

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Analyse automatisée des documents pour le Compte des flux physiques des matières plastiques

par Oladayo Ogunnoiki et Alexandre Istrate

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Analyse automatisée des documents pour le Compte des flux physiques des matières plastiques

Oladayo Ogunnoiki et Alexandre Istrate¹

Résumé

Le Compte des flux physiques des matières plastiques (CFPMP) vise à améliorer l'analyse environnementale et économique en suivant les flux des matières plastiques dans l'économie canadienne. Afin de rationaliser ce processus complexe, le projet a tiré parti des techniques avancées de traitement du langage naturel (TLN), dont les grands modèles de langage (GML), pour automatiser la classification sectorielle et la synthèse des répercussions de la pandémie de COVID-19 à partir des rapports d'entreprises. L'intégration de modèles d'apprentissage automatique et de méthodes de génération augmentée d'information contextuelle (GAIC) a permis de réduire considérablement la charge de travail manuel, d'améliorer l'efficacité de l'analyse des données et d'obtenir des aperçus de meilleure qualité.

Mots clés : TLN; GAIC; GML.

1. Introduction

1.1 Description

Le **Compte pilote des flux physiques des matières plastiques (CFPMP)** est un compte économique et environnemental qui estime les flux des matières plastiques dans l'économie canadienne. Le compte présente des estimations annuelles par catégorie de produit, par type de résine, ainsi que par province et territoire. Dans le cadre de ce programme, les analystes s'appuient sur un éventail diversifié de rapports annuels provenant de diverses entreprises et organisations. De telles sources de données de rechange sont essentielles pour mener des recherches approfondies et vérifier les activités qui sont pertinentes au compte. Cependant, l'examen manuel d'un corpus substantiel de documents PDF est laborieux et inefficace, et demande beaucoup de temps aux analystes.

En utilisant de nouvelles techniques d'apprentissage automatique pour réduire cette charge de travail, le programme devrait bénéficier d'une efficacité accrue, permettant aux analystes d'extraire des aperçus inestimables à partir d'un plus grand nombre de sources de données, avec une précision et une efficacité accrues.

1.2 Objectifs

Dans le cadre de l'analyse de sources de données de rechange sur les flux de matières plastiques, constituées de divers rapports provenant de différentes entreprises, le projet visait à tirer parti des techniques avancées de traitement du langage naturel (TLN) afin d'atteindre deux objectifs clés :

Classification sectorielle : Utiliser des algorithmes de TLN pour classer les clients mentionnés dans les rapports des entreprises dans des secteurs distincts, à savoir : résidentiel, commercial, institutionnel, industriel et construction. Ce processus de classification permet de bien comprendre les diverses parties prenantes participant aux activités du flux des matières plastiques, facilitant ainsi une analyse plus détaillée et désagrégée.

¹Oladayo Ogunnoiki, Statistique Canada, oladayo.ogunnoiki@statcan.gc.ca; Istrate Alexandre, Statistique Canada, alexandre.istrate@statcan.gc.ca;

Synthèse des répercussions de la pandémie de COVID-19 : Mettre au point des algorithmes pour résumer automatiquement les répercussions de la pandémie de COVID-19 sur les activités liées à l'industrie des matières plastiques des entreprises. Il s'agit d'extraire les renseignements pertinents des rapports afin de fournir des aperçus concis, mais informatifs des répercussions de la pandémie sur la dynamique du flux des matières plastiques, notamment en ce qui concerne les taux de collecte et de recyclage et les perturbations logistiques.

L'objectif final était d'élaborer un pipeline algorithmique capable d'atteindre ces objectifs de manière transparente et de générer des documents contenant les résultats synthétisés.

