Hierarchical Topic Modeling (T4)

_____

Groupe 21

Maxime Tchibozo
Alexandre Lanvin
Antoine Habis
Dimitri Theveneau

Rappel : description et problématisation du sujet

“Les « gilets jaunes », le symptôme d’une France fracturée”. Voici ce que titrait *le Monde* en novembre dernier. Des ronds-points aux Champs-Elysées, le mouvement s’intensifie et gagne progressivement en violences. D’une part, le gouvernement dénonce des actions inexcusables, et illustre l’opposition entre les camps à travers la loi anti-casseurs. De leur côté, les gilets jaunes déclarent, à travers Julie Tissier, que "La première violence, c'est la violence institutionnelle". Le débat se rapproche de la partie de tennis, où chaque champion renvoie la balle à l’autre. S’attaquant aux symboles de richesses comme le restaurant le Fouquet’s sur les Champs Elysées, ou à ses propres membres s'engageant dans la politique de manière traditionnelle (i.e. en formant une liste) , le mouvement se décrit comme celui de ceux aux fins de mois difficiles, de ceux des campagnes, des classes moyennes face aux élites urbaines.

Il semble ainsi pertinent de traiter les données en ayant en tête ces divisions. C’est pourquoi nous pouvons orienter notre travail sur l’étude des disparités des réponses, en fonction de certains paramètres, et **se demander si les données du Grand Débat National témoignent d’une fracture en France**.

Notre sujet consiste ainsi dans une large mesure à recenser les thèmes et intérêts chers aux différentes régions, mais aussi les classer suivant les différents types de communes (grande ou petite ville, village, …) afin de se questionner sur le problème de la rupture géographique. Une éventuelle perspective pourrait enfin être de prolonger la problématique en étudiant les disparités éventuelles suivant la classe sociale des participants. Cela reviendrait donc à arriver à déterminer la classe sociale d’un individu si ses réponses sont suffisantes. Mais cela n’est pas prioritaire, et reste une perspective. La question de la faisabilité de cette dernière se pose d’ailleurs.

Description des données

Pour notre travail, nous avons décidé de nous restreindre à l’étude d’une sélection de questions. Cela paraît être un bon compromis entre temps d'exécution et pertinence des résultats, c’est-à-dire que les données ne sont pas trop lourdes de telle sorte que nous puissions faire tourner nos algorithmes dans des temps raisonnables, tout en ayant un output interprétable et significatif dans le cadre de notre étude.

Concrètement, nous avons choisi une question, qui nous semblait la plus pertinente: *Pour quelle(s) politique(s) publique(s) ou pour quels domaines d'action publique, seriez-vous prêts à payer plus d'impôts ?.* Le choix de cette question nous vient d’une part de la conclusion du travail de la CEPREMAP sur la question “Quels sont les soutiens des Gilets Jaunes?” [2] sur l’importance de la question de l’impôt, et a été confortée au cours de notre travail par le décision du gouvernement actuel qui, suite au grand débat, a pris la décision de baisser l’impôt sur le revenu.

Etant donné le grand nombre de réponses à chaque question (ici : 45 588 mots), le preprocessing demeure adéquat. On peut toutefois noter qu’on ne traite de fait pas toutes les données, là est un biais éventuel. Nos résultats sont ainsi moins généralisables et universels, mais il semble qu’ils nous permettent tout de même de traiter le sujet avec pertinence.

Représentation des données

Comme nous le préciserons dans le paragraphe qui suit, les données de chaque question sont mises sous forme de matrice que l'on appelle matrice des occurrences. Nous reviendrons dessus par la suite. A partir de chaque question, après le preprocessing, nous regroupons sous la forme d'un dictionnaire les mots utilisés.

Les topics sont représentés sous forme de liste de mots qui les caractérisent. Par exemple, on peut avoir comme topic ['pay', 'transit', 'écolog', 'aid', 'trop'] que l'on pourrait ensuite interpréter comme le thème de la transition écologique. On a, de la même manière, recours à des dictionnaires pour représenter les pourcentages d'intérêt à un topic de la part d'un département. Par ailleurs, le bruit dans les réponses aux départements sont regroupés dans un département que l'on nomme "0". Par exemple, si quelqu'un a répondu que son code postal est 985366562, alors on indique que son département est 0.

