Nantes Université

Master 2 : ECAP

NLP

Analyse des tendances des publications scientifiques dans le domaine de l'économétrie en 2023

Bouedo Théo

July 14, 2024

Sommaire

Chapitre I

Introduction

Le Natural Language Processing, ou traitement automatique du langage naturel, est un domaine clé de l'intelligence artificielle et de la linguistique informatique qui se concentre sur l'interaction entre les ordinateurs et le langage humain. Le but principal du NLP est de permettre aux ordinateurs de comprendre, interpréter, manipuler et reproduire le langage humain de manière à ce qu'ils puissent effectuer des tâches spécifiques utiles telles que la traduction automatique, la reconnaissance vocale, et la génération de texte. Le NLP combine des techniques provenant de divers domaines tels que la linguistique, l'informatique, et l'intelligence artificielle pour aborder des problèmes complexes de compréhension et de production du langage. Les applications du NLP sont vastes, couvrant des systèmes de réponse aux questions, des assistants personnels intelligents, des interfaces de programmation, et l'analyse de sentiments, pour n'en nommer que quelques-unes.

Dans le cadre de ce projet, nous nous consacrons à l'exploration des tendances des publications scientifiques en économétrie pour l'année 2023, avec pour objectif de décrypter les thèmes et sujets dominants au sein des recherches récentes. En exploitant les techniques avancées du traitement automatique du langage naturel (NLP), nous analysons d'importantes quantités de données textuelles issues de publications académiques, collectées via l'API Open ALEX. Ce corpus de textes, rigoureusement nettoyé pour exclure tout contenu non pertinent ou non anglophone, sert de fondement à notre exploration thématique. Nous utilisons le Latent Dirichlet Allocation (LDA) pour identifier et classer les principaux topics, révélant ainsi les structures sous-jacentes et offrant une vue d'ensemble des discours prévalents dans le domaine. En complément, le modèle BERT est déployé pour approfondir notre analyse sémantique, permettant une compréhension contextuelle des textes qui transcende la simple extraction de topics pour s'attarder sur des relations et des structures sémantiques plus subtiles et complexes.

Chapitre II

Topic Modeling Simple

II.1 Traitement et présentation des données

Dans le cadre de notre projet d'analyse des tendances en économétrie pour l'année 2023, nous avons exploité l'API OpenAlex pour accéder à un large corpus de publications scientifiques. OpenAlex est une interface de programmation d'application robuste et ouverte qui offre un accès exhaustif aux métadonnées académiques à travers des millions de documents scientifiques. Cette API permet une exploration approfondie de diverses caractéristiques des travaux de recherche, telles que les titres, les DOI, les auteurs, et les mots-clés des résumés.

