

Analyse de réponses à un questionnaire : exemple de pré-traitement d'une question ouverte

www.adjectif.net/spip/spip.php

Le cas du Mooc Efan-2

jeudi 14 janvier 2016 par [Mehdi Khaneboubi](#)

Mots-clés

- [Analyse de données](#)

Les réponses brutes

Le jeudi deux avril 2015 on comptait 580 réponses au questionnaire. Parmi ces réponses, on disposait de 544 réponses à la première question : "Pouvez-vous nous préciser ce qui motive principalement votre inscription à ce MOOC ?". Pour commencer il est utile de parcourir les réponses, en voici une dizaine tirées au sort.

- "Je suis enseignante dans le secondaire. Je dois faire évoluer ma pédagogie en y intégrant les outils numériques. Ce MOOC est l'opportunité d'apprendre à le faire"
- "pour mieux intégrer les outils numériques dans ma pratique de l'enseignement des sciences SVT"
- "Ce MOOC constitue un premier pas pour moi vers cette manière d'enseigner"
- "Trouver des réponses à des interrogations sur des pratiques pédagogiques"
- "l'usage de la technologie dans l'enseignement suivant une approche connectiviste"
- "Améliorer mes pratiques pédagogiques et mieux intégrer le numérique dans l'enseignement"
- "Découvrir ce que signifie enseigner avec le numérique. Connaître ces méthodes. Enrichir mes connaissances et en garder des supports. Je ne travaille pas dans le domaine de l'éducation ou la formation, mais je m'y intéresse"
- "Me familiariser avec les nouveaux enjeux de l'enseignement à distance ainsi qu'avec les nouvelles technologies utilisables"
- "Je souhaite approfondir mes connaissances dans le domaine des technologies éducatives et l'utilisation des TIC dans l'enseignement. Avoir, si possible, un feedback sur mes réalisations et les projets. Mieux connaître la communauté enseignante qui croit en le potentiel pédagogique des TIC et collaborer avec d'autre enseignants"
- "compléter ma préparation au concours du capes de documentation"

On voit que les messages ont des formes très différentes. Certains sont très rédigés, d'autres beaucoup plus elliptiques. Certains présentent des éléments de contexte personnel, tandis que d'autres répondent strictement à la question. Examinons 10 autres réponses toujours tirées au hasard.

- "Enseignante en SHS dans le supérieur, je m'intéresse aux pédagogies alternatives - en particulier à la conception de MOOCs"
- "Je souhaite découvrir "de l'intérieur" ce qu'est un MOOC et, du même coup, travailler sur un sujet qui m'intéresse"
- "Module ESPE obligatoire"
- "Je suis enseignante de formation mais j'ai depuis 3 ans des fonctions de coordination qui m'éloignent du terrain. Cela dit, je souhaite garder un lien avec la classe et je ne voudrais pas être déconnectée des évolutions techniques en termes d'enseignement. Et avoir l'occasion de réfléchir à ses pratiques"

professionnelles en compagnie de spécialistes n'arrive pas si souvent !" _

- "Je souhaite faire évoluer ma pédagogie et me lancer dans l'enseignement numérique"
- "Ayant à travailler sur une formation à mettre en place à distance, je souhaite pouvoir me former aux outils et à la pédagogie. Quoi de mieux qu'un MOOC !!!"
- "APPRENDRE À APPRENDRE AVEC LE NUMÉRIQUE, J'AI BESOIN AUSSI DE PIQUES DE RAPPEL SUR LA FORMATION CONTINUE AVEC DES ADULTES"
- "Je veux apprendre à créer des MOOCs"
- "La possibilité d'améliorer ses connaissances relativement facilement, sans perturber son emploi du temps"
- "j'enseigne avec le numérique et souhaite faire évoluer mes pratiques" Là encore on voit une assez grande variété dans les messages.

