Le Machine Learning pour coder dans une nomenclature : l'enquête Associations

Florian Lécrivain (Insee)

Séminaire Big Data de l'Insee (14 janvier 2020)

Insee

Mesurer pour comprendre

01 Contexte et objectif

L'enquête Associations

- Objectifs : dénombrer le nombre d'associations actives en France, connaître leur(s) domaine(s) d'activité, évaluer la participation associative, connaître leurs ressources...
- Champ : ensemble des associations relevant de la loi de 1901
- Première édition réalisée en 2014 (sur l'année de référence 2013).
- La seconde édition a été lancée fin 2019 (sur l'année de référence 2018).

Base de sondage et échantillonnage

- Base de sondage constituée à partir :
  - du Répertoire National des Associations (RNA)
  - du répertoire Sirène
- Échantillon stratifié selon :
  - **Le domaine d'activité**
  - L'age de l'association
  - Le nombre de salariés (pour les associations employeuses)

Plusieurs nomenclatures pour le domaine d'activité

- 3 nomenclatures différentes pour l'activité entre le répertoire Sirène, le RNA et l'enquête.
- On dispose avec le RNA de l'objet social de l'association, texte libre qui décrit ses activités :

Notre association a pour objet l'enseignement spécialisé de la musique, danse et théâtre.
Réaliser produire diffuser et promouvoir des projets culturels et des créations artistiques.
Organiser les loisirs de la collectivité dans son ensemble. Favoriser la pratique de l'éducation physique et sportive. Renforcer la solidarité morale des habitants.
Développement connaissance pratique du sport en général et en particulier des sports de combat ainsi que le maintien des traditions propres à chaque discipline.
Nous sommes une association de commerçants elle sert à animer le centre ville par des festivals commerciaux qui fait gagner à nos clients des lots.

Objectif

- **Objectif** : exploiter l'objet social afin de coder le domaine d'activité de l'association selon la nomenclature propre à l'enquête.
- Expérimentation réalisée en collaboration avec l'unité SSP Lab de l'Insee de janvier à mai 2019.
- Tirage de l'échantillon en mai, en mettant à profit le domaine d'activité codé avec ces travaux.

02 Les données utilisées

Les données labellisées

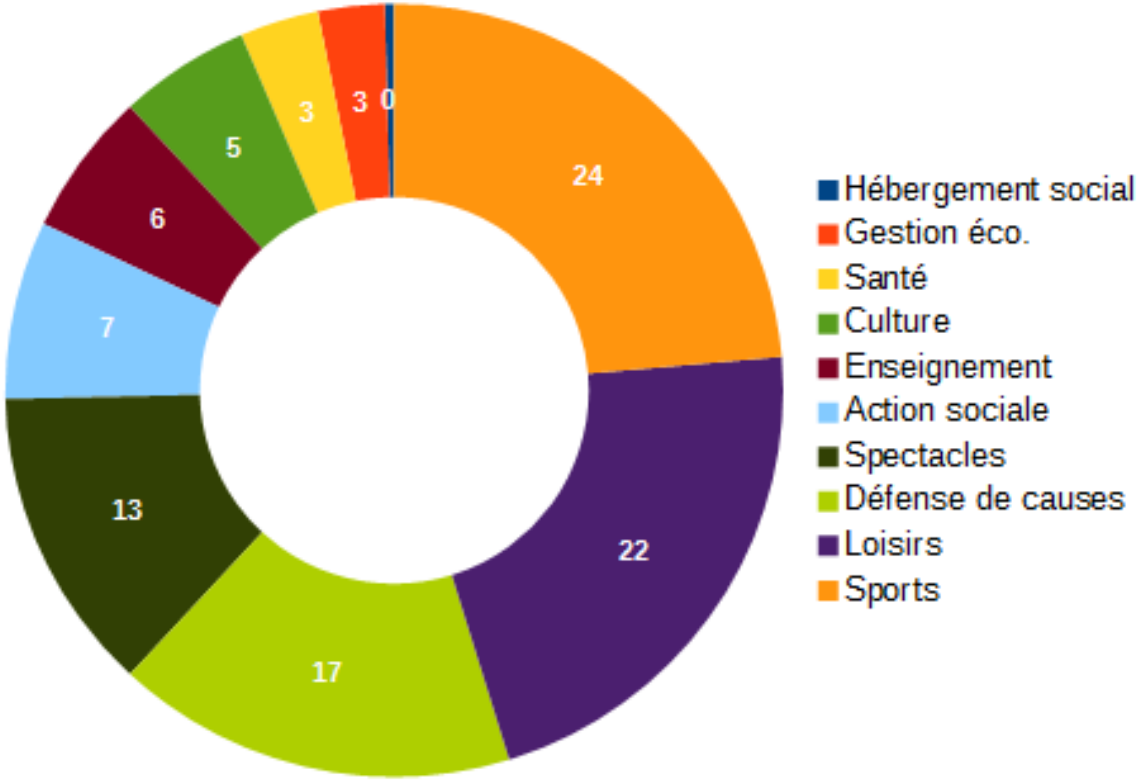
- Les réponses de la première enquête Associations fournissent, après appariement avec le RNA, un jeu de données labellisées de **17 200 observations**.

