Rapport projet n°2: Advanced Machine Learning for Big Data and Text

Table des matières

Membres de l’équipe : Axel de Cuniac, Rémi de Ferrieres et Vadim Macher

Introduction

Le but étant de synthétiser du texte provenant de différentes sources, notre méthodologie a été de rendre le code générique de façon à toutes questions et tout fichier disponible. La première étape du projet a été de nettoyer notre dataset. Dans le cadre de notre projet, il a fallu séparer les colonnes contenant du texte et garder celle qui contenaient les réponses des participants au grand débat.

Préprocessing

Concernant le pre-processing, plusieurs étapes ont été nécessaires, comme le fait de retirer les caractères vide de sens, de mettre tout le texte en minuscule, d'enlever la ponctuation et les accents, de retirer les stops-words, de retirer les mots courts, de lemmatizer, de retirer les nombres de notre analyse, de filtrer les mots par rapport à leur TF-IDF.

En ce qui concerne les stop words nous nous sommes basés sur une librairie appelé stop-words qui avait pour avantage d'avoir deux fois plus de stop words que nltk. Cependant il nous a semblé restait beaucoup de stop words qui n'étaient pas filtré et nous avons donc rajouté nos propres mots à la liste :

["assez","comme","depuis","dessus","vers","si","jamais","toujours","le","la",'les',
'etre','plus','sur','sous','contre','non','ni','et','donc','mais','ou','celui','celle','tous','quelque','un',
'une','abord','jusqu','afin','surtout','apres','non','il','elle','ils','elles','nous','vous','etc']

Pour lemmatizer le texte, une des difficultés a été de lemmatizer le texte en français. En effet, nltk, l'outil utilisé dans le cadre de notre cours de NLP, ne fonctionne pas en français. En remplacement de nltk pour le lemmatizer on a trouvé les packages spacy et FrenchleffLemmatizer. Une autre complication a été de trouver un package qui puisse tourner sur colab. FrenchleffLemmatizer fonctionne très bien seulement il a besoin d'avoir le tag du mot pour fonctionner de façon optimale. Ayant essayé de trouver un pos taguer en français qui fonctionnerai sur colab, nous avons trouvé : StanfordNLP mais cette solution a besoin de java pour fonctionner, TreeTagger mais cette solution ne fonctionnerait pas sur colab et RNNTagger qui ne fonctionnerait pas sur colab.

Par ailleurs nous avons mis en place un stemmer qui provient de snowball de nltk. Cependant nous ne l'avons pas gardé dans notre pré-processing final car l'analyse des données et l'interprétation des résultats devenait beaucoup plus complexe.

Visualisation des données

Une autre partie importante du projet a été de créer des data visualisations pour visualiser les mots les plus fréquemment utilisés.

Pour cela nous avons utilisé deux méthodes pour comprendre rapidement les mots les plus importants dans les réponses :

Tout d'abord nous récupérons le top 10 des mots les plus fréquents dans nos jeux de données pré-processés. Cela permet d'avoir un aperçu global des sujets abordés les plus fréquemment.

Pour avoir une analyse plus fine nous avons aussi mis en place un WordCloud qui va conserver les mots les plus importants et leur donner une taille en fonction de leurs importances. Voici un exemple obtenu :

Analyse des données

Pour montrer les différentes méthodes adoptées nous nous reposerons sur l'étude de la question suivante : « Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ? » qui nous semblait intéressante car complétement ouverte et avec des champs variés.

1. LDA :

Nous avons mis en place l'algorithme de LDA en premier lieu sur nos données pré-processés car il nous semblait intéressants de pouvoir distinguer les sujets de manière aussi claires.
Le LDA permet de créer des groupes (topics) non-observés. Ces groupes viennent de la présence de similitude dans les données et va pouvoir tirer quelques thèmes possibles et différents dans l'ensemble du texte.

Le modèle LDA prend 2 paramètres. Un pour définir le nombre de groupes que le modèle doit générer. L'autre pour le nombre de mots à montrer par groupes en choisissant les mots les plus pertinent.

Il faut donc trouver la meilleure combinaison de nombre de groupes/nombre de mots par groupes.

Pour cela on s'est référé au bar chart affichant les x mots les plus fréquent de notre jeu de donnée et leur nombre d'apparition (cf. Bar chart dans "Données récoltées"). Elle nous a permis de faire plusieurs hypothèses sur les valeurs à choisir pour les 2 paramètres du modèle LDA.
Voici les résultats obtenus :

Topic #0: important littoral lies cause plastique ocean

Topic #1: climatique crue environement distinction relie sauvegard

Topic #2: dereglement pouvoir voiture homme agir moyen

Topic #3: lier dechet pollution probleme terre humain

Topic #4: global planete general probleme bien ecologiqu

Topic #5: biodiversite certain erosion point agricole intensif

Topic #6: energie quatre mondial animal cite agriculture

Topic #7: espece disparition maitrisee paire vegetales air

Après avoir essayé différents nombres de topics et de mots, nous sommes arrivés sur le résultat ci-dessus. Il est intéressant car il relève très rapidement les différents sujets des réponses à cette questions qui sont très variés : la pollution de l'eau (#1), le dérèglement et la faute de l'homme de manière plus ou moins directe (#2 #3 #4), les problèmes énergétiques et de l'agriculture (# 5 #6) ainsi que la disparition des espèces animales et végétales (#7)

2. Word Embedding :

L'objectif de cette méthode était de créer un word embedding pour visualiser graphiquement les mots. Pour cela nous étions tout d'abord basé sur nos données pré-processés. Ces données ont servi à entrainer un modèle de word2vec issue de gensim. Les vecteurs issus de word embedding étant de taille 100 nous les avons réduits à l'aide du TSNE pour pouvoir les afficher en deux dimensions.

Seulement il s'est avéré qu'en affichant tous les mots du corpus, le graphique était illisible et contenait beaucoup de bruits. Suite à cela nous avons mis en place un filtre basé sur le score de tf-idf des mots pour n'entrainer le modèle que sur les « meilleurs » mots du tf-idf.

