Johann de Soyres
Henri de Montalembert

Adv. Machine Learning:
Rapport Projet 2

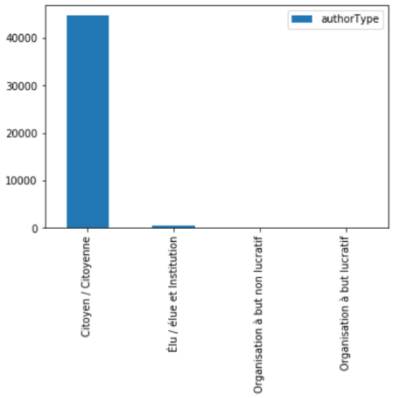
Thème choisi :

Transition écologique

Dataset :

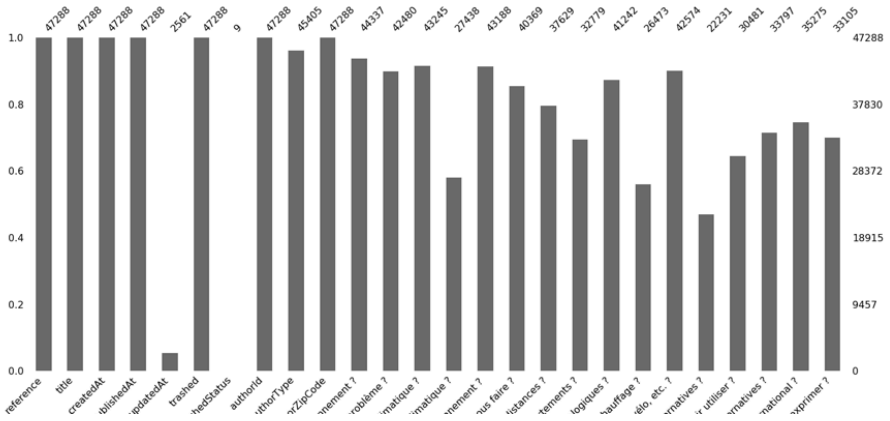
http://opendata.auth6f31f706db6f4a24b55f42a6a79c5086.storage.sbg5.cloud.ovh.net/2019-02-06/LA_TRANSITION_ECOLOGIQUE.csv

Le grand débat national a été une grande réussite sur le thème de l'environnement. En effet, 47 288 personnes ont voulu donner leur avis sur cet enjeu. Pour la plupart, ce sont des citoyens ou des citoyennes. Cependant, certains élu ou organisation ont participé à ce débat.



	authorType
Citoyen / Citoyenne	44716
Élu / élue et Institution	458
Organisation à but non lucratif	179
Organisation à but lucratif	52

En moyenne, 73 % de ces individus ont répondu et 60 % ont répondu à l'ensemble de toutes les questions. Cela, nous montre l'importance de la transition écologique pour les français.