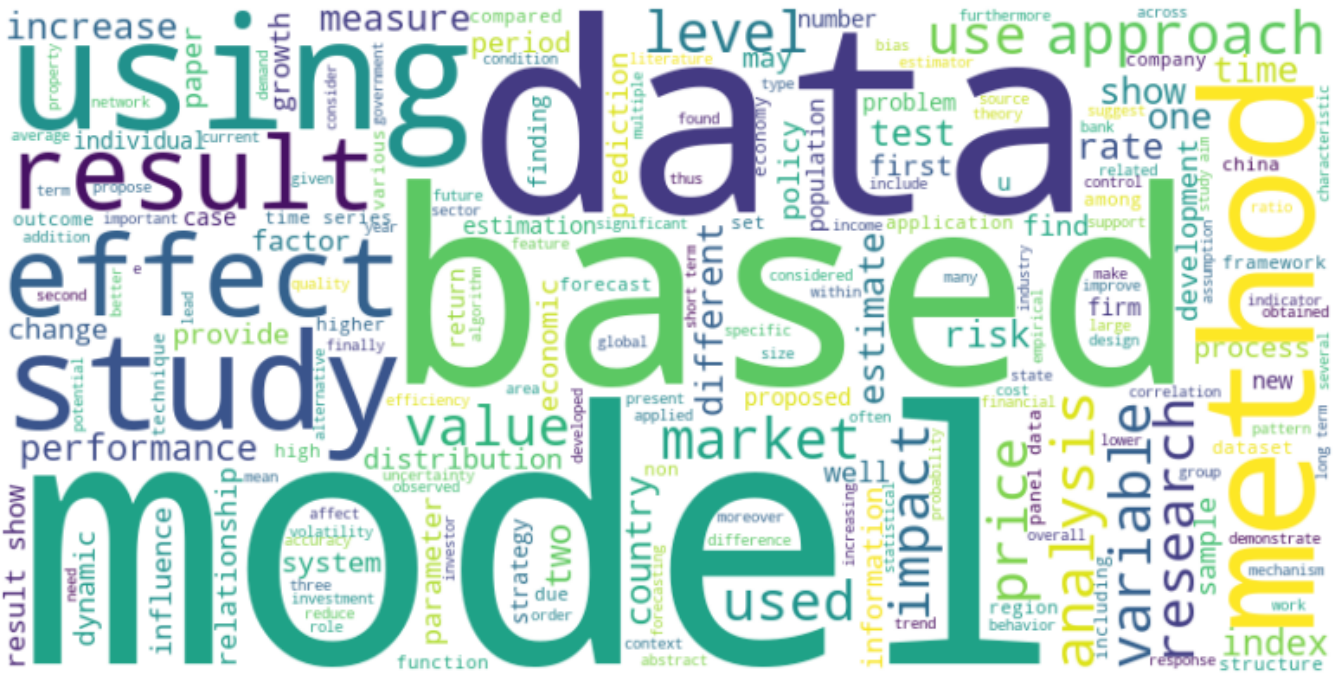
Pour notre analyse, nous avons initialement extrait plus de 75 000 entrées de l'API, chacune incluant le titre, l'année de publication, le DOI (Digital Object Identifier, un identifiant numérique unique pour les publications scientifiques qui permet un accès pérenne à ces documents), et les mots-clés du résumé. Le DOI est particulièrement crucial pour la traçabilité et l'accès aux travaux de recherche, tandis que les mots-clés du résumé offrent un aperçu condensé et pertinent des thèmes abordés, fournissant ainsi une base plus riche pour l'analyse sémantique que le titre seul.

Après avoir collecté les données, nous avons procédé à un nettoyage minutieux, en ne conservant que les articles en anglais et en excluant ceux dont les résumés étaient indisponibles. Ce processus de filtrage nous a permis de réduire le jeu de données à 65 000 articles. Toutefois, afin de minimiser les coûts de calcul, particulièrement élevés lors de l'utilisation de techniques de NLP avancées comme BERT (Bidirectional Encoder Representations from Transformers), qui nécessite des ressources computationnelles substantielles pour traiter le langage naturel, nous avons décidé de limiter notre étude à un échantillon représentatif de 5 000 articles.

Cette taille d'échantillon garantit un équilibre entre la profondeur analytique et l'efficience computationnelle.

Le traitement des données a inclus des étapes de prétraitement textuel telles que la conversion en minuscules, la suppression de la ponctuation et des chiffres, la tokenisation, l'élimination des mots vides (stopwords), et la lemmatisation. Ces étapes ont été essentielles pour préparer les textes à une analyse plus poussée via le topic modeling et l'analyse sémantique profonde, nous permettant de capter avec précision les nuances et les relations complexes dans les données textuelles des publications en économétrie.

Figure II.1 – Panorama des concepts clés dans les publications d'Économétrie de 2023



Ce graphique en forme de nuage de mots représente visuellement les termes les plus fréquemment utilisés dans un échantillon de 5000 articles académiques publiés en 2023 dans le domaine de l'économétrie. Chaque mot dans le nuage est dimensionné proportionnellement à sa fréquence d'apparition dans les résumés des articles, offrant une perspective instantanée des sujets et des concepts qui dominent la recherche actuelle en économétrie.

Les termes tels que "model", "market", "data", et "risk" sont particulièrement prédominants,

soulignant l'accent mis sur les modélisations complexes, l'analyse des marchés, la gestion des risques et l'utilisation intensive des données dans les études économétriques. Le mot "study" ressort également comme un point central, ce qui reflète la nature académique des travaux analysés. Cette visualisation met en lumière la diversité des approches et des thèmes abordés dans la recherche récente, des questions de "forecast" (prévision) aux enjeux de "performance" et "efficiency" (efficacité), indiquant une riche variété de domaines d'intérêt et de méthodologies employées.

