le graphique 30 suivant représente le score de cohérence suivant le nombre de thèmes. Cette courbe commence à décroître à partir de l'abscisse 20, c'est donc la valeur de 20 qui sera retenue comme nombre de thèmes dans la modélisation. Le score de cohérence correspondant est de 0,4614 indiquant une cohérence moyenne entre les sujets.

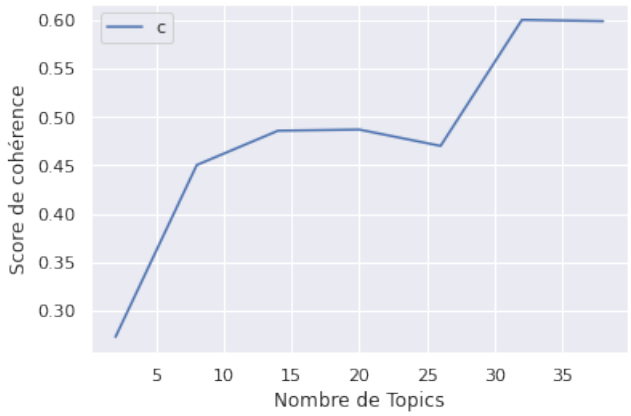


FIGURE 30 – Score de cohérence suivant le nombre de thèmes

L'histogramme suivant 31 représente pour chacun des 20 sujets, le nombre de propositions qui y sont classées. Il en ressort une distribution assez déséquilibrée. En effet, le thème 13 apparait comme englobant la majorité des thématiques (plus de 50%) ; Le reste se répartissant de façon presque équivalente entre les autres propositions.

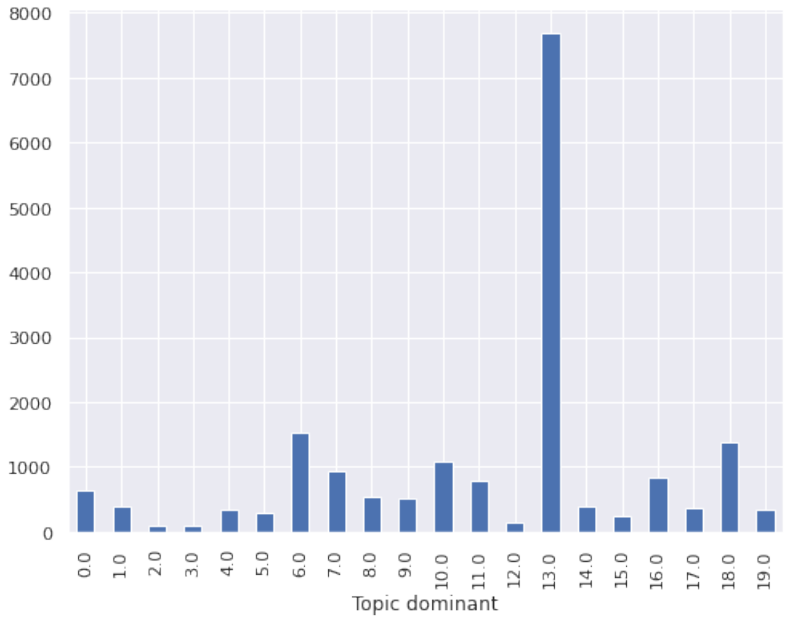


FIGURE 31 – Distribution des propositions selon le thème - LDA

Le graphique ci-dessous va nous permettre d'analyser les mots portés par les thèmes prédominants.

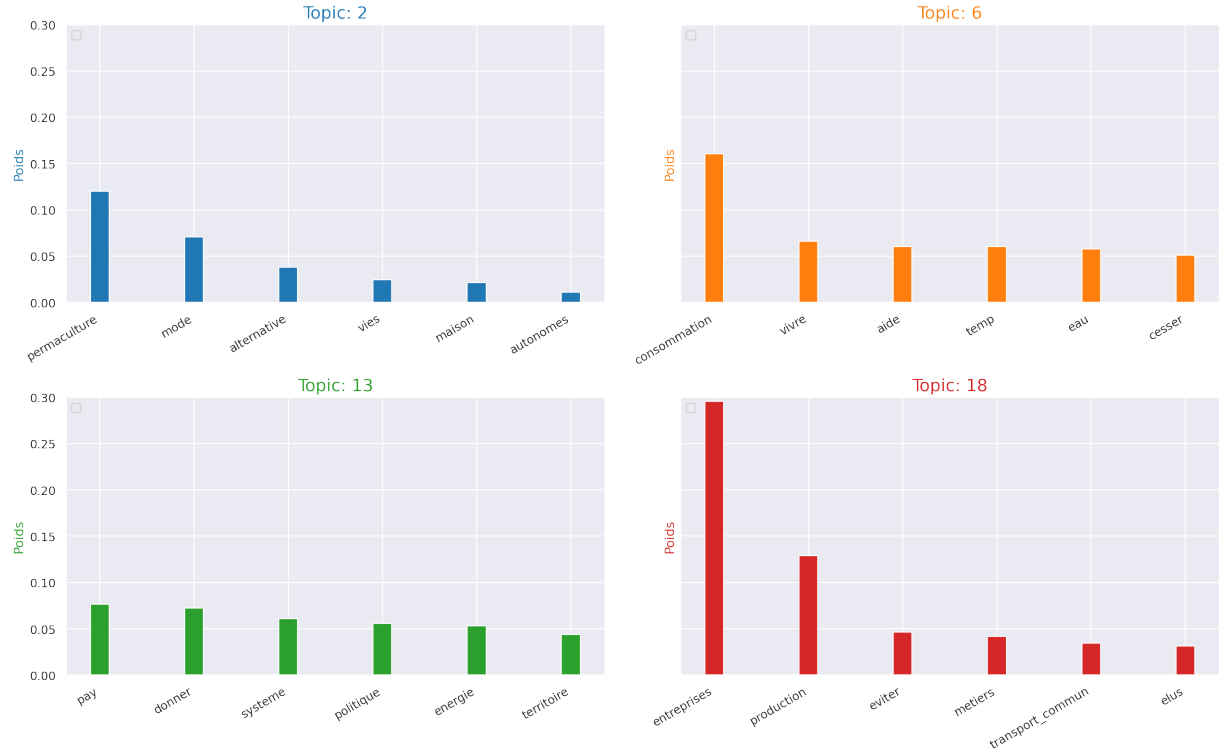


FIGURE 32 – poids des mots clés par thème

Les mots obtenus pour chacun des thèmes, donnent une idée, bien que pas assez flagrante,