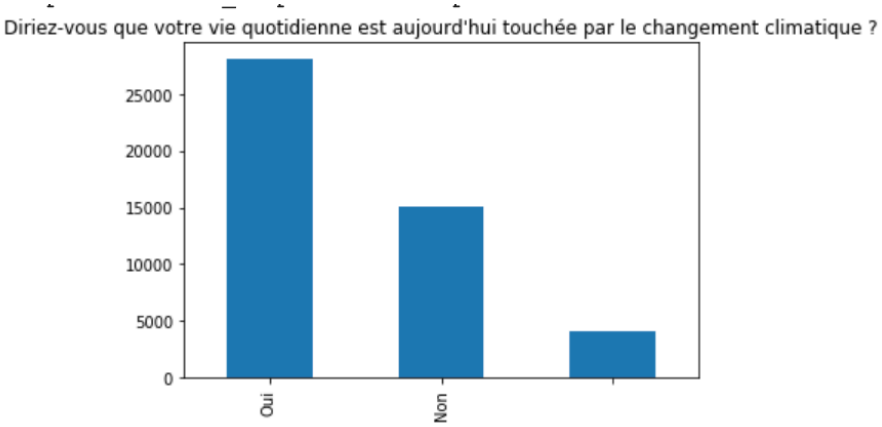


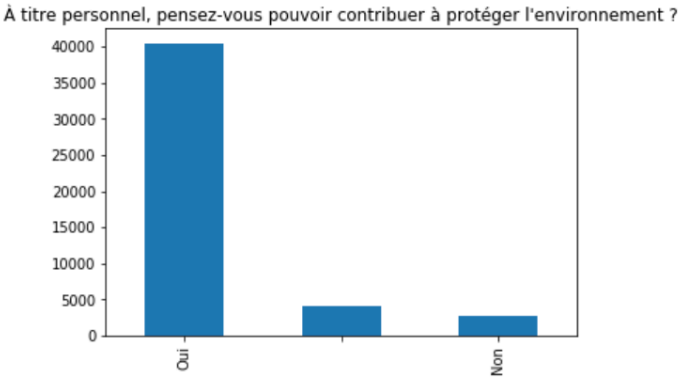
Le grand débat sur ce thème a intéressé de façon homogène toute la France. En effet, une petite analyse sur le zip code nous montre que pratiquement tous les départements ont participé et que les plus les plus grandes villes regroupe le plus de participants.

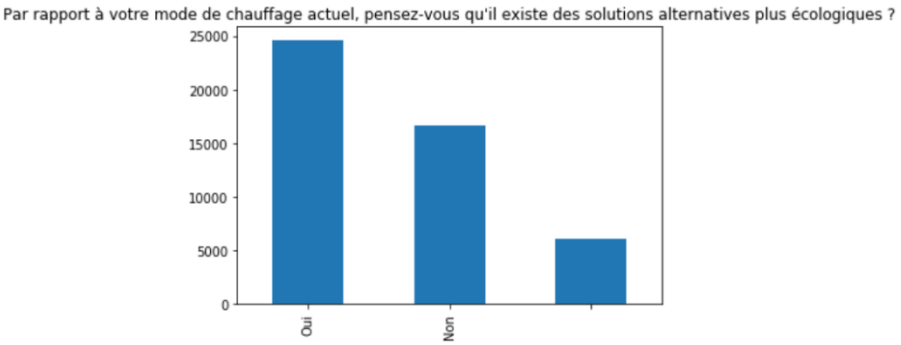
	authorZipCode
75	2573
69	1902
13	1683
31	1625
92	1562
33	1521
78	1518

I. Questions fermées

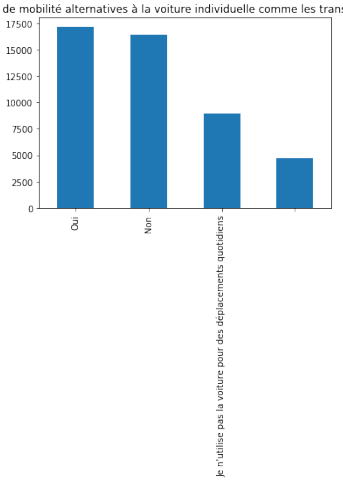
Les questions fermées ont été plutôt simple à analyser. Pour chaque question nous avons afficher la répartition des réponses. La réponse sans label correspond aux valeurs NAN qui étaient présentes dans le dataset pour la question donnée.







Avez-vous pour vos déplacements quotidiens la possibilité de recourir à des solutions de mobilité alternatives à la voiture individuelle comme les transports en commun, le covoiturage, l'auto-partage, le transport à la demande, le vélo, etc. ?



II. Questions ouvertes

Afin d'analyser les questions ouvertes nous avons utilisé plusieurs techniques qui sont les suivantes :

- Analyse des mots les plus représentés
- Résumé des réponses des individus
- LDA
- T-SNE

Pour permettre l'application de ces techniques, certains pré-traitements ont été nécessaires. Pour chaque question on a regroupé les réponses dans une seule et même variable. On se retrouve donc avec la structure de donnée suivante :

	question	responses
0	Que faudrait-il faire selon vous pour apporter...	Les problèmes auxquels se trouve confronté l'e...
1	Si oui, de quelle manière votre vie quotidienn...	Pollution de l'air, pollution de nos aliments,...
2	Si oui, que faites-vous aujourd'hui pour proté...	En consommant autrement, en vivant autrement. ...
3	Qu'est-ce qui pourrait vous inciter à changer ...	développer les transports en commun , Aménag...
4	Quelles seraient pour vous les solutions les p...	Plus de transports publics dans les petites co...