  → usage possible de techniques d'apprentissage supervisé

Les domaines d'activité de l'enquête

- 64 modalités regroupées en **10 grands domaines**

Figure 1 – Répartition des associations selon leur principal domaine d'activité
(en 2013, source : Enquête Associations 2014)



Les variables explicatives retenues

- En plus de l'**objet social**, 3 autres variables corrélées au domaine d'activité sont utilisées :
  - Le statut employeur ou non de l'association
  - La taille de la commune d'implantation de l'association
  - L'age de l'association

Traitement de l'objet social

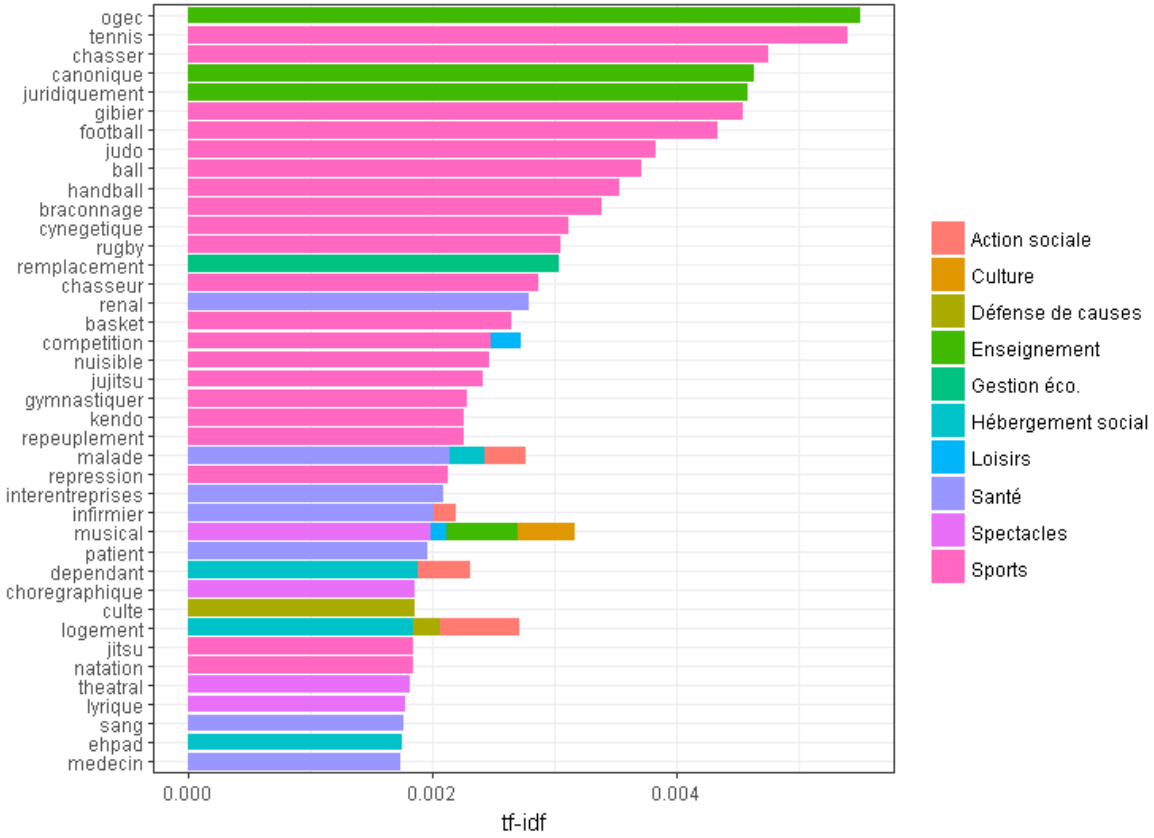
- Nettoyage des données textuelles
  - Suppression de la casse, de la ponctuation, des caractères spéciaux, etc.
  - Suppression des termes qui n'apportent aucune information (*Stopwords)*
  - Substitution des chiffres par le terme « number »
- Lemmatisation (dictionnaire Morphalou) → réduction de la dimension
- Améliorations apportées au fil de l'eau :
  - Prise en compte des fautes d'orthographe (ou de frappe) les plus courantes
  - Ajouts et corrections de certains lemmes

