Approche non-supervisée pour la création d'un Référentiel Sémantique

Lydia Khelifa Chibout [1], Manuele Kirsch Pinheiro[2]

1. Centre Scientifique et Technique du Bâtiment (CSTB)
Lydia.CHIBOUT@cstb.fr
2. Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
Manuele.Kirsch-Pinheiro@univ-paris1.fr

RESUME. Les organisations font face à des défis sans précédent en matière de gestion et de structuration de leurs connaissances. La capacité à extraire, organiser et utiliser des informations pertinentes à partir de vastes collections de documents est devenue un facteur clé pour l'efficacité opérationnelle et la prise de décision éclairée. Cependant, l'identification des sources de connaissances nécessaires et la construction de bases de connaissances appropriées représentent une tâche ardue et chronophage. Cet article aborde ces défis en exploitant des techniques de traitement du langage naturel (NLP) et des modèles de langage de grande taille (LLM), afin de faciliter la création et l'enrichissement de vocabulaires spécialisés à des fins de gestion des connaissances. Nous explorons l'application de techniques non supervisées de clustering combinées à l'extraction de mots-clés avec des techniques de NLP pour l'aide à la construction de vocabulaires spécialisés répondant à la nature multidisciplinaire du CSTB, centre de recherche scientifique français spécialisé dans le bâtiment. Nous détaillons ici l'approche proposée, les résultats de nos expérimentations au CSTB, ainsi que le processus de validation humaine utilisé pour évaluer ces résultats.
Mots-clés : Extraction de mots-clés, clustering, identification du vocabulaire, construction basée sur les connaissances, gestion des connaissances.

ABSTRACT. The exponential growth of digital information has exposed organizations to unprecedented challenges in managing and structuring their knowledge repositories. The ability to extract, organize, and use relevant information from large collections of documents has become a critical factor for operational efficiency and informed decision-making. However, identifying necessary knowledge sources and building appropriate knowledge bases represents a cumbersome and time-consuming task. In this paper, we address these challenges by leveraging advanced Natural Language Processing (NLP) techniques and Large Language Models (LLMs), to facilitate the creation and enrichment of vocabularies for knowledge management purposes. We explore the application of clustering techniques combined with NLP-driven keyword extraction to support the construction of specialized vocabularies that address the multidisciplinary nature of the content at CSTB, a French scientific research center specialized on buildings. We provide a detailed overview of the proposed approach, present the results of our experiments, and describe the human validation process used to evaluate these results.
KEYWORDS: keywords extraction, clustering, vocabulary identification, knowledge-based construction, knowledge management

1. Introduction

La gestion de connaissances est devenue au fil des années un élément clé pour les organisations. Celle-ci, comme l'ensemble de notre société, est caractérisée par une certaine infobésité, avec la production toujours croissante de documents et des données. Dans ce contexte, la construction de bases documentaires efficaces au sein des organisations est devenue une tâche cruciale pour une bonne gestion de connaissances. En mettant en place des systèmes bien définis pour stocker, récupérer et indexer les documents, les organisations peuvent s'assurer que les informations pertinentes sont facilement disponibles pour ceux qui en ont besoin, permettant ainsi une prise de décision rapide et éclairée (Maharjan, 2020), (Morse, 2000). Cet accès simplifié à l'information facilite le partage des connaissances et la collaboration, favorisant un environnement dynamique propice à l'amélioration continue des processus (Maharjan, 2020), (Narazaki et al, 2020). Ceci est d'autant plus vrai dans un environnement multidisciplinaire, comme le CSTB, Centre Scientifique spécialisé dans le bâtiment, dont les projets couvrent des domaines multiples, tels que les sciences de l'environnement, la biodiversité, l'acoustique, les matériaux de construction et la santé dans les bâtiments.