Nous avons obtenu le résultat suivant :

Comme on peut le voir, il n'émerge aucun cluster apparent au sein de ces données. Nous pensons que cela est dû au fait qu'il y a encore trop de bruit dans les données pour ce genre de méthode. Pour pallier cela nous avons cherché à pos-tagger les mots pour n'extraire que les noms, qui nous semble plus porteur de sens dans ce cas-là. Seulement nous n'avons pas trouvé de pos-tagger en français qui puisse fonctionner sur nos machines.

3. TextRank :

Pour concaténer l'ensemble des informations issues de nos données pré-processés, nous nous sommes dit qu'ils pouvaient être intéressant d'utiliser des méthodes de résumé pour extraire les informations les plus importantes. Pour cela nous nous sommes basés sur l'algorithme TextRank qui dans notre cas va extraire les commentaires les plus pertinents selon lui.

Voici le résultat obtenu :

"A ceux cités, il faudrait ajouter la pollution et la dégradation des sols, aussi bien par l'industrie, l'agriculture, les collectivités et les particuliers.Les dérèglements climatiques (crue, sécheresse).Un manque d'action pour la protection de l'environnement dans tous les domaines de la société au bénéfice du profit.Les dérèglements climatiques (crue, sécheresse).Les dérèglements climatiques (crue, sécheresse).L'évitement constant de nos dirigeants à démarrer VRAIMENT la transition écologique .La pollution de l'air.La pollution de l'air et de l'eau (qualité des nappes phréatiques) et les perturbateurs endoctriniens (dont les pesticides et les additifs alimentaires).Les dérèglements climatiques (crue, sécheresse).Pas de transition.La biodiversité et la disparition de certaines espèces.Tous ces sujets sont essentiels et interconnectés dans leur origines, c'est l'ensemble qui doit être traité de front.La biodiversité et la disparition de certaines espèces.plutôt la concentration des pollutions de l'air sur les métropoles, la pollution des eaux par les pesticides, l'exploitation des terres rares par des acides polluants, externalisation des déchets toxiques.LA PROTECTION ANIMALE.La pollution de l'air dans les métropoles et autres grandes agglomérations (> 100 000 habitants en France).il y en a pas un plus important que les autres, c'est l'ensemble et ils sont liés entre eux .l'enjeu énergétique, charbon, nucléaire, pétrole....La biodiversité et la disparition de certaines espèces.Les dérèglements climatiques (crue, sécheresse).fortes préoccupation sur beaucoup de points difficiles à hiérarchiser entre la pollution sous toutes ses formes (air, eau, sol, mer, plastique ...), le réchauffement climatique et la disparition des espèces sauvages.la réduction des espaces végétaux et en particulier la déforestation .Les dérèglements climatiques (crue, sécheresse).Les dérèglements climatiques (crue, sécheresse).La pollution de l'air.La pollution de l'air.le bruit.Les dérèglements climatiques (crue, sécheresse).les lobbys.La pollution de l'air.Les 4 items sont importants et il y en a d'autres.Agir contre le réchauffement global du aux émissions de gaz à effet de serre est l'enjeu prioritaire pour que la planète reste durablement habitable.quand j'achète une voiture neuve c'est mois qui suis montré du doigt, comme pollueur hors c'est le garage qui me vend ce véhiculei me .Considérer que la planète est faite pour fournir toujours plus d'énergie.La biodiversité et la disparition de certaines espèces.l'impact de l'être humain.Tous les problèmes touchant l'environnement sont concrets et ne doivent pas êtres pris à la légère..La biodiversité et la disparition de certaines espèces.Les dérèglements climatiques (crue, sécheresse).La pollution de l'air.La biodiversité et la disparition de certaines espèces.La pollution de l'air.Les dérèglements climatiques (crue, sécheresse).Non respect des normes environnementales pour les entreprises .Les seules actions de l'homme sont le principal problème."

Comme on peut le constater cette méthode est assez intéressante car elle permet de faire ressortir les grands thèmes de cette question en évoquant la biodiversité, la pollution, la disparition des espèces, etc. Cette méthode est cependant très longue pour créer un résumé (environ 40 minutes par question) et les résultats ne sont pas tout de suite compréhensible comme cela pourrait être le

cas pour un LDA par exemple. Cependant elle a pour avantage de ne nécessiter aucun pré-processing.

4. Extraction de KeyWord

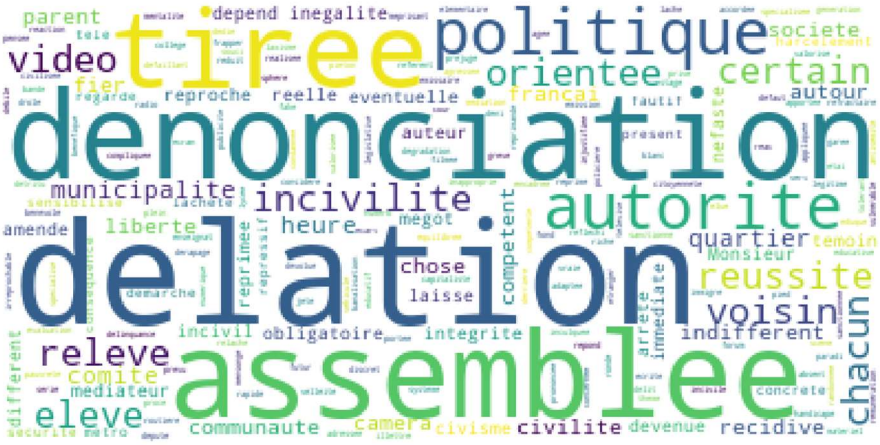
Nous avons cherché aussi en se basant sur TextRank d'extraire les mots clés des réponses pour définir les thèmes importants. J'ai eu l'occasion de tester cette méthode sur un projet PI² pour extraire des informations de phrase de tailles normale et cette méthode s'est révélé très efficace. Malheureusement ici la taille du jeu de données est telle que même via google collab le code ne pouvait pas être exécuté car trop gourmand en RAM. Nous avons donc abandonnée cette méthode.