Bien que le nuage de mots offre une visualisation directe des termes fréquents dans les études récentes en économétrie, son utilité est limitée en termes de profondeur et de spécificité. Cette méthode ne permet pas de déceler les relations entre les mots ou de regrouper les concepts en thèmes cohérents, donnant lieu à un panorama intellectuel large et éparpillé. Face à cette dispersion, l'utilisation du Latent Dirichlet Allocation (LDA) devient essentielle. LDA pallie les faiblesses du nuage de mots en classant les termes en topics structurés et compréhensibles, ce qui permet d'obtenir un aperçu plus riche et plus ordonné des tendances et des thèmes sous-jacents dans le corpus de données.

II.2 Latent Dirichlet Allocation (LDA)

Conçu par Blei, Ng, et Jordan en 2003 [1], LDA est fondé sur l'hypothèse que les documents sont des mélanges de topics, où un topic est défini comme une distribution sur un vocabulaire fixe. Ce modèle probabiliste est utilisé principalement pour découvrir la structure latente des collections de documents, offrant ainsi une méthode robuste pour l'analyse sémantique de vastes corpus textuels.

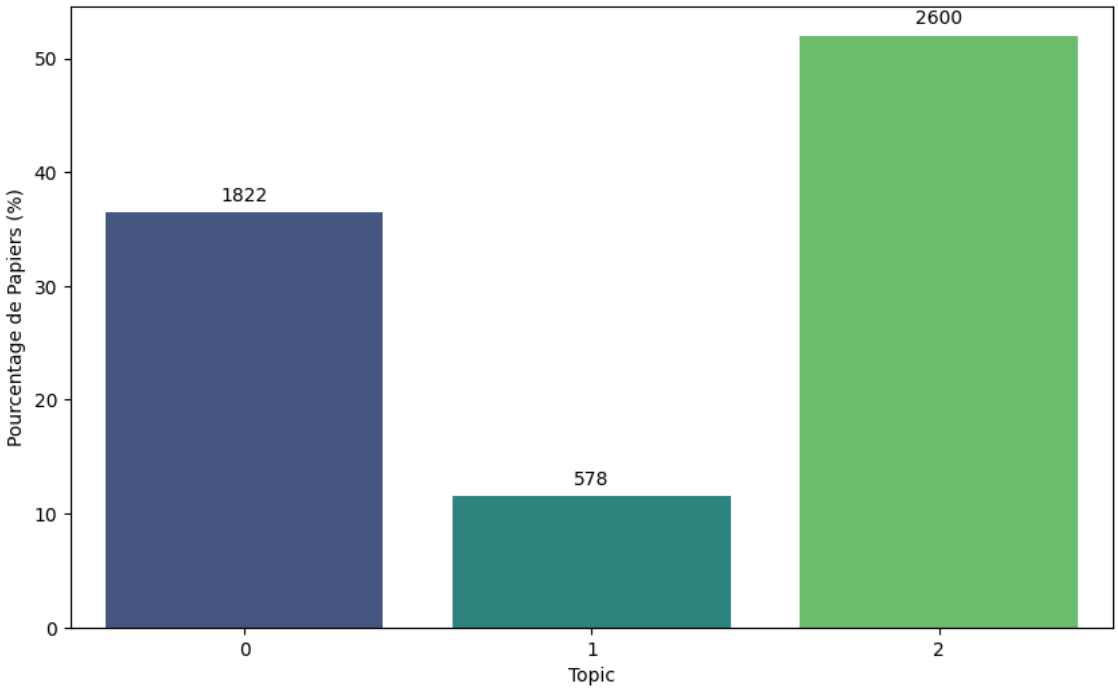
Le processus de génération des documents selon LDA est ainsi conceptuellement simple : pour chaque mot dans un document, un topic est d'abord tiré de la distribution de topics du document, puis un mot est choisi à partir de la distribution de mots de ce topic. Cette approche générative permet non seulement de comprendre comment un document est structuré mais aussi de faire des inférences sur des documents non vus en estimant leur distribution de topics.

Pratiquement, LDA a trouvé des applications dans de nombreux domaines au-delà de l'analyse de texte, comme le filtrage collaboratif, la classification automatique de documents et même en bio-informatique, où il aide à catégoriser des structures dans des données génétiques complexes. Cependant, le choix du nombre de topics et des paramètres du modèle reste un aspect critique qui nécessite une attention méticuleuse pour équilibrer la complexité du modèle et la qualité de l'interprétation des résultats.