Cependant, identifier les sources de ces connaissances nécessaires représente une tâche ardue et chronophage. Or la capacité à identifier facilement les informations pertinentes permet de prendre de meilleures décisions, de résoudre les problèmes plus efficacement et de contribuer plus efficacement aux objectifs organisationnels. Un vocabulaire bien structuré facilite non seulement la récupération des documents pertinents, mais améliore également l'efficacité globale des processus de gestion de l'information et de prise de décision. L'extraction de mots-clés est une étape clé de ce processus, car elle permet l'identification automatique des termes essentiels qui encapsulent le contenu principal des documents et aident les lecteurs à comprendre rapidement l'idée du contenu. En extrayant des mots-clés de plusieurs documents, il devient possible d'identifier des termes partagés entre plusieurs disciplines. Ces mots-clés communs agissent comme des ponts entre des domaines distincts, mettant en évidence des zones de chevauchement conceptuel ou d'intérêts partagés. Ce processus favorise la collaboration interdisciplinaire en fournissant un point de départ clair pour des projets et discussions conjoints. Par exemple, au sein du CSTB, l'identification de termes partagés tels que « durabilité » ou « modélisation des données » pourrait connecter les équipes de recherche en sciences de l'environnement et en informatique, permettant le développement de solutions innovantes interdisciplinaires. Cette approche rationalise l'intégration des connaissances et encourage la synergie entre les équipes multidisciplinaires.

Ces mots-clés directement liés au contexte du document constituent ainsi un référentiel sémantique. Pour y parvenir, les techniques d'extraction de mots-clés non supervisées ont attiré l'attention, en particulier dans les contextes où les ensembles de données étiquetés sont indisponibles ou impraticables à créer. Les récents progrès en traitement du langage naturel (NLP), y compris le développement de grands modèles de langage (LLM) tels que BERT (Devlin et al, 2019) et GPT (Brown et al, 2020), offrent des opportunités prometteuses pour améliorer la qualité et la précision

de l'extraction de mots-clés. Ces modèles, grâce à leur compréhension contextuelle approfondie, peuvent capturer des relations nuancées entre les mots et les concepts dans le texte, fournissant une base solide pour les approches non supervisées. Des études telles que celles (Mihalcea et Tarau, 2004) ou encore (Wan et Xiao, 2008) ont établi la base de l'extraction de mots-clés non supervisée. Plus récemment, BERTopic (Grootendorst, 2022) a émergé comme un outil efficace pour analyser et résumer de grands corpus, les rendant hautement pertinents pour la construction de référentiels sémantiques dans des environnements riches en connaissances.

Dans cet article, nous combinons les méthodes non supervisées avec des LLMs pour extraire des mots-clés d'un corpus documentaire, dans une démarche cohérente contribuant à la construction d'un vocabulaire spécialisé pour soutenir la récupération de documents et le partage des connaissances. L'objectif est de rendre la recherche d'informations plus précises et pertinentes. Le regroupement de documents est crucial pour les chercheurs engagés dans des recherches interdisciplinaires sur divers sujets. Notre approche propose l'extraction de mots-clés non-supervisée après le regroupement de documents textuels, ce qui améliore considérablement la découverte d'informations utiles et contribue à la compréhension et à la recherche d'information par les utilisateurs. Nous illustrons notre approche sur un ensemble de documents bilingues et multidisciplinaires du CSTB, dont les mots-clés identifiés ont été soumis à un panel d'experts du domaine à des fins d'évaluation.

Cet article est structuré comme suit : la section 2 présente les travaux connexes sur les méthodes d'extraction de mots-clés supervisées et non supervisées. La section 3 détaille nos approches proposées, suivie de la présentation des expérimentations réalisées au CSTB (section 4). La section 5 discute des résultats et du processus d'évaluation et de validation. Enfin, la section 6 conclut l'article.

2. Etat de l'art

L'organisation d'une base documentaire efficace présente de nombreux avantages aux organisations, quel que soit leur secteur d'activité (Laihonen et al., 2023 ; Jain, 2012 ; Yao-Sheng, 2007). Dans chacun de ces cas, un accès efficace aux connaissances explicites s'est traduit directement par une amélioration des performances, une innovation accrue et une efficacité globale augmentée. Pour cela, l'identification des mots-clés pertinentes représente une étape clé pour la construction de ces bases. Différents travaux dans la littérature traitent l'identification et l'extraction de mots-clés citons par exemples les méthodes traditionnelles qui reposent sur des analyses statistiques, telles que TF-IDF (Salton et Buckley, 1988), et linguistiques, comme la lemmatisation et l'extraction de syntagmes nominaux (Delamaire et al, 2019). Toutefois, elles peinent souvent à capturer la richesse sémantique des textes complexes. Des algorithmes non supervisés, tels que TopicRank (Bougouin et al., 2013), introduisent une approche par graphe pour structurer les mots-clés autour de thèmes cohérents ou encore Khelifa et al. (2012) qui proposent de structurer les mots en graphe de topics en gardant leurs contextualisations dans le texte à travers les dimensions sémantiques