du thème général porté par le thème. En effet, à la lecture du graphe, on peut voir qu'une fois les mots mis en commun, le thème 13 référerait aux affaires étatiques, le thème 18 a tout ce qui a trait aux activités de production incluant les entreprises et le thème 6 des aspects plus sociaux. La figure 50 en annexe a permis de zoomer sur les propositions contenues dans le thème 13. Comme on pouvait s'y attendre, le thème est beaucoup trop général et renferme des propositions relatives à des thématiques, parfois tout à fait opposées. Nous explorons par la suite une nouvelle méthode afin d'obtenir des thèmes moins englobants.

4.2.3 Résultats du Non-Negative Matrix Factorization (NMF)

Suite aux résultats pas assez convaincants du LDA, le NMF a été compilé. Tout d'abord les propositions ont été transformées sous forme de vecteur à l'aide du Tfidf afin d'alimenter le modèle. Les étapes de préprocessing demeurent les mêmes que celles effectuées pour le LDA (tokénisation, lemmatisation et suppression des stopwords incluant certains verbes ou adverbes n'apportant pas d'informations concrètes). Nous avons sélectionné 20 classes qui correspondent à un score de cohérence de 0.51 ; Le tableau ci-dessous présente les classes obtenues avec les sujets dominants :

	numero_topic	Topic
0	0.0	produits locaux alimentaires tva importés saison prix
1	1.0	revenu universel base minimum salaire vivre inconditionnel
2	2.0	transport commun gratuité marchandises gratuits train polluants
3	3.0	réduire drastiquement pollution déchets inégalités télétravail nombre
4	4.0	circuit court producteurs distribution alimentaires locaux alimentation
5	5.0	consommation viande mode animaux changer énergie biens
6	6.0	agriculture bio biologique pesticide agriculteurs permaculture vers
7	7.0	entreprises télétravail aide grandes salariés obliger polluantes
8	8.0	production relocaliser locale alimentaire médicaments nécessité stratégiques
9	9.0	arrêter animaux élevage intensif penser sauvages construire
10	10.0	travail temp repenser travailler partager vivre semaine
11	11.0	limiter déplacements avion nombre maximum voyage télétravail
12	12.0	place remettre taxe mise système verre carbone
13	13.0	santé éducation environnement enfants dès école système
14	14.0	plastique emballages usage unique déchets verre consigne
15	15.0	taxer fortement transaction financières dividendes kérosène avion
16	16.0	service public civique moyens obligatoire biens remettre
17	17.0	consommer local produire saison manger bio niveau
18	18.0	économie locale circulaire réelle solidaire monnaies proximité
19	19.0	villes grandes centre voitures espaces végétaliser cyclables

FIGURE 33 – Thèmes issus du modèle Nmf

Les différents sujets semblent concerner des thématiques assez spécifiques bien que l'on peut constater globalement que les questions de l'environnement, consommation, production locale englobent la plupart. C'est d'ailleurs ce que matérialise l'histogramme ci-dessous :