L'analyse par Latent Dirichlet Allocation (LDA) de notre corpus de papiers en économétrie a abouti à l'identification de trois topics principaux, chacun capturant un ensemble distinct de thèmes et de concepts. Le choix de restreindre l'analyse à trois topics a été guidé par un désir d'équilibrer la granularité et la généralisabilité : trop de topics auraient pu diluer les thèmes principaux en sous-groupes trop spécifiques, tandis que trop peu auraient masqué des nuances importantes entre les sujets de recherche. Plusieurs tests ont été effectués avec un nombre de topics variant entre deux et six, et il est apparu que trois était le nombre optimal pour une analyse équilibrée.

Le graphique ci-dessous illustre la distribution des papiers parmi les trois topics identifiés. Le Topic 0 représente 35%, le Topic 1 seulement 11%, et le Topic 2 une majorité de 54%, mettant en évidence la prépondérance de certaines thématiques au sein de notre corpus. Cette répartition est également quantifiée par le nombre de papiers attribués à chaque topic, indiqué au-dessus de chaque barre, fournissant une mesure concrète de l'intérêt et de l'engagement dans chaque domaine thématique exploré.

Figure II.2 – Distribution des topics



II.3 Présentation des topics

Table II.1 – Topics et mots associés

Topic	Mots associés
Topic 0	country, effect, growth, panel, study, impact, economic, policy, firm, relationship, development, finding, result, level, economy
Topic 1	price, stock, market, return, asset, volatility, company, forecasting, forecast, financial, model, investor, portfolio, exchange, ratio
Topic 2	model, method, data, distribution, time, based, approach, estimate, parameter, study, proposed, paper, two, function, process

Topic 0 : Dynamique Économique et Impact des Politiques

Ce topic couvre un large éventail de recherches sur l'impact des politiques économiques et des facteurs de croissance sur les pays et les entreprises. Les termes tels que "country", "economic", et "policy" soulignent une focalisation sur l'analyse des politiques, tandis que "effect", "growth", et "impact" indiquent une concentration sur les résultats et les évolutions économiques.

Topic 1 : Marchés Financiers et Modélisation de Prévisions

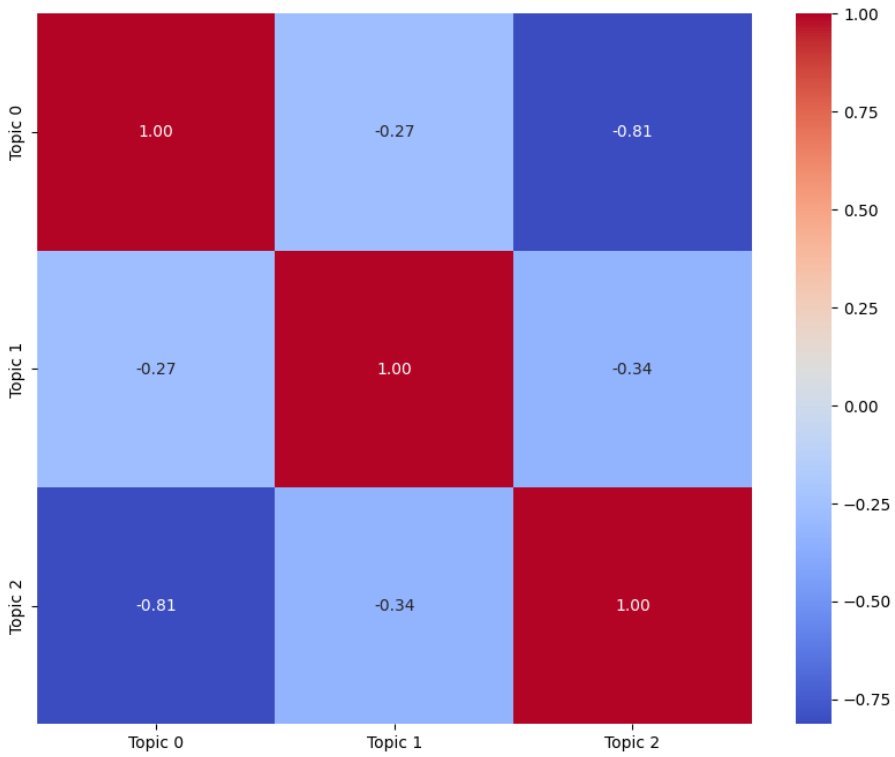
Ce topic se concentre sur la modélisation des marchés financiers et les prévisions de tendances économiques. Des mots comme "price", "stock", "market", et "forecasting" démontrent une orientation vers l'analyse et la prédiction des mouvements du marché financier, tandis que "volatility", "asset", et "portfolio" s'alignent sur les études de risque et de comportement d'investissement.