telles que la région, le temps, la discipline/le domaine ou encore la langue. Abilhoa et De Castro (2014) proposent une méthode d'extraction de mots-clés pour les collections de tweets qui représente les textes sous forme de graphes et applique des mesures de centralité pour trouver les sommets pertinents (mots-clés). Hasan et al. (2018) proposent un système qui extrait un nombre spécifique de termes clés des documents pour identifier le contenu principal d'un texte. Les données sont collectées à partir de différentes sources telles que des livres et des journaux. Diverses techniques bien connues de l'apprentissage automatique, comme le SVM, la régression logistique ou l'arbre PAT-tree, ont été utilisées pour extraire les mots-clés. Bisht (2022) quant à lui a évalué différentes méthodes d'extraction de mots-clés basées sur la distribution spatiale et a proposé une mesure basée sur la fréquence, la fréquence inverse des documents, la variance et l'entropie de Tsallis, dont les résultats ont mis en évidence le fait qu'il n'existe pas de méthode parfaite. Ahadh et al. (2021) ont proposé une approche automatisée, semi-supervisée et indépendante du domaine pour analyser les rapports d'accidents. Étant donné un ensemble de sujets de classification définis par l'utilisateur et la littérature du domaine telle que des manuels, des glossaires et des articles Wikipédia, la méthode peut identifier des mots-clés spécifiques au domaine et les regrouper en sujets avec une implication minimale d'experts. Ces mots-clés et sujets peuvent ensuite être utilisés à diverses fins de fouille de données, y compris la classification. Cependant, ces méthodes nécessitent un nombre élevé de documents étiquetés comme exemples d'entraînement. Les approches plus récentes ont exploré l'utilisation des techniques d'apprentissage profond pour améliorer la précision et l'efficacité de l'extraction de mots-clés, démontrant le potentiel des réseaux neuronaux à apprendre des motifs complexes dans les textes et à identifier les mots-clés pertinents (Umair et al., 2024).

Par ailleurs, le clustering de documents, avec des techniques telles que K-means ou DBSCAN (Ester et al., 1996), permet d'améliorer la contextualisation des mots-clés en regroupant des documents similaires, mais l'évaluation de la qualité du clustering repose uniquement sur des indices comme le Silhouette Score (Rousseeuw, 1987). Les grands modèles de langage (LLMs) tels que BERT (Devlin et al., 2019) ou GPT-4 (OpenAI, 2023) offrent de grandes capacités pour comprendre le contexte sémantique profond, améliorant significativement la pertinence et la diversité des mots-clés extraits (Liu et al., 2019). Leur usage en classification multi-label (Tsoumakas & Katakis, 2007) optimise également la cohérence des résultats. Zhou et al. (2023) ont expérimenté l'utilisation de ChatGPT pour l'extraction de mots-clés, obtenant un ensemble représentatif et opérationnel pour la recherche scientifique.

On observe dans la littérature une tendance vers des approches supervisées, avec des mots-clés qui ne sont pas toujours regroupés en topic, et une confrontation au regard d'experts humains qui n'est pas toujours mise en avant. Or cette évaluation par des experts nous semble essentielle, notamment dans le cadre d'environnement multidisciplinaires, comme le CSTB. Combiner clustering et LLMs optimise la pertinence des mots-clés et réduit le bruit dans les données. Cette approche est particulièrement efficace pour des tâches telles que l'indexation intelligente, la

recherche d'information (Manning et al., 2008), la veille technologique et la recommandation de contenu personnalisée.

Dans cet article, nous proposons une approche non-supervisée permettant l'analyse d'un large volume de documents non étiqueté, combinant les approches de clustering de documents et d'extraction de mots de clés avec les LLM dont les résultats ont été présentés à un panel d'experts pour évaluation et validation.