En plus de ça les pré-traitements habituels (tokenization, stop words, lemmatization et suppression de la ponctuation) ont été effectué sur chacune des réponses. Nous avons gardé une copie de cette structure sans les pré-traitements pour le résumé des réponses des individus. En effet nous avons observé que pour cette technique il était préférable de ne pas appliquer de pré-traitements afin d'obtenir un bon résultat.

1) **Analyse des mots les plus représentés :**

<u>Quel est aujourd'hui le problème le plus concret dans le domaine de l'environnement ?</u>

Ressortir les mots les plus important tout en retirant les STOPSWORD nous permet de faire ressortir les problèmes les plus concret. Pour une meilleure visualisation, nous avons représenté ces mots dans un nuage de mots. La disparition des espèces, la crue, la sècheresse ainsi que la pollution sont les domaines ou les français sont les plus inquiets. L'utilisation de gensim nous a permis de créer un résumé des réponses des individus. Cette technique a été celle qui nous a permis de mieux analyser ces questions.

2) **Résumé des réponses des individus**

<u>Que faudrait-il faire selon vous pour apporter des réponses à ce problème ?</u>

Les français ont apporté beaucoup de réponse à cette question. En effet, ils ont proposé de nombreuses idées qui sont les suivantes :

- Favoriser le réutilisable
- Se passer des climatisations
- Les prix des transports en commun
- Réduire le nombre de publicités
- Avantager le chauffage écologie
- Sortir de la société de consommation
- Réduire la consommation d'énergies fossiles
- Développer l'emploi local
- Limiter les importations de produits manufacturés, de fruits, de légumes
- Produire local
- Surtaxer les pollueurs

<u>Y a-t-il d'autres points sur la transition écologique sur lesquels vous souhaiteriez vous exprimer ?</u>

- Location des batteries pour les voitures électriques qui ajoutent un prix fixe qui revient supérieur au prix d'un plein
- Supprimer sur certains produits alimentaires qui sont bons pour l'environnement et la santé

3) LDA :

Pour faire ressortir les thèmes principaux de toutes du questionnaire sur la transition écologique, nous avons utilisé le modèle **Latent Dirichlet Allocation** (LDA). Ce modèle permet de découvrir des structures thématiques cachées dans des vastes archives de documents. Pour appliquer ce modèle, nous avons tokenizer l'agrégat des réponses pour chaque question. Nous avons fixé le nombre de thèmes à 4 et le nombre de mots par thème à 5. Les thèmes générés sont les suivants :

- `(0, '0.018*"plus" + 0.008*"transport" + 0.008*"a" + 0.006*"moins" + 0.006*"pollution"')`
- `(1, '0.019*"plus" + 0.010*"transport" + 0.008*"a" + 0.007*"commun" + 0.005*"chauffage"')`
- `(2, '0.047*"transport" + 0.032*"commun" + 0.017*"vélo" + 0.011*"demande" + 0.010*"covoiturage"')`
- `(3, '0.011*"plus" + 0.007*"a" + 0.006*"faire" + 0.005*"transport" + 0.005*"chauffage"')`

On remarque qu'un même mot se retrouve dans plusieurs thèmes comme « transport » ou « commun ». Les thèmes générés ici ne sont pas bien distincts. De plus malgré les pré-traitements effectués, des stops words persistent.

4) T-SNE :

T-SNE est (**t-distributed stochastic neighbor embedding**) est une technique de réduction de dimension pour la visualisation de données. Notre but ici était de représenter les réponses a la question **« Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ? »** dans un espace a deux dimensions. Nous avons d'abord récupéré la colonne correspondante dans le dataset initial. Apres avoir nettoyer chacune des réponses (tokenization, stop words, lemmatization, suppression de la ponctuation et remplacement des NAN par ' ') on a construit un corpus. Chaque élément du corpus est une réponse, cette réponse est stockée sous forme de liste de mots.