Topic 2 : Méthodes Statistiques et Analytiques Avancées

Les mots-clés de ce topic révèlent une concentration sur les méthodes quantitatives avancées pour l'analyse de données. "Model", "method", "data", et "distribution" suggèrent des recherches centrées sur la modélisation statistique et l'analyse de données complexes, tandis que "estimate", "parameter", et "process" se rapportent à l'application de ces méthodes à des problèmes spécifiques de recherche.

La matrice de corrélation présentée ci dessous démontre les interconnexions complexes entre trois topics principaux identifiés grâce à l'analyse LDA de notre corpus de recherche en économétrie. Cette visualisation aide à comprendre les relations sous-jacentes entre des sujets de recherche variés, éclairant les dynamiques de collaboration et de divergence conceptuelle.

Figure II.3 – Corrélation entre les différents topics



Dynamique Économique et Impact des Politiques (Topic 0) est légèrement en opposition avec Marchés Financiers et Modélisation de Prévisions (Topic 1), comme l'indique la corrélation négative de -0.27. Cela suggère que, bien que les deux topics partagent un contexte économique général, leurs approches et focalisations thématiques divergent, le premier se concentrant davantage sur les impacts macroéconomiques des politiques, tandis que le second se penche sur les aspects plus techniques et prédictifs des marchés financiers.

La relation entre Dynamique Économique et Impact des Politiques (Topic 0) et Méthodes Statistiques et Analytiques Avancées (Topic 2) révèle une divergence encore plus marquée, avec une forte corrélation négative de -0.81. Cette forte anti-correlation peut refléter un écart significatif entre les approches macroéconomiques et les méthodologies statistiques avancées, indiquant que les études sur l'impact des politiques économiques tendent à opérer indépendamment des techniques analytiques plus fines et centrées sur les données.

Enfin, Marchés Financiers et Modélisation de Prévisions (Topic 1) et Méthodes Statistiques et Analytiques Avancées (Topic 2) montrent également une relation négative modérée (-0.34), suggérant que les approches prédictives des marchés financiers et les méthodes statistiques avancées pourraient être appliquées dans des contextes distincts, avec peu de chevauchements dans leurs applications ou leurs bases théoriques.

Chapitre III

Analyse sémantique avancée

III.1 Embeddings BERT

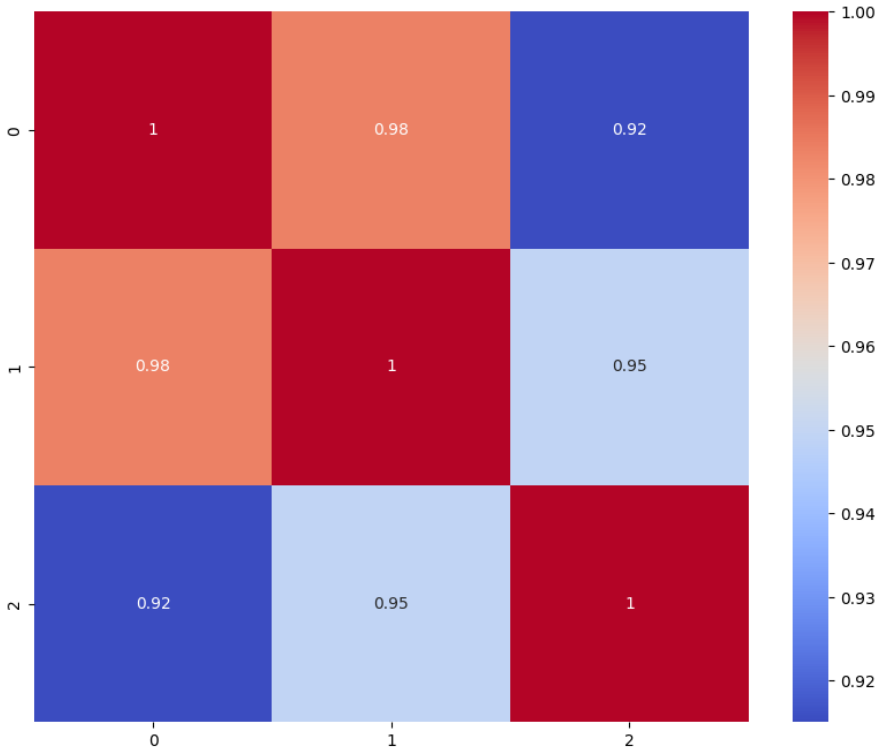
L'analyse sémantique avec les embeddings BERT représente une avancée majeure dans le traitement automatique du langage naturel (TALN), en offrant une compréhension profonde du contexte des mots au sein de leur environnement textuel. BERT (Bidirectional Encoder Representations from Transformers) est un modèle de pré-entraînement basé sur l'architecture des transformers, conçu pour pré-traiter de vastes quantités de texte et générer des vecteurs de caractéristiques (embeddings) qui capturent des subtilités linguistiques complexes.