3. Contribution

Dans cet article, nous proposons d'utiliser des techniques de traitement du langage naturel (TAL) pour extraire des mots-clés significatifs à partir de documents textuels et de les regrouper en topic. Le processus suit une chaîne structurée conçue pour optimiser les tâches de modélisation de sujets et de regroupement. Cette approche proposée représente une méthode non supervisée et modulaire capable de gérer un large corpus de documents multidisciplinaires non étiquetés. Un aperçu de ces différentes étapes du pipeline est présenté dans la Figure 1 :

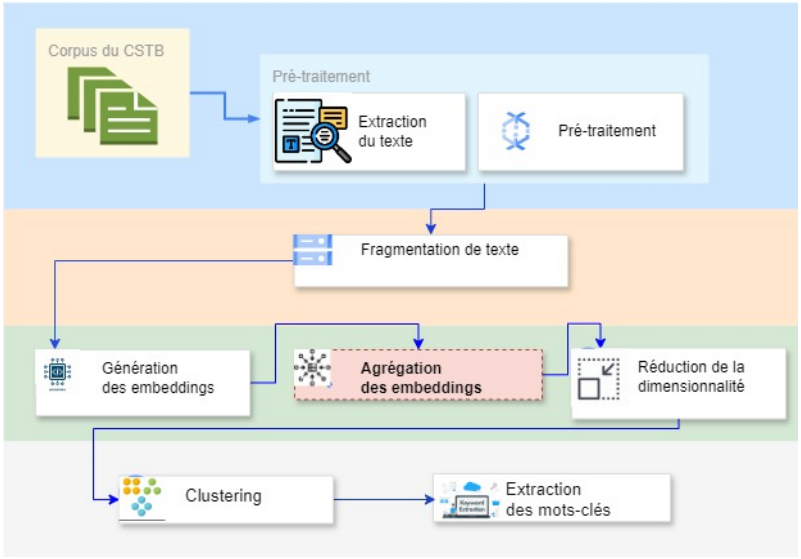

Figure 1. Approches de clustering de documents et de chunks

1. Extraction du texte et pré-traitement du corpus documentaire ;
2. Fragmentation de texte (*text chunking*) : les documents d'entrée sont divisés en segments plus petits et cohérents appelés *chunks* pour faciliter leur représentation et de permettre une analyse plus précise ;

3. Génération *d'embeddings* et réduction de la dimensionnalité : des modèles basés sur des *transformers* sont utilisés pour convertir les segments de texte (*chunks*) en vecteurs qui encapsulent la signification sémantique et pour réduire la dimensionnalité des *embeddings* tout en préservant leur structure ;
4. L'agrégation des *embeddings* est une étape obligatoire pour l'approche de clustering de documents. Eependant elle n'est pas utilisée dans le clustering par *chunks*. En effet, deux approches ont été envisagées pour le clustering, une approche uniquement basée sur les documents et une basée sur les *chunks*, offrant toutes les deux des perspectives différentes sur les documents analysés ;
5. Réduction de dimensionalité permet de réduire le nombre de dimensions (variables) tout en gardant les caractéristiques principales de chaque *embedding* ;
6. Clustering : les *embeddings* dont la dimensionalité a été réduite sont regroupés, pour identifier efficacement les régions denses et séparer le bruit ;
7. Extraction de mots-clés extrait des termes/mots clés et pour améliorer l'interprétabilité des sujets à partir des clusters en utilisant des techniques telles que c-TF-IDF et CountVectorizer.

Ces différentes étapes sont expliquées dans ce qui suit. L'ensemble de l'approche a été expérimentée sur une grande base de documents scientifiques et techniques du CSTB concernant le bâtiment, le génie civil, la qualité de l'air dans les bâtiments, les expérimentations acoustiques et d'autres domaines liés à la construction. Ces documents sont disponibles pour les chercheurs et les experts du CSTB, qui ont pu évaluer les mots-clés suggérés par l'approche proposée.

3.1. Extraction de texte et pré-traitement du corpus

La phase de prétraitement est une étape nécessaire pour n'importe quel processus d'analyse de données. Dans notre cas, celle-ci se traduit par l'extraction de texte à partir de documents PDF et leur prétraitement. Le format PDF est largement utilisé pour le partage de documents en raison de la présentation cohérente qu'il offre sur différents appareils. Cependant ce format présente des défis pour l'extraction automatisée de texte. En effet, les documents PDF ne sont pas conçus pour la manipulation de texte, et le contenu peut varier considérablement en complexité, incluant du texte brut, des tableaux, des images, et même des documents scannés contenant du texte sous forme d'images. Le prétraitement du texte est ainsi effectué afin de s'assurer que les données sont propres et prêtes pour une analyse ultérieure. Ce prétraitement consiste d'abord au *nettoyage* et à la *normalisation* du texte. Les caractères spéciaux tels que les guillemets typographiques et les apostrophes sont remplacés par leurs équivalents plus simples (par exemple, les guillemets courbes remplacés par des guillemets droits). De plus, les caractères de nouvelle ligne ***\n*** sont supprimés, et les espaces excessifs sont gérés en séparant et en concaténant les mots. Ensuite, tous les caractères numériques sont supprimés à l'aide d'expressions régulières. Par ailleurs, afin de réduire le bruit du texte et de focaliser l'analyse sur les éléments potentiellement significatifs, la *lemmatisation* et l'*étiquetage des parties*