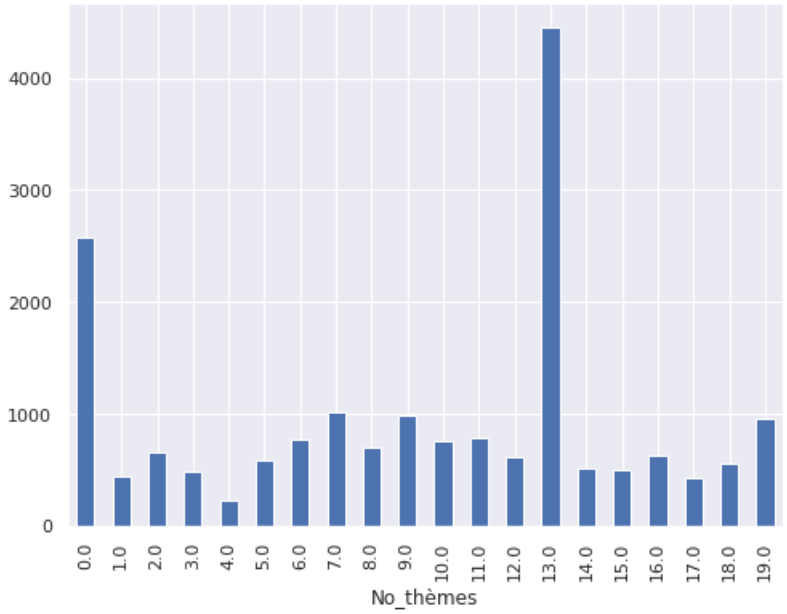


FIGURE 34 – Distribution des propositions selon le thème - Nmf

Les propositions touchant les aspects sociaux tels que l'environnement, l'éducation prédominent. Il en est de même pour celles ayant trait à la production locale. Afin d'explorer davantage les résultats, nous avons fait ressortir quelques propositions associées à des thématiques. Bien qu'il s'agisse d'un échantillon peu représentatif, les propositions ci-dessous sont cohérentes avec le thème et n'apparaissent pas totalement opposées dans leurs idées intrinsèques.

```
Topic 5 : consommation viande mode animaux changer énergie biens
---
Il faut prioriser les modes de vies alternatives, comme la permaculture, maison autonomes...
Il faut trouver des solutions face aux inégalités d'accès aux ressources alimentaires, faire face aux disparités de consommation .
Il faut être plus exigeant sur la qualité et quantité la nourriture : provenance, mode culture (huile de palme, avocat) viande, poisson...
Il faut penser local et agir localement, créer du lien dans une consommation saine et durable de chaque instant
Il faut modifier l'offre de biens de consommation et ne délocaliser que les productions critiques dans les régions et départements français
--------------------------------------------------------------------------------
Topic 7 : entreprises télétravail aide grandes salariés obliger polluantes
---
Il faut imposer un cahier des charges plus strict aux entreprises (production, récupération, réparation, recyclage), bref cercle vertueux.
Il faut que l'Etat travaille plus en coopération avec les associations environnementales.
Il faut rendre légale l'aide à mourir dans la dignité. Toute personne doit avoir le droit de cesser de vivre quand bon lui semble.
Il faut diminuer les bénéfices versés aux actionnaires.
Il faut demander des comptes sur les aides attribuées aux entreprises
--------------------------------------------------------------------------------
Topic 13 : santé éducation environnement enfants dès école système
---
Il faut gérer les ressources terrestres avec intelligence et non à profit.
Il faut permettre à chacun.e d'avoir un accès gratuit à des soins primaires de qualité axés sur la prévention et la prise en charge globale
Il faut éduquer sur l'intérêt, pour notre santé et la santé de notre planète, des régimes alternatifs (flexitariens, végétariens, vegans)
Il faut réglementer très strictement les allégations santé sur les aliments exotiques qui viennent de loin: chia, goji, quinoa...
Il faut investir massivement dans l'éducation afin de donner aux citoyens les moyens intellectuels de penser le changement.
--------------------------------------------------------------------------------
Topic 18 : économie locale circulaire réelle solidaire monnaies proximité
---
Il faut que l'industrie change pour s'engager réellement dans l'économie circulaire : proposer des produits et des services durables !
Il faut soutenir l'économie réelle (et non le système financier), redévelopper les productions sur le territoire et l'économie circulaire
Il faut éduquer nos enfants à l'économie circulaire et à la gestion des ressources naturelles.
Il faut ne tenir compte que de l'économie réelle et bannir la finance
Il faut créer des monnaies locales pour inciter la population à acheter localement et que cela participe au développement du territoire
```

FIGURE 35 – Extrait de propositions par thèmes - Nmf

4.2.4 Résultats du Nearest Neighbors search

Tel que mentionné précédemment, la méthode du Nearest Neighbors Search s'applique sur des vecteurs. C'est pourquoi nous allons appliquer l'algorithme de la recherche des plus proches voisins sur trois représentations vectorielles différentes dans le but comparer les résultats et retenir la méthode qui a le mieux fonctionné sur nos données.

```
# Nearest Neighbors Search
from sklearn.neighbors import NearestNeighbors
knn = NearestNeighbors(n_neighbors=10, metric='cosine')
knn.fit(features)
```

Le paramètre *features* changera en fonction de la méthode de vectorisation utilisée, et contiendra la forme vectorisée de l'ensemble des propositions. Le nombre de voisins recherchés pour chaque proposition a été fixé à 10, mais peut être ajustée selon les besoins. La métrique choisie ici est la distance Cosinus car c'est celle qui est la plus utilisée dans le domaine de l'analyse de textes. Par exemple, dans le cas du TF-IDF, les indices des 10 plus proches voisins des deux premières propositions sont :

```
knn.kneighbors(features[0:2], return_distance=False)
```

```
array([[    0,  7076,  2681, 15214,   247,   848, 13952, 12213, 12255,
        10433],
       [    1,  4110, 12985, 11859, 10714, 11416, 13432, 13223, 14880,
         3648]], dtype=int64)
```

Il est aussi possible d'afficher la liste des distances entre les propositions en passant le paramètre *return_distance* à *True*. Ainsi, on se basera sur la liste des indices obtenus pour afficher les propositions voisines.

Le tableau ci-dessous est un récapitulatif des résultats obtenus à l'aide de la recherche des plus proches voisins.

Récapitulatif du Nearest Neighbors Search	
Méthodes de vectorisation	**Commentaire des résultats** [1]
TF-IDF	- Résulats intéressants. - Regroupe les propositions qui possède les mêmes mots sans considérer le contexte dans lequel les mots sont employés.
Word2vec	- Bons résultats. - Regroupe les propositions en prenant en compte le contexte des mots.
Doc2vec	- Résultats moins pertinents qu'avec les deux autres méthodes. - Certaines propositions regroupées ensemble ne semblent pas avoir de point commun.

Globalement, le TF-IDF combiné à la recherche des plus proches voisins peut fonctionner, mais il semble plus adéquat dans les problématiques où on cherche à regrouper des phrases selon les mots qu'elles contiennent sans porter une attention particulière de la phrase. Cela n'étant pas le cas pour nous, le meilleur choix (meilleur et non parfait) serait donc la combinaison du Word2vec (la proposition étant représentée par le vecteur correspondant à la moyenne des mots la contenant) avec la recherche des plus proches voisins.
Enfin, bien que cette démarche ne permette pas de générer un tag pour une proposition donnée, elle permet de déterminer les propositions qui lui sont similaires. Ainsi, cela pourrait éventuellement servir de base pour la création d'un programme (voir d'une application) qui attribuera à chaque nouvelle proposition entrée sur le site, le tag de son plus proche voisin.

1. Voir les figures 45 à 49 en annexes pour illustration de chacun des commentaires.

Conclusion

Depuis novembre 2013, le monde est confronté au défi sans précèdent que représente la maladie du Covid-19, les communautés et les économies dans le monde entier tentent tant bien que mal de ralentir voir stopper la propagation du virus. Les avancées importantes réalisées dans la lutte contre cette pandémie dont principalement la découverte d'un vaccin, laissent envisager que cette pandémie n'en a plus pour longtemps.