Description du modèle

Notre modèle est fondé sur un procédé d'analyse sémantique latente (LSA). Ce dernier consiste à analyser les relations entre un ensemble de documents (les réponses d'un utilisateur à une question) et les termes qu'ils contiennent en produisant un ensemble de thèmes liés aux documents et aux termes. Pour cela, on se utilise une matrice dite matrice des occurrences qui est creuse (i.e. sparse). Ses lignes correspondent à un mot, et ses colonnes à un document, et de cette manière on obtient dans chaque case la fréquence d'un mot dans un document. Plus précisément, on a la relation suivante:

$M = U\Sigma V$ avec avec M la matrice d'occurrences de taille $m * n$, U une matrice de taille $m * k$, Σ diagonale de dimension k et Vde taille $k * n$.

k représente le nombre de topics que nous avons considérés. C'est un paramètre fixé avant l'exécution de la LSA. U représente la matrice d'appartenance d'un mot à un topic, c'est-à-dire la case i,j de cette matrice indique si le topic j inclut le mot i (par exemple si le topic que l'on renommerait a posteriori "impôt" inclut le mot "taxe"). Σ est la matrice diagonale d'importance de chaque topic. Enfin, V représente la matrice d'occurrences de chaque topic dans les réponses, de la même manière que la matrice M d'entrée représentait celle d'occurrences des mots dans les réponses.

Séparation des données

Après les transformations matricielles opérées par les procédés décrits précédemment, on obtient des matrices dont les valeurs ne sont plus nécessairement 0 ou 1 mais des réels (caractérisant l'importance d'un mot dans un topic ou celle d'un topic dans un

document). A partir de ceci, le but est de bien séparer les données de telle sorte à déterminer sans trop d'erreur si une réponse traite ou non d'un topic ou encore si un mot représente un topic. Pour cela, on fixe un **seuil d'acceptation**: si la valeur correspondant à la case i,j de la matrice d'arrivée, c'est-à-dire correspond à l'appartenance ou non du i-ème topic à la j-ième réponse, dépasse ce seuil, alors ou considère que le mot caractérise ce topic (sinon, logiquement, on considère qu'il n'est pas inclu dans le topic).

Pour cela, il faut donc que les données soient bien séparées. Typiquement, on aimerait bien que les valeurs "rejetées" (i.e. celles pour lesquelles on considère que le mot caractérise le topic) soient proches de 0, et les celles "acceptées" proches de 1. Pour cela, le nombre de topics joue un rôle important. On ne pas en choisir un nombre trop grand car sinon les valeurs de la matrice seraient du même ordre de grandeur, mais il faut en choisir suffisamment pour que les réponses soient pertinentes et intéressantes. Ainsi, on agit de la même manière que lorsqu'on prune un arbre : on choisit dans un premier temps un k grand, puis on observe la pertinence des topics trouvés. Si les résultats ne semblent pas pertinents, c'est-à-dire si l'on reconnaît pas de thèmes clairs, ou si certains semblent s'entremêler, alors on réduit k. On répète l'opération jusqu'à trouver des thèmes satisfaisants.

Validation des résultats:

Ce paragraphe est fondé sur l'article 'TOWARDS BETTER INTEGRATION OF SEMANTIC PREDICTORS IN STATISTICAL LANGUAGE MODELING' écrit par Noah Coccaro* and Daniel Jurafsky. [0]

Pour valider notre algorithme, il faut étudier la cohérence des résultats. Toutefois il paraît difficile de trouver une expression mathématique vérifiant si une phrase parle bien de tel ou tel thème sans le faire à la main. Nous avons trouvé sur l'article ci dessus, une proposition d'expression visant à contrôler la pertinence de l'algorithme.

$$LSA\ Confidence\ _i = 1 + \sum_{j=1}^{ndocs} \frac{P(i|j)log(P(i|j))}{log(ndocs)}$$

Avec: $P(i|j) = \frac{Nombre\ de\ termes\ i\ dans\ le\ document\ j}{Nombre\ de\ terme\ i\ dans\ tous\ le\ document}$

On peut toutefois noter que cette formule est limitée car elle ne prend pas en compte la valeur de k associée aux nombre de thèmes ciblés.

Nous trouvons comme valeur de LSA Confidence moyenne autour de 0.999. Cela témoigne de la pertinence du modèle choisi.

Temps d'exécution:

Le temps d'exécution du preprocessing pour une question donnée est de 22 secondes.