du discours (POS) sont appliqués directement aux mots-clés et à l'extraction de bi-grammes. La *lemmatisation* permet de réduire les mots à leur forme de base ou racine. Par exemple, "running" devient "run" et "better" devient "good". Cette étape est cruciale pour s'assurer que les différentes formes d'un mot sont traitées pareillement lors de l'analyse. L'étiquetage des parties du discours, quant à elle, est utilisée pour identifier les verbes dans les mots-clés extraits, qui sont ensuite exclus pour se concentrer uniquement sur les noms et les adjectifs qui sont plus susceptibles de contribuer à la modélisation des sujets et à l'extraction des termes clés. Enfin, l'extraction de bi-grammes est réalisée.

3.2. Processus de segmentation du document

La segmentation du document correspond au processus de division d'un grand corpus de texte en segments plus petits et plus gérables, communément appelés *chunks*. L'objectif principal est de s'assurer que chaque fragment reste dans les limites d'entrée d'un modèle (par exemple, les limites de *tokens* dans les modèles de langage), tout en préservant suffisamment de contexte pour maintenir sa signification. La méthode utilisée ici est la division de texte basée sur les caractères, dans laquelle un long texte est divisé en fragments en fonction d'un nombre spécifique de caractères. Afin d'améliorer la continuité entre les fragments, un chevauchement de texte (*text overlapping*) a été introduit. Celui-ci implique d'inclure une petite portion du fragment précédent au début du fragment suivant, permettant ainsi une meilleure rétention du contexte à travers les fragments.

3.2 Vectorisation et réduction de dimensionnalité

La vectorisation, ou encore *embedding,* est le processus de conversion du texte en vecteurs numériques qui capturent sa signification sémantique. Ces vecteurs forment les modèles sur lesquels les techniques de clustering sont appliquées. Les méthodes principales pour créer des *embeddings* sont le Word2Vec (Churh, 2017), BERT (Feng et al, 2020) et Sentence-BERT. Word2Vec, développé par Google, utilise deux approches principales : le Continuous Bag of Words (CBOW), qui prédit un mot basé sur son contexte, et le Skip-gram, qui prédit les mots de contexte à partir d'un mot donné. Ces modèles sont entraînés sur de grands corpus de texte, produisant des *embeddings* qui capturent les relations entre les mots. BERT (Bidirectional Encoder Representations from Transformers) est un modèle basé sur les transformateurs qui génère des *embeddings* sensibles au contexte. Pour les tâches nécessitant des *embeddings* au niveau des phrases ou des documents, des méthodes comme Sentence-BERT (SBERT) ou Doc2Vec sont utilisées. SBERT affine BERT pour produire des *embeddings* de phrases sémantiquement significatives, comparables par similarité cosinus. Ces techniques ont diverses applications en traitement du langage naturel, dont la recherche de similarité, le clustering de documents, et la récupération d'information.

La réduction de dimensionnalité est un processus clé en apprentissage automatique, visant à réduire le nombre de caractéristiques (ou dimensions) d'un jeu