Face à cela, les chercheurs du monde entier s'interrogent sur la manière dont l'économie doit se reformer pour faciliter la reprise mais surtout quelles sont les enseignements à retenir de la pandémie. Cette question cruciale suscite aussi bon nombre d'avis au niveau des citoyens.
En France, la plateforme Make.org a recueilli environ 18 681 propositions concernant une multitude de thématiques sur la question. L'objectif de notre étude était de modéliser les thématiques issus des divers propositions faites par les citoyens.

La première partie de notre étude consistait à effectuer une modélisation supervisée des propositions en prenant comme variable cible les tags manuellement saisis par les agents de Make.org. L'idée étant qu'une fois un modèle obtenu, les agents de make.org n'auraient besoin de tagger manuellement qu'une partie des propositions ; le modèle supervisé permettant de tagger le reste. Dans ce contexte, nous avons exploré différentes possibilités en considérant aussi bien différents classifieurs (KNN, NB multinomial), types de vectorisation (TfIdf, BoW) ou encore types de classification textuelle (multi label, multi classe). Il en ressort qu'une modélisation multi classe effectuée sur la base du bayésien naïf et associé à une vectorisation de type BoW a été le plus concluant avec une précision 71%. En explorant les matrices de confusion, nous pensons que le balancement déséqulibré des tags (en faveur de la protection de l'environnement) peut avoir eu un impact sur la précision des modèles.

La seconde partie consistait à effectuer une modélisation non supervisée de ces propositions dans l'idée pour make.org de ne plus avoir besoin de tagger manuellement les propositions car le modèle le ferait automatiquement (admettant biensur, une perte évidente de précision). Encore ici différentes perspectives ont été exploré car les modèles étant très sensibles à l'étape de prétraitement. Il en ressort globalement, des résultats assez différents du méthode à l'autre mais néanmoins le modèle NMF semble produire les résultats les plausibles. La principale difficulté dans cette partie réside dans la comparaison des différentes méthodes.

Au regard de ces résultats, nous pouvons proposer dans le cadre de la mise en place d'un outil d'automatisation des tags au sein de Make.org, une démarche supervisée avec une approche multi classe et incorporant le bayésien naif multinomial en classifieur et le TfIdf comme moyen de vectorisation préalable. La modélisation non supervisée restant plus immédiate mais moins précieuse et nécessitant un prétraitement très poussé.

