Je propose ici le résumé d'une communication l'appel à communication du CFIES de Grenoble

Axes formation continue professionnelle ou enseignement de la statistique à l'ère de la science des données

L'ESSOR D'UNE FORMATION EN STATISTIQUE TEXTUELLE

Bénédicte Garnier, *Institut National d'Etudes Démographiques (INED)*, F-75020 Paris, France.
benedicte.garnier@ined.fr

Résumé.

L'essor du « Big data » et des données textuelles collectées sur le web a démultiplié les corpus et la demande pour apprendre à explorer ces nouveaux textes. En parallèle, l'offre de logiciels « gratuits /libres» s'est rapidement élargie pour offrir la possibilité de les analyser.

Dès 2010, dans le cadre des formations internes de l'Institut National d'Etudes Démographiques (Ined), j'ai conçu une formation en statistique textuelle dont l'objectif était de permettre aux participants d'analyser des réponses à des questions ouvertes ou traiter des entretiens liés aux enquêtes. Aujourd'hui, cette formation a aussi pour objet de proposer une méthode exploratoire pour analyser ces « nouveaux » corpus, ce qui entraine un accroissement de la demande pour ce type de formations en interne et en externe.

Le déroulé pédagogique de ces formations alterne des périodes de cours magistraux, et de mise en pratique sur ses propres corpus afin d'alimenter la réflexion sur la finalité des analyses de ce type de données et de sensibiliser aux règles de sémiologie graphique indispensables pour toute présentation efficace. (Pourquoi ? Pour qui ? Comment ?).

Cette communication présentera le contenu de ma formation dont l'objectif est de fournir sur une période courte (2 jours) les outils indispensables pour construire des statistiques sur des textes, intégrer le vocabulaire de la méthode et ses mesures avec pour objectif de bien la comprendre afin de présenter les résultats qui ne semblent pas, pour un public non averti, sortir d'une « boite noire ».

Je commencerais ma présentation par l'historique de ma démarche et ferais un retour sur cette expérience de formation en analyse exploratoire très demandée. Enfin, je reviendrais sur ses atouts et ses limites.

Mots-clés. Formation continue, corpus, lexicométrie, statistique textuelle, sémiologie graphique

Abstract

The emergence of the “Big data” and the textual data collected on the web multiply corpuses and the demand to learn how to explore this new material has never been so strong. On the other hand, the offer of free softwares to analyze them is quickly growing.

From 2010, as part of trainings at Institut National d’Etudes Démographiques (INED), I developed a course in textual statistics which aim was to enable the analysis of open questions in surveys, life stories or interviews related to surveys. Today the aim of the course is to propose an exploratory tool to analyze these new corpuses, which created an increase demand for this kind of courses inside or outside the institute.

The structure of this training alternates between lectures and technical applications. The participants are asked to use their own text data in order to stimulate their reflexion about the reasons of the production of the analysis on these kind of “special” data and to present the rules of graphic semiology, essential skill for an effective presentation (Why ? For who ? How ?).

The content of the course will be presented with the aim to give in a short time period (2 days) the essential tools to provide textual statistics, assimilate the vocabulary and the figures particularly to understand the method to be able to present it to a non-specialized public.

I will start my communication with how I developed this approach and I will then reflect on this highly-demanded course in exploratory analysis

Finally, I will discuss of the assets and limits of my approach and of my training experience.

Keywords. Trainings, corpus, lexical analysis, textual statistics, graphic semiology

1 Introduction

Ingénieure en statistique dans un service d'appui à la recherche, je propose depuis longtemps des ateliers et formations internes en statistique textuelle. Dès 1995, des chercheurs de l'Ined ont utilisé ces méthodes pour analyser des réponses à des questions ouvertes dans un questionnaire d'enquête (Guerin Collomb) ou des entretiens passés auprès d'enquêtés ciblés en complément d'une enquête quantitative (Bonvalet). Aujourd'hui la statistique textuelle est très en vogue et les demandes de formation se sont démultipliées.

2. Nouvelle méthode ?

La statistique textuelle regroupe plusieurs approches de la statistique : lexicométrie, analyse factorielle, classification, calcul de spécificités qui ont émergé depuis la 1ère moitié du XXe siècle (cf Lebart CFIES de 2010) :

- statistique lexicale en 1944 (Yule),
- analyse multidimensionnelle lexicale en 1960 (Benzecri),
- stylométrie en 1964 (Mosteller et Wallace)
- la méthode « Alceste » en 1983 (Reinert).