de données tout en conservant un maximum d'informations pertinentes (Anowar et al., 2021). Plusieurs algorithmes sont couramment utilisés à cet effet, dont UMAP (Uniform Manifold Approximation and Projection), particulièrement adapté à la visualisation de données à haute dimension. Il préserve davantage la structure globale des données par rapport à t-SNE, ce qui le rend efficace aussi bien pour la visualisation que pour l'apprentissage du manifold (Leland et al., 2018). T-SNE (t-Distributed Stochastic Neighbor Embedding) est une autre technique non linéaire très populaire pour visualiser des données complexes. Elle repose sur la conversion des similarités entre les données en probabilités conjointes et cherche à minimiser la divergence de Kullback-Leibler entre les espaces de haute et de basse dimension (Van Der Matteen and Hinton, 2018). Enfin, le PCA (Principal Component Analysis) est une méthode linéaire de réduction de dimensionnalité, qui projette les données sur les directions maximisant la variance. Elle est largement utilisée pour réduire la complexité des ensembles de données tout en préservant le plus possible la variabilité (Abdi et Williams, 2010). Les résultats de ces processus sont fortement influencés par la langue des documents. Mélanger des documents en différentes langues peut avoir un impact négatif sur ces résultats. Cependant, les documents dans différentes langues sont, à nous jours, couramment présents dans les organisations, ce qui doit être pris en compte lors des étapes d'*embedding* et de la réduction de la dimensionnalité. Ainsi, lors des expérimentations réalisées au CSTB, l'ensemble de documents utilisé a pu être divisé en deux sous-ensembles, composés respectivement de documents en anglais et en français. Un modèle distinct adapté à chaque langue a été utilisé pour chacun des deux sous-ensembles.

3.3 Les approches de clustering

Au cours de la phase de *clustering*, nous avons évalué le type de clustering le plus approprié, en tenant compte du fait que les documents sont multidisciplinaires et qu'un seul article peut aborder des questions transversales couvrant plusieurs domaines. Afin d'identifier l'approche optimale à notre expérimentation, nous avons exploré deux méthodes de clustering : une basée sur les documents et une autre basée sur les *chunks*. Ces deux méthodes ont été testées pour déterminer la stratégie la plus efficace pour l'extraction de mots-clés à intégrer dans le vocabulaire.

3.3.1 Approche de clustering de document avec les LLM

L'approche de clustering par document met l'accent sur le clustering de documents entiers plutôt que de segments individuels. Pour ce faire, une méthode d'agrégation est utilisée pour fusionner les *embeddings* des segments appartenant à un même document en une représentation unifiée. Cet *embedding* agrégé est ensuite utilisé pour le regroupement, permettant d'identifier les similitudes entre les documents dans leur ensemble. Différentes méthodes d'agrégation peuvent être employées pour combiner les *embeddings* individuels des chunks en une seule représentation pour chaque document, telles que l'agrégation par moyenne ou celle par somme. La méthode d'agrégation par moyenne calcule la moyenne des *embeddings* de tous les chunks au sein d'un document. Chaque *embedding* de chunks

est un vecteur de grande dimension, et l'agrégation par moyenne crée un vecteur unique en moyennant les *embeddings*. Cette approche aide à créer une représentation équilibrée du document, dans laquelle chaque chunks contribue de manière égale à *l'embedding* final du document. Par ailleurs, l'agrégation par la somme combine les *embeddings* en additionnant les vecteurs de tous les segments au sein d'un document. Les deux méthodes permettent d'agréger les *embeddings* au niveau des segments en une représentation au niveau du document qui capture la signification globale de celui-ci.

3.3.2 Approche de clustering de chunks

L'approche de clustering par *chunk* consiste à regrouper de plus petits segments (*chunks*) d'un document plutôt que le document entier. En divisant le document en fragments et en les regroupant individuellement, cette méthode capture des motifs plus localisés et des nuances thématiques au sein du texte.

3.4 Extraction des mots-clés et des topics

Pour construire un vocabulaire spécialisé dans une organisation multidisciplinaire telle que le CSTB, nous avons besoin de mots-clés qui caractérisent efficacement les documents, facilitant ainsi l'accès aux connaissances qu'ils contiennent. Cette extraction de mots-clés et de sujets (topics) se déroule en deux étapes :

- **Vectorisation** : le texte de chaque document est transformé en une matrice de fréquence des termes (par document). Chaque entrée de cette matrice indique combien de fois un mot spécifique apparaît dans chaque document ;
- **Calcul du c-TF-IDF** (ou TF-IDF basé sur les classes) : une fois les clusters formés, le c-TF-IDF (Xu & Wu, 2014) est calculé pour chaque cluster. Cela fournit une pondération des termes basée sur leur importance relative au sein de ce cluster.