Annexes

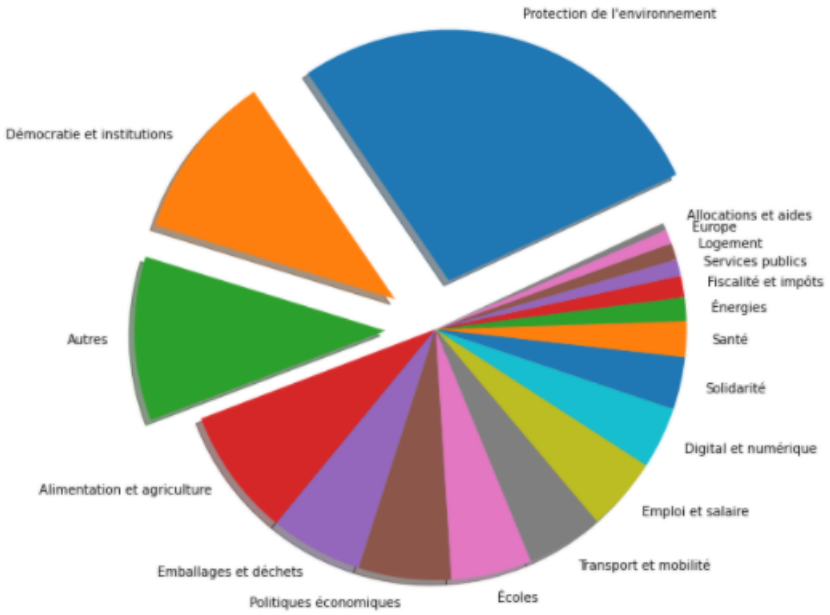


FIGURE 36 – Thématiques auxquelles s'intéressent les jeunes

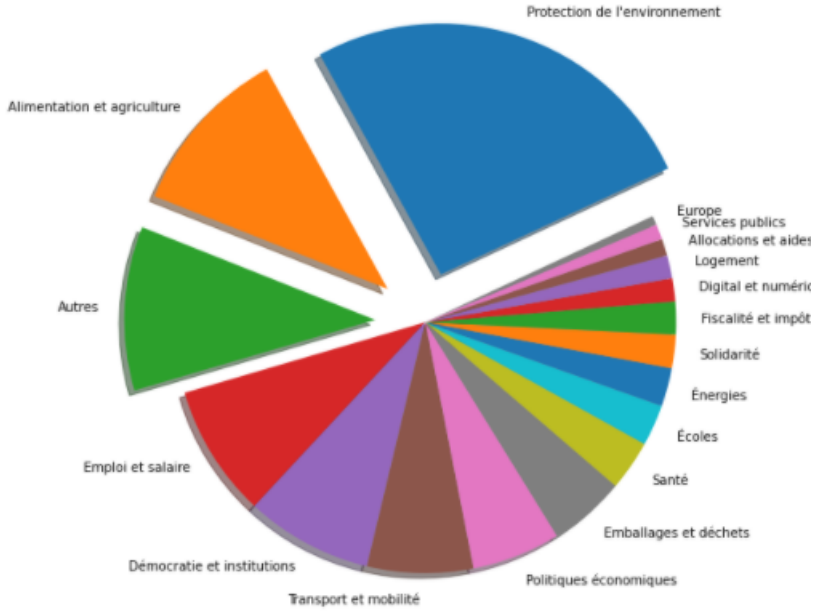


FIGURE 37 – Thématiques auxquelles s'intéressent les jeunes actifs

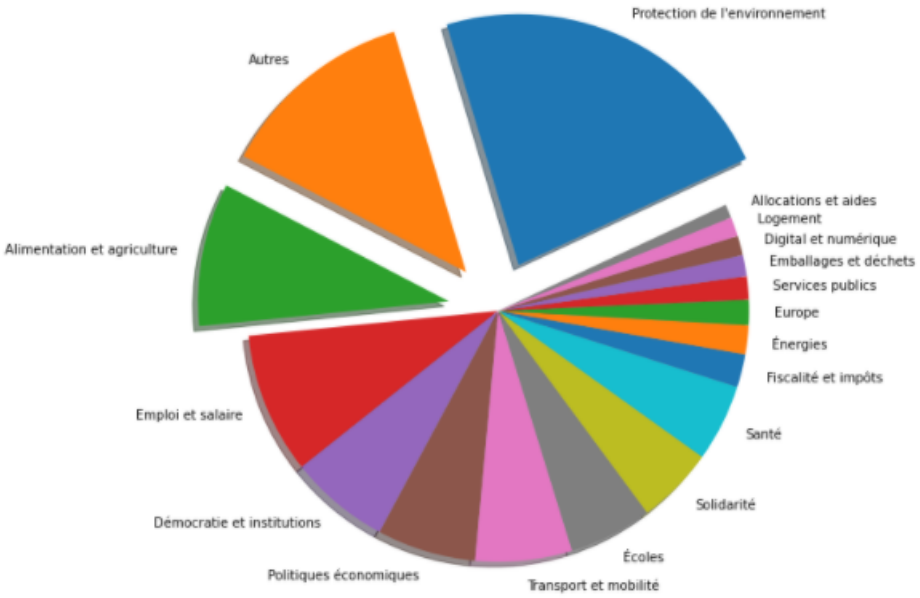


FIGURE 38 – Thématiques auxquelles s'intéressent les jeunes retraités

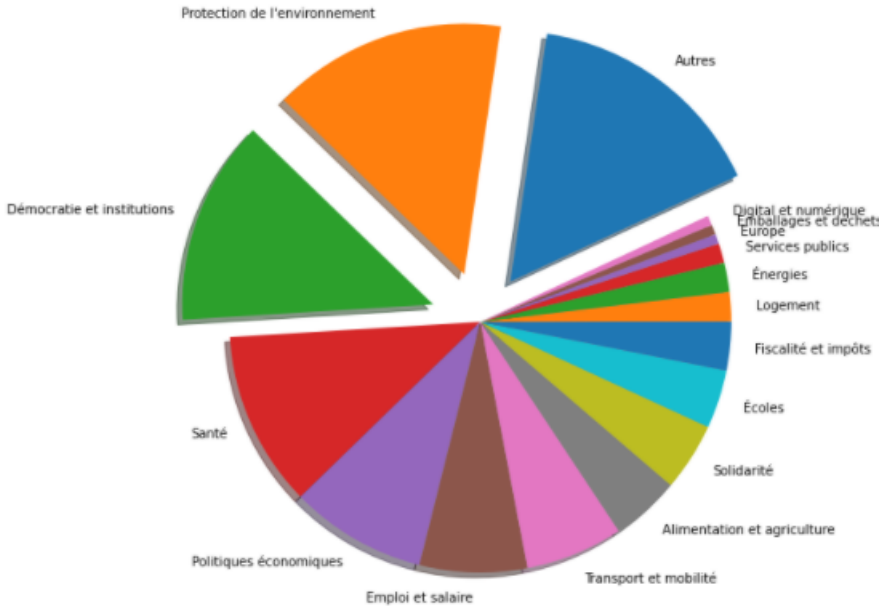


FIGURE 39 – Thématiques auxquelles s'intéressent les retraités

_Alimentation et agriculture_x	_Allocations et aides_x	_Digital et numérique_x	_Démocratie et institutions_x	_Emballages et déchets_x	_Emploi et salaire_x	_Europe_x	_Fiscalité et impôts_x	_Logement_x	_Politiques économiques_x	_Protection de l'environnement_x	_Santé_x	_Services publics_x
0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...

FIGURE 40 – Modalités des tags sous forme de dummy variable

TABLE 4 – Nombre de propositions par groupe avec le TF-IDF

Thème	**Effectif**
0	1907
1	382
2	9718
3	851
4	1001
5	208
6	521
7	717
8	349
9	788
10	231

