Comment agir ensemble pour une mode plus responsable ?
Analyse par NLP

Hugo BOUTTES, Théo DAMPERON, Bachir SABO, Laurent SPILLEMAECKER
Tutrice : Anne-Cécile GAY

Contexte

"make.org" est une plateforme de consultation citoyenne sur laquelle les utilisateurs peuvent émettre des propositions sur des sujets de société. Nous nous sommes concentré sur le sujet de la mode responsable. Le but de ce projet est d'automatisé la labellisation des propositions avec le NLP.

Objectifs

<u>Méthode de classification supervisée :</u>
Classer les nouvelles propositions à partir des 12 différents labels choisis au préalable.
critère d'évaluation ► precision, recall, accuracy (F1-score)
<u>Méthode de classification non supervisée :</u>
Déterminer des groupes de propositions sur la base de similitudes.
critère d'évaluation des clusters ► règle du coude, silhouette

Méthodologie

Étape 0 : texte brut
'Il faut interdire aux marques de jeter des vêtements neufs.'

Étape 1 : tokenisation
On isole les mots de la proposition
["il","faut","interdire","aux","marques", "de","jeter,"des","vêtements","neufs"]

Étape 2 : retrait des stopwords
On enlève les mots "inutiles".
["interdire","marques","jeter", "vêtements","neufs"]

Pré-processing

Étape 3.1 : racinisation
On ne garde que la racine des mots.
["interdir","marqu","jet", "vet","neuf"]

Étape 3.2 : lemmatisation
On "simplifie" les mots.
["interdire","marque","jeter", "vetement","neuf"]

Normalisation

WordsCloud du top 30 des mots les plus fréquents après lemmatisation

Étape 4.2
Méthode de vectorialisation utilisée:
Fasttext

Étape 4.1.2 : TF-IDF
On pondère chaque mot en fonction de sa fréquence d'apparition dans un dictionnaire précis.
[0.2, 0.2, 0, 0.2, 0.2, 0.2] avec le dictionnaire
{"interdir","jet","lou","marq","neuf","vet"}

Étape 4.2
Méthode de vectorialisation utilisée:
Fasttext, Word2vec, Doc2vec, DLA

Vectorialisation

Méthode supervisée
Nous avons utilisés 5 méthodes de modélisation ici :
► **Bayésien naïf**,
► **Fasttext**,
► **Régression logistique**,
► **SVM** (machine à vecteurs de support)
► **Random forest**.

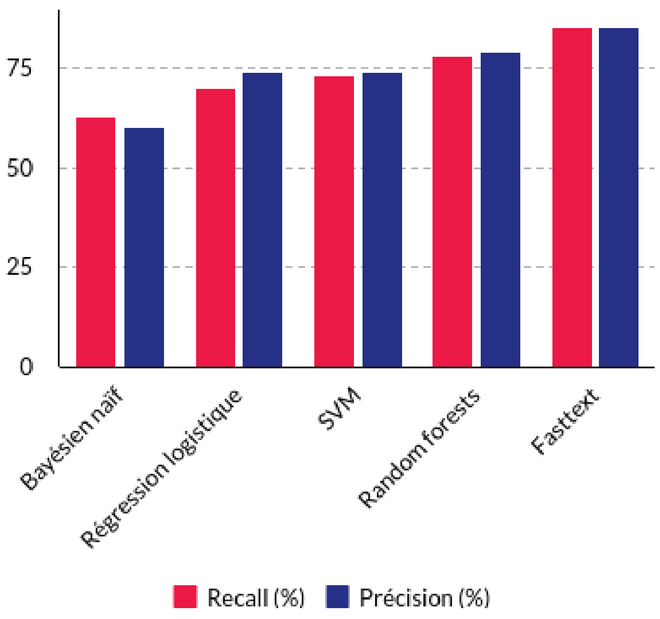
Méthode non supervisée
► Nous avons utilisé plusieurs méthodes de clustering ici :
HDBSCAN, **Kmeans**, et le **Spectral clustering.**
Les vecteurs utilisés pour ces méthodes en inputs sont des fasttext et les word2vec.

► Nous avons aussi testé un algorithme de clustering utilisant comme input des vecteurs issus de la **LDA**. Il s'agit en fait d'une modélisation thématique : cela permet de choisir directement les thèmes de nos différents clusters.