4. Expérimentation

Nous avons testé notre approche sur un large corpus de documents multidisciplinaires non étiquetés qui a été divisé en deux parties en fonction de la langue (français et anglais). Des modèles spécifiques pour les *embeddings* et les techniques de traitement du langage naturel ont été appliqués à chaque partie. L'objectif de cette expérimentation est triple : i) effectuer un clustering non supervisé de documents et de chunks ; ii) extraire des mots-clés en suivant les deux approches de clustering ; et iii) évaluer les résultats pour identifier la meilleure approche, les mots-clés et les sujets pour la construction du vocabulaire du CSTB.

4.1 Description du corpus documentaire

La plupart des documents disponibles au CSTB sont des rapports de recherche et d'évaluation publiés en format PDF entre 2000 et 2024, couvrant différents domaines scientifiques tels que le génie civil, la sécurité incendie, la santé et la qualité de l'air dans les bâtiments. Le choix de la période et des documents réside dans le fait que les documentalistes chargés de leur indexation au CSTB, et qui sont mobilisés dans le cadre de la validation des résultats, ont une très bonne connaissance de ce corpus (ces sont eux qui nous ont transmis ce corpus).

Ce corpus contient un total de 6627 documents, dont 3279 en français, 3055 en anglais et 293 documents jugés inexploitables. Les causes empêchant l'exploitabilité de ces documents sont nombreuses : documents scannés sans traitement OCR, rendant leur contenu non consultable ; des documents sont mal encodés, causant des problèmes de lisibilité et d'accessibilité ; et des PDF complètement vides, n'offrant aucune donnée utilisable. Pour les documents en français, l'analyse a identifié 2 808 fichiers uniques et 419 groupes de doublons, où chaque groupe contient des fichiers identiques réduits à un seul représentant. De même, pour les documents en anglais, il y a 2265 fichiers uniques et 716 groupes de doublons. Cette catégorisation détaillée et l'identification de doublons améliorent l'efficacité de la gestion des documents et aident à rationaliser l'analyse ultérieure en s'assurant que les données redondantes n'encombrent pas l'ensemble de données. Les résultats soulignent l'importance du prétraitement et de l'amélioration de la qualité des documents afin de garantir une meilleure utilisabilité dans la gestion des connaissances.

4.2 4.2 Modèles et techniques utilisés

L'expérimentation a été faite sur un ordinateur équipé d'un processeur Intel i5 de 5^{e} génération et de 16 Go de RAM DDR4 combiné à l'environnement de développement virtuel gratuit Google Colab pour son type d'exécution GPU.

Pour des contraintes techniques liées à l'infrastructure, notre choix s'est porté sur le modèle d'*embedding all-MiniLM-L6-v2* (Wenhui et al., 2020) et paraphrase-multilingual-MiniLM-L12-v2 (Ciancone et al., 2024), des modèles d'*embedding* par LLM qui appartiennent à la famille des modèles MiniLM, qui sont des versions allégées de BERT. Le MiniLM est conçu pour offrir des *embeddings* de haute qualité tout en étant plus léger et plus rapide que les modèles BERT. Concernant la réduction de la dimensionnalité, nous avons opté pour *UMAP* (McInnes et al., 2018), qui est rapide et préserve à la fois les structures de données locales et globales. Le choix de cette méthode est fortement lié au modèle de clustering HDBSCAN. Notre choix s'est porté sur ce modèle car il ne nécessite pas de spécifier le nombre de clusters à l'avance et gère bien les densités variables (Malzer & Baum, 2020). De plus, connaître la hiérarchisation des sujets sera très utile pour l'intégration des résultats dans le référentiel sémantique. L'extraction de mots-clés quant à elle, a été réalisée en utilisant *CountVectorizer* combiné avec *c-TF-IDF* (Xu & Wu, 2014), mettant en avant l'importance spécifique des clusters. En considérant l'approche

basée sur le clustering, nous avons appliqué une agrégation par moyenne sur les deux langues afin de fusionner les segments de chaque document.

5. Résultats and Evaluation

5.1 Processus d'évaluation et de validation

Compte tenu de l'importance des mots-clés extraits (candidats au vocabulaire spécialisé) et du rôle central de ce référentiel dans la stratégie de gestion des connaissances du CSTB, une validation humaine des résultats a été privilégiée. Ce processus d'évaluation s'appuie sur l'expertise métier et l'expérience approfondie des intervenants, mobilisés spécifiquement pour cette tâche.

