Analyse d'une consultation citoyenne

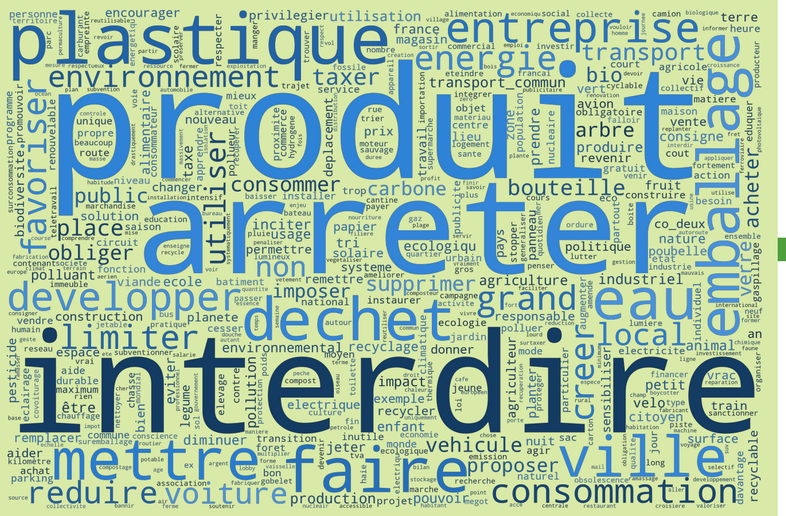
« *Comment agir ensemble dès maintenant pour l'environnement ?* »

La problématique

Plusieurs centaines de milliers de citoyens ont participé en 2020 sur *Make.org* à une concertation sur l'environnement et les manières d'agir. Des milliers de propositions ont été faites. Nous avons eu pour objectif de déterminer les thématiques principales de celles-ci par l'intermédiaire du NLP : Le *Natural Language Processing*.

La consultation en quelques chiffres :

- **9 501** propositions
- **4 469** auteurs
- **540 596** participants



Le pré-traitement

Objectif : Découper chaque proposition en mots et ne considérer que les informations importantes.

Exemple :

« boycotter les fast-food éthiquement discutables (exemples : Burger king, KFG, etc…) »

Normalisation (accents + majuscules)	boycotter les fast-food ethiquement discutables (exemples : burger king, kfc, etc…)
Suppression des *stop words*	boycotter fastfood ethiquement discutables exemples burger king kfc
Lemmatisation	boycotter fastfood ethique discuter exemple burger king kfc

Pour aller plus loin

Spacy (Python), Lemmatizer

Glossaire

Stop word : Terme apportant peu de sens à une phrase (ex : de, ?, où)

Lemmatisation : Transformer en une forme plus générique les mots ex : 'aimeras' et 'aimeriez' deviennent 'aimer'

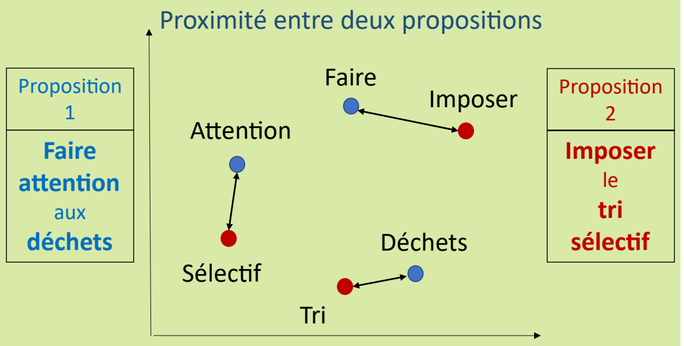
NLP : *Natural Language Processing*, Domaine de l'informatique où l'on étudie échanges textuels et oraux.

L'*embedding*

Objectif : Transformer les propositions en données numériques et les représenter graphiquement afin de pouvoir étudier leur sens et leur proximité.

Solution : 2 méthodes de représentation :

- le *word embedding* (ou « plongement de mots »)
- le *doc embedding* (ou « plongement de documents »)



Sur le graphique : 2 propositions de 4 mots, « aux » et « le » sont des stop words

Deux propositions sont proches lorsque leurs mots sont proches en sens (côté-à-côté sur le graphique).

Pour aller plus loin

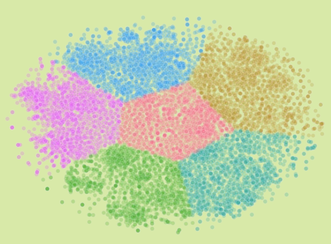
Word embedding : GloVe, Word2vec, (CBOW, Skip-gram), Hypothèse distributionnelle
Évaluation : dictionnaire BATS, similarité cosinus, RMSE
Doc embedding : distance WMD, TF, TF-IDF, valeur moyenne, MDS

La classification

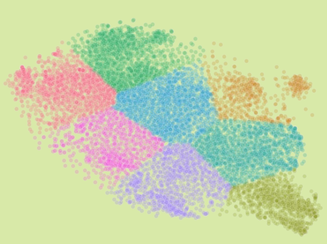
Objectif : Former des familles de propositions semblables.

Solution : Les méthodes de classification

Résultats obtenus avec trois *word embeddings* différents

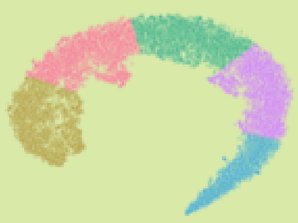


6 groupes (GloVe)
(modèle retenu)



8 groupes (Skip-Gram)

5 groupes (CBOW)



Pour aller plus loin

Réduction de dimension : t-SNE
Clustering : K-means, GMM, HDBSCAN
Coefficient de silhouette, Calinski-Harabasz,DBCV

Résultat final

On obtient finalement **6 classes** de propositions : chacune d'entre elles est representée dans la fleur au centre.

Alimentation	Démographie Finance	Recyclage & emballage	Transport	Eau & sécheresse Végétalisation	Produits technologiques Déchets

Année universitaire 2021-2022 | **Tutrice** : Anne-Cécile Gay | **Responsable du module** : David Audenaert
Étudiants : Anne Sophie Le Bayon, Hugo Miccinilli, Tangi Tassin | Groupe n°03