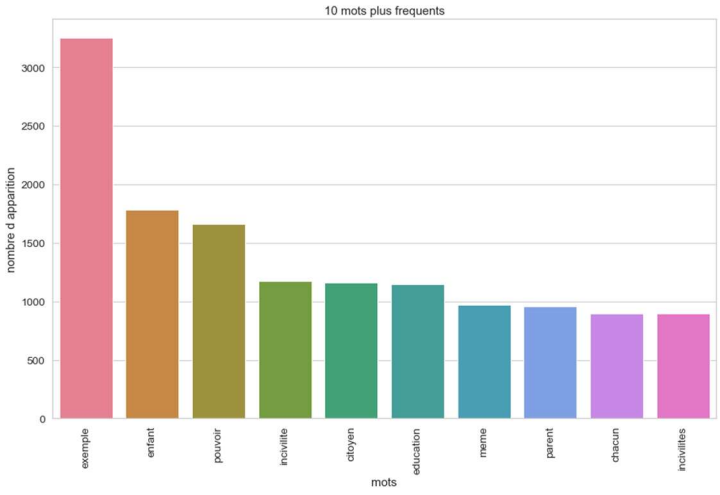
Données récoltées

Par ailleurs, nous avons traité différentes questions et sujets. Voici les résultats que nous avons obtenus par thème et par question:

1. Démocratie et citoyenneté

a) "Quel pourrait être le rôle de chacun pour faire reculer les incivilités dans la société ?"



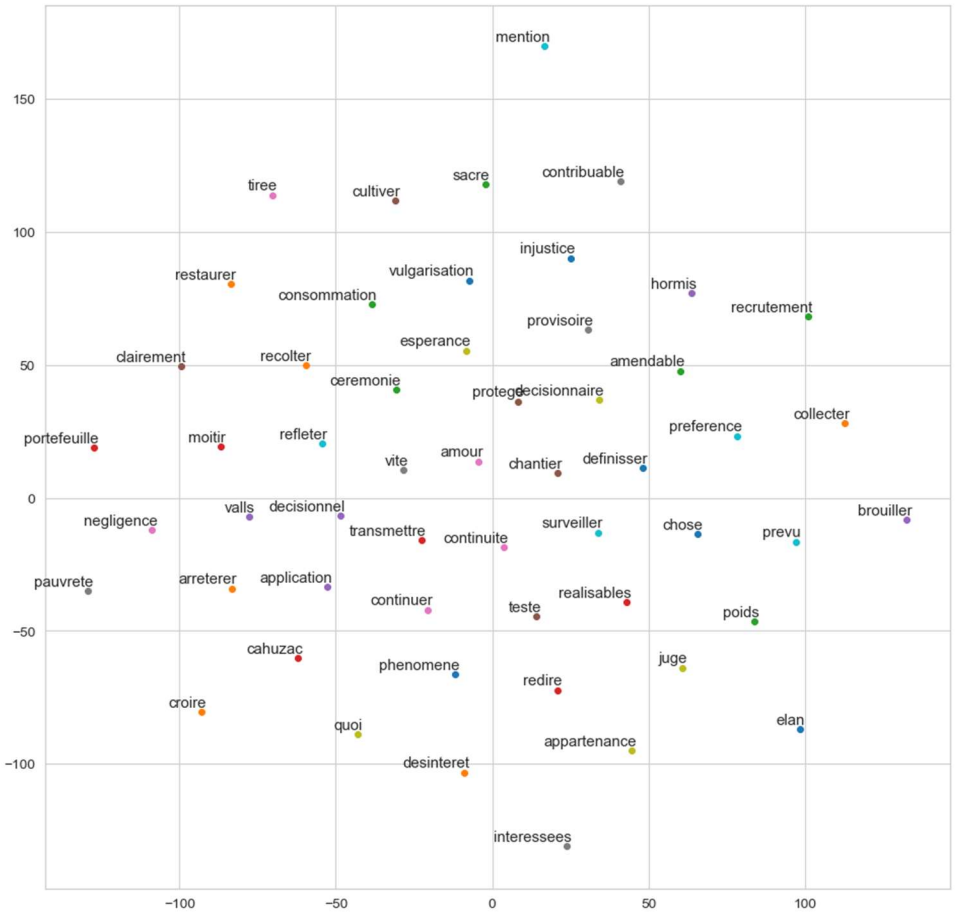


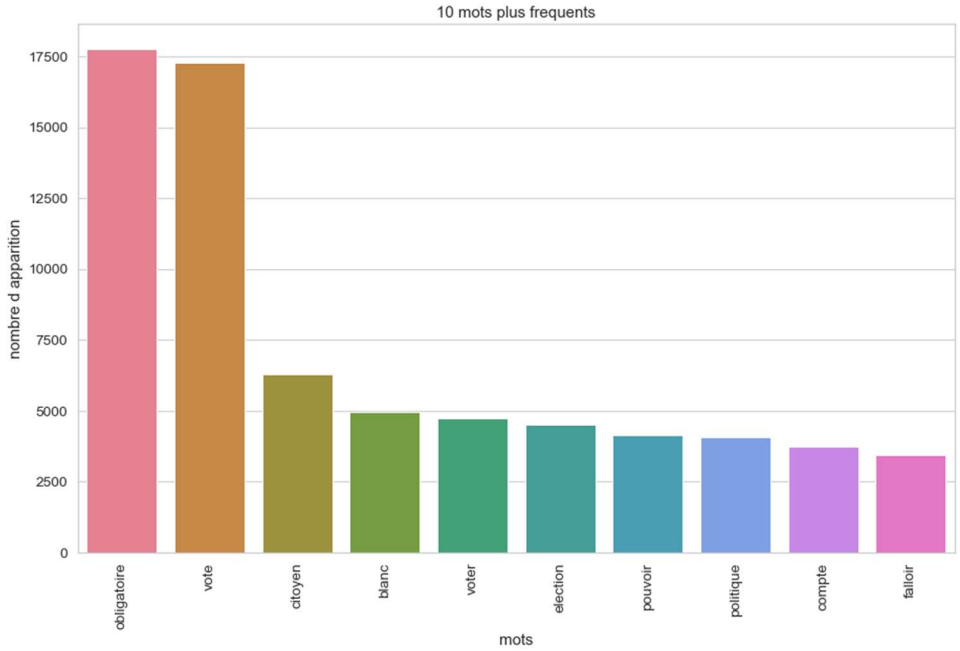
Topics found via LDA:

Topic #0:
pouvoir eduquer dire aujourd justice hui
Topic #1:
exemplarite bien reponse public dialogue autrui
Topic #2:
respect respecter police droit laisser savoir
Topic #3:
exemple enfant exemplaire citoyen mieux montrer
Topic #4:
chacun regle jeune donner role niveau
Topic #5:
incivilite meme denoncer responsable lorsque question
Topic #6:
education parent ecole comportement civique changer
Topic #7:
incivilites auteur peur pouvoir prendre ordre

Résultats obtenus : Les principaux sujets relevés sur cette question sont : l'exemplarité nécessaire des pouvoirs publics, l'éducation des citoyens et la notion de dénoncer punir les incivilités.

b) “ Que faudrait-il faire pour valoriser l'engagement citoyen dans les parcours de vie, dans les relations avec l'administration et les pouvoirs publics ? »



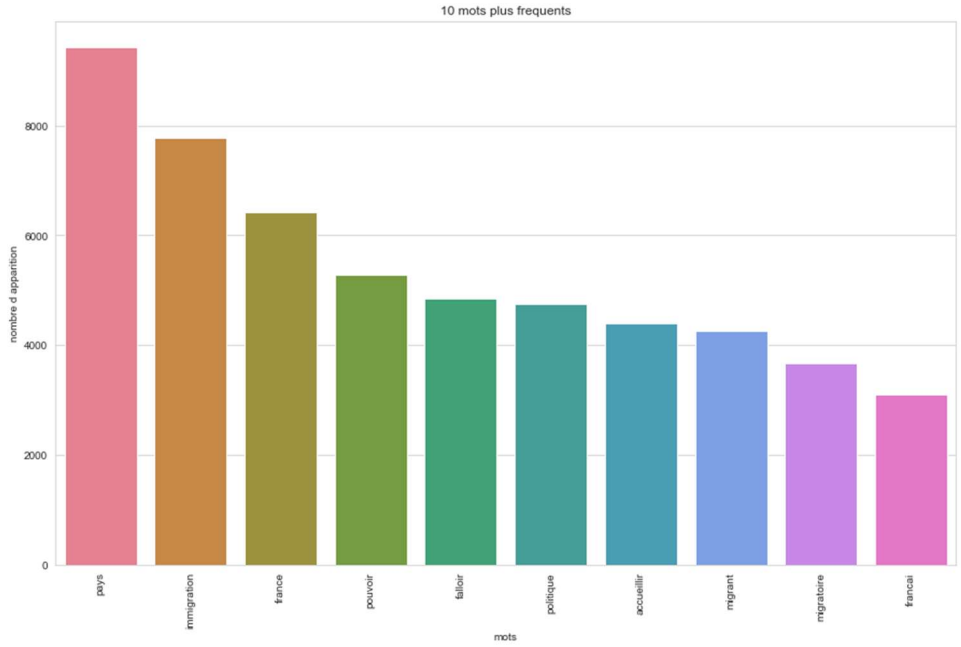


Topics found via LDA:

Topic #0:
pouvoir referendum citoyen programme promesse gens
Topic #1:
vote election voter candidat pouvoir scrutin
Topic #2:p
politique chose voter pouvoir rien citoyen
Topic #3:
citoyen electeur participation faible falloir candidat
Topic #4:
voter droit pouvoir citoyen civique vote
Topic #5:
obligatoire vote blanc compte election prendre
Topic #6:
citoyen confiance politique falloir participer francais
Topic #7:
rendre obligatoire voter election internet falloir

Résultats obtenus : Les principaux sujets relevés sur cette question sont : Rendre obligatoire le vote blanc, le faible pouvoir du citoyen vis-à-vis de son vote, rendre possible le vote par internet et rendre le vote obligatoire.

c) “ Que pensez-vous de la situation de l'immigration en France aujourd'hui et de la politique migratoire ? Quelles sont, selon vous, les critères à mettre en place pour définir la politique migratoire ? “



Topics found via LDA:
Topic #0:
migrant migration migratoire pays droit situation
Topic #1:
immigre francai france langue politique integration
Topic #2:
pays integrer refugie migrant origine aide
Topic #3:
pouvoir accueillir france migrant pays immigration
Topic #4:
pays falloir pouvoir france accueillir accueil
Topic #5:
politique quota asile falloir integration immigration
Topic #6:
aujourd hui france choisir immigration droit
Topic #7:
immigration besoin pays asile france fonction

Résultats obtenus : Les principaux sujets relevés sur cette question sont : L'intégration et l'accueil migrants/réfugiés et l'instauration de quota. On peut tout de même noter le vocabulaire très violent qui ressort de cette question avec « mécréant » ou « voyou ».

d) Conclusion

Parmi l'ensemble des propositions relevés par Emmanuel Macro dans le résumé sur Reddit, nous avons pu capter un certain nombre de mesures :