Le nuage de mots

Une première étape dans le traitement, consiste à produire un lexique, c'est-à-dire la liste des mots de vocabulaire utilisés dans l'ensemble des messages. On peut ensuite compter le nombre d'apparitions de ces termes et représenter l'ensemble sous la forme d'un nuage de mots. Il existe des sites web spécialisés dans cette tâche comme le célèbre wordle.net mais ce genre d'instrument est à manier avec précaution car, comme tous les services en ligne pseudo-gratuits, on remet des données à un tiers. Il est donc toujours préférable d'utiliser des logiciels hors ligne.

En faisant un nuage de mots avec l'ensemble des messages pour lesquels on laisse le logiciel piloter l'analyse, on obtient la figure suivante. Vous remarquerez que tous les mots figurent en minuscules que les chiffres, les accents et les signes de ponctuations ont été enlevés par le logiciel.

vocabulaire	nombre d'apparitions
nouvelles	75
connaissances	74
souhaite	72
apprendre	69
pratique	62
cours	54
faire	54
...	...

On voit de façon assez précise le vocabulaire majoritairement employé. Il serait surtout nécessaire d'aller consulter les messages dans lesquels figurent les mots et voir si il n'est pas possible d'en faire des paquets "à la main", d'identifier des régularités.

Lexique et questions fermées

Le nuage de mots global ne permet pas de distinguer le profil des répondants. Il est donc intéressant de croiser le vocabulaire des questions ouvertes avec les réponses à des questions fermées : on pourrait par exemple examiner si les répondants âgés de moins de 43 ans et de plus de 43 ans [1] emploient un vocabulaire différent, notamment en produisant un tableau comme celui-ci :

	age < 43 ans	age > 43 ans
numérique	115	104
mooc	85	74
enseignement	83	61
pratique	57	47
plus	46	44
enseignant	47	42
formateur	19	41
professionnel	46	40
connaissance	42	38
...

On y voit que le mot "numérique" [2] a été employé 115 fois par des répondants de moins de 43 ans et 104 fois par les autres (pour 203 de moins de 43 ans et 303 autres qui ont répondu à la question). Il est possible sur ce type de tableau d'appliquer un test statistique [3] qui nous permet de dire qu'on ne trouve pas de différence significative entre le vocabulaire employé par les uns et par les autres.

En revanche, on va trouver des différences statistiques dans le vocabulaire employé par les hommes et les femmes. En interprétant le tableau ci-dessous, on peut dire que les femmes ont significativement plus utilisé les mots *souhaite*, *pédagogie*, *utiliser*, *apprendre*, etc. alors que les répondants hommes ont tendance à mentionner les mots *compétence*, *mooc*, *enseignement*, *enseignant*, etc. On peut aussi remarquer que les femmes

emploient les mots *apprendre*, *utiliser* et *utilisation* qui évoquent une dimension opérationnelle. Les hommes font référence à des *compétences* et à la série *enseigner*, *enseignement*, *enseignant*, ce que l'on peut considérer comme faisant référence à des potentialités : *améliorer ses compétences d'enseign-ant/-ement*. C'est une piste qu'il serait intéressant d'éprouver en allant consulter les réponses intégrales dans lesquelles figurent ces mots.

	homme	femme
souhaite	-1.9553707	1.5792899
pedagogie	-1.9471549	1.5726543
utiliser	-1.9131664	1.5452028
apprendre	-1.7072771	1.3789127
technologie	-1.3706497	1.1070296
former	-1.1160046	0.9013610
utilisation	-1.0405808	0.8404435
eleve	-1.0189690	0.8229884
...
ameliorer	0.9171081	-0.7407187
enseignant	0.9891713	-0.7989218
enseigner	1.1839815	-0.9562637
enseignement	1.2135174	-0.9801189
autres	1.4875530	-1.2014486
mooc	1.5433664	-1.2465273
competence	1.6827867	-1.3591326