Représentation *bag of words*

- Après mise en forme des N objets sociaux, on compte l'occurrence de chaque terme (vocabulaire de taille V) dans chaque descriptif → on obtient une **matrice document-terme de taille N x V**
- Cette matrice est complétée de la même manière avec les **bigrammes**.
- Au final on obtient 16 200 lemmes et 55 800 bigrammes différents, soit 72 000 variables (de type indicatrices) → sélection de variables
- Les travaux sont menés en utilisant une **représentation simplifiée de l'objet social dite *bag of words***, où seule l'occurrence d'un terme est prise en compte.

Aperçu des termes les plus discriminants

Figure 2 - Termes les plus discriminants (selon le score tf-idf)



03 Apprentissage supervisé

Cadre des travaux d'apprentissage supervisé

- **3 familles de modèles testées** : Forêts aléatoires, *Extrem Gradient Boosting* (XGBoost) et machines de vecteurs à support (SVM)
- Optimisation des hyper-paramètres de chaque algorithme
- 5 types de sélection de variables testées (entre 28 et 6000 variables)
- Répartition des données entre jeu d'apprentissage et jeu test (80/20)

Cadre des travaux d'apprentissage supervisé

- Indicateurs de performance retenus :
  - Au niveau de chaque domaine
    - Précision
    - Rappel
    - **Score F1** : moyenne harmonique entre la précision et le rappel
  - Au niveau global
    - Précision globale et rappel global : moyenne pondérées des indicateurs au niveau domaine
    - **Score F1 global** : moyenne harmonique entre la précision globale et le rappel global
- Meilleures performances obtenues avec XGBoost (score F1 global de 68.6 %), mais les Forêts aléatoires (67 %) et les SVM (67.1 %) obtiennent des performances proches

Performances

- Sélection des lemmes et des bigrammes présents dans au moins 5 descriptifs → **score F1 global de 68.6 %**.

Tableau 1 – Performances du modèle XGBoost

	F1-score	*Précision*	*Rappel*
Ensemble	**68.6**	**68.7**	**68.5**
Action sociale	67.8	63.9	72.1
Culture	46.6	61.1	37.7
Défense de causes	58.9	64.7	54
Enseignement	70.5	68.2	72.9
Gestion éco.	67.2	72.7	62.5
Hébergement social	60.7	61.7	59.8
Loisirs	58.5	54.5	63.1
Santé	64.1	71	58.5
Spectacles	79.6	78.4	80.7
Sports	82.3	82.7	81.9