2. Méthodologie

2.1 Données

L'ensemble des données comprend les rapports annuels provenant d'un large éventail d'entreprises et d'organisations prenant part à la gestion des flux de matières plastiques, principalement stockés sous forme de documents PDF textuels. Ces rapports annuels contiennent des résumés opérationnels et financiers, ainsi que des renseignements pertinents sur leurs activités liées à l'industrie des matières plastiques. Au total, l'ensemble de données comprenait plus de 500 documents totalisant plus de 20 000 pages. Étant donné la nature annuelle du corpus, plusieurs rapports visant différentes années de 2015 à 2023 provenant de la même organisation ont été inclus. Pour mener à bien la tâche de classification sectorielle, nous avons utilisé toutes les années disponibles, tandis que nous nous sommes concentrés sur les rapports produits à partir de 2020, soit environ 112 rapports, pour synthétiser les répercussions de la pandémie de COVID-19.

2.2 Outils et techniques d'analyse

2.2.1 Base de données de vecteurs

Une base de données de vecteurs est l'un des composants standard utilisés dans les projets exploitant les grands modèles de langage. Elle sert à stocker les textes et les représentations vectorielles qui saisissent leur signification sémantique. Les documents sont tout d'abord divisés en parties, pour lesquelles les représentations vectorielles sont calculées par un modèle d'intégration, puis sont stockées ensemble dans cette base de données. La base de données fournit non seulement le stockage, mais aussi des algorithmes d'extraction efficaces basés sur la recherche de similarités sémantiques. Un utilisateur peut alors extraire d'un document les parties qui ont la signification la plus similaire à une requête.

Cette approche résout le problème créé par la taille limitée de la fenêtre contextuelle des grands modèles de langage. Au lieu d'alimenter un modèle avec un document PDF entier comprenant des centaines de pages, seul un petit sous-ensemble de texte pertinent est utilisé.

Pour ce projet, nous avons stocké les documents PDF dans une base de données de vecteurs appelée Chroma en raison des possibilités supplémentaires d'interrogation des métadonnées qu'elle offre, ce qui nous permet par exemple de limiter la requête à certaines années ou à certains noms d'entreprise.

2.2.2 Génération améliorée par récupération d'information : Compression contextuelle

La génération améliorée par récupération d'information (GARI) est une technique d'intelligence artificielle conçue pour améliorer les réponses générées par les grands modèles de langage (GML). Elle fournit au GML des données qu'il n'a pas vues pendant son entraînement. Elle se compose de deux éléments clés : un modèle de récupération et un générateur. Le modèle de récupération interagit avec la base de données de vecteurs pour trouver les données les plus pertinentes pour une requête. Le générateur, quant à lui, utilise les données renvoyées par le modèle de récupération pour exécuter des tâches spécifiques telles que la synthèse, la classification, etc. Parmi les exemples de modèles de récupération de GARI, on peut citer EnsembleRetriever, MultiQueryRetriever et la compression

contextuelle. La technique de modèle de récupération de GARI sélectionnée pour ce projet est la compression contextuelle.

La compression contextuelle détaille la grande quantité de renseignements disponibles à la phase de récupération et la condense en un ensemble de renseignements plus petit et plus pertinent que le générateur (résuméur) peut utiliser efficacement (Hou et coll., 2024) (Lo et coll., 2023). Cela est nécessaire, car il arrive que le modèle de récupération renvoie des renseignements qui sont suffisamment proches sur le plan sémantique, mais qui ne sont pas pertinents pour la tâche définie dans la requête. Le compresseur résout ce problème en filtrant les renseignements non pertinents.