Contrairement aux modèles antérieurs, BERT analyse les mots dans leurs contextes bidirectionnels (à la fois à gauche et à droite du mot cible), ce qui lui permet de saisir les significations contextuelles variées des mots. Par exemple, dans les phrases "Il est allé à la banque pour pêcher" et "Il est allé à la banque pour obtenir un prêt", le mot "banque" a des significations radicalement différentes que BERT peut distinguer efficacement grâce à son contexte bidirectionnel.

L'utilisation des embeddings BERT pour l'analyse sémantique enrichit notablement les études de corpus textuels en économétrie. En capturant la nuance contextuelle des mots et des phrases, BERT ouvre la voie à des interprétations plus nuancées et précises des données textuelles que les approches classiques comme LDA, qui se concentrent principalement sur la fréquence des termes sans prendre en compte le contexte de leur apparition.

III.2 Analyse des relations inter-topics

Figure III.1 – Similarités entre les topics



La matrice de similarité présentée ci-dessus, générée à partir des embeddings BERT, diffère conceptuellement d'une matrice de corrélation traditionnelle. Alors que la matrice de corrélation mesure directement les relations linéaires entre les topics, la matrice de similarité se concentre sur la proximité sémantique entre les embeddings des topics. Cette nuance est cruciale car elle permet de capter des nuances plus subtiles et des relations conceptuelles qui peuvent ne pas être strictement linéaires, offrant ainsi une vision plus riche des dynamiques thématiques.

Dynamique Économique et Impact des Politiques (Topic 0) et Marchés Financiers et Modélisation de Prévisions (Topic 1) : Avec une similarité très élevée de 0.98, cette proximité indique que bien que les deux topics puissent sembler distincts par leur focus—le premier sur l'impact macroéconomique et le second sur la modélisation financière—ils partagent en fait une base sémantique commune. Cela peut être dû à l'utilisation partagée de méthodologies quantitatives pour analyser les effets des politiques économiques et les prévisions des marchés financiers.

Dynamique Économique et Impact des Politiques (Topic 0) et Méthodes Statistiques et Ana-

