Cinquante ans de résumés d'Economie et Statistique

Fifty Years of Abstracts in Economie et Statistique

Julie Djiriguian* et François Sémécurbe*

Le traitement du langage naturel, véritable boîte à outils d'analyse textuelle, est de nos jours couramment utilisé pour explorer le contenu de divers textes. On en propose ici, à l'occasion des 50 ans de la revue *Économie et Statistique* (puis *Economie et Statistique / Economics and Statistics*), une application aux résumés des 2 184 articles « académiques » qui y ont été publiés depuis 1969 (voir encadré). Quels sont les mots dont la fréquence est la plus élevée ? Quelles thématiques sous-jacentes suggèrent-ils et ces thématiques ont-elles changé au cours des années ?

À l'issue de pré-traitements (encadré), nous obtenons un ensemble de 181 572 mots pour les 50 années. Une représentation sous forme de nuage de mots permet de mettre en évidence les mots les plus fréquents (figure I).

Figure I
Nuage de mots sur le corpus des résumés de 1969 à 2019

Note de lecture : emploi est le terme le plus fréquent dans l'ensemble du corpus des résumés d'Économie et Statistique (avec 2 176 occurrences sur 181 572 mots).
Source : résumés des articles académiques, *Économie et Statistique* (1969-2016) et *Economie et Statistique / Economics and Statistics* (2017-2019).

Codes JEL / JEL Classification : C38, C63
Mots-clés : analyse textuelle, traitement du langage naturel, modèle de thème, allocation latente de Dirichlet
Keywords: text analysis, natural language processing, topic modeling, Latent Dirichlet Allocation

** Insee, SSP Lab (julie.djiriguian@insee.fr ; francois.semecurbe@insee.fr)*

Citation : Djiriguian, J. & Sémécurbe, F. (2019). Fifty Years of Abstracts in the Journal Economie et Statistique. *Economie et Statistique / Economics and Statistics*, 510-511-512, 7–11. https://doi.org/10.24187/ecostat.2019.510t.1999

Dans sa globalité, cette représentation du vocabulaire des résumés illustre d'abord le caractère généraliste de la revue. Le mot dont la fréquence relative est la plus élevée est 'emploi', puis, par fréquence relative décroissante, les mots 'entreprise' et 'ménage'.

Les mots les plus fréquents sur cinquante ans le sont aussi, naturellement, par décennie, et la trilogie 'emploi', 'entreprise', 'ménage' se confirme, dans un ordre variable jusque dans la décennie 2000 et avec des éclipses – 'ménage' dans la décennie 2000, 'entreprise' dans la dernière décennie (figure II). La variabilité est beaucoup plus grande pour les mots dont la fréquence relative est plus faible.

Figure II
Nuages de mots par décennie

Source : résumés des articles académiques, *Économie et Statistique* (1969-2016) et *Economie et Statistique / Economics and Statistics* (2017-2019).

La plus grande constance est celle du mot 'emploi'. Il serait toutefois aventureux d'interpréter cette dominance comme le signe d'une « spécialisation » des articles publiés dans la revue. On peut voir plutôt ce mot comme un terme pivot, autour duquel peuvent s'articuler de nombreux angles d'analyse de l'activité économique au niveau macro ou, au niveau micro, des comportements et de la situation des acteurs, entreprises et ménages. En tirant un peu plus le fil, on peut aussi y voir le reflet d'une préoccupation pour l'emploi presque permanente depuis la fin des années 1970, qui en ferait soit le sujet d'intérêt, soit le point

d'entrée, de nombreux articles ; et en tirant encore un peu, rappeler que l'enquête 'emploi', l'une des plus anciennes des enquêtes de l'Insee couvrant la population d'âge actif, est l'une des sources les plus mobilisées par les travaux publiés dans la revue.

Quantifier les mots les plus fréquents n'est évidemment pas suffisant pour décrire les contenus d'un ensemble de textes – d'autant qu'il ne s'agit que de « mots » et pas de « mots-clés », et qu'ils sont considérés indépendamment les uns des autres. Les méthodes de modélisation thématique permettent de dégager des associations, en analysant simultanément l'ensemble des mots qui constituent un texte. Pour tenter d'aller un peu plus loin, nous utilisons ici l'allocation latente de Dirichlet (voir encadré), qui repose sur une modélisation probabiliste. Cette méthode est fréquemment utilisée pour interpréter des thèmes sous-jacents à partir du groupe de mots qui les caractérisent. Notons toutefois que, comme toute analyse textuelle, cette méthode s'appuie sur des hypothèses fortes et des choix (notamment les pré-traitements réalisés) qui conditionnent le résultat, et que l'identification – ou l'interprétation – de thèmes à partir des seuls mots associés à ceux-ci peut s'avérer délicate.