Performances

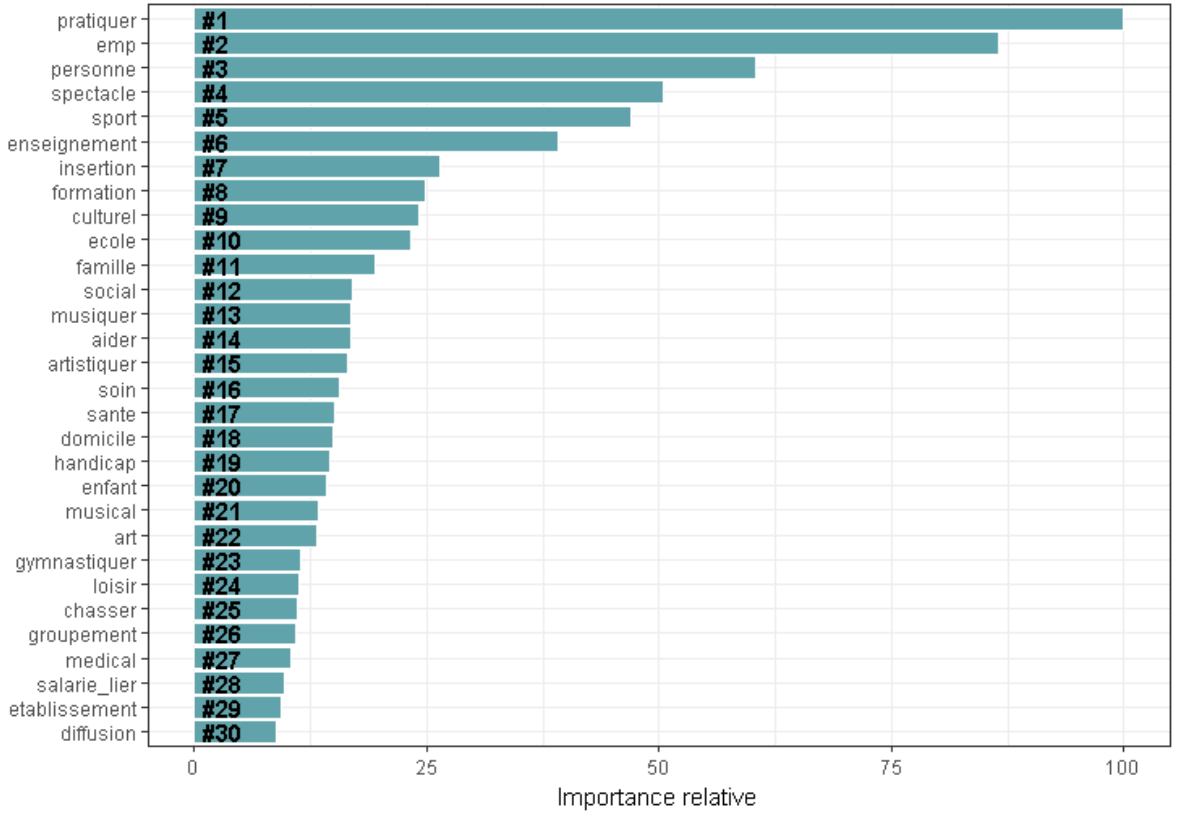
Figure 3 – Matrice de confusion associée à la précision

- Mauvaises prédictions : 28 % des unités classées par le modèle dans l'hébergement social appartiennent en réalité à l'action sociale.

Termes les plus influents

- Variables les plus importantes :
  « pratiquer », statut employeur ou non,
  « personne »,
  « spectacle » et « sport »

Figure 4 – Variables les plus influentes



04 Limites et pistes d’amélioration

Limites et pistes d'amélioration

- Des performances réduites pour certains domaines, en particulier la culture
- Une lemmatisation qui peut encore être améliorée, avec par exemple la prise en compte de la fonction grammaticale qui aurait permis de traiter les homonymies
- Utilisation possible de représentations plus complexes, telles que le *word embedding*
- Utilisation possible d'autres modèles d'apprentissage supervisé (réseaux de neurones par exemple), voir de combinaisons de différents algorithmes

Merci de votre attention !

florian.lecrivain@insee.fr

Retrouvez-nous sur :

insee.fr