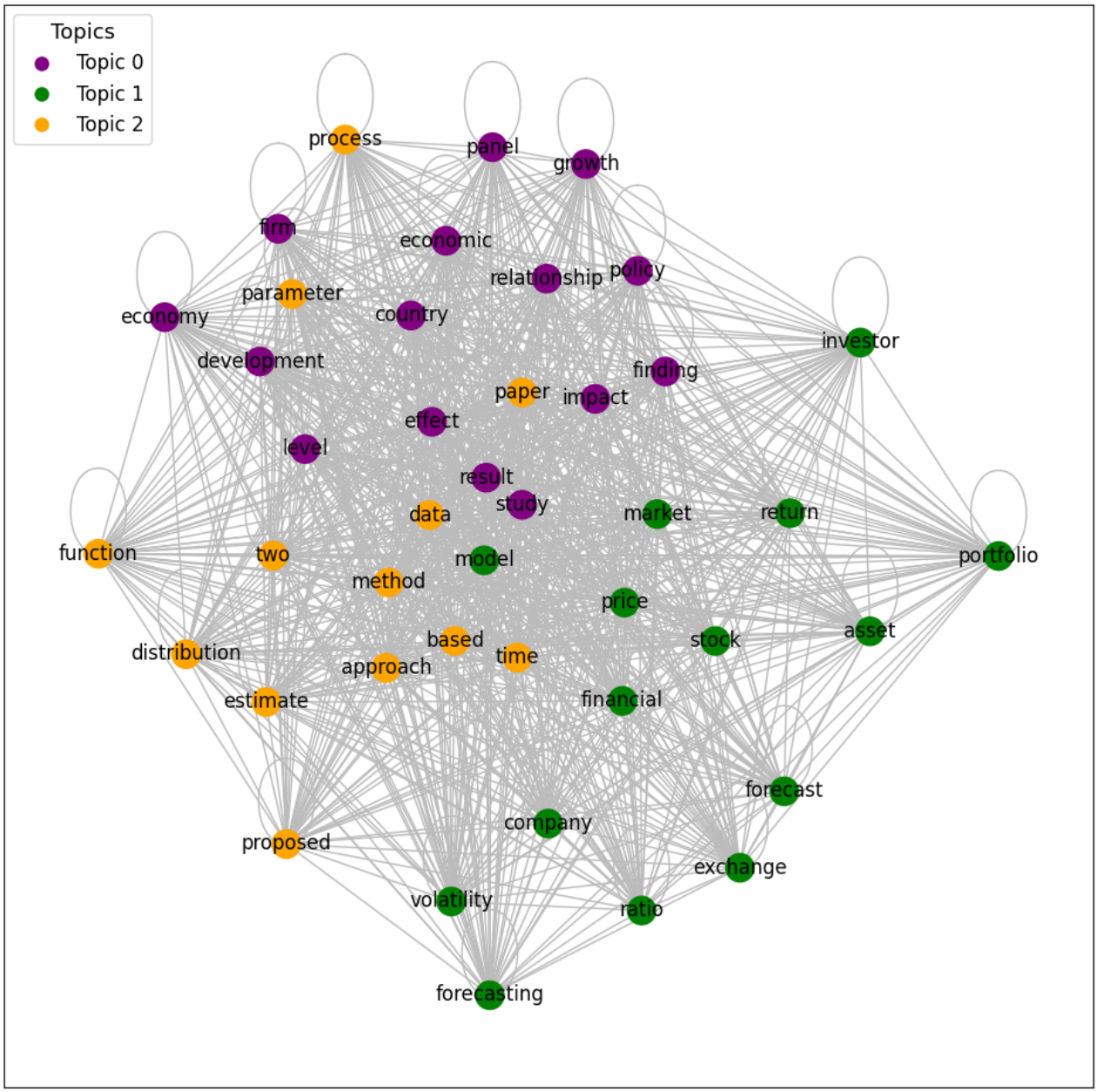
lytiques Avancées (Topic 2) : Le score de similarité de 0.92 suggère une forte corrélation positive, soulignant que les recherches sur les politiques économiques emploient souvent des méthodes statistiques et analytiques avancées pour évaluer leurs impacts. Cela indique une intégration substantielle des techniques de modélisation et d'analyse de données dans l'étude des effets macroéconomiques.

Marchés Financiers et Modélisation de Prévisions (Topic 1) et Méthodes Statistiques et Analytiques Avancées (Topic 2) : Avec un score de 0.95, cette forte similarité révèle que les modèles prédictifs pour les marchés financiers reposent fortement sur des méthodes statistiques et analytiques, affirmant que l'analyse quantitative est au cœur des études financières. Cela peut refléter une utilisation croisée de techniques avancées pour la modélisation du marché et la prévision financière.

Ces relations montrent non seulement la cohésion interne entre les sujets de recherche, mais aussi comment différents domaines peuvent se chevaucher significativement dans leurs approches méthodologiques et thématiques. Cette interconnectivité accentue la multidisciplinarité dans la recherche économique, où les méthodes quantitatives forment un pont entre les analyses des politiques, les études de marché et les approches statistiques.

III.3 Interconnexion et Multidisciplinarité

Figure III.2 – Réseau des 15 principales cooccurrences de mots par topic



Le réseau de co-occurrences présenté ci dessus illustre les interactions sémantiques entre les mots-clés centraux de trois domaines distincts en économétrie, révélant ainsi la complexité et l'interdépendance des sujets au sein de ce champ académique. Cette structure révèle non seulement les focalisations internes de chaque topic mais aussi leur interaction, soulignant ainsi la multidisciplinarité et la synergie méthodologique dans le domaine de l'économétrie.