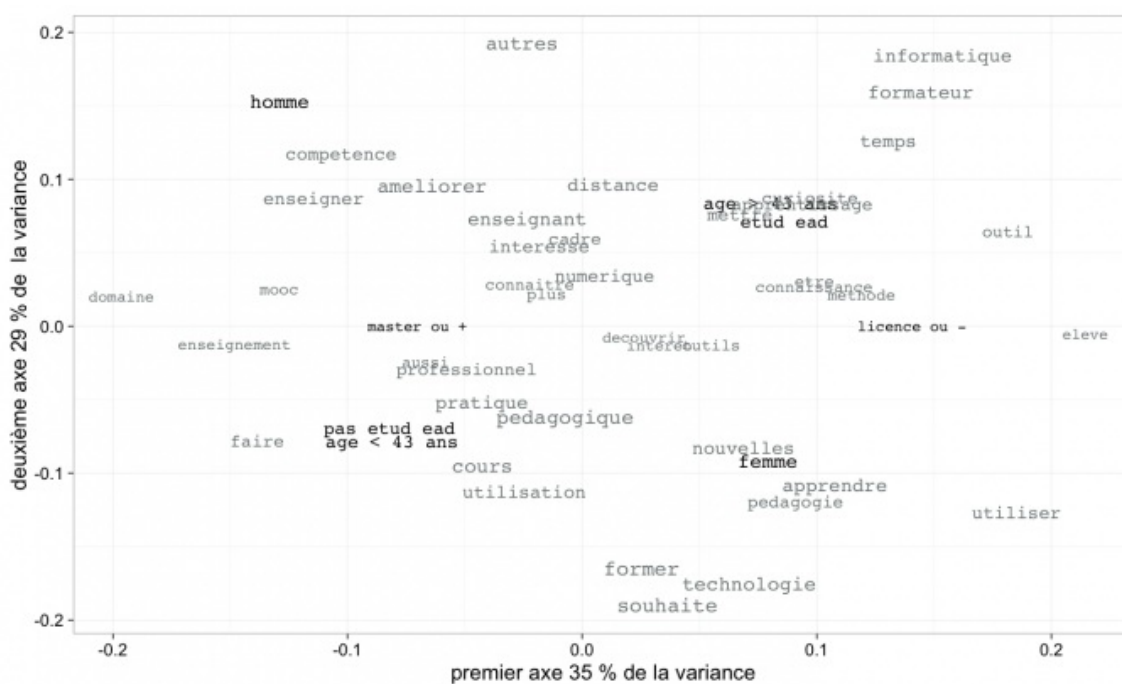
Le tableau lexical des questions

Pour affiner un peu, on peut maintenant chercher à employer une technique présentée par Philippe Cibois avec les commentaires de visiteurs sur un site marchand pour le film *Bienvenue chez les Ch'tis* [4]. Les étapes de la méthode sont consultables dans un article un peu plus ancien mais fort clair [5]. Il s'agit essentiellement de résumer la structure des attractions et des répulsions entre les lignes et les colonnes d'un tableau dont les premières lignes sont les suivantes :

	age < 43 ans	age > 43 ans	homme	femme	licence ou -	master ou +	etud ead	pas etud ead
numerique	115	104	94	125	78	114	111	105
mooc	85	74	75	84	39	106	69	91
enseignement	83	61	66	78	34	93	59	83
pratique	57	47	39	66	32	68	46	57
plus	46	44	40	50	34	46	40	50
enseignant	47	42	41	48	29	55	49	40

formateur	19	41	25	35	22	32	35	25
professionnel	46	40	35	52	27	51	36	48
...

Ce tableau se lit ainsi : le mot "numérique" à été employé 115 fois par des répondants âgés de moins de 43 ans et 104 fois par des gens âgés de plus de 43 ans, 94 fois par des hommes et 125 fois par des femmes, 78 fois par des titulaires d'un diplôme de licence ou moins, etc. À partir de ce tableau nous allons produire une représentation graphique (figure suivante) qui présentera les attractions entre les modalités des réponses aux questions fermées (en colonnes) et le vocabulaire des réponses à la question ouverte (en ligne).