TABLE 5 – Nombre de propositions par groupe avec le Word2Vec

Thème	Effectif
0	1663
1	1261
2	1375
3	579
4	1624
5	125
6	1333
7	467
8	1405
9	630
10	1160
11	1540
12	202
13	560
14	1241
15	905
16	323
17	280

```
5 propositions assignée au cluster  0
Il faut indiquer, sur les aliments, s'ils ont été produits à l'aide de pesticides de synthèse
Il faut modifier durablement nos habitudes de consommation en privilégiant les produits locaux et bio pour soutenir nos producteurs
Il faut accompagner les agriculteurs vers la polyculture pour réduire les polluants et faciliter le circuit court
Il faut développer les monnaies complémentaires et permettre de payer ses impôts avec pour relocaliser la production
Il faut revenir à l'essentiel, bien manger et arrêter de consommer des nouvelles technologies à outrance !!
_____________________________________________________________________________________________________________________
5 propositions assignée au cluster  1
Il faut envisager sérieusement la croissance 0
Il faut interdire la chasse, "exutoire de la violence" (Théodore Monod) et favoriser le retour des loups, lynx, castors...
Il faut impérativement sauver les abeilles.
Il faut faire cotiser les machines et l'IA pour  la retraite et pour le chômage (les machines remplacent les hommes qui ne cotisent plus)
Il faut que les journalistes soient tenus d'annoncer leur salaire en début d'article ou d'émission pour que le citoyen sache qui lui parle
_____________________________________________________________________________________________________________________
5 propositions assignée au cluster  2
Il faut réapprendre à être plus autonomes : cultiver la terre, un jardin, cuisiner soit même, apprendre à coudre, faire son pain, réparer...
Il faut Que les GAFAM paient leurs impôts en France
Il faut planter des arbres, des haies, selon une articulation locale/nationale pour nous préparer au réchauffement, et que chacun s'implique
Il faut interdire le lobbying des industries chimiques, pharmaceutiques, pétrolières, nucléaires etc.. qui tuent le vivant
Il faut pénaliser fortement les déchets alimentaires pour tout commerce vendant de la nourriture: supermarchés, restaurants, boulangeries...
_____________________________________________________________________________________________________________________
5 propositions assignée au cluster  3
Il faut instaurer un revenu universel dont les modalités restent à définir
Il faut instaurer une taxe sur toutes les transactions financières, pour permettre la mis en place d'un revenu universel réaliste.
Il faut assurer un revenu de base à tous les citoyen.ne.s, afin que chacun.e dispose de temps pour construire ensemble le monde d'après
Il faut mettre en place un revenu universel, permettant à chacun de ne plus avoir à se soucier de sa survie.
Il faut rémunérer tout agent à juste prix avec une même base pour tous , le revenu universel.
_____________________________________________________________________________________________________________________
5 propositions assignée au cluster  4
Il faut supprimer le "je" dans les écoles au profit du "on".
Il faut instaurer un système de santé gratuit pour tous
Il faut remettre la taxe carbone en place
Il faut sortir de la priorité économique et financière. Mettre au minimum à l'égal l'humain
Il faut rendre sa place à la nature
```

FIGURE 41 – Extrait de classes (0-4) obtenues par le Kmeans combiné au Word2Vect

```
5 propositions assignée au cluster  5
Il faut fournir des masques gratuitement pour protéger la vie
Il faut mettre en avant d'autres valeurs que celle de l'ère industrielle et donc réinventer un mode de vie plus humble.
Il faut un "revenu d'existence" dès la naissance évalué en fonction du coût de la vie. Chacun apporte une valeur ajoutée a la société.
Il faut demander des comptes sur les aides attribuées aux entreprises
Il faut mieux accompagner les créateurs de richesse et de lien social (commerces, associations) dans les petits communes
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  6
Il faut un revenu universel pour tous
Il faut créer un revenu universel
Il faut un revenu universel de base
Il faut réfléchir au revenu universel.
Il faut mettre en place un revenu universel.
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  7
Il faut rétablir l'ISF.
Il faut des outils de consultation encore plus performants que Make.org
Il faut surtout, s'attaquer à la corruption, supprimer la spéculation des marchés pour répartir la richesse, le reste viendras tout seul.
Il faut des élections présidentielles au jugement majoritaire pour choisir le meilleur candidat, plutôt que le moins mauvais.
Il faut que le législatif soit le fruit d'une concertation populaire et non d'énarques issus du même moule, trop éloignés de la vraie vie.
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  8
Il faut un salaire minimum comme une retraite minimum permettant à chacun de vivre correctement.
Il faut un revenu universel grâce à l'ISF et l'augmentation du prix des loisirs destructeurs de l'environnement des personnes fortunées.
Il faut offrir un revenu minimum à chacun et valoriser les mi temps pour que l'on s'épanouisse de manière professionnelle et personnelle
Il faut que les salaires des professeurs soient revalorisés.
Il faut ,dans le cadre de la PAC, donner des moyens aux paysans de se protéger face à la volatilité des prix afin de leur garantir un revenu
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  9
Il faut favoriser le local, la ville où le canton pour réfléchir aux enjeux de demain (mobilité, alimentation, emplois, santé...)
Il faut privilégier fortement les productions locales (taxation spécifique).
Il faut définitivement arrêter l'enfouissement des déchets.
Il faut interdire à la construction la totalité du foncier non encore construit et le réserver à l'agriculture, foresterie et biodiversité.
Il faut enfin faire quelque chose de « durable »: arrêter de s'inquiéter de l'Homme mais plutôt de la haute mer..
```

FIGURE 42 – Extrait de classes (5-9) obtenues par le Kmeans combiné au Word2Vect

```
5 propositions assignée au cluster  10
Il faut créer un système de taxe ou d'incitation financière pour privilégier les circuits courts et locaux
Il faut interdire la chasse
Il faut augmenter les taxes sur les produits qui seront des déchets polluants en fin de vie
Il faut protéger et respecter la vie des animaux.
Il faut aider l'agriculture bio
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  11
Il faut repenser la mobilité pour réduire drastiquement l'utilisation du véhicule particulier
Il faut réduire les déplacements en avion au profit du train (et développer les trains de nuits)
Il faut apprendre à tous la culture des fruits et légumes et le respect de la saison de leur production
Il faut augmenter le nombre de petits hôpitaux locaux.
Il faut relocaliser la production de médicaments en France, valoriser les systèmes alimentaires locaux (maraîchers, agriculteurs) .
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  12
Il faut apprendre à faire pousser des fruits et légumes à l'école.
Il faut relancer l'emploi en développant de nouvelles compétences en lien avec l'écologie et en créant de nouveaux diplômes.
Il faut imposer aux cantines scolaires et municipales un minimum de 90% de produits locaux
Il faut mettre en place le vote blanc.
Il faut augmenter les salaires des métiers essentiels (comme ceux qui nous font vivre pendant le confinement): médecine, éducation, éboueurs
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  13
Il faut  mettre en place un salaire minimum universel
Il faut donner le revenu universel à tous de 18 ans à la fin de la vie.
Il faut mettre en place un revenu universel afin que chacun puisse vivre normalement.
Il faut mettre en place le Revenu Universel (1000 euros par personnes à partir de 18 ans) en supprimant toutes les autres aides.
Il faut instaurer un revenu maximum universel pour stopper les escrocs
_____________________________________________________________________________________________________________________________
5 propositions assignée au cluster  14
Il faut encourager les circuits courts.
Il faut favoriser les circuits courts
Il faut  revenir à une production et à une consommation  locale pour les produits de premières nécessités
Il faut privilégier les produits locaux
Il faut développer l'économie circulaire et imposer les circuits courts
```

