Analyse des contributions déposées sur l'espace participatif de la concertation citoyenne sur la vaccination

Avner Bar-Hen

5 Novembre 2016

Introduction

L'un des buts du plan d'action pour la rénovation de la politique vaccinale (présenté le 12 janvier 2016) est de renforcer la confiance des Français dans la vaccination, en répondant de façon transparente à leurs inquiétudes et préoccupations. Dans ce but, un espace participatif a été ouvert entre le 15 septembre 2016 et le 14 octobre 2016 afin de permettre à chaque citoyen d'exprimer son opinion et de présenter des propositions, individuelle ou collective, pour faire évoluer la politique vaccinale en France.

11844 contributions longues d'au plus 2000 caractères ont été proposées. Suite aux échanges avec Santé Publique France, le but de ce document est ici de présenter une *première* analyse des 10435 contributions validées par le modérateur.

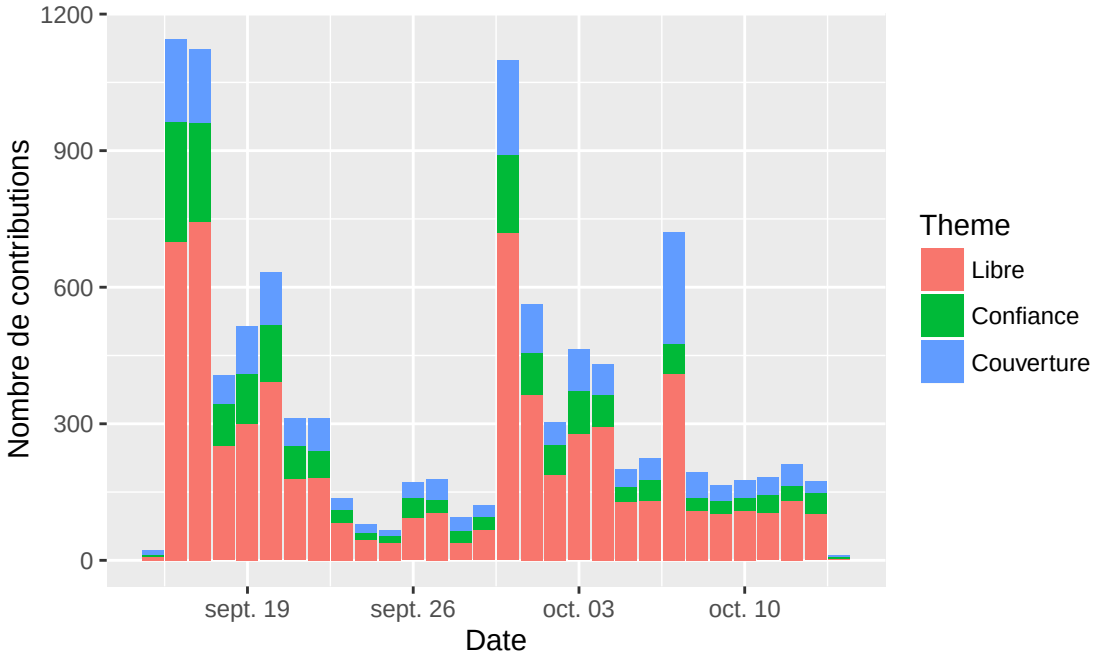
355 contributions sont présentées comme des contributions d'associations. Dans la mesure où ce champ est auto-déclaratif, il ne sera pas pris en compte dans l'analyse.

Le comité d'orientation proposait aux internautes trois possibilités pour contribuer : - Adresser ses questions aux pouvoirs publics, exprimer un avis ou un ressenti sur la vaccination au sens large : cette question a suscité 6404 contributions. - Faire des recommandations pour améliorer la confiance dans la vaccination : cette question a suscité 1980 contributions. - Faire des recommandations pour améliorer la couverture vaccinale : cette question a suscité 2051 contributions.