2.3 Processus d'automatisation

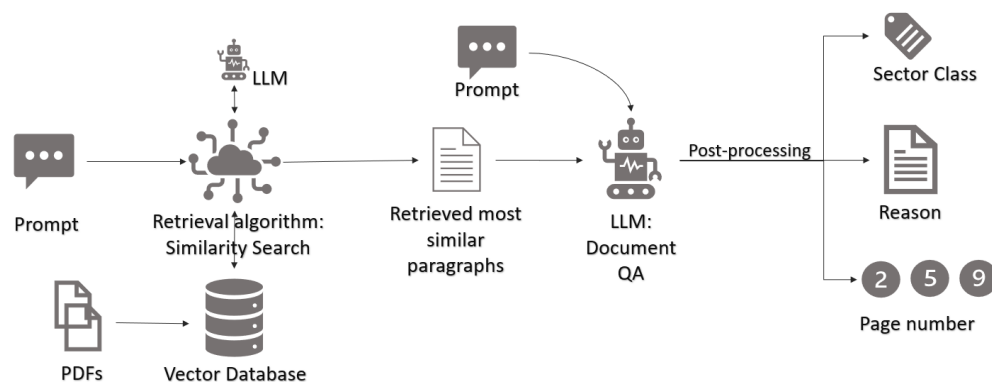
2.3.1 Classification sectorielle

Le pipeline de classification sectorielle utilise une combinaison d'algorithmes de recherche, de modèles linguistiques et de techniques de post-traitement pour parvenir à identifier avec précision et efficacité les secteurs. Le processus est décrit ci-dessous :

1. **Requête d'initialisation** : Au début du pipeline, une requête spécifique est formulée pour interroger l'algorithme de récupération. Par exemple, une requête typique pourrait être ce qui suit : « Ce document contient-il des clients résidentiels? »
2. **Conversion et récupération des données vectorielles** : La requête formulée est vectorisée, ce qui la convertit en une représentation numérique adaptée au traitement informatique. L'algorithme de recherche utilise ensuite la capacité de recherche de similarité de la base de données de vecteurs pour identifier les paragraphes du document qui correspondent étroitement à la requête donnée.
3. **Réponses générées par le GML** : Lors de la récupération des paragraphes pertinents, les documents récupérés et la requête initiale sont intégrés et fournis en tant qu'entrée au GML. L'objectif du GML est de générer une réponse intelligible à la requête initiale en fonction du contexte extrait des documents récupérés.
4. **Post-traitement et extraction de résultats** : La réponse générée par le GML subit une étape de post-traitement visant à extraire les renseignements pertinents. Cette étape comprend l'identification des secteurs mentionnés dans le document, la justification de l'identification et l'indication du numéro de la page d'où provient la conclusion.
5. **Itération et compilation** : L'ensemble du processus se répète pour tous les documents de l'ensemble de données. Les résultats de chaque itération, y compris les secteurs identifiés, les raisons et les références de pages, sont compilés dans un document Excel en vue d'une analyse et d'un examen plus approfondis.

La figure 2.3.1-1 illustre le processus de recherche et de traitement des renseignements à l'aide d'un GML. En partant de la gauche, le processus commence par une requête, suivi d'un algorithme de recherche qui effectue une recherche de similarité dans une base de données de vecteurs. Les paragraphes les plus similaires sont ensuite récupérés et traités par le GML pour répondre aux questions sur les documents. Le post-traitement a lieu avant que les résultats finaux, qui comprennent la classe de secteur, la raison et le numéro de page, ne soient déterminés.

Figure 2.3.1-1
Représentation visuelle du pipeline de classification sectorielle.



2.3.2 Classification sectorielle : Modèle génératif

Dans le pipeline de classification sectorielle, GPT-3.5 a été utilisé comme modèle génératif de base, offrant des capacités avancées de traitement et de compréhension de données textuelles complexes. Sa compréhension avancée du langage naturel a permis de détecter et de classer de façon nuancée de la terminologie et du contexte propres à l'industrie, ce qui a permis d'assigner les textes aux secteurs correspondants. L'application de GPT-3.5 a joué un rôle clé dans l'efficacité du traitement, en réduisant le temps nécessaire à la catégorisation manuelle des données tout en maintenant un niveau élevé de rappel.