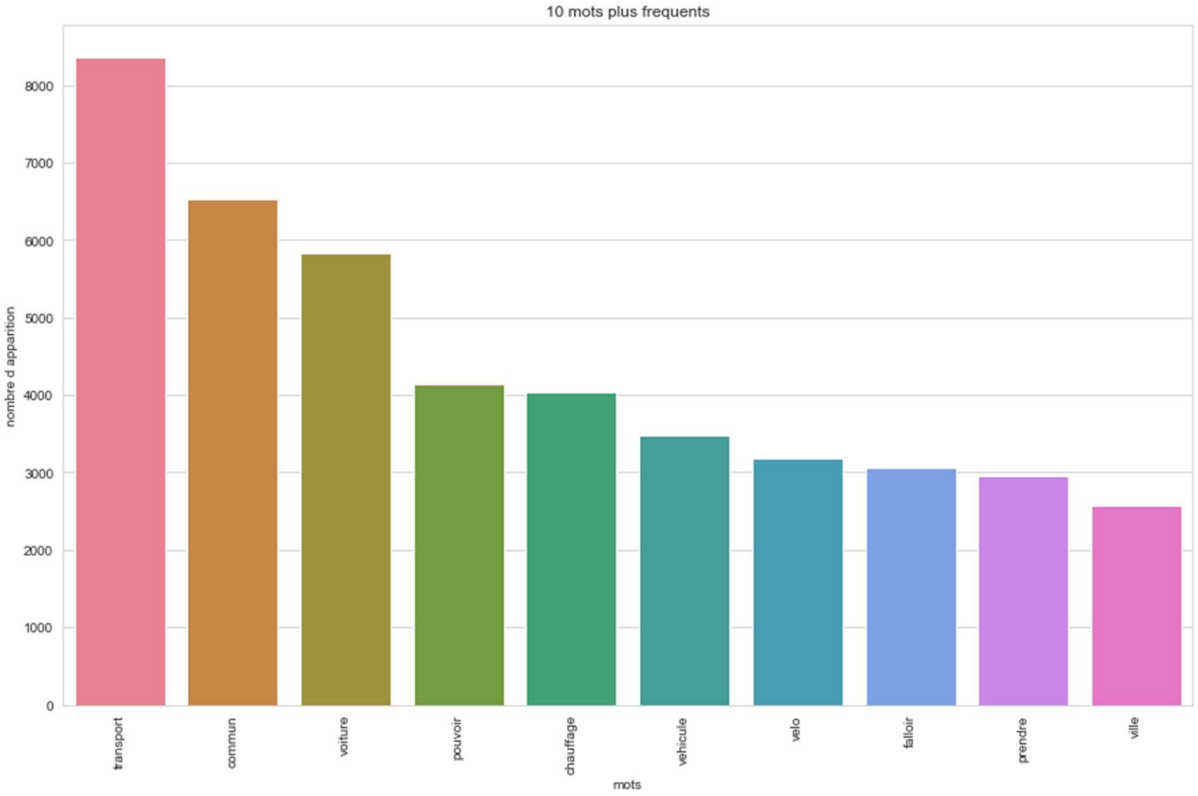
- Pas de vote obligatoire
- Pas de reconnaissance du vote blanc, qui n'est que "l'agrégation des refus"

- Renforcer le droit d'interpellation des élus, à l'échelle locale

Les questions que nous avons traité pour le thème Démocratie et citoyenneté ne sont pas les principaux relevés dans le résumé sur Reddit (les incivilités ne sont pas traités par exemple) mais permettent d'avoir un aperçu global des thèmes des réponses.

2. La Transition écologique

a) « Qu'est-ce qui pourrait vous inciter à changer vos comportements comme par exemple mieux entretenir et régler votre chauffage, modifier votre manière de conduire ou renoncer à prendre votre véhicule pour de très petites distances ? »

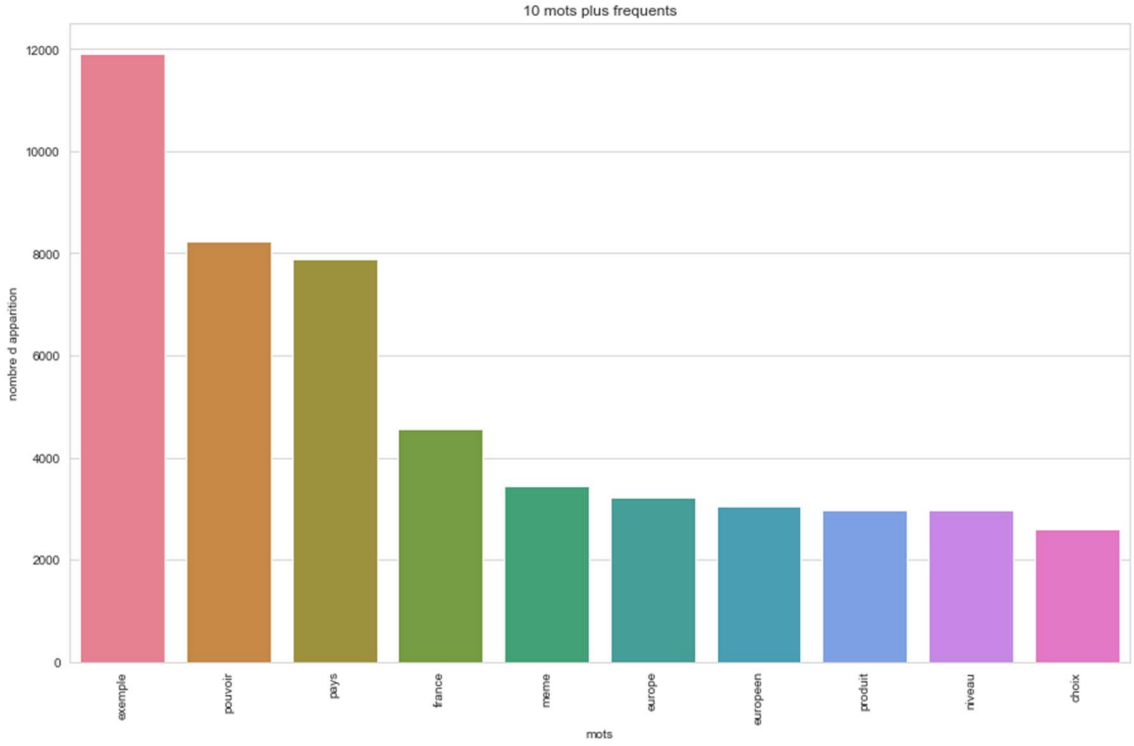


Topics found via LDA:

Topic #0:
voiture velo electriqu achat dejer consommation
Topic #1:
etat public exemple comportement citoyen pouvoir
Topic #2:
commun prendre transport collectif voiture changement
Topic #3:
voiture aller changer financiere falloir aujourd
Topic #4:
vehicule chauffage isolation financier distance voiture
Topic #5:
transport ville piste cyclable commun velo
Topic #6:
prix aide fiscal pouvoir economie moyen
Topic #7:
deja energie chauffage pouvoir chaudiere cout

Résultats obtenus : Les principaux sujets relevés sur cette question sont : L'amélioration des transports en commun et de moyens de déplacement comme le vélo et le prix du chauffage et des transports.