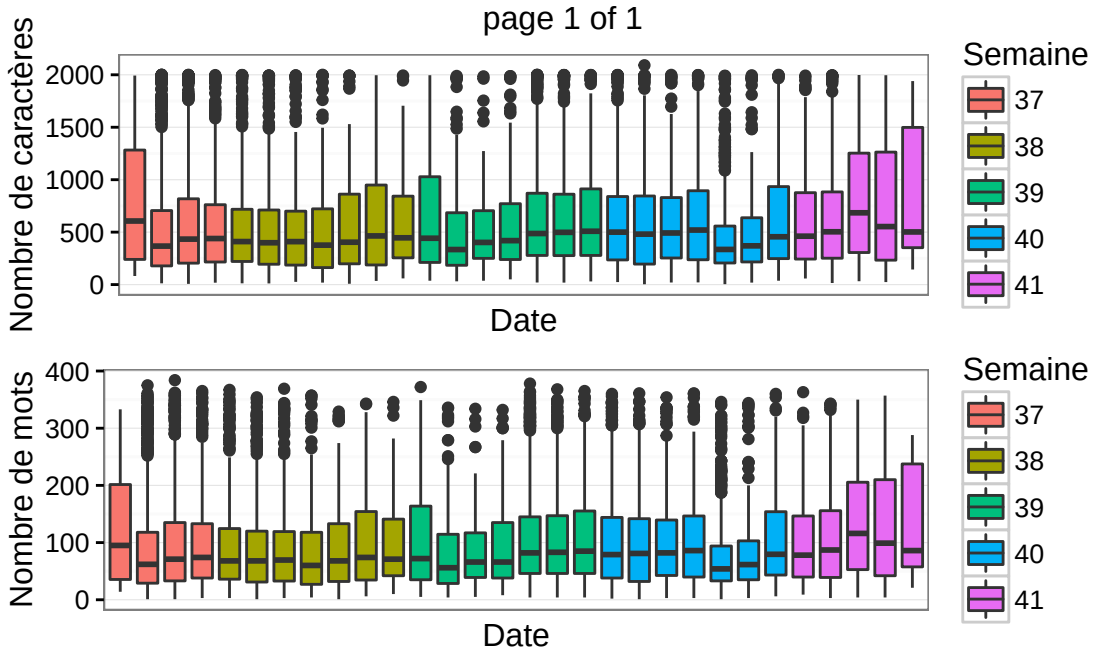
Il y a donc surtout des contributions libres émanant de particuliers.

Répartition des contributions sur le mois

Le forum a connu une très forte activité au début et un tassement régulier jusqu'au 29 septembre. Puis il y a eu une reprise d'activité début octobre qui s'est estomptée par la suite. Notons un pic inexpliqué le 7 octobre.



La longueur moyenne des messages (que l'on considère le nombre de mots ou de caractères) reste constante tout au long de la concertation. Il n'y a pas de dérive vers des messages en majorité plus longs ou plus courts au cours de la concertation.