FIGURE 43 – Extrait de classes (10-14) obtenues par le Kmeans combiné au Word2Vect

```
5 propositions assignée au cluster  15
Il faut proposer des alternatives aux pesticides, accompagner les agriculteurs dans ces changements en garantissant un bon niveau de vie.
Il faut mettre un quota par personne pour les déplacements en avion
Il faut éduquer dès le plus jeune âge au respect de la nature et à prendre soin des autres
Il faut conditionner les aides aux entreprises à des obligations environnementales et sociales.
Il faut apprendre aux enfants à respecter le vivant, à régénérer les sols et à réparer a l'école
--------------------------------------------------------------------------------------------------------------
5 propositions assignée au cluster  16
Il faut redonner de la valeur au travail et ne reconstruire qu'avec ceux qui partagent.
Il faut mettre en place le RIC (Référendum d'Initiative Citoyenne)
Il faut développer les nouveaux métiers, pour anticiper les bouleversements qui seront liés à une perte de croissance dans certains secteurs
Il faut augmenter très nettement (fois 10 ou plus) les taxes d'habitation et foncières pour les maisons secondaires
Il faut passer à la semaine de 4 jours et ainsi continuer à voir nos familles et nous occuper de nos proches.
--------------------------------------------------------------------------------------------------------------
5 propositions assignée au cluster  17
Il faut imposer un cahier des charges plus strict aux entreprises (production, récupération, réparation, recyclage), bref cercle vertueux.
Il faut mettre en place un green new deal à l'échelle mondiale
Il faut une transition agricole vers la polyculture, plus robuste et résiliente que la monoculture intensive
Il faut utiliser les eaux de pluies pour les toilettes et machines à laver
Il faut pouvoir écourter de moitié le mandat d'un président qui n'agirait pas pour le bien planétaire, le progrès écologique.
```

FIGURE 44 – Extrait de classes (15-17) obtenues par le Kmeans combiné au Word2Vect

```
---------------------------------------------------------------------------------------------------
Il faut que les allocations familiales soient accessibles du 1er enfant jusqu'au 4ème
Il faut que les allocations familiales soient distribuées dès le premier enfant
Il faut envisager l'arrêt des allocations familiales au bout du quatrième enfant accueilli dans des conditions économiques insu
ffisantes
Il faut, en France, limiter les allocations familiales à 2 enfants.
Il faut supprimer les allocations familiales et le quotient familial qui bénéficie aux familles aisées. Un crédit d'impôt pour
chaque enfant
Il faut limiter les allocations familiales aux deux premiers enfants
Il faut supprimer/réduire fortement les allocations familiales pour stopper l'augmentation de la population
Il faut limiter les allocations familiales au 3ème nouvel enfant, tant que l'équilibre courbe des âges/chômage ne sera pas à
l'équilibre
Il faut donner moins d'allocations familiales aux familles de quatre enfants et plus pour la planète.
Il faut limiter le nombre d'enfants par femme à 3 par une incitation financière. Plus d'allocation familiale à partir du 4ème e
nfant.
---------------------------------------------------------------------------------------------------
Il faut valoriser l'humain et non le capitalisme. C'est peut être le moment de changer de modèle économique, instaurer le salai
re universel
Il faut instaurer le salaire universel
Il faut instaurer le salaire à vie (et non le revenu universel).
Il faut changer le capitalisme.
Il faut changer de modèle économique
Il faut changer le système économique qui repose sur le capitalisme.
Il faut instaurer un salaire universel
Il faut cesser le capitalisme libéral, changer complètement de système économique
Il faut changer de modèle : sortir du capitalisme et aller vers de la décroissance
Il faut un salaire universel
---------------------------------------------------------------------------------------------------
```

FIGURE 45 – Extrait de classes obtenues par le NNS combiné au TF-IDF avec des propositions bien classées

```
Il faut gérer les ressources terrestres avec intelligence et non à profit.
Il faut utiliser les technologies numériques, dont l'intelligence artificielle, pour nous aider à créer des solidarités et non
du profit.
Il faut repenser le système éducatif en l'axant sur l'apprentissage, sur l'intelligence collective et l'intelligence émotionnel
le
Il faut décentraliser les décisions et développer l'intelligence collective
Il faut une économie compatible avec les ressources naturelles et non au détriment de la planète
Il faut redonner du pouvoir et de l'autonomie aux régions afin qu'elles puissent gérer ses ressources et richesses au mieux.
Il faut fonctionner avec les méthodes d'intelligence collective, apprendre la méditation,....
Il faut, notre planète ayant des ressources limitées, gérer d'urgence la démographie et notre manière de consommer
Il faut faire gérer les ressources planétaires par une organisation internationale scientifique au dessus de nos organisations
politiques
Il faut augmenter les surfaces terrestres et marines protégées afin de conserver les habitats et la biodiversité
---------------------------------------------------------------------------------------------------
```