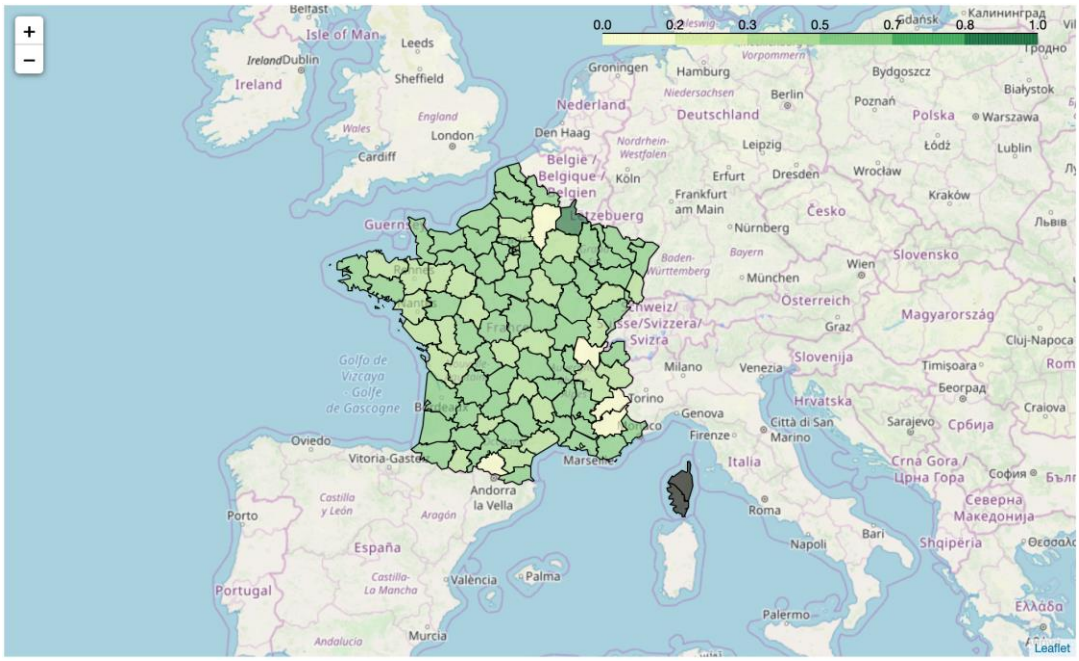
Le temps d'exécution de LSA est de 19 secondes.
L’affichage de la carte est instantané.

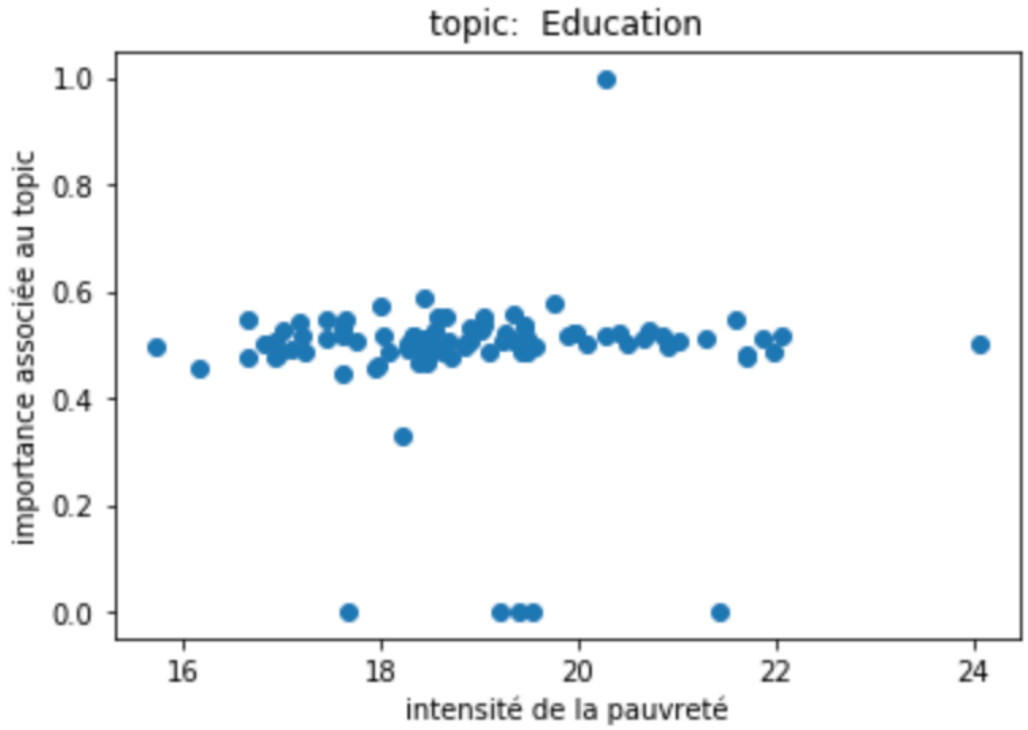
La durée d'exécution est relativement peu élevée ce qui permet d’observer les résultats sur diverses questions rapidement et justifie le choix d’une seule question.