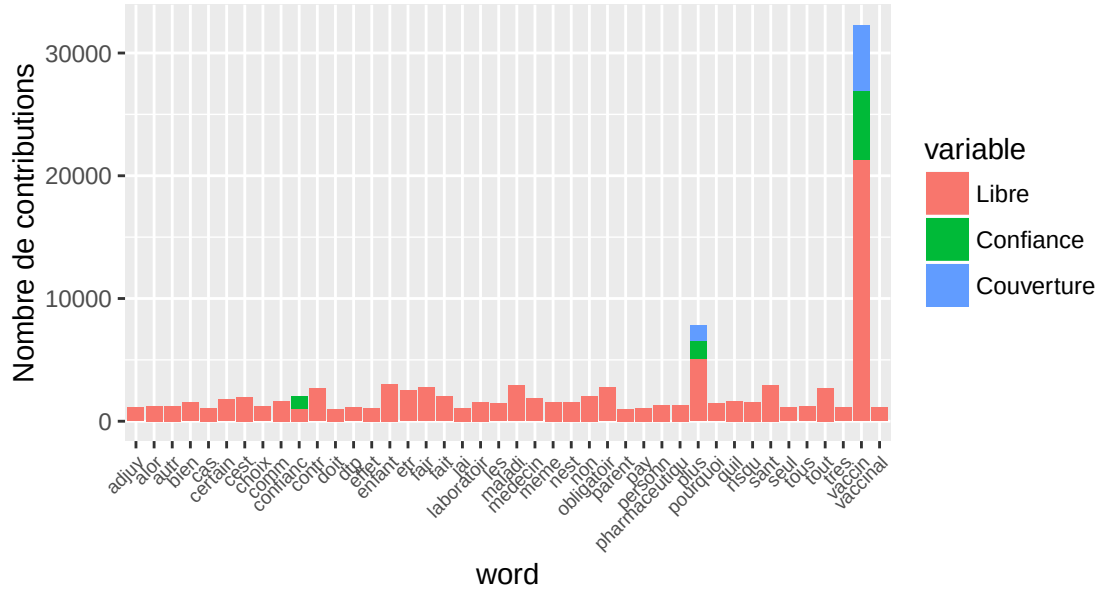


Analyse fréquentielle des termes utilisés

Pour commencer nous avons supprimé les accents, les nombres et les signes de ponctuation. De même nous avons extrait les radicaux (Lemmatisation) afin de rassembler les différentes variations d'un même mot.

Les mots les plus fréquents proviennent des contributions libres. On a donc clairement l'impression que

les gens ont plus envie de s'exprimer sur le sujet de la vaccination que de proposer des réponses aux deux questions thématiques. Ce point est important et sera précisé plus bas.



Les mots "vaccins" et "plus" sont les plus fréquents mais les réponses varient beaucoup selon les questions. Le mot "plus" est à prendre ici dans un sens positif et non dans un sens négatif (ne plus). Le tableau ci-dessus donne, de façon ordonnée, les mots les plus fréquents pour chacune des questions.

	Question 1	Question 2	Question 3	Total
rang 1 :	vaccin	vaccin	vaccin	vaccin
rang 2 :	plus	plus	plus	plus
rang 3 :	enfant	confianc	fair	sant
rang 4 :	sant	fair	sant	fair
rang 5 :	maladi	sant	vaccinal	enfant
rang 6 :	obligatoir	tout	medecin	maladi
rang 7 :	fair	obligatoir	obligatoir	tout
rang 8 :	tout	maladi	maladi	obligatoir
rang 9 :	contr	medecin	tout	etr
rang 10 :	etr	non	enfant	contr
rang 11 :	non	etr	etr	non
rang 12 :	fait	enfant	couvertur	medecin
rang 13 :	cest	laboratoir	contr	fait
rang 14 :	medecin	fait	non	cest
rang 15 :	certain	risqu	fait	comm

Association des termes les plus fréquents

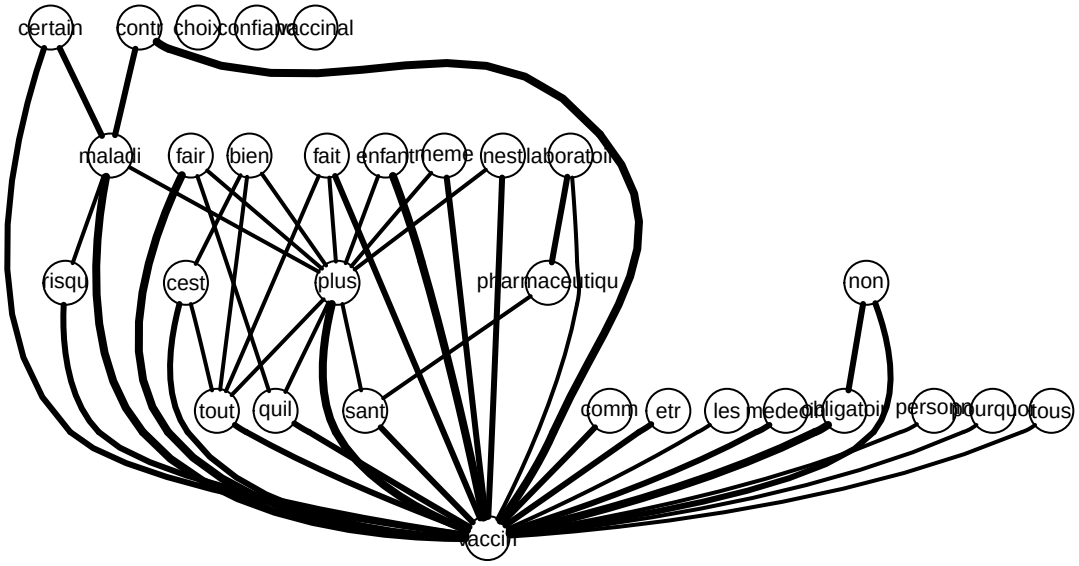
Le partitionnement de données (*data clustering* en anglais) est une des méthodes statistiques d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité). . En clair, le *clustering* cherche à faire des classes telle que : - les différences intra-classe soient minimales pour obtenir des clusters - les différences inter-classe soient maximales afin d'obtenir des sous-ensembles bien différenciés.