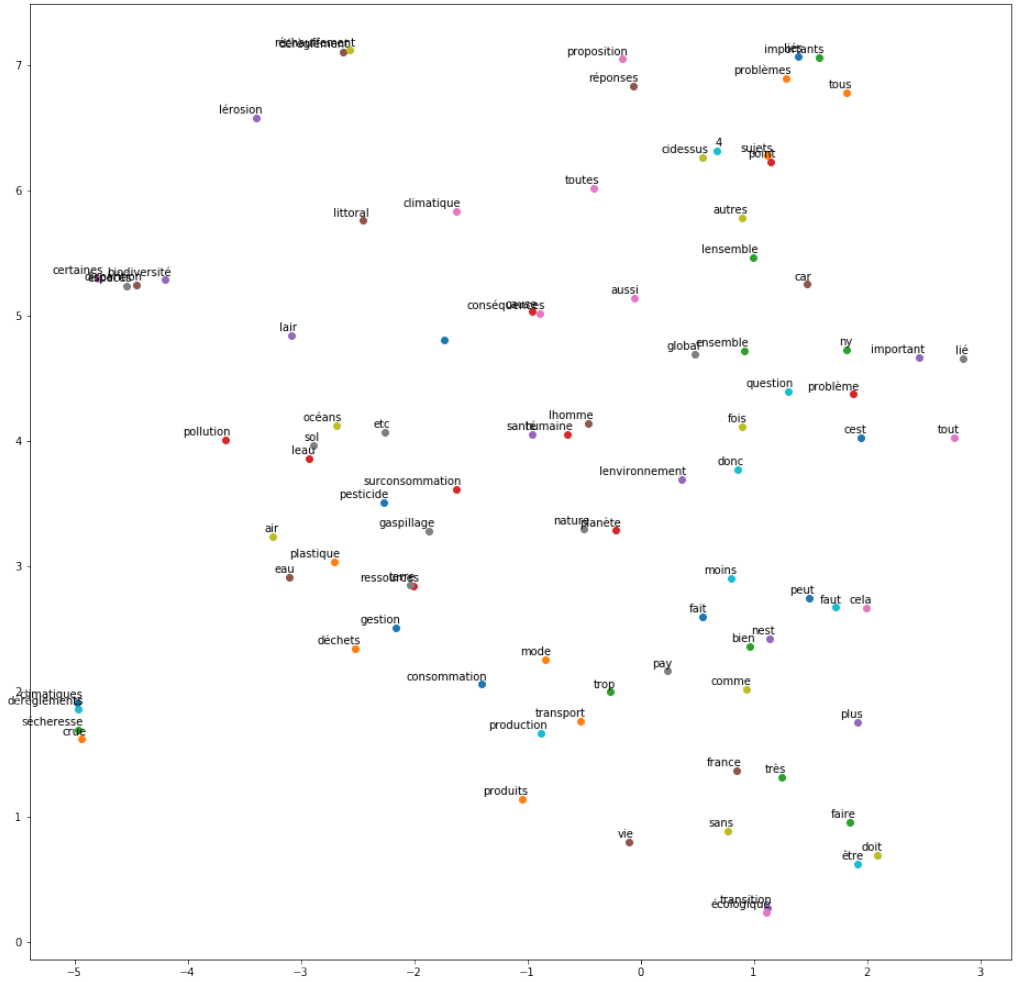
```
[['létat', 'melle'],

['dérèglements', 'climatiques', 'crue', 'sécheresse'],

[''],

['pollution', 'créée', 'lhomme', 'règle', 'générale'],

['pollution', 'lair'],

['désinformation','problèmes','environnementaux','lobby','vegan','relayés','médias'],

…

['voir','problèmes','ensemble','pourquoi','vouloir','tout','catégorise','avancer','différents','front']]
```