Résultats & conclusion

Méthode supervisée

Comparaison de la qualité des différentes méthodes d'apprentissage supervisé



VS

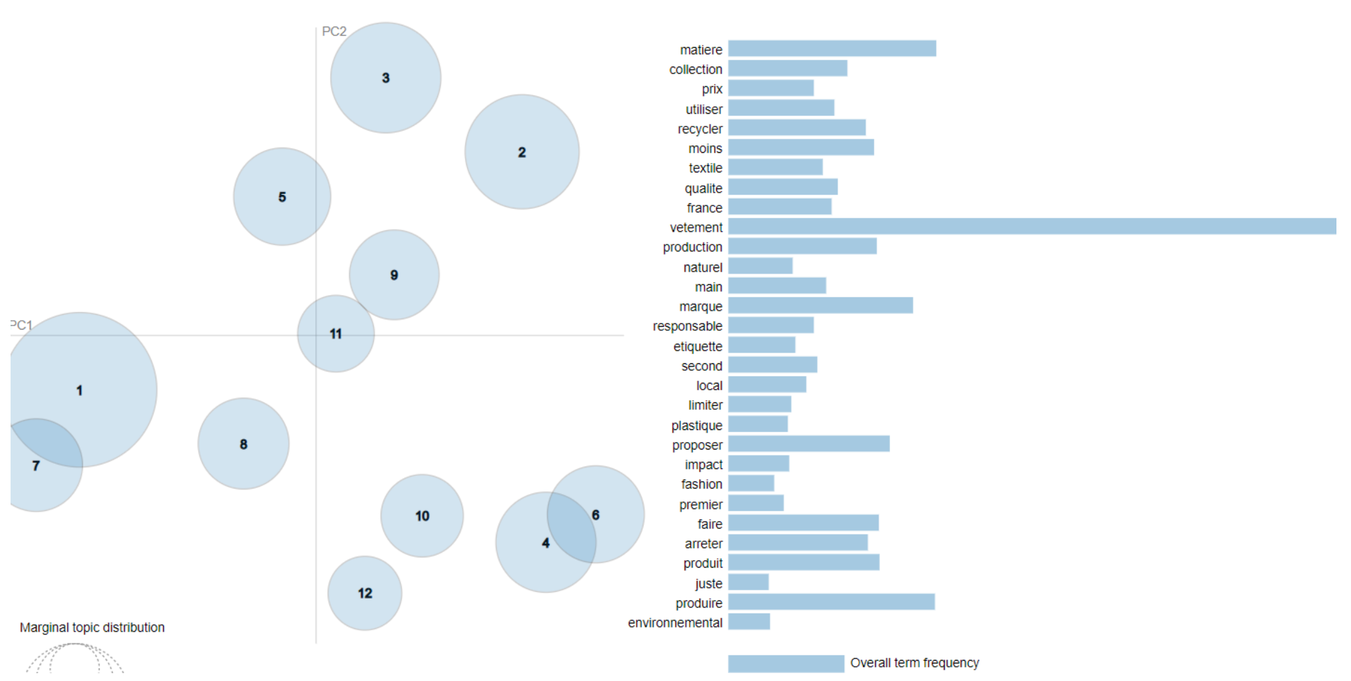
On cherche à évaluer la qualité d'une modélisation à l'aide de deux indicateurs. Le premier est le "recall", le ratio entre les vrais positifs et toutes les observations de la classe. Le deuxième est la "précision", le ratio entre les vrais positifs et les positifs.

Méthode non supervisée

► La méthode du coude nous donne 4 et 6 cluster respectivement pour la modélisation word2vec et fasttext
► score de silhouette de 0.6 avec la modélisation word2vec & 0.35 avec le fasttext.

Le word2vec discrimine mieux nos clusters. Il permet des clustering plus rapide (2 à 5 fois plus rapide que pour le fasttext).

► Clustering avec une modélisation thématique LDA (nombre de thèmes imposé : 12)



Conclusion

Pour la modélisation supervisée, nous avons retenu les modèles random forrest et fasttext. Pour la modélisation non supervisée la méthode LDA a permis de faire ressortir les différents thèmes abordés dans notre corpus.
Pour répondre à la problématique, une modélisation supervisée permet de bien classer les nouvelles propositions, néanmoins la modélisation thématique (non supervisée) fait ressortir des nouveaux thèmes parfois plus pertinent. (Fast-fashion pour le topic9).
Une combinaison judicieuse de ces deux méthodes permet de meilleurs résultats.