Ici nous allons utiliser des données correspondant aux contributions validées et nous allons utiliser les termes les plus fréquents comme caractéristiques. Le calcul de la distance entre les contributions se base sur la matrice termes-documents qui recense le nombre ou la fréquence d'apparition des termes dans chaque contribution.

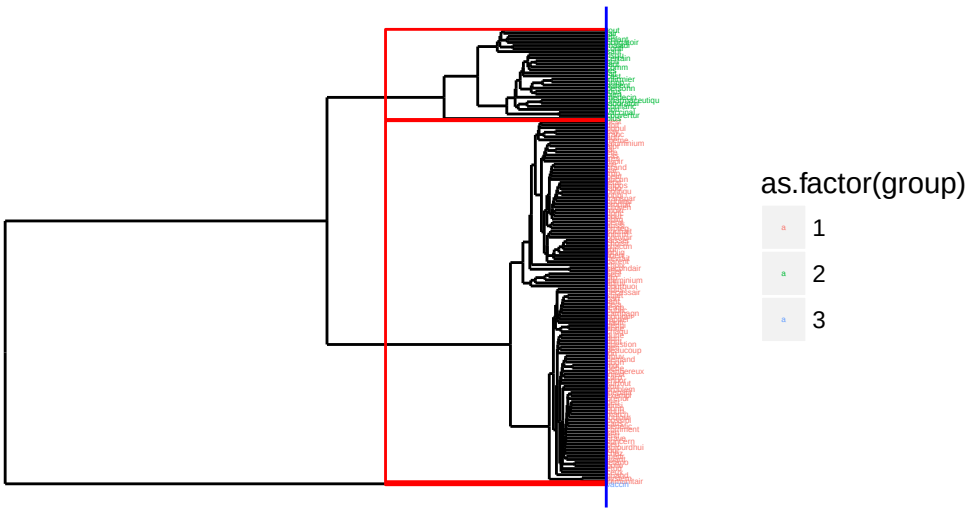
La distance la plus couramment utilisée entre deux termes est la distance du cosinus qui compte le nombre de co-occurence de ces deux termes divisés par le produit de leur occurence. La distance du cosinus est toujours comprise entre -1 et +1 et peut être reliée intuitivement avec la notion de corrélation de la manière suivante :

- plus deux documents sont similaires et plus ils auront tendance à utiliser les mêmes termes. Les termes seront alors corrélés, ce qui se traduit par un cosinus proche de +1 ou -1 ;
- inversement, deux documents non similaires auront tendance a avoir des termes décorrélées. Autrement dit, ils seront "orthogonaux" géométriquement et leur cosinus sera proche de 0.

Le graphe reprend les 30 termes les plus fréquents. Deux termes sont reliés si leur distance cosinus est supérieur à 0.2. La plus forte association est obtenue avec les termes "vaccins", "obligatoire" et "non". Le mot "plus", rappelons-le, est plutôt associé à des choses positives (et non pas "ne plus").



Regardons les 150 mots les plus fréquents. En utilisant la distance cosinus, iIl est possible de construire une matrice de distance et ensuite un arbre de classification (critère de Ward).