On va ensuite appliquer le modèle Word2Vec qui va convertir ce corpus en une liste de 89 mots. Le modèle choisit automatiquement les mots les plus représentatifs du corpus. Chaque mot devient un vecteur de taille 100.

```
model["proposition"]
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: DeprecationWarning: Call
  """Entry point for launching an IPython kernel.
array([-9.15320367e-02,  3.42665941e-01,  8.02600682e-01, -3.07124764e-01,
        1.02689765e-01,  6.97112605e-02,  7.16224611e-01,  1.32602811e-01,
       -5.03182411e-01, -4.72031385e-01, -6.18528306e-01,  4.12149578e-01,
       -1.28785729e-01,  2.10017711e-02, -2.10247561e-03,  2.18637735e-01,
       -5.25150113e-02, -1.29674718e-01, -1.06485590e-01, -2.08779782e-01,
       -3.06450784e-01, -3.20786297e-01,  4.96977985e-01, -2.13105112e-01,
        1.00345835e-01, -3.23908508e-01, -4.51208919e-01,  1.38890922e-01,
       -2.24548087e-01,  1.19879343e-01,  8.16268176e-02,  1.15384050e-01,
        3.95958275e-02, -2.70038337e-01, -8.42142925e-02,  2.86038697e-01,
        1.14188172e-01, -3.35753530e-01,  3.49220961e-01, -1.24098204e-01,
       -2.84833014e-01,  2.69076675e-01, -2.60396134e-02, -1.37321070e-01,
       -3.67489457e-01, -7.92706683e-02, -4.28822994e-01,  1.46129414e-01,
       -2.84632057e-01,  2.50555295e-02, -3.18931453e-02, -1.71147853e-01,
       -1.24434948e-01,  2.93476671e-01,  4.51309502e-01, -3.00845414e-01,
        2.44474541e-02,  4.90390579e-04,  8.03812370e-02,  2.43325844e-01,
       -1.58493787e-01,  6.38563931e-02, -7.29261100e-01, -3.77558142e-01,
        3.79984170e-01,  3.66662562e-01, -3.19984257e-02, -6.60669282e-02,
       -1.00412883e-01, -1.05655260e-01,  1.16740085e-01,  5.06301187e-02,
        1.96248014e-02,  4.12126668e-02,  1.50890052e-01,  1.06165342e-01,
       -1.79793090e-01, -3.83318424e-01, -4.94054593e-02, -3.51032764e-01,
       -4.71791059e-01,  3.11054468e-01, -4.66138460e-02,  5.71201146e-01,
       -1.97179317e-01,  1.26988649e-01,  1.38423875e-01,  2.07814034e-02,
       -2.63865858e-01,  1.10781156e-01, -3.52379680e-01,  1.95396453e-01,
       -4.74626273e-01,  3.16499472e-01,  5.56768954e-01,  1.85644394e-03,
        3.22667181e-01,  1.75561123e-02,  3.01949948e-01,  2.21599452e-02],
      dtype=float32)
```

Grace au T-SNE on va représenter ces 89 mots, initialement dans un espace a 100 dimensions, dans un espace a 2 dimensions. Finalement on obtient la figure suivante :



On peut ainsi voir les mots les plus représentatifs des réponses a la question **« <u>Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ?</u> ».** On peut également voir les mots que le modèle word2vec considère comme similaire (porteur du même sens, apparait dans le même contexte, etc.).

III. Conclusion

En résumé nous avons applique 4 techniques sur les questions ouvertes. L'analyse des mots les plus représentés, le résumé des réponses des individus et le modèle T-SNE s'appliquaient par question et ont permis d'avoir une vue d'ensemble des réponses. Les résultats sont bons et cohérents avec les questions. Enfin le modèle LDA a permis de faire ressortir les thèmes de toutes les réponses. Les résultats sont cette fois moins intéressant puisque les thèmes ne sont pas directement identifiables. De plus nous avons réalisé plusieurs visualisations des réponses aux questions fermées. Grâce à cette analyse, nous nous sommes rendu compte que le gouvernement n'a pas correctement écouté les Français. L'écologie semble être plus une priorité pour les français que pour le gouvernement. Ces derniers ont annoncé très peu de réforme comparé aux propositions des français.