2.3.3 Synthèse des répercussions de la pandémie de COVID-19

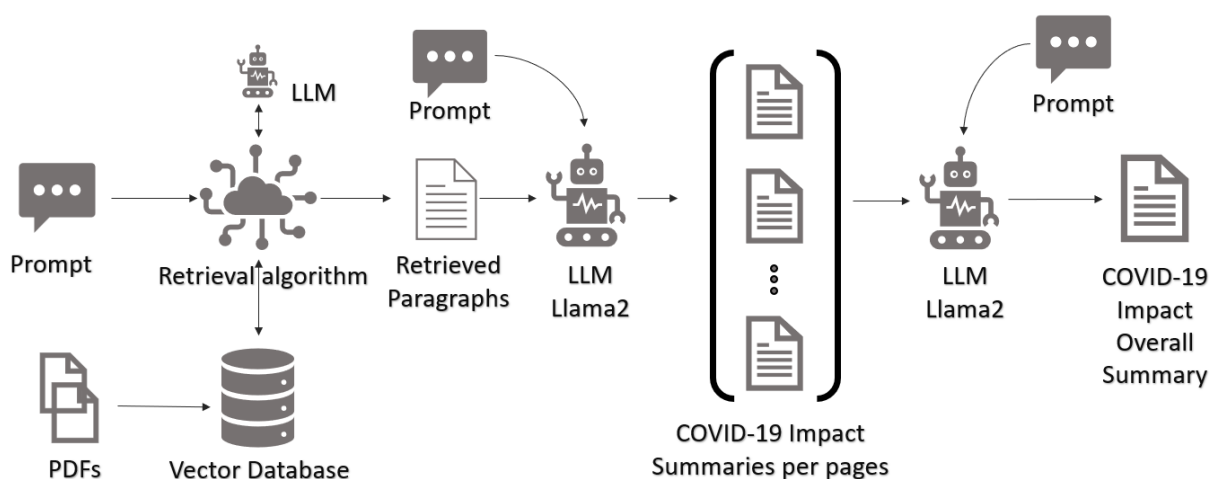
Cette section présente une vue d'ensemble du pipeline illustré dans la figure 2.3.3-1 ci-dessous, qui est conçu pour générer des résumés des répercussions de la pandémie de COVID-19 dans les documents. Le pipeline s'appuie sur des algorithmes de recherche, des modèles linguistiques et des techniques de post-traitement pour produire des résumés concis et informatifs. Le processus est décrit ci-dessous :

1. **Requête d'initialisation** : Le pipeline commence par la formulation d'une requête spécifique conçue pour interroger l'algorithme de recherche. Par exemple, une requête typique pourrait être la suivante : « Quelles sont les répercussions de la pandémie de COVID-19? ». Dans notre cas, nous avons utilisé la description de la pandémie et de ses conséquences sur Wikipédia.
2. **Conversion et récupération des données vectorielles** : Après la formulation de la requête, celle-ci est convertie par le modèle d'intégration en une représentation vectorielle adaptée au traitement informatique. L'algorithme de recherche utilise la capacité de recherche de similarité de la base de données de vecteurs pour identifier les pages du document qui correspondent étroitement au sujet de la COVID-19 et aux effets de la pandémie.
3. **Réponses générées par le GML - Résumés par page** : Lorsque les pages pertinentes sont récupérées, elles sont intégrées à une nouvelle requête et fournies en tant qu'entrée au GML. L'objectif du GML est de générer un résumé des répercussions de la COVID-19 dans la page consultée.
4. **Résumé et agrégation** : Tous les résumés de page, ainsi qu'une nouvelle requête, sont intégrés sous forme de vecteurs. Ces intégrations sont ensuite introduites dans le GML pour produire un résumé global

- des répercussions de la pandémie de COVID-19 sur l'organisation. Le résumé est organisé par thème (notamment, les perturbations opérationnelles, les impacts financiers, les impacts sur les taux de collecte et de recyclage), chaque thème ayant plusieurs puces, avec des références au numéro de page où l'information peut être trouvée.
5. **Stockage et documentation** : Les deux résumés, le résumé des répercussions par page et le résumé global, sont compilés et stockés dans un document Word afin d'en faciliter l'accès et la consultation. L'analyste peut ainsi facilement se référer au résumé global, à une page mentionnée dans un sujet et, si nécessaire, consulter le document source.
 6. **Itération** : Les étapes susmentionnées sont répétées de manière itérative pour tous les rapports produits après 2020 dans l'ensemble de données afin de garantir une couverture complète et une synthèse précise des répercussions de la pandémie de COVID-19.