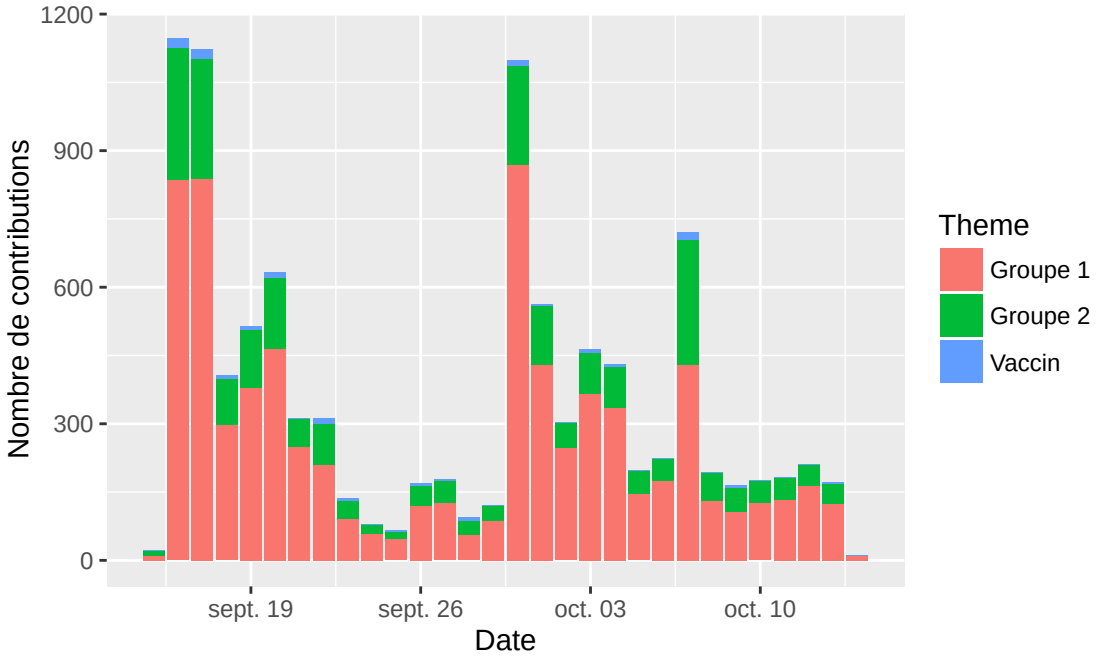


Pour aider à l'interprétation du graphe ci-dessus, regardons les mots les plus fréquents pour chacun des groupes (tableau ci-dessous). Il semble en ressortir trois groupes. D'un côté le mot vaccin, de l'autre des termes opposés au caractère obligatoire de la vaccination, enfin un troisième paquet concernant la vaccination des enfants.

	Groupe 1	Groupe 2	Groupe 3
rang 1 :	meme	plus	vaccin
rang 2 :	nest	sant	
rang 3 :	pourquoi	fair	
rang 4 :	choix	enfant	
rang 5 :	autr	maladi	
rang 6 :	alor	tout	
rang 7 :	effet	obligatoir	
rang 8 :	adjuv	etr	
rang 9 :	seul	contr	
rang 10 :	cas	non	
rang 11 :	dtp	medecin	
rang 12 :	pay	fait	

Regardons si les clusters évoluent dans le temps

Il y a plus de contributions libres quand il y a un pic de contributions et le mot vaccin est de moins en moins majoritaire dans les contributions. . .