Cette méthode exigeant de fixer *a priori* le nombre de thèmes, nous l'avons fixé à trois. À l'issue des diverses estimations réalisées pour l'ensemble des résumés, on obtient des associations de mots que l'on va, par commodité, dénommer par leur mot le plus fréquent – qui renvoie (forcément, puisqu'à la base se trouve le même « stock » de mots) à l'un ou l'autre des trois mots qui apparaissaient les plus fréquents dans la figure I :

- un thème dit « entreprise », qui évoque le vocabulaire de l'activité économique au sens le plus standard, plutôt macroéconomique :

entreprise / croissance / secteur / france / industriel / marché / pays / production / industrie / activité / emploi / économique / prix / investissement / économie / travail / développement / taux / productivité / produit / cours / demande / baisse / structure / commerce / politique / extérieur / petit / étranger / terme / compte / capital / région / service / productif / expliquer / intérieur / coût / équipement / fortement…

- un thème dit « ménage » qui associe plutôt des mots du vocabulaire des revenus et des conditions de vie :

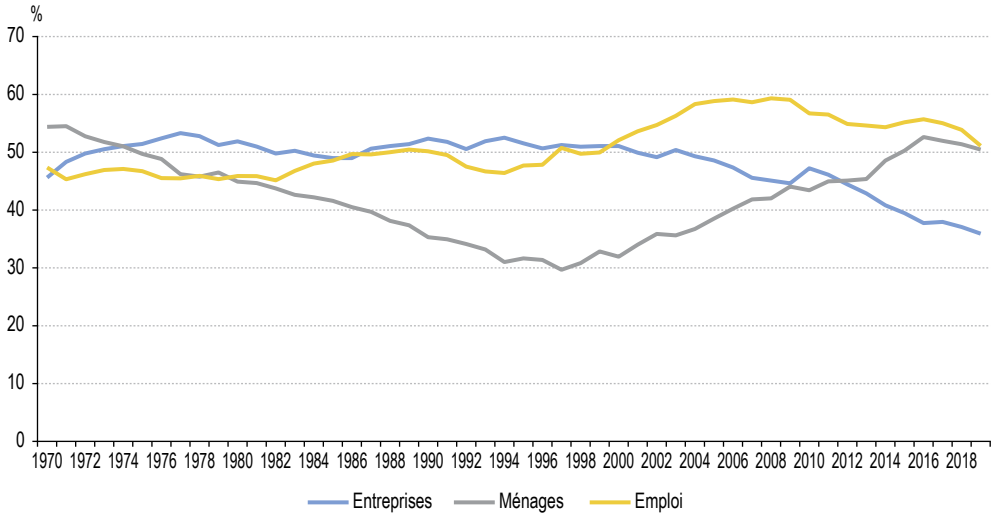
ménage / enquête / revenu / compte / économique / statistique / social / modèle / france / terme / méthode / financier / logement / question / estimation / politique / travail / information / consommation / taux / principal / système / vie / prix / insee / aide / national / public / approche / conduire / coût / prendre / comportement / base / individuel / étudier / dépense / population / évaluation / économie…

- un thème dit « emploi », où l'on retrouve les catégories de l'analyse microéconomique du marché du travail :

emploi / travail / jeune / femme / âge / actif / salarié / chômage / professionnel / activité / catégorie / social / homme / temps / ouvrier / vie / enfant / salaire / durée / cadre / population / enquête / famille / supérieur / marché / occuper / formation / familial / âgé / secteur / profession / augmenter / cours / écart / génération / travailler / diplôme / chômeur / revanche / diminuer…

Pour finir, on peut représenter le poids des « thèmes » au fil des années, comme la proportion de résumés contenant au moins trois des principaux mots de chaque thème (figure III). La présence dans les résumés des mots associés au thème « emploi » tend à s'élever à partir des années 1980, puis de façon plus marquée du milieu des années 1990 à la fin des années 2000. Celle des mots du thème « entreprise » au contraire décroît à partir des années 2000. Enfin, la présence des mots du thème « ménage » présente un aspect plus singulier, avec une baisse jusqu'à la seconde moitié des années 1990, puis une remontée d'ampleur équivalente ensuite.

Figure III
Part des résumés contenant au moins l'un des trois principaux mots de chaque thème



Note : le total est supérieur à 100% car le même mot peut être présent dans plusieurs thèmes.
Source : résumés des articles académiques, *Économie et Statistique* (1969-2016) et *Economie et Statistique / Economics and Statistics* (2017-2019).

* *
*

Nous arrivons ici aux limites de l'exercice proposé dans ce court article, qui n'avait d'autre visée qu'illustrative. En retirer une interprétation demanderait un investissement beaucoup plus important… qui devra attendre, car si 2 184 articles et 181 572 mots semblent « beaucoup », c'est un corpus de taille modeste pour la mise en œuvre des techniques mobilisées ici. □

ENCADRÉ – **Méthodologie**

L'analyse textuelle rassemble l'ensemble des méthodes visant à extraire et analyser l'information contenue dans des textes. Elle peut être mobilisée sur des données d'origine très variée (textes administratifs, décisions juridiques, échanges sur les réseaux sociaux, etc.) pour en faire apparaître les thèmes sous-jacents, en analyser le sentiment, ou encore pour prédire une variable (cf. Anzovino *et al*., 2018 ; Wu *et al*., 2018 ; Xing *et al*., 2018). Les données textuelles sont par nature non structurées ; toute démarche d'analyse textuelle commence donc par une étape de préparation du texte qui vise à le nettoyer et à le transformer en données numériques exploitables. C'est sur ces données numériques que porte ensuite l'analyse statistique proprement dite.