Ces méthodes permettent, au moyen de méthodes statistiques et sans a priori, de faire émerger le contenu (thématique) d'un corpus ou de répondre à la question *Qui dit quoi* . Elles sont aujourd'hui très utilisées en Text mining (notamment en machine learning) mais ne sont pas nouvelles. Toutefois, l'utilisation de textes et de mots s'est développée dans le monde scientifique ainsi que l'émergence d'outils de visualisation comme les nuages ou graphes de mots.

Nouveaux outils ?

Ces méthodes ont été implémentées dans des logiciels quand l'informatique a permis de faire des calculs sur les tableaux lexicaux de très grande dimension (en termes de nombre de lignes et de colonnes) ; ainsi on peut évoquer, dès les années 1980, SpadT, Alceste ou Lexico.

L'arrivée de logiciels gratuits/libres, comme R avec le package TM en 2008, a contribué à élargir le nombre d'utilisateurs potentiels et le développement d'applications.

Nouvelles données ?

Le volume des textes n'est pas non plus une nouveauté, on peut encore citer L. Lebart qui évoque des banques de données textuelles comme la base Frantex regroupant les « Trésors de la langue française » d'E. Brunet dès 1981.

Cependant, l'accessibilité à des bases de données en ligne comme le site data.gouv.fr et les techniques de scapping (aspiration) ont permis un accès « gratuit » et simple à des corpus relativement récents.

Nouveaux utilisateurs/ nouvelles utilisations

L'accès aux formations en méthodes quantitatives ou en statistiques s'est rapidement étendu. De nombreux métiers, aujourd'hui traitent du chiffre. Ceci élargit le public pouvant avoir accès à des méthodes de statistiques, notamment textuelles. Internet rend presque tout accessible immédiatement, générant des corpus de textes qu'il est tentant d'analyser. Enfin, nous faisons face aujourd'hui à de nouveaux enjeux : des textes et données qui restaient non traités hier qui sont maintenant sujets à analyses et questions. Aujourd'hui, on se doit de tout analyser sans parfois rien n'y trouver.

3- Faire pratiquer pour questionner

Dans le cadre de mon activité d'ingénieur dans un service d'appui à la recherche, j'ai mis en œuvre la méthode sur des corpus liés à des enquêtes Ined (réponses à des questions ouvertes, mots associés, entretiens) auprès de chercheurs, en utilisant des logiciels dédiés. Par la suite, j'ai été amenée à transmettre ces compétences acquises aux autres chercheurs de l'Ined dans le cadre

de formations internes dès 2010. J'ai ainsi proposé cette formation régulièrement et la demande n'a cessé de croitre en fonction de l'émergence de « nouvelles » données et d'outils de plus en plus accessibles. Cette formation m'a ensuite été demandée « hors de murs ». Aujourd'hui, ma formation s'adresse à multiples publics : des chercheurs de différentes disciplines (comme la sociologie, la géographie, l'histoire, …), des étudiants de master en sociologie ou en statistique ou encore de chargés de recherche dans le secteur privé.

Si les méthodes statistiques sous-jacentes de la statistique textuelle ne sont pas nouvelles pour des statisticiens ou des sociologues « quantitatifs » une partie du vocabulaire est issue de la collaboration des statisticiens avec des linguistes dans leur genèse (corpus, lemme, occurrence, hapax).

La partie théorique de la formation est simple, avec peu de formules mathématiques, pour permettre de se concentrer sur ce que représente la donnée textuelle, les indicateurs issus de l'analyse et ce que l'on peut interpréter. Chaque étape progressive de l'analyse (lexicométrie, calcul de spécificités, analyse factorielle, classification) est suivie d'une mise en œuvre avec un logiciel « gratuit » qui sert surtout à comprendre le corpus, les calculs et les aides à l'interprétation fournis afin de répondre à une problématique.

Lors de la mise en pratique de chaque étape, chacun peut travailler sur des données qu'il a lui-même collecté et dont il connait (ou pas) les limites (données aspirées du web, issues d'une enquête, titres d'articles, …). Ce pas-à-pas permet de comprendre que le corpus est fabriqué de données textuelles et leur métadonnées (caractéristiques des locuteurs, lieu, dates, …).