FIGURE 46 – NNS combiné au TF-IDF : propositions classées ensemble ayant des mots communs mais pas de sens commun

```
--------------------------------------------------
Il faut que chaque enfant puisse être éduqué afin de prendre soin du végétal et des animaux sauvages et domestiques
Il faut que chaque loi en cours de vote soit validée par un conseil de l'environnement indépendant pour une harmonie Homme, ani
mal, nature
Il faut sensibiliser à la nécessité du respect de la faune et de la flore pour notre survie en liberté
Il faut vivre en harmonie et respecter la nature, s'inspirer de sa perfection en développant, dans tous les domaines, le biomim
étisme
Il faut consigner les boîtes de boisson, afin de ne plus les jeter dans la nature
Il faut organiser des vacances pour la nature.
Il faut dispenser des cours sur les espèces animales et leur mode de vie dès la maternelle et pendant tout le cursus scolaire
Il faut laisser la nature tranquille
Il faut sacraliser les Arbres : interdiction d'y toucher, couper une branche, de les taguer, de leur manquer de respect, et tou
te la Nature
Il faut enseigner les disciplines psychiques et corporelles à l'école et arrêter avec ces matières qui ne servent à rien dans l
a vie.
--------------------------------------------------
```

FIGURE 47 – Exemple de classes obtenues par le NNS combiné au Word2vec

```
--------------------------------------------------
Il faut que chaque enfant puisse être éduqué afin de prendre soin du végétal et des animaux sauvages et domestiques
Il faut apprendre aux enfants à reconnaître les plantes et les animaux
Il faut un ministère de la nature comme il y a un ministère de la culture. Un ministère de la biodiversité et de la conditio
n animale
Il faut intégrer le bien-être animal dans l'éducation de nos enfants.
Il faut réapprendre dans les écoles ce qu'est le monde du vivant, ses bénéfices vitaux pour l'espèce humaine, et notre devoi
r de respect
Il faut redéfinir ce qui est polluant et pourquoi afin d'adapter leur gestion. Les plantes ont-elles polluées la terre ?
Il faut interdire les marchés d'animaux sauvages et vivants sur toute la planète. Contrôler activement les prises de pêche e
t sensibiliser.
Il faut informer les enfants sur les conditions d'élevage et d'abattage des animaux.
Il faut interdire le travail des enfants partout dans le monde ! Préparer les nouvelles générations par une éducation respon
sable et pacifié
Il faut créer une école de la nature pour tous, car la respecter est la seule façon de sauver la planète et toutes les espèc
es, dont l'homme
--------------------------------------------------
```

FIGURE 48 – Exemple de classes obtenues par le NNS combiné au Doc2vec

```
--------------------------------------------------
Il faut que les allocations familiales soient accessibles du 1er enfant jusqu'au 4ème
Il faut apprendre à vivre avec le Covid-19, nous serons plus forts et mieux préparés pour la prochaine épidémie
Il faut enseigner diverses langues étrangères (pas uniquement l'anglais), dès l'école maternelle et primaire, avec des ensei
gnants formés
Il faut apprendre à nos enfants et l'instaurer dans le programme scolaire à consommer différemment, se reconnecter à notre b
elle planète
Il faut que le jardinage et la cuisine fasse partie de l'apprentissage scolaire des enfants comme les maths !
Il faut que l'école nous apprenne les choses de la vie : cuisiner, s'entraider, construire soi même ses outils pour mieux vi
vre.
Il faut permettre aux parents d'être avec leurs enfants en arrêtant la course à la productivité qui ne permet que l'augmenta
tion des déchets
Il faut éduquer les enfants à réfléchir par eux-mêmes, leur parler des émotions, leur enseigner le sens de la vie et le resp
ect de toute vie
Il faut confiner les enfants malades après le covid tout en aidant fiscalement les parents pour la garde à domicile
Il faut apprendre à coopérer et vivre ensemble
--------------------------------------------------
```

FIGURE 49 – NNS combiné au Doc2vec : propositions classées ensemble n'ayant pas forcément de sens commun

	No Topic_dominant	Mots clés	Propositions
13	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut investir massivement dans l'éducation afin de donner aux citoyens les moyens intellectuels de penser le changement.
30	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut, pour redresser le pays, adopter les modèles allemands, coréens et suédois, et se lancer dans la production nationale et l'export.
38	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut renforcer la cohésion sociale et environnementale en uniformisant la politique: indicateurs, plan d'actions à l'échelle européenne.
44	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut soutenir l'économie réelle (et non le système financier), redévelopper les productions sur le territoire et l'économie circulaire
45	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut rétablir l'ISF.
48	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut que les organismes mondiaux de protection des droits pour animaux sauvages & domestiques fusionnent car l'union fait la force.
55	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut supprimer les voitures tel que les SUV, polluantes, imposantes, gourmandes en énergie et prenant de la place sur la route.
60	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut faire travailler les prisonniers contre rémunération ou les aider à trouver une formation , pour une réinsertion plus rapide.
82	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut interdire les airbnb
88	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut mettre la créativité comme discipline scolaire. Créativité = transformation, innovation, entraide, solutions.
93	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut interdire l'obsolescence programmée
98	13.0	pay, donner, systeme, politique, energie, territoire, renforcer, action, penser, protection	Il faut mettre en place le vote blanc.

FIGURE 50 – Propositions associées au topic 13 - Modèle LDA

Références

[1] *Différents types de distances utilisées dans l'apprentissage automatique*. URL : https://ichi.pro/fr/differents-types-de-distances-utilisees-dans-l-apprentissage-automatique-259968051799106.

[2] *Effectuez des plongements de mots (word embeddings)*. URL : https://openclassrooms.com/fr/courses/4470541-analysez-vos-donnees-textuelles/4855006-effectuez-des-plongements-de-mots-word-embeddings.

[3] Izzet Fatih Senturk METIN BILGIN. *Sentiment Analysis on Twitter data with Semi-Supervised Doc2Vec*. URL : https://www.researchgate.net/publication/320829283_Sentiment_analysis_on_Twitter_data_with_semi-supervised_Doc2Vec.

[4] *Word embedding*. URL : https://fr.wikipedia.org/wiki/Word_embedding.