Résultats trouvés par notre algorithme:

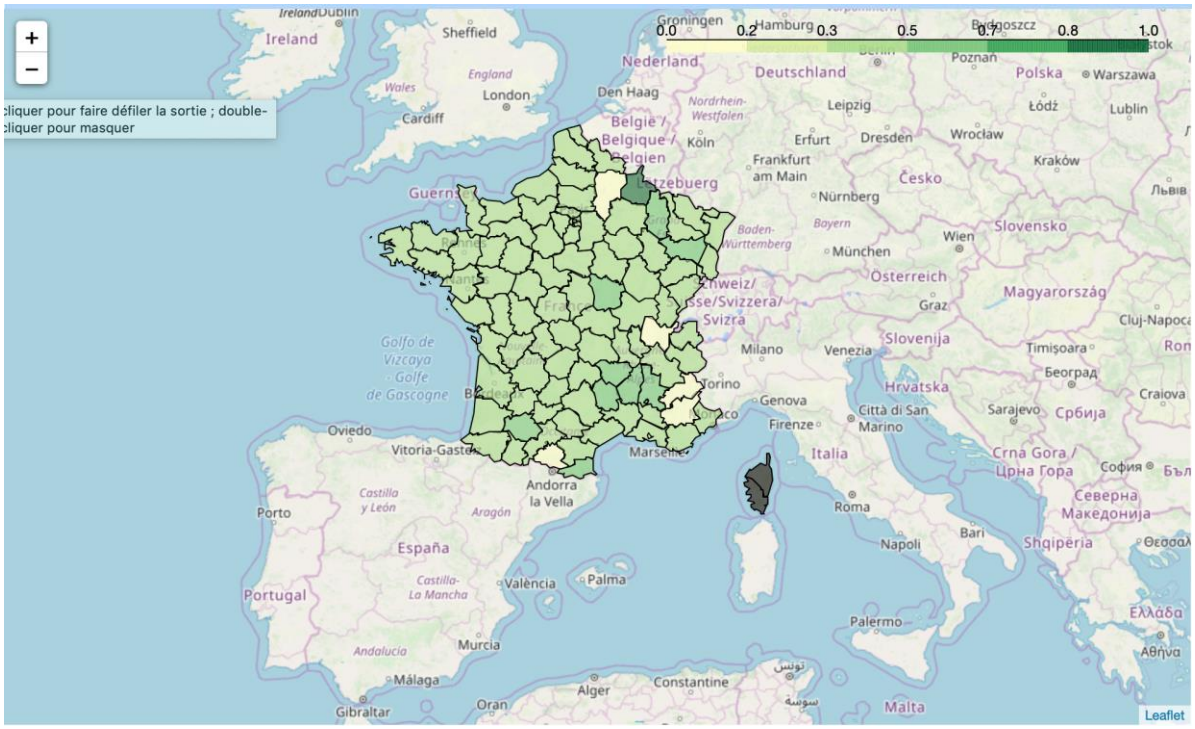
Question : *Pour quelle(s) politique(s) publique(s) ou pour quels domaines d'action publique, seriez-vous prêts à payer plus d'impôts ?*

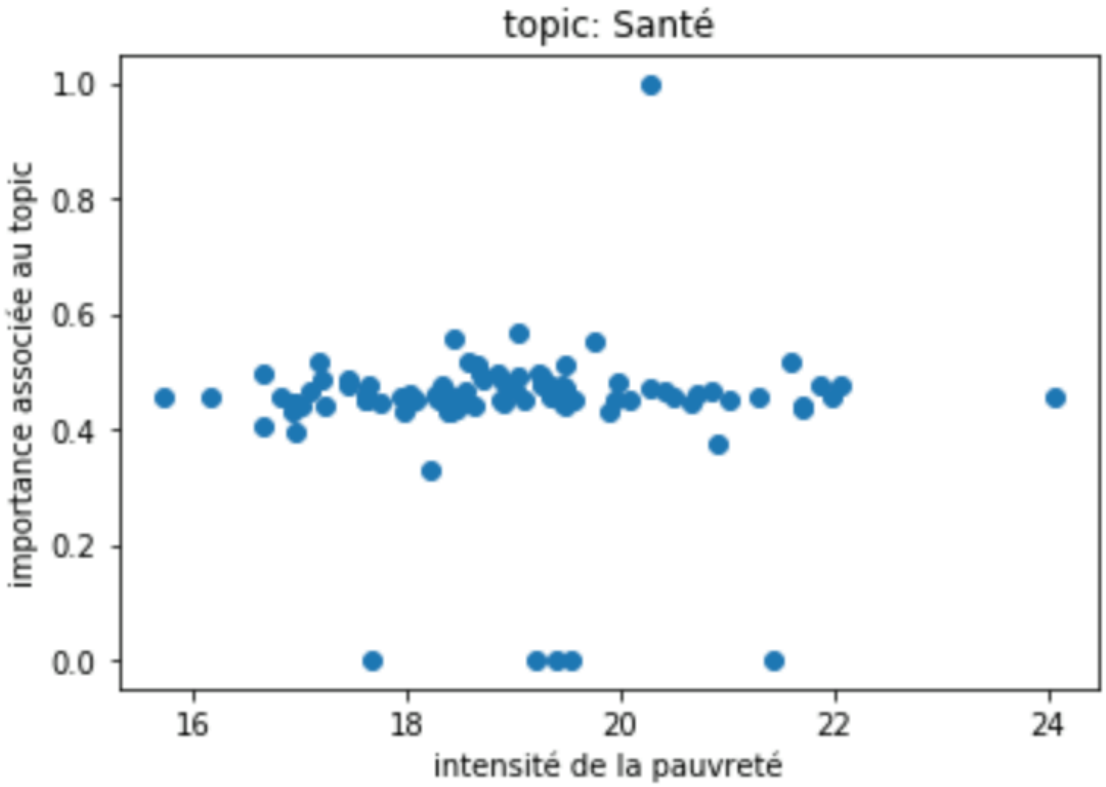
Thème de l’éducation:



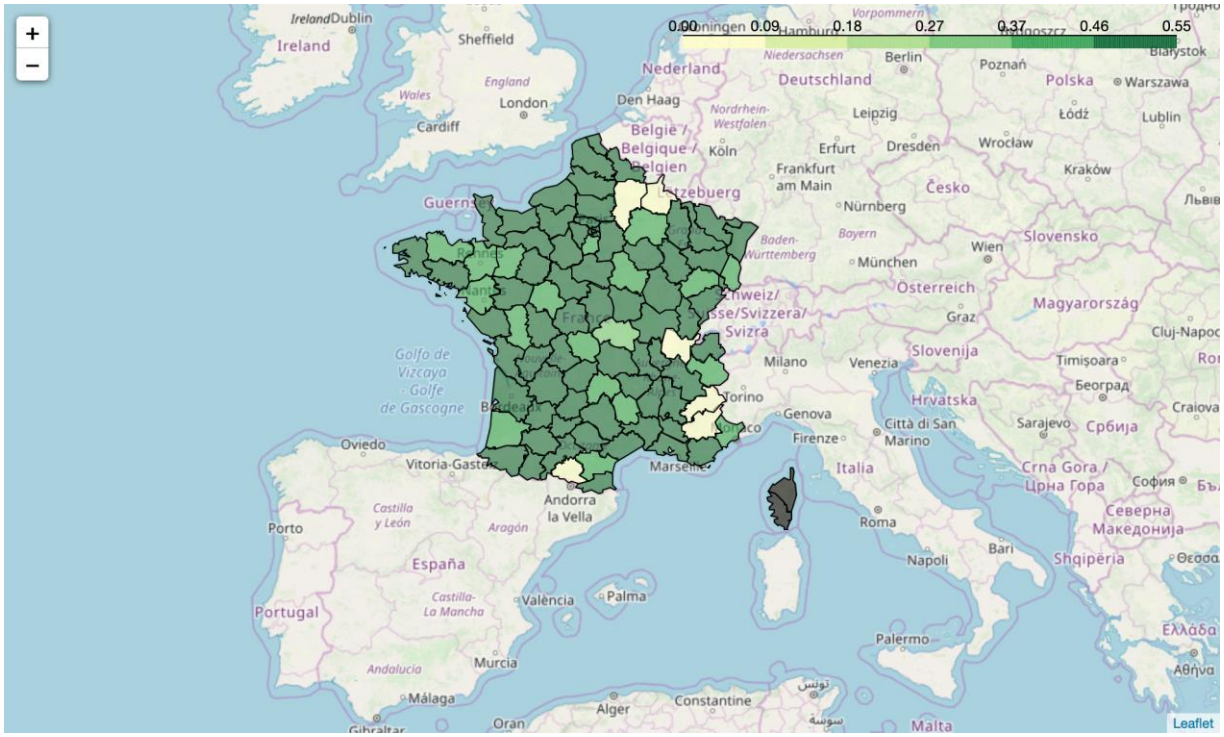


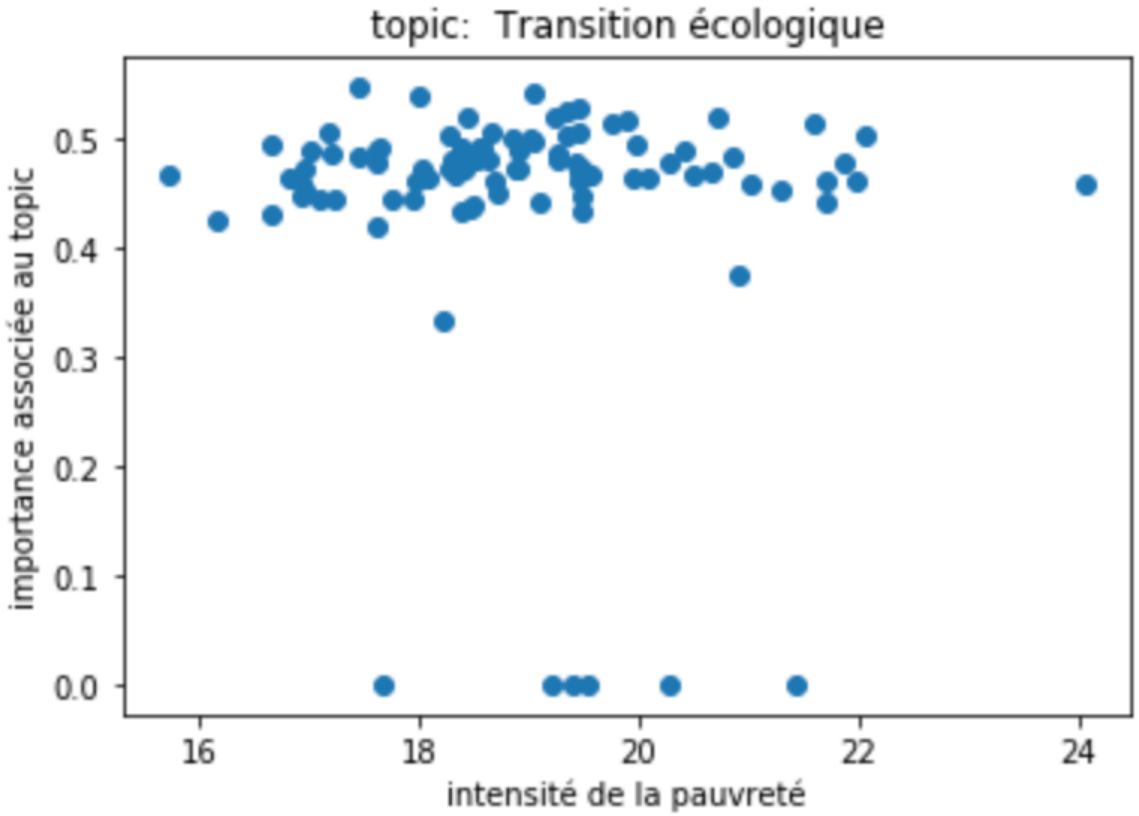
Thème de la Santé:





Thème de la transition écologique:





Interprétation

Les graphiques sont fondés sur nos propres résultats ainsi que ceux d'un rapport de l'INSEE de 2011 (aucun rapport plus récent concernant les richesses dans les départements

n’a été publié). Dans ce rapport [1] figure un tableau représentant l’intensité de la pauvreté en fonction du département.

L'intensité de la pauvreté selon l’INSEE (ou « poverty gap ») est un indicateur qui permet d'apprécier à quel point le niveau de vie de la population pauvre est éloigné du seuil de pauvreté. L'INSEE mesure cet indicateur comme l'écart relatif entre le niveau de vie médian de la population pauvre et le seuil de pauvreté.
Formellement, il est calculé de la manière suivante : ***(seuil de pauvreté- niveau de vie médian de la population pauvre) / seuil de pauvreté***. Plus cet indicateur est élevé et plus la pauvreté est dite intense, au sens où le niveau de vie des plus pauvres est très inférieur au seuil de pauvreté.

Les courbes et cartes ci-dessus indiquent une représentation globalement uniforme sur la France. En effet, il ne semble pas y avoir de corrélation apparente entre le niveau de pauvreté d’un département et les réponses à la question. Les nuages de points décrivent approximativement une droite horizontale. De la même manière, en superposant la carte de la médiane du niveau de vie par département (voir Figure 1. ci-dessous) et celles présentées ci-dessus, on ne semble pas pouvoir trouver de corrélation explicite.

Toutefois, il faut tout de même remarquer que l’Aisne, l’Ariège, les Alpes de Haute-Provence, Les Hautes-Alpes et l’Ain ne semblent pas près à dépenser plus d’impôts quels que soient les thèmes qui ressortent. Il se trouve, comme en témoigne la carte de l’intensité du mouvement des Gilets Jaunes, que ces départements sont niches de Gilets Jaunes auxquels ils portent un fort soutien.