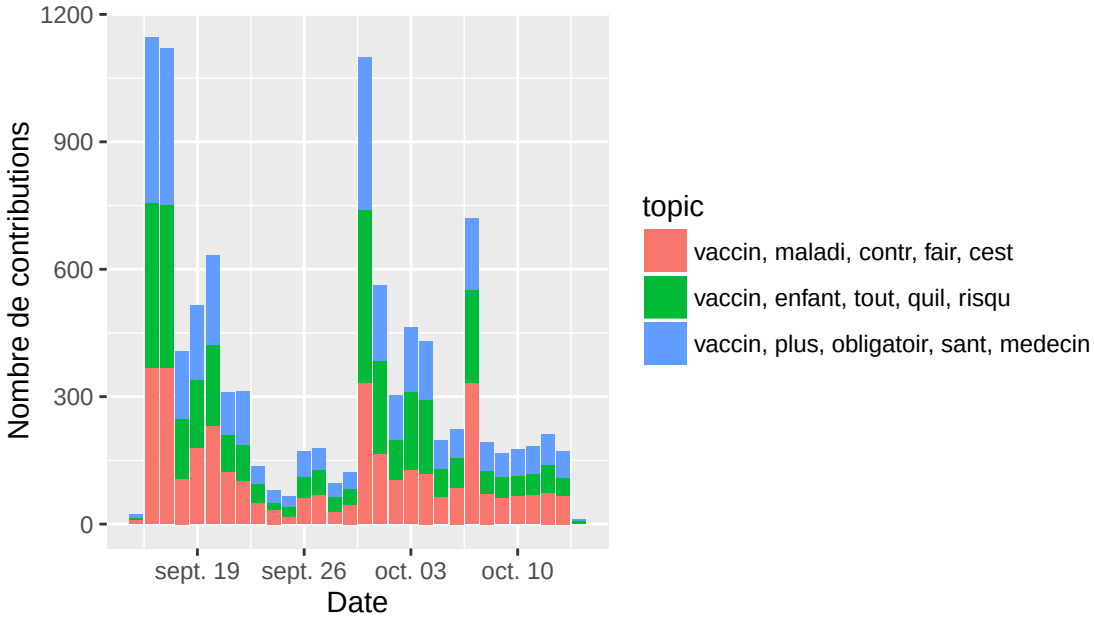


Regardons les thèmes (topics)

L'analyse des *topics* utilise un modèle probabiliste génératif qui permet de décrire les contributions. chaque contribution sera représentée par un « sac de thèmes/topics », à partir desquels les termes des contributions ont été choisis. Le but est donc de reconstruire les thèmes et d'associer les termes des contributions aux thèmes (*Latent Dirichlet Association*)

De tels modèles vont aussi permettre de prendre en compte des problèmes récurrents du Text Mining, la « polysémie » (possibilité pour un terme d'appartenir à plusieurs thèmes) et la « synonymie » (différents termes appartiennent au même thème). L'idée est de faire ressortir des structures thématiques cachées dans les contributions.

Trois thèmes apparaissent dans les contributions. Globalement, les contributions en rouge sont favorables à la vaccination et concernent les enfants. Les contributions en bleu sont critiques envers le côté obligatoire et le corps médical. Les contributions en vert sont contre la vaccination "à cause" des labos pharmaceutiques. L'hostilité à la vaccination apparaît dominante, mais il faut aller un peu plus loin. On note que les pics d'activité correspondent aux gens contributions critiques...



Analyse des sentiments

L'analyse de sentiment (parfois appelée *opinion mining*) essaye de définir les opinions, sentiments et attitudes présentes dans un texte. Nous avons ici utilisé une version simpliste qui consiste à prendre un dictionnaire positif et un dictionnaire négatif (ici issus du Data Science Lab). Un texte est défini comme négatif (resp. positif) si il a plus de termes négatifs que de termes positifs (et inversement). Un texte avec un score nul donne un texte neutre (souvent sans mots positifs ou négatifs). On voit une majorité de textes négatifs pour la question 1 et positifs pour les questions 2 et 3. Il n'y a pas d'évolution dans le temps

	Question 1	Question 2	Question 3
Négatif	2895	775	708
Neutre	1043	338	388
Positif	2466	867	955

Pour conclure, l'espace de contributions a eu un très grand succès. La question ouverte, placée en première place, a laissé court à des contributions critiques envers la vaccination. Ces critiques expriment un rejet envers le caractère obligatoire des vaccins, une défiance envers les laboratoires pharmaceutiques ainsi qu'une défiance envers le corps médical. Les deux questions ciblées permettent d'obtenir des contributions plus positives. Le côté positif de la vaccination obligatoire est surtout associé à la vaccination des enfants. Les délais très courts n'ont pas permis de faired'entreprendre une analyse sémantique des contributions.