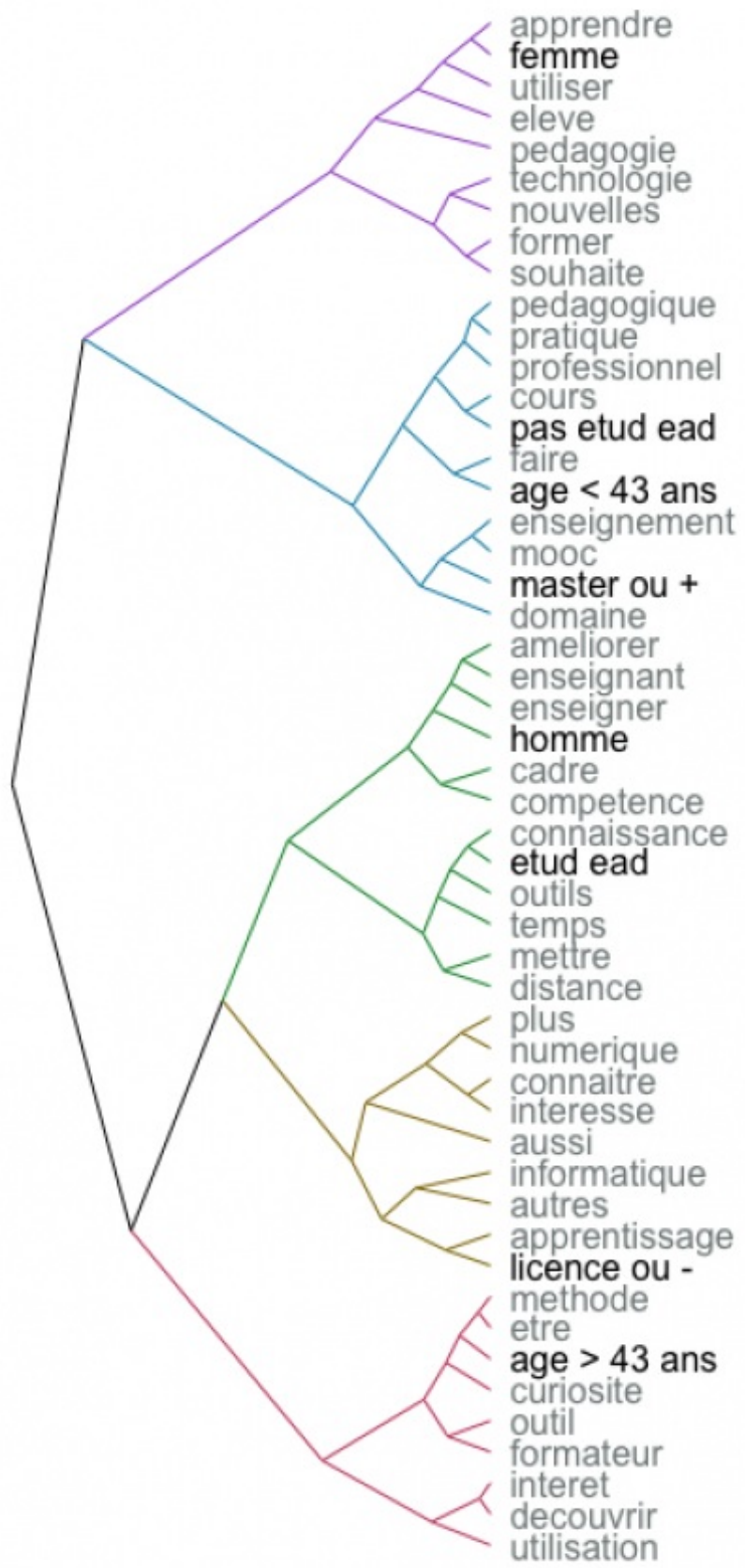
On peut dire, pour simplifier, que ce graphique résume 64 % (35+29) de l'information (variance ou inertie) contenue dans le tableau précédent. Ce qui est signifiant sur cette figure, ce sont les oppositions et les proximités entre les mots de vocabulaire (en gris sur la figure) et les réponses aux questions fermées (en noir). On voit que figurent exactement sur l'axe horizontal les modalités *licence ou moins* et *master ou plus*. On peut dire qu'à droite les répondants titulaires d'un diplôme de niveau licence ou moins sont proches des répondants ayant employé les mots *élève*, *utiliser*, *outil*, *informatique*, *formateur* et *temps*. À gauche de l'origine, on dira que les répondants titulaires d'un master ou plus, sont proches de ceux ayant utilisés les mots *améliorer*, *professionnel*, *pratique* ou notamment *compétence*. Pour les modalités correspondant au genre, dont on a déjà parlé dans la section précédente, on retrouve grosso modo les mêmes éléments. On pourrait reproduire l'analyse avec les autres modalités d'âge et le suivi ou non d'un enseignement à distance (*etud ead*).