De cette manière, cela semble détromper l’idée que les idées portées par les Gilets Jaunes sont celles de la France des fins de mois difficiles. Les soutiens de ces idées semblent trouver des partisans chez tous milieux sociaux, chez tous les types de revenus. Comme l’indique les notes de la CEPREMAP [2], les Gilets Jaunes trouvent des soutiens chez tous les partis politiques, y compris En Marche, où ils sont tout de même à peu près 15% à déclarer soutenir “plutôt” les Gilets Jaunes. C’est à peine moins que pour les partisans de Marine le Pen. La rupture semble plus venir de certains départements particuliers, comme l’indique les cinq cités précédemment. On note ainsi quelques différences géographiques.

Toutefois, il est possible de noter que si l’Ile de France se dit globalement en défaveur des Gilets Jaunes (voir *Figure 0.* ci-dessous), elle ne semble pas faire avec leurs idées pour autant. Cela n’est pas forcément étonnant dans la mesure où le département était le lieu de rassemblement des manifestations du mouvement, et l’agacement peut d’avantage porté sur les dégradations ou les méthodes que sur les thèmes traités. De même, la diagonale du vide, pourtant réputée à faveur des GJ, ne semble pas non plus pour autant contraster avec le reste du territoire. De cette manière, si l’on retrouve certaines dissonances chez certains départements, on peut conclure qu’il n’est pas possible, sur la base de nos résultats, de parler de fracture sociale ou de fracture territoriale. Comme l’indique les notes de la CEPREMAP, la fracture est celle du bien-être, et se retrouve partout en France. Il semble de même qu’il y ait une volonté globale de changement, ce qui est cohérent avec le bouleversements politiques

récents, comme la non-présence des partis traditionnels au second tour de la Présidentielle ou encore la montée des votes extrêmes.

Pour ce qui est de la transition écologique, il est important de souligner que cette question est le point de départ du mouvement des gilets jaunes. Elle est donc capitale pour comprendre l'origine du "soulèvement". L'écologie est un sujet très controversé comme en témoigne le dernier graphe qui affiche une dispersion beaucoup plus importante que les deux précédents. L'article de la CEPREMAP affirme même que "un tiers des soutiens des Gilets jaunes disent refuser une réduction du niveau de vie pour améliorer l'environnement, un tiers y est favorable et le dernier bloc est indifférent." De cette manière, il n'est pas surprenant que les résultats de ce thème soient ceux avec la plus grande variance, comme en témoigne la courbe. Toutefois, la carte de France est beaucoup plus foncée que pour les deux autres thèmes, ce qui montre que le sujet est revenu souvent dans les réponses et qu'il semble animé particulièrement les débats. La présence de ce thème dans les réponses à notre question confirme la pertinence du choix de cette dernière.