Figure 2.3.3-1

Représentation visuelle du pipeline de la génération de résumés Il trouve le contenu lié à la COVID-19 dans le document et génère une synthèse des répercussions.



2.3.4 Développement itératif du pipeline

Afin de trouver les bonnes requêtes pour orienter le GML vers la production de rapports utilisables par l'analyste, de nombreux échanges avec les experts en la matière se sont avérés nécessaires. Étant donné qu'il n'existe pas de mesure capable d'indiquer automatiquement au scientifique de données quand un résumé est utile, il a fallu procéder à un examen approfondi des résultats du modèle et obtenir une rétroaction à leur sujet. Ensuite, pour résoudre les problèmes soulevés par l'analyste, le scientifique de données a dû essayer différentes techniques d'ingénierie de requêtes et variations aux différentes étapes du pipeline. Ce processus s'est poursuivi jusqu'à ce que les analystes soient convaincus que les rapports constitueront une source d'information utile pour leurs recherches.

2.3.5 Synthèse de la pandémie de COVID-19 : Modèles génératifs

Le modèle GPT 3.5 d'Azure Open-AI ainsi que le modèle en source libre Llama2-13B hébergé sur nos propres serveurs ont été testés pour ce pipeline. Les experts en la matière ont dû choisir entre ces deux modèles, en se basant sur leur évaluation de la qualité et de l'exhaustivité des résumés. Le modèle Llama-2 a été choisi, car il a permis de mieux résumer les rapports et de les organiser par thèmes. L'équipe a ensuite poursuivi ses efforts d'ingénierie de requêtes pour améliorer la qualité des résumés.

Le modèle en source libre s'est avéré tout à fait capable d'exécuter une tâche telle que la synthèse, mais l'utilisation de Llama-2 a posé des problèmes distincts par rapport à la solution d'Open-AI :

- Fenêtre contextuelle plus petite : avec seulement 2 000 jetons textuels, la quantité de renseignements que le pipeline peut traiter est limitée. Les documents comportant de nombreuses pages semblent donner lieu à des résumés moins précis.
- Techniques de requête spécifiques : ces modèles sont plus sensibles aux petites modifications de la requête, des éléments tels que la ponctuation ou l'utilisation de majuscules améliorant les résultats. Dans l'ensemble, il semble plus difficile d'amener le modèle à se conformer par rapport à GPT3.5, en particulier lorsque la requête devient plus longue en raison de la quantité de renseignements contenus dans le document PDF.
- Temps d'inférence : le modèle requiert nettement plus de temps pour produire des résultats, ce qui augmente le temps nécessaire pour trouver des requêtes adéquates.

3. Résultats

3.1 Évaluer la qualité de la classification sectorielle à l'aide d'un rappel

La classification sectorielle est un problème de classification multiétiquettes avec cinq classes possibles : résidentiel, commercial, institutionnel, industriel et construction. Une organisation peut recueillir des matières plastiques auprès de clients dans n'importe laquelle de ces catégories, les renseignements nécessaires étant répartis dans les rapports.

Pour évaluer la qualité des résultats, l'expert en la matière avait déjà étiqueté une centaine de rapports, en identifiant les secteurs manuellement. Pour évaluer la qualité du classificateur, nous utiliserions généralement une mesure telle que la moyenne pondérée des scores F1 pour évaluer le rendement d'un modèle. Toutefois, dans notre cas, les performances sur l'ensemble étiqueté semblaient assez médiocres. Après enquête, nous avons réalisé que le rappel était très élevé (92 %), alors que la précision était assez faible. En fait, le modèle prédisait des classes supplémentaires par rapport au traitement manuel.

Nous avons demandé à l'expert en la matière d'analyser un échantillon d'une douzaine de prédictions, en utilisant les justifications fournies par le modèle ainsi que les pages identifiées pour vérifier ses résultats. Dans pratiquement tous les cas, après avoir vérifié la source, l'expert a approuvé le modèle, reconnaissant que les étiquettes n'avaient pas été identifiées auparavant, mais qu'elles devaient être incluses.