Dynamique Économique et Impact des Politiques (Topic 0) Ce topic, représenté en violet, intègre des mots comme "economy", "policy", "impact", et "development", suggérant une concentration sur l'étude des politiques macroéconomiques et leur effet sur la croissance et le développement. Les connections entre ces mots et ceux d'autres topics indiquent que l'évaluation des politiques économiques peut souvent s'appuyer sur des prévisions financières (lien vers "forecasting" du Topic 1) et des analyses statistiques poussées (lien vers "data" et "method" du Topic 2), montrant l'intégration des techniques quantitatives dans l'analyse des impacts politiques.

Marchés Financiers et Modélisation de Prévisions (Topic 1) Les mots clés tels que "market", "stock", "forecasting", "financial", et "price" dominent ce cluster vert, accentuant l'importance des modèles prédictifs et des analyses de tendances dans les marchés financiers. Les liens avec le Topic 2 via "forecasting", "model", et "method" soulignent l'utilisation robuste des méthodes statistiques et analytiques avancées pour soutenir les prédictions de marché, révélant une forte interdépendance avec les techniques quantitatives avancées.

Méthodes Statistiques et Analytiques Avancées (Topic 2) Ce topic, visualisé en jaune, rassemble des termes comme "model", "data", "method", "estimate", et "distribution", illustrant un focus sur la modélisation statistique et l'analyse de données complexes. Ces mots forment des ponts conceptuels avec les discussions sur les impacts des politiques économiques et les modélisations des marchés financiers, témoignant de leur rôle central dans la fourniture d'outils analytiques et de modèles pour diverses branches de la recherche économique.

Cette analyse du réseau de co-occurrences dévoile l'entrelacement profond des domaines d'étude en économétrie, où les méthodes quantitatives lient les analyses des politiques économiques, les études des marchés financiers, et les approches statistiques avancées.

Chapitre IV

Conclusion

Ce projet a mis en lumière la capacité du traitement automatique du langage naturel (NLP) à structurer et à analyser d'importants volumes de données textuelles académiques. En se concentrant sur les tendances des publications en économétrie de 2023, nous avons pu identifier des thèmes prédominants et les relations subtiles entre eux grâce à des outils avancés tels que LDA pour le topic modeling et BERT pour l'analyse sémantique approfondie. Les résultats obtenus ont démontré non seulement la diversité des sujets de recherche dans le domaine de l'économétrie mais aussi les interactions complexes entre différents aspects de cette discipline.

Nous avons constaté que, bien que le NLP offre des informations précieuses, les approches actuelles pourraient être encore affinées. Par exemple, le modèle LDA, bien qu'efficace pour identifier les grands thèmes, pourrait bénéficier d'une calibration plus fine pour mieux discerner les nuances entre des sujets similaires. De même, l'usage des embeddings BERT, bien que révolutionnaire pour saisir le contexte linguistique, nécessite une exploration plus poussée pour optimiser la compréhension des liens entre les topics à un niveau plus granulaire.

Une voie d'amélioration pourrait être l'intégration de techniques de deep learning plus sophistiquées qui pourraient permettre une analyse encore plus nuancée des textes. Par exemple, l'application de réseaux neuronaux convolutifs (CNN) ou de réseaux de neurones récurrents (RNN) pourrait améliorer la modélisation des séquences textuelles pour une meilleure prédiction et classification des données textuelles.

Pour l'avenir, il serait pertinent d'explorer comment ces techniques de NLP peuvent être appliquées dans d'autres domaines de la recherche académique pour faciliter la synthèse des connaissances et la découverte de nouvelles informations. Par ailleurs, l'exploration de l'impact

de l'évolution rapide de l'IA sur les méthodes de recherche traditionnelles en économétrie pourrait ouvrir de nouvelles avenues pour la modélisation économique et la prévision politique.

En conclusion, ce projet a non seulement affirmé la valeur du NLP dans le domaine académique mais a aussi mis en lumière des domaines prometteurs pour des recherches futures, contribuant ainsi à l'avancement de la compréhension et de l'application de l'intelligence artificielle dans la recherche économique.

Bibliographie

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

Liste des tableaux

Table des figures

Table des matières