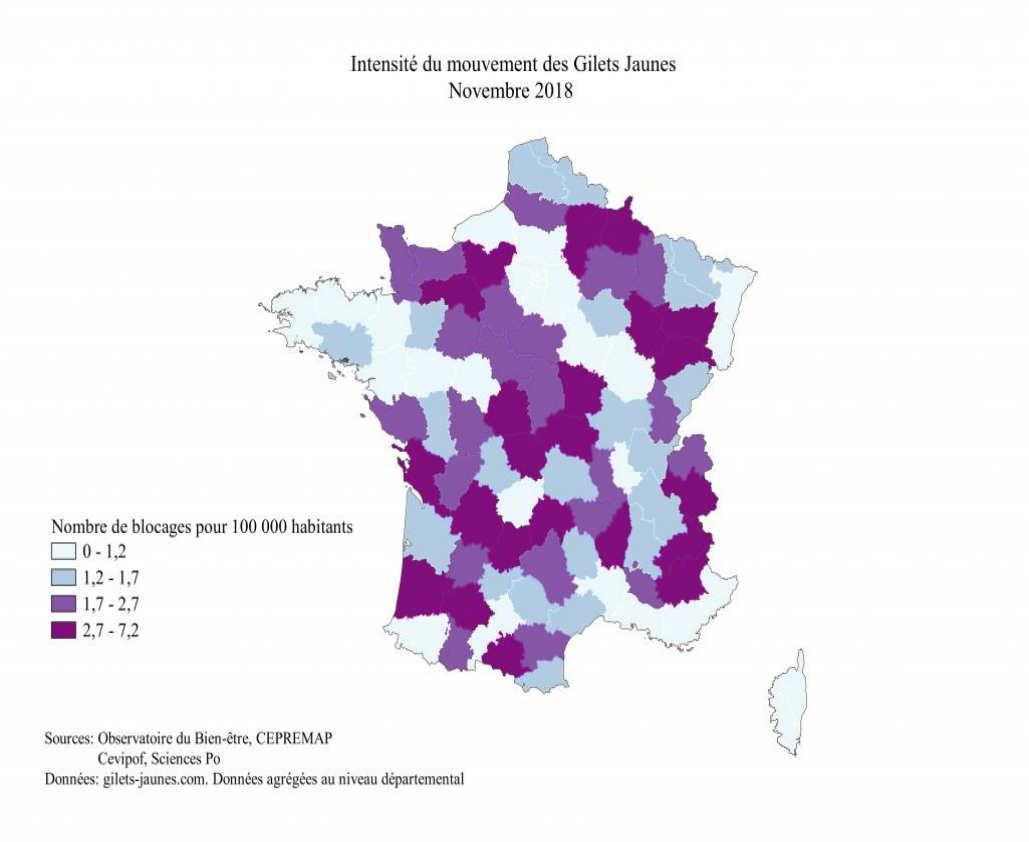
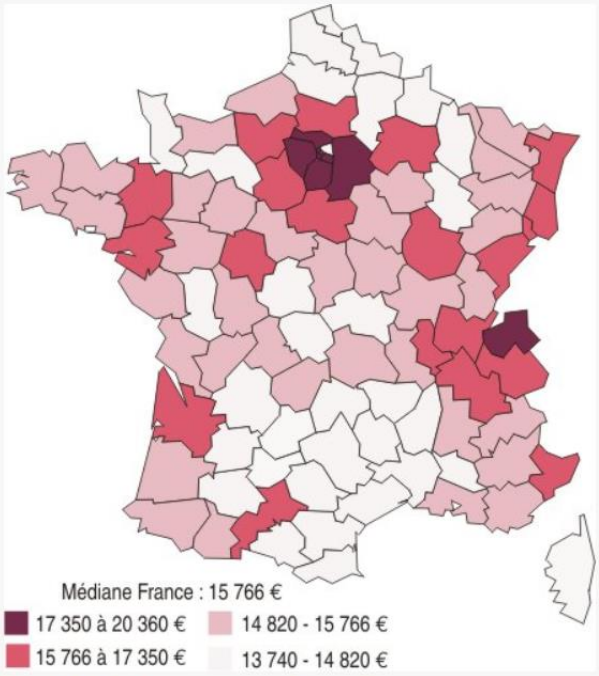


Figure 0.

Carte 1 – Médiane du niveau de vie par département



Les deux départements de la Corse ont été regroupés (sources).
Source : revenus disponibles localisés 2004, Insee-DGI.

Figure 1.

Conclusion

Quelle que soit l'issue du mouvement des Gilets Jaunes, l'initiative du Grand Débat ne peut être que saluée par la communauté scientifique spécialisée dans le traitement automatique de langage naturel.

Le Grand Débat nous fournit une vaste base de données de textes, et nous permet – en plus de tester nos outils – de recenser l'avis des français.

A travers notre projet, nous avons cherché à démontrer la pertinence de ces outils à travers trois axes :

- Le preprocessing et la tokenization
- L'optimisation de paramètres et l'extraction de topics via LSA
- La mise en corrélation du niveau de vie, de l'intensité du mouvement, et des topics mentionnés

Ces étapes nous ont permis d'aboutir à plusieurs conclusions (souvent contre-intuitives) vis-à-vis des millions de données du Grand Débat :

- Les idées portées par les Gilets Jaunes **ne sont pas uniquement celles de la France des fins de mois difficiles**, elles se retrouvent dans pratiquement tous les départements, indépendamment du niveau de vie.
- Il n'est **pas possible, sur la base de nos résultats, de parler de fracture sociale** ou de fracture territoriale.
- Le thème de l'écologie est parmi ceux qui a suscité les plus vives réactions, pourtant, c'est également le **thème qui présente la plus grande variance dans réponses**. Il semblerait donc qu'il n'y ait aucun consensus à la fois localement et à l'échelle nationale sur ce sujet.

Nous vous invitons enfin à interpréter ces résultats apportés par nos outils, en étudiant directement les scripts, méthodes, et cartes interactives disponibles dans notre notebook.

Bibliographie :

[0] ‘TOWARDS BETTER INTEGRATION OF SEMANTIC PREDICTORS IN STATISTICAL LANGUAGE MODELING’ écrit par Noah Coccaro et Daniel Jurafsky.

[1] Rapport de l’INSEE sur l’intensité de la pauvreté en france https://www.insee.fr/fr/statistiques/1895072

[2] Note de l’Observatoire du Bien-être n°2019-03 : Qui sont les Gilets jaunes et leurs soutiens ? http://www.cepremap.fr/2019/02/note-de-lobservatoire-du-bien-etre-n2019-03-qui-sont-les-gilets-jaunes-et-leurs-soutiens/#Les_soutiens_aux_Gilets_jaunes