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



Topics found via LDA:
Topic #0:
pays montrer europeen choix ecologiqu exemplaire
Topic #1:
exemple lecon plutot appliquer fonctionne place
Topic #2:
pouvoir france meme pays monde falloir
Topic #3:
transport energie marchandise camion taxer hydrogene
Topic #4:
politique donner partager europeens engagement choix
Topic #5:
produit europe importation respecter taxe pays
Topic #6:
niveau pouvoir environnement france politique europeen
Topic #7:
bien propre france chose etat francai

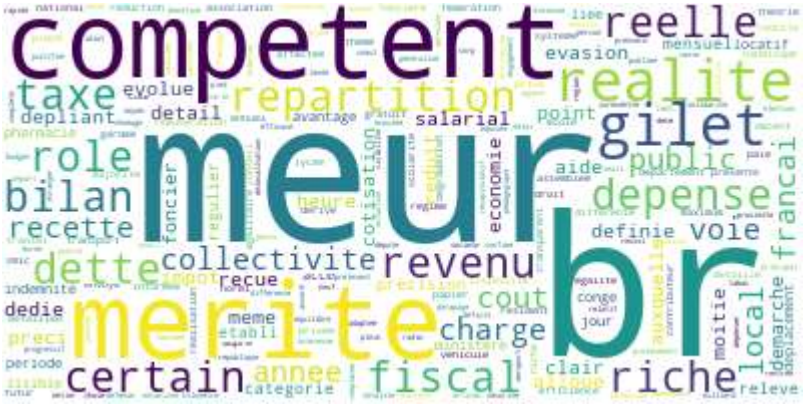
Résultats obtenus : Les principaux sujets relevés sur cette question sont : la France doit être un exemple vis-à-vis de l'écologie, taxer les transports au niveau Européen et mieux gérer les transports européens en taxant certains véhicules polluants et laisser place à des véhicules plus propres comme l'hydrogène.

