Amélie Sciberras
Zitian Gao
Enrique Gómez
Cristian Rodríguez Rodríguez
Víctor Nàcher Castellet

Projet SD-TSIA210

Étude des données du Grand Débat National.

Sujet T3: Supervised learning for location prediction from the text content.

Description du choix du sujet et objectifs

Nous souhaitons réaliser une prédiction sur la localisation des réponses en fonction du contenu des contributions du grand débat. Nous avons d'abord pensé que nous focaliser sur le thème de la transition écologique était peut-être l'un des meilleurs moyens de prédire la localisation de la réponse car c'est un sujet dans lequel le lieu de vie a une grande influence. Par ailleurs, nous pensons trouver peu de variations dans les réponses qui viennent de grandes villes françaises. Nous pensons que les habitants de Lyon et de Toulouse sont globalement affectés de la même manière par la transition écologique et les mesures qui sont prises à ce sujet en France. L'utilisation de la voiture dans des agglomérations à même densité de population nous semble être similaire, comme le sont les problématiques de pollution de l'air. Cependant, nous pensons que le regard sur la transition écologique peut être très différent dans une zone peu peuplée (et donc moins sujette à la pollution de l'air et moins bien desservie en transports).

Nous pensons donc que prédire la densité de la population de la zone de la réponse en fonction du contenu texte de celle-ci peut être intéressant et permettrait de mesurer les contrastes qu'il existe entre les zones urbaines et rurales sur le thème de la transition écologique en France. Enfin, nous souhaitons réaliser une prédiction en fonction du contenu texte (comme demandé dans l'intitulé du sujet T3), donc nous étudierons plutôt les contributions que les questionnaires rapides

Projet : Prédire la densité de population du lieu de rédaction de la réponse en fonction du contenu textuel de celle-ci.

Avancement et objectifs.

Il nous semble ambitieux, et assez dépourvu de sens finalement, de prédire une densité de population précise en fonction du contenu des réponses. Nous avons donc décidé de faire deux clusters dans un premier temps : l'un avec des villes plutôt denses en population et l'autre avec des villes à faible densité de population.

1) **Prétraitement.**

La première étape du prétraitement consiste à lire le fichier .csv en brut et choisir les colonnes utiles, notamment les questions, le code postal et le titre de chaque entrée (cf. load_raw_data). On en fait une matrice avec comme ligne chaque questionnaire remplie et comme colonne chaque question. Puis on traite chaque colonne selon sa catégorie. En effet, les questions à réponses binaires (cf. preprocessing_yesno) sont traitées différemment que celles à choix multiple (cf. preprocessing_categories) et que celles à réponse libre (cf. preprocessing_text). Ce dernier traitement comporte le filtrage et la compte des mots. Pour ceci, on a utilisé les fonctions fournies dans le notebook (french-nltk.ipynb) de Christopher M. Church (cf. get_nltk_text, no_accents, filter_stopwords,sort_dictionary). Finalement, on regroupe toutes les colonnes dans une seule matrice (cf. preprocessing_raw_data)

Une fois les questions sont traitées, on fixe de ground truth. On cherche à différencier les grandes villes de petites villes en fonction de leur densité et les assigner une étiquette (cf find_zip_codes_by_town, city_village_classifier). Par exemple, les grandes villes recevront l'étiquette "1" et les petites recevront "-1".

Dans un deuxième temps, on compte le nombre total des mots par question pour toutes les entrées (cf. word_count_by_question). Finalement, on sélectionne un nombre fixe des mots les plus utilisées par question. Pour chaque entrée et question, on vérifie le nombre d'occurrences des réponses dans l'array des mots les plus utilisées (cf. get_most_used_words)

2) Extraction des features.

La question de l'extraction des features est assez importante.
Dans le questionnaire, il y a 4 types de questions, pour chaque type de question nous avons
*Questions réponse Oui/non
* QCM avec plusieurs choix possibles
*QCM avec un seul choix possible
*Champ libre.

Input :

*Questions réponse Oui/non : -1 pour Non, 1 pour Oui et 0 sans réponse.
* QCM avec un seul choix possible : 0 pour pas de réponses et des entiers de 1 au nombre de réponses pour les autres choix
*Champ libre, titre ou QCM avec plusieurs choix possibles : Nous pensons déterminer, pour chaque question, les 20 mots les plus utilisés dans toutes les réponses (après preprocessing), puis nous choisirons les mots qui présentent la plus grande cross-entropy pour chaque question (c'est à dire ceux qui sont les plus discriminants pour dire déterminer à quel cluster appartient la réponse).

Finalement nous souhaitons obtenir cela :
Les colonnes correspondent aux réponses aux questions et au titre (les features) , les lignes correspondent chacune à un questionnaire et les colonnes à une question chacune. Les questions avec du texte donnent plusieurs colonnes (une par mot sélectionné).

Oui/Non	QCM	Text Q1		TextQ2	
0	0	f1	f2	f1	f2
1	5	0	0.1	0	0.3
−1	3	0.2	0	0	0
1	2	0	0	0.3	0.01

f1 = fréquences de ce mot dans la réponse.
Chaque ligne correspond à une contribution.
Selon les résultats nous choisiront les 10 ou les 50 mots les plus discriminants.

Output : vecteur avec l'étiquette du cluster de densité de population.

3) Prédiction.

Pour le moment, nous avons décidé de travailler avec deux clusters. Il est clair, avec les features extraites avec la méthode précédente, que certaines d'entre elles sont plus importantes que les autres. Nous pensons utiliser la méthode des Random Forest pour classer les features les plus importantes afin de limiter nos coûts de calcul et de prédire mieux. De plus trouver les features les plus importantes nous semble ce qui est le plus intéressant compte tenu du problème : nous pouvons savoir quels sont les mots qui traduisent le plus les différences d'urbanisation par question.

Ensuite nous pensons comparer les résultats de Random Forest et d'un réseau neuronal pour prédire l'appartenance au cluster.

Perspectives.

Pour le moment le projet de deux clusters est assez grossier, si nous avons le temps et que cela ne nous semble pas pertinent, nous pourrons essayer de faire d'autres clusters.

Nous pensons mettre en forme nos résultats en présentant les résultats de la prédiction et réaliser une carte des mots qui sont les plus discriminants dans telle ou telle question selon la provenance de la réponse.