Cela signifie que le travail d'étiquetage manuel a omis un nombre important de vrais positifs dans les documents. Étant donné que le client ne disposait pas de ressources supplémentaires pour vérifier manuellement un échantillon suffisamment important pour calculer un score F1 significatif, nous avons utilisé le rappel comme mesure de la qualité de la classification. Cette mesure permet d'évaluer le nombre de fois où le modèle a identifié les étiquettes que l'expert a également identifiées.

3.2 Indicateurs de rendement clés

Les pipelines modernisés et leurs résultats ont permis d'obtenir les résultats suivants :

- **Efficacité améliorée** dans l'analyse de rapports sur les répercussions de la pandémie de COVID-19 :
 - Localisation du contenu pertinent
 - Synthèse des incidences par thème
- **Amélioration de l'efficacité** dans l'identification des secteurs d'intérêt dans les rapports
- **Augmentation du nombre de rapports** pouvant être analysés par rapport à l'approche manuelle
- Production de **résumés aux fins de référence rapide** pendant le travail de validation itératif
- **Identification de nouvelles étiquettes sectorielles qui n'ont pas été relevées** lors de l'étiquetage manuel

Tableau 3.2-1
Aperçu des IRC et des mesures

IRC et mesures	
Diminution du temps de traitement – localisation des répercussions de la pandémie de COVID-19	70 %
	(économie de 14 à 32 jours)
Diminution du temps de traitement – identification sectorielle	97 %
	(économie d'environ 40 jours)
Rappel de la classification sectorielle	92 %
Nombre de fichiers supplémentaires examinés	x 3

Comme le montre le tableau 3.2-1, nous avons choisi le rappel comme mesure, car il était important de déterminer toutes les instances positives possibles, même s'il fallait pour cela inclure de faux positifs. Les résultats produits sur la base d'un rappel élevé peuvent ensuite être examinés plus en détail.

En plus de ces mesures, les analystes ont souligné la valeur ajoutée que ces rapports apporteront aux projets futurs. Ils disposeront désormais d'une référence leur permettant de comprendre rapidement un document PDF entier et de déterminer quels documents contiennent des renseignements importants, puis d'accéder directement aux pages pertinentes dans ces documents.

4. Conclusion

Le projet a atteint ses objectifs de création de gains d'efficacité par la réduction du travail manuel et d'amélioration de la qualité des aperçus analytiques. Il a démontré que les GML peuvent être utilisés efficacement pour des tâches de classification et de synthèse de textes, facilitant ainsi l'utilisation de données alternatives dans le CFPMP.

Bibliographie

Hou, H., Ma, F., Bai, B., Zhu, X., et Yu, F. (2024), "Enhancing and accelerating large language models via instruction-aware contextual compression", arXiv preprint arXiv:2408.15491.

Lo, K., Chang, J. C., Head, A., Bragg, J., Zhang, A. X., Trier, C., Anastasiades, C., August, T., Authur, R., Bragg, D., Bransom, E., Cachola, I., Candra, S., Chandrasekhar, Y., Chen, Y.-S., Cheng, E. Y.-Y., Chou, Y., Downey, D., Evans, R., Fok, R., Hu, F., Huff, R., Kang, D., Kim, T. S., Kinney, R., Kittur, A., Kang, H., Klevak, E., Kuehl, B., Langan, M., Latzke, M., Lochner, J., MacMillan, K., Marsh, E., Murray, T., Naik,

A., Nguyen, N.-U., Palani, S., Park, S., Paulic, C., Rachatasumrit, N., Rao, S., Sayre, P., Shen, Z., Siangliulue, P., Soldaini, L., Tran, H., van Zuylen, M., Wang, L. L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., Hearst, M. A., et Weld, D. S. (2023), “The Semantic Reader project: Augmenting scholarly documents through AI-powered interactive reading interfaces”, arXiv preprint arXiv:2303.14334.

Verma, S. (2024), “Contextual Compression in Retrieval-Augmented Generation for Large Language Models: A Survey”, Published in arXiv: 2409.13385 [cs.CL], <https://arxiv.org/abs/2409.13385>