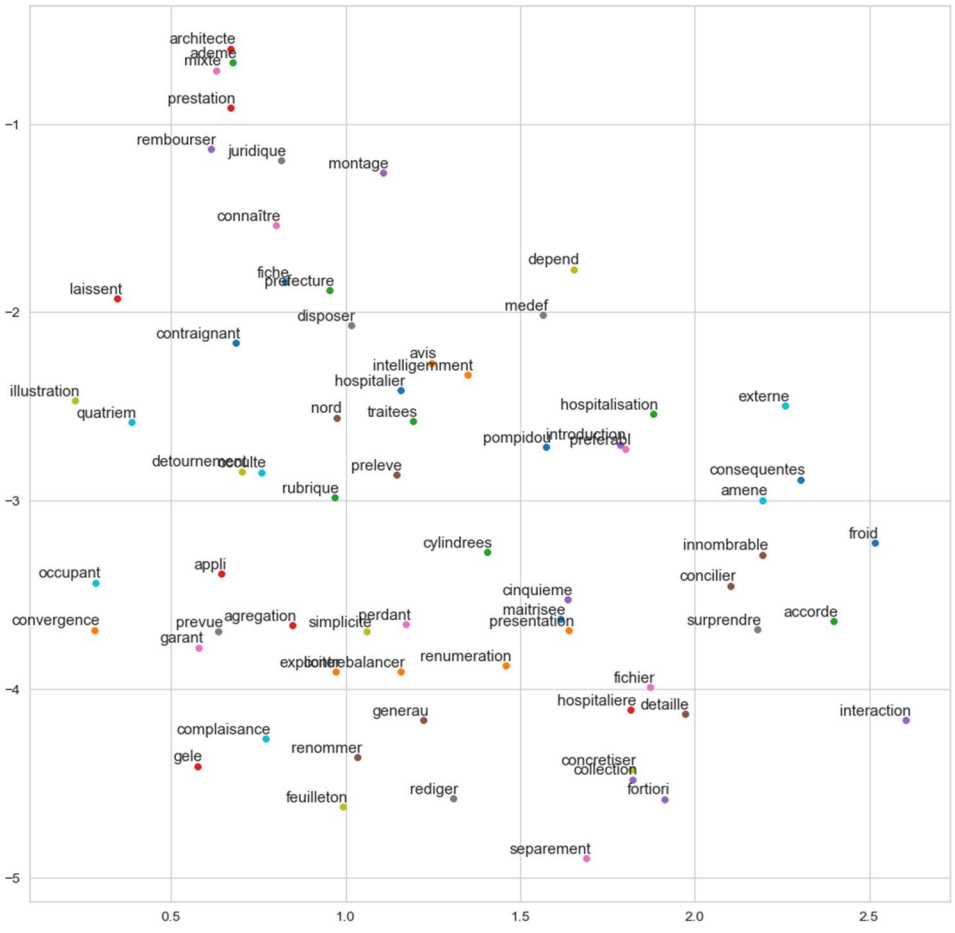
c) Conclusion

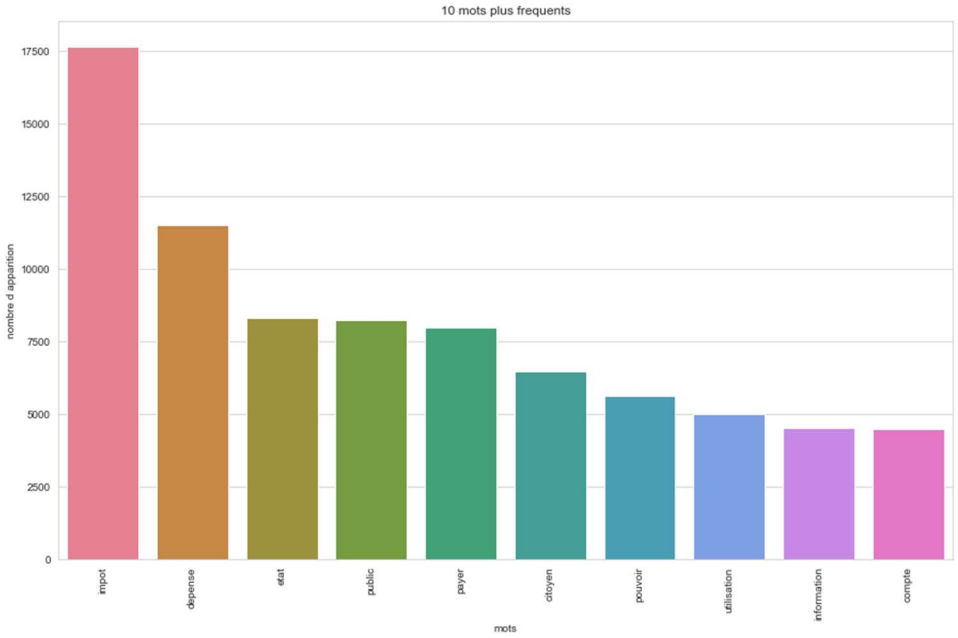
D'un point de vue écologique les déclarations de Macron sont assez vagues avec des annonces pour prendre des mesures d'ici 2025. Cependant on peut voir à travers les réponses aux différentes questions une envie d'aider la planète et de développer l'écologie avec si possible une aide de l'Etat dans ce sens.

3. La fiscalité et les dépenses publiques

a) “Quelles sont toutes les choses qui pourraient être faites pour améliorer l'information des citoyens sur l'utilisation des impôts ?”





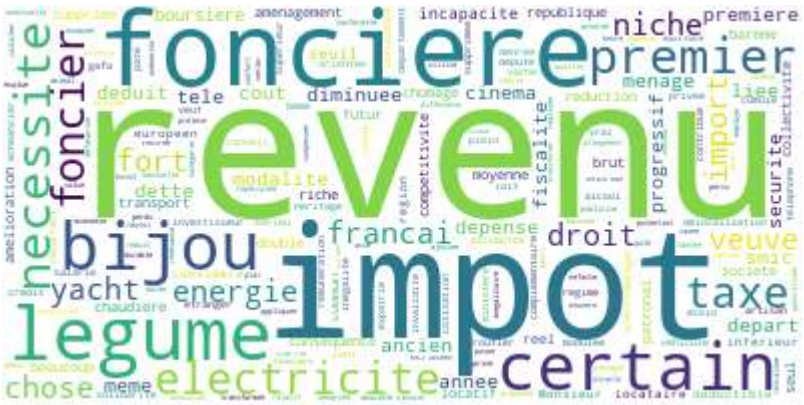


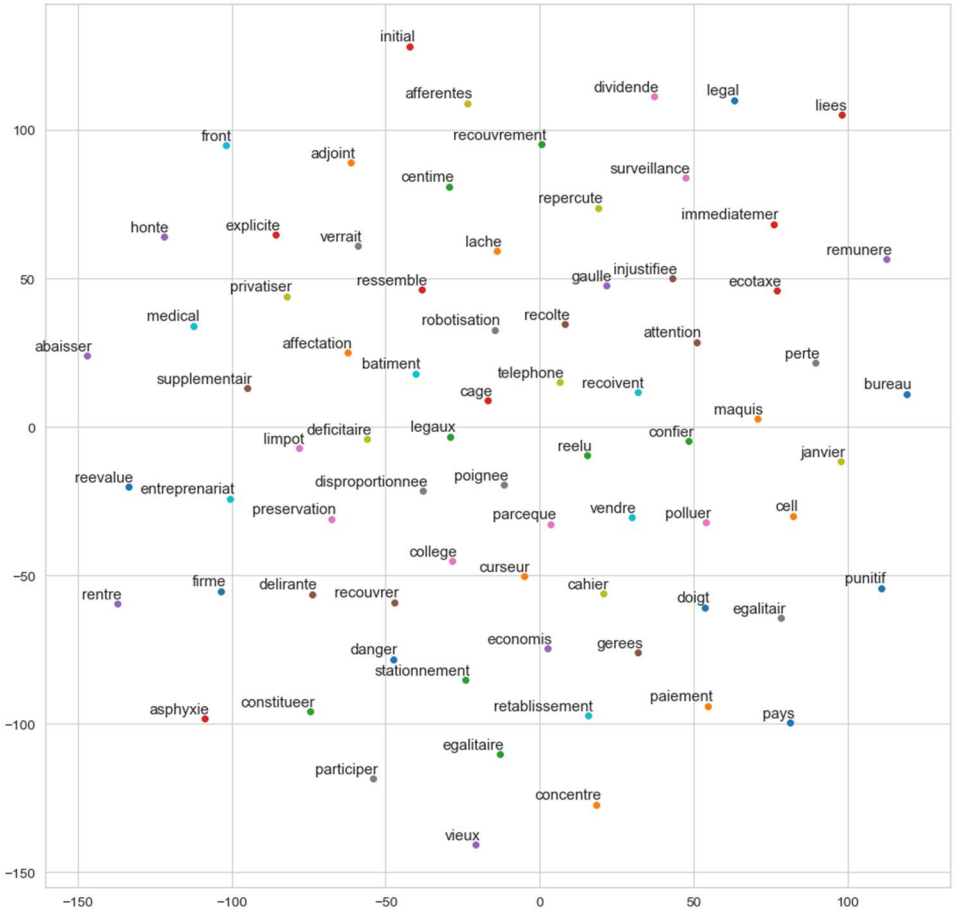
Topics found via LDA:

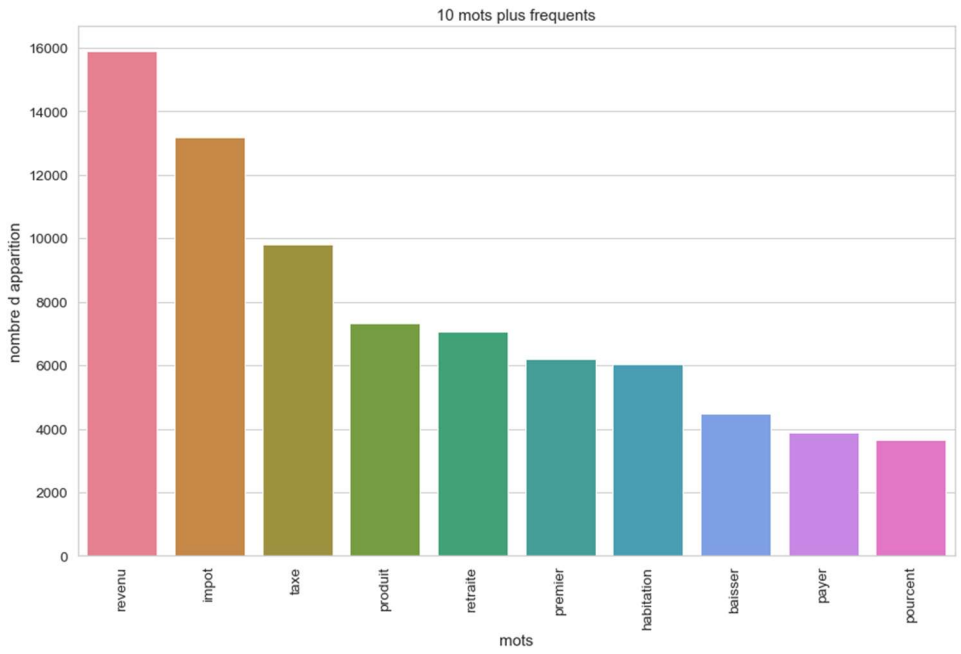
Topic #0:
impot utilisation budget internet depense annee
Topic #1:
retraite pourcent taxe revenu pouvoir social
Topic #2:
depense salaire fonctionnaire avantage frais public
Topic #3:
impot fiscal payer pouvoir revenu pourcent
Topic #4:
depense information compte site citoyen impot
Topic #5:
public service cout impot etat ecole
Topic #6:
etat impot depense annuel taxe commune
Topic #7:
payer impot falloir aller servir gens

Résultats obtenus : Les principaux sujets relevés sur cette question sont sur l'usage des impôts. Cependant, on peut remarquer que cela n'est pas en adéquation avec la question qui porte sur l'information de l'usage des impôts. Concernant l'usage des impôts différents avis sont émis comme l'utilisation des impôts pour la commune ou pour les salaires des fonctionnaires.

b) “Quels sont selon vous les impôts qu'il faut baisser en priorité ?”







Topics found via LDA:

Topic #0:
taxe habitation fonciere supprimer suppression local
Topic #1:
social baisser injuste societe travail salarial
Topic #2:
produit premier necessite carburant alimentaire local
Topic #3:
foncier permettre societes consommation premiere charge
Topic #4:
impot falloir baisser entreprise fiscal pouvoir
Topic #5:
retraite salaire pourcent etat aujourd euro
Topic #6:
revenu moyen tranche impot payer pourcent
Topic #7:
succession impot travail bien payer droit

Résultats obtenus : Les principaux sujets relevés sur cette question sont : Les différents types d'impôts et leurs nécessités, le souhait d'être enlevé notamment la taxe foncière et la justice fiscale.

c) Conclusion

Dans les annonces faites par Emmanuel Macron faites sur Reddit, on peut voir qu'il parle de justice fiscale notamment une baisse d'impôts pour ceux qui payent l'IR et la baisse d'impôts pour les entreprises est un sujet que notre algorithme a réussi à capter. Une des inquiétudes perçues est le salaire des fonctionnaire, chose à laquelle Emmanuel Macron a répondu. Une des limites de notre travail est le caractère très précis des recommandations faites quand Emmanuel Macron évoque des sujets plus globaux.

Conclusion global

Parmi toutes les méthodes que nous avons mis en place, celle qui s'est révélé le plus efficace a été le LDA car elle donne rapidement un aperçu de l'ensemble des sujets. La méthode de résumé de TextRank est aussi très efficace mais nécessite beaucoup de temps et de pré-processing pour ordonner les données obtenues.

Nous avons été surpris par l'avantage que propose un wordcloud d'un point de vue de la visualisation des données car il met en avant les différents mots clés une fois les données preprocessés.

Pour le word embedding, nous pensons que notre méthode n'a pas été suffisamment fine pour garder les mots les plus importants mais que nous aurions pu parvenir à un résultat très intéressant.

De plus, il s'est cependant avéré plus complexe de traiter des données en français qu'en anglais comme pour le premier projet. Les librairies sont moins performantes et moins nombreuses et il faut parfois les compléter comme pour les stops words.

Pour le futur, nous aimerions affiner notre pré-processing pour le word embedding en essayant de pos-tagger les mots et de filtrer plus finement via le tf-idf (avec un seuil supérieur et un seuil inférieur au lieu de simplement un seuil inférieur). Pour le résumé, il aurait pu être intéressant de lancer un LDA sur un résumé issu de TextRank pour affiner encore le LDA sur les réponses les plus pertinentes. Il aurait aussi pu être intéresser d'extraire les mots clés issues des résumés de textrank, dans la même optique.