Les observations statistiques d'une analyse textuelle sont des textes, appelés « documents » (ici les résumés). Chaque document est décomposé en éléments (*tokens*), mots, ponctuation, association de plusieurs mots (n-grams) si nécessaire. Les pré-traitements consistent d'abord à supprimer les éléments non informatifs – ponctuation, chiffres, et *stopwords.* Les *stopwords* sont des mots non significatifs qui apparaissent dans l'intégralité du corpus étudié ; certains sont « évidents » (par exemple, les conjonctions de coordination) mais d'autres demandent un arbitrage, inévitablement subjectif. L'ensemble des termes restant après cette étape constitueront les « variables » de l'analyse. Ces étapes de pré-traitement sont souvent fastidieuses et peuvent impliquer des choix arbitraires ou *ad hoc.* L'analyse textuelle en aval est très sensible à ces choix.

Les termes informatifs sont alors normalisés pour les rendre comparables : harmonisation de casse, correction d'orthographe, « lemmatisation ». La lemmatisation consiste à trouver la forme neutre d'un mot : par exemple, un verbe conjugué se retrouve, après cette opération, à l'infinitif. Cette opération est complexe, car elle nécessite notamment de désambiguïser les cas d'homonymie. Documents et « variables » peuvent ensuite être représentés par une matrice numérique où chaque ligne mesure →

Encadré (suite)

pour un document donné le nombre d'occurrences (ou une autre mesure : le codage binaire – présence/absence – est classique) de chaque mot / « variable » de l'ensemble du vocabulaire retenu au sein de chaque document. La matrice obtenue est souvent de grande dimension (il y a plus de mots / de colonnes que de documents/lignes) et *sparse* (beaucoup de 0). Elle peut être analysée à l'aide de diverses méthodes statistiques.

L'interprétation d'un texte résulte de l'association des mots (Hapke *et al*., 2019). Pour examiner ces associations, nous avons mis en œuvre ici une analyse relevant du *Topic Modelling* appelée Allocation latente de Dirichlet (*Latent Dirichlet Allocation*, LDA, cf. Blei *et al*., 2003). Il s'agit d'un modèle probabiliste génératif, qui estime par des méthodes d'inférence bayésienne (Bayésien variationnel, échantillonnage de Gibbs, etc.) à partir des mots observés dans les documents, le poids des thèmes dans chaque document et les distributions des mots caractéristiques de chaque thème. Cette méthode exige de fixer *a priori* le nombre de thèmes.

La LDA repose sur des hypothèses fortes qu'il est nécessaire de rappeler. Tout d'abord, l'estimation des paramètres des distributions des mots pour chaque thème et de celles des thèmes dans un même document débute par une initialisation aléatoire : deux initialisations différentes peuvent engendrer deux structures thématiques différentes. Ensuite, la LDA, comme une large partie des méthodes d'analyse textuelle, repose sur l'hypothèse selon laquelle l'ordre des mots n'a pas d'impact (on parle d'approche « sac de mots », *bag-of-words*). Sous cette hypothèse, les documents sont découpés en listes non ordonnées de mots. Comme les mots sont aussi déterminants pour l'interprétation des thèmes, les pré-traitements sont là aussi cruciaux.

L'analyse présentée ici porte sur l'ensemble des résumés des articles académiques publiés dans la revue entre 1969 et 2019. Préalablement aux pré-traitements des textes, nous avons écarté 764 articles « non académiques » : la revue publiait en effet jusque dans les années 1970 la présentation de résultats d'enquête, des panoramas territoriaux, ou d'autres petits articles d'information qui ont ensuite disparu (ou ont donné lieu à des publications dans d'autres collections de l'Insee). Les introductions générales de numéros n'ont pas non plus été retenues. Restent 2 184 résumés qui contiennent 432 000 mots.

Les pré-traitements ont principalement consisté à supprimer les chiffres et *stopwords* contenus dans les résumés et « lemmatiser » les mots. À cette fin, nous avons utilisé la libraire *Spacy* sous Python qui reconnaît la fonction grammaticale des mots dans un texte, ce qui est plus efficace que l'usage d'un simple dictionnaire. Aux *stopwords* proposés par *Spacy*, nous avons ajouté des mots *ad hoc* en fonction des résultats obtenus dans les traitements statistiques (par exemple, le mot « année », qui produit des liens non significatifs entre des résumés). À l'issue de ces pré-traitements la base contient 2 184 résumés et 181 572 mots.

Références

Anzovino, M., Fersini, E. & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems,* pp. 57–64. Springer, Cham.

Blei, D & Ng, A. Y. & Jordan M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Hapke, H. M., Lane, H. & Howard, C. (2019). *Natural language processing in action*. Manning.

Wu, J. T., Dernoncourt, F., Gehrmann, S. ... & Celi, L. A. (2018). Behind the scenes: A medical natural language processing project. *International journal of medical informatics*, 112, 68–73.

Xing, F. Z., Cambria, E. & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1), 49–73.