Le problème de cette figure est double. D'abord elle ne prend pas en compte l'intégralité de l'inertie du nuage de points contenu dans le tableau d'origine. Ensuite, alors que ce n'est pas visible sur la figure, toutes les modalités et termes n'y sont pas bien représentés : pour connaître la qualité de la représentation de chaque item, il est nécessaire de consulter d'autres éléments chiffrés de l'analyse qui amèneraient à alourdir ce billet.

La classification automatique

En revanche, il est possible de représenter les attractions et les répulsions entre modalités et mots de

vocabulaire en faisant appel à une autre représentation graphique (figure suivante). Cela comporte notamment l'intérêt de classer automatiquement les lignes et les colonnes du tableau et de faire des groupes automatiques ainsi que de prendre en compte l'ensemble de l'information contenue dans le tableau (l'inertie ou la variance du nuage de points).



On voit sur cette figure, nommée dendrogramme ou arbre de classification, 5 groupes identifiés par des couleurs. Dans chaque groupe on trouve au moins une modalité de réponse aux questions fermées et les mots de vocabulaires. On voit que les répondants âgés de plus de 43 ans (figurant dans le groupe en rouge en bas du graphique) ont plutôt employé un vocabulaire évoquant la découverte, la curiosité, l'intérêt.

Les termes employés par ceux âgés de moins de 43 ans (groupes bleu clair), proches du vocabulaire des répondants n'ayant pas suivi de formation à distance (*pas etud ead*) et de celui des plus diplômés (*master ou +*) suggèrent une dimension *professionnelle avec enseignement, pédagogique, cours* et opérationnelle avec les mots *pratique* et *faire*.

Les femmes (groupe violet en haut de la figure) ont employé des termes qui font plus référence à *apprendre, former, utiliser*, elles évoquent les *élèves* et les *technologies*. Les hommes (en vert) sont proches des répondants ayant déjà suivi un enseignement à distance (*etud ead*) et mentionnent des *compétences* (qui est un terme particulier notamment parce que son emploi en éducation est récent) et des *connaissances*, une idée d'amélioration, le mot *enseignant* et le verbe *enseigner*.

Enfin les répondants les moins diplômés (en marron) sont proches de ceux qui emploient le terme *informatique* et *apprentissage*.

Perspectives

Cette méthode d'analyse automatique permet d'esquisser des tendances grossières et d'effectuer des regroupements. On peut ainsi considérer qu'un groupe de répondants, plutôt jeunes, plutôt diplômés d'un master ou d'un doctorat dans l'enseignement en présence se détache à partir du vocabulaire employé. Leur motivation pour participer au Mooc est probablement fondée sur une dimension professionnelle en éducation mais aussi opérationnelle. Les participants hommes font appel à des termes suggérant une amélioration de compétences et de connaissances. Cela revêt une dimension de besoin ou de nécessité, par opposition au groupe des femmes, des plus de 43 ans et des moins diplômés qui expriment davantage des notions de souhait, d'intérêt ou d'apprentissage.

Pour être tout à fait rigoureux, il serait maintenant nécessaire d'aller confronter ces premières interprétations aux textes d'origine et de se lancer dans une analyse sémantique "à la main" de ces réponses. L'analyse automatique aura tout de même permis d'orienter le travail futur et de gagner du temps.