Cette démarche soulève plusieurs écueils. D'une part, quand on traite des données textuelles, la mise en situation permet de s'interroger sur l'interprétation des « mots », trop rapidement sortis du texte : il est indispensable de faire des allers-retours entre mots pris dans leur contexte (concordances), interprétation et choix des données à analyser (analyse d'un sous-corpus de répondantes femmes par exemple). D'autre part, pouvoir utiliser des métadonnées, parties prenantes de l'analyse, peut mettre en lumière la « pauvreté » de données comme c'est le cas par exemple de l'analyse de blogs pour lesquels on dispose très peu d'informations sociodémographiques. Enfin, il faut d'abord comprendre comment ont été constituées les données (garbage in→garbage out). Sans connaissance des données initiales, les explications/interprétations des résultats sont moins évidentes.

La dernière partie de la formation est consacrée à quelques présentations rapides de mise en œuvre par les stagiaires permettant de mettre en lumière la nécessité de suivre des règles de présentation de résultats. Le nuage de mots par exemple est un modèle : il est généré rapidement sans titre ni légende. Il offre l'opportunité d'aborder l'utilisation de variables visuelles et de faire une introduction sur la sémiologie graphique afin de faciliter la lecture de résultats et s'adapter au public. Ce constat peut s'étendre à toute présentation de résultats ou graphiques statistiques.

4 Conclusion

Au-delà de la présentation de méthodes et de logiciels dédiés, ces formations donnent lieu à des discussions sur les analyses exploratoires, la constitution des fichiers et le temps consacré à la préparation des données. La difficulté ne réside pas dans la nouveauté mais dans la nécessité de cerner son sujet, poser sa problématique et faire des choix. Elle rappelle le besoin de faire un travail de fouilles et de description simple des données avant de pouvoir le compléter par des analyses plus poussées. Il faut apprendre à ne pas se perdre dans les données et les analyses.

Pour cela, seule la pratique permet de se confronter à la réalité de ces nouvelles

données ouvertes ou libres (« open data ») et aux difficultés techniques qu'elles engendrent (taille des données, comparabilité des systèmes d'exploitation, des formats) et le risque de générer des résultats limités, triviaux, et pas assez contextualisés pour répondre à une problématique de recherche.

Bibliographie

- Benzecri J.-P., 1973 – *L'analyse des Données* (tome 1 et 2). Dunod, Paris
- Bertin J., 2005, *Sémiologie graphique. : Les diagrammes - les réseaux, - les cartes*, EHESS
- Garnier B., Guérin-Pace F., 2010 - *Appliquer les méthodes de la statistique textuelle*, Ceped, les clefs pour, Paris (http://www.ceped.org/fr/publications-ressources/editions-du-ceped-1988-2012/les-clefs-pour/article/appliquer-les-methodes-de-la)
- Guérin-Pace F. 1997. La statistique textuelle : un outil exploratoire en sciences sociales. In: *Population*, (4), Ined. pp. 865-887, Ined. Paris
- Lebart L., Salem A. 1994. *Statistique textuelle*. Paris, Dunod, 342 p.
- Vautier, C. (dir.) 2015. Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles. Volume 11, numéro 1, l'analyse de données textuelles informatisée, Prise de parole, p. 15-461
- Bonvalet C., Gotman A., Grafmeyer Y., Bertaux-Wiame I., Le Bras H., Maison D.,1999, La famille et ses *proches : l'aménagement des territoires,* Travaux et documents N°143, Ined
- Collomb Ph., Guérin-Pace F. 1998. Les contours du mot « environnement » : enseignements de la statistique textuelle Espace Géographique, L'espace géographique, 41 (1), p. 41-52 (1)
- Tufféry S. Data Mining et Statistique décisionnelle. (4^{e} Ed) Technip
- Reinert M. 1983, Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. Cahiers de l'Analyse des Données, 3, pp. 187-198

Sur la toile

http://lexicometrica.univ-paris3.fr/ (actes des JADT)
http://textometrie.ens-lyon.fr/ (projet Textometrie)
http://tal.univ-paris3.fr/wakka/wakka.php?wiki=Glossaire (Glossaire de Statistique Textuelle)
http://www.dtmvic.com/ (site de Ludovic Lebart)