Le panel d'experts se compose de : *(i)* Un ingénieur en génie civil qui a une double casquette et qui contribue à l'indexation documentaire (15 ans d'expérience au CSTB) ; *(ii)* Une chercheuse spécialisée en santé et confort des bâtiments (40 ans au CSTB) ; *(iii)* Deux documentalistes experts (28 et 20 ans d'expérience chacun en indexation de documents techniques et gestion des contenus) et occasionnellement un veilleur en technologique de l'information (14 ans d'expérience au CSTB).

Ce comité, déjà en charge de l'animation des ateliers de construction du référentiel sémantique du CSTB, a appliqué une approche d'analyse fondée sur : *(i)* La pertinence thématique des mots-clés pour leur intégration futur dans le référentiel sémantique ; *(ii)* Leur adéquation aux enjeux multidisciplinaires de l'organisme ; *(iii)* Leur potentiel d'interopérabilité avec les systèmes existants.

Les missions de ce comité sont, en fonction des résultats, de guider le choix des meilleures méthodes de clustering et des techniques de NLP, et la nomination des sujets ou topics correspondants aux clusters résultants pour identifier les domaines scientifiques. Le processus de validation est cyclique et itératif. Il s'est organisé en quatre réunions de travail (de 2 heures chacune). Au cours de celles-ci, les résultats du *clustering* et de l'extraction ont été présentés, et des exigences/recommandations ont été établies pour les mots-clés. Parmi celles-ci, le comité a établi 3 **critères de sélection** : *(i)* Les mots-clés ne doivent pas être des verbes ; *(ii)* Ils ne doivent pas contenir de chiffres, de noms de villes ou de pays ; *(iii)* Ils doivent se composer d'un-grammes ou de bi-grammes. Le comité a également défini des indications sur la **métrique d'évaluation** : Le nombre de documents par clusters pour décider d'intégrer ou non les mots-clés candidats au référentiel sémantique, permettant ainsi d'analyser la distribution des documents par thématique.

5.2 Résultat de l'approche de clustering de documents

Documents en anglais

Dans notre expérimentation menée sur 2265 documents en anglais, nous avons obtenus six clusters distincts. Chaque topic représente un cluster, caractérisé par les mots-clés extraits et le nombre de documents qu'il contient. Comme illustré dans la

Figure 2, nous pouvons voir qu'en utilisant l'extraction de bi-grammes, le topic numéro 3 contient les bi-grammes « *wind speed* » et « *wind tunnel* » que le comité de validation identifie comme des mots-clés importants, se référant à des expériences menées avec la soufflerie Jules Verne, une installation de recherche pour les évaluations techniques au CSTB. Le comité a trouvé que les bi-grammes fournissent des informations plus significatives par rapport aux simples mots-clés. Nous pouvons voir que la plupart des documents contiennent les mots « *building, constructions* » et des mots connexes, ce qui est normal compte tenu du contexte de tous les documents.

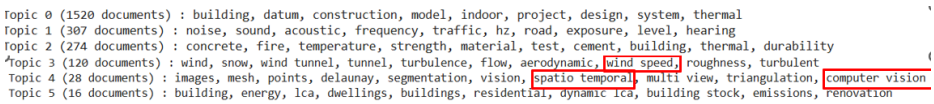

Figure 2. Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des documents

Documents en français

A partir de 2808 documents en français, nous avons obtenus six clusters distincts. Comme pour l'expérimentation sur des documents en anglais, les mots bi-gramme ont été considérés comme plus pertinents et couvrant mieux certains domaines comme « la gestion patrimoine » et le « vide sanitaire ».

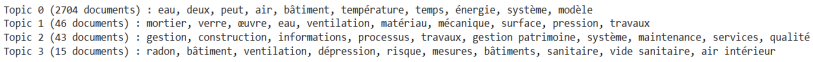

Figure 3: Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des documents

5.3 Résultats de l'approche de clustering des chunks

Documents en anglais

Le même corpus de documents en anglais a été testé avec l'approche basée sur les *chunks*. Comme illustré dans la Figure 4, cette approche offre 90 clusters et plus de mots-clés. Les domaines sont plus largement couverts et plus étendus, garantissant qu'aucun domaine abordé par le CSTB n'a été négligé. Les bi-grammes extraits fournissent également une richesse sémantique accrue, selon les experts, similaire à la première approche. Le comité a en effet considéré que ces résultats étaient plus riches et plus proches de la nature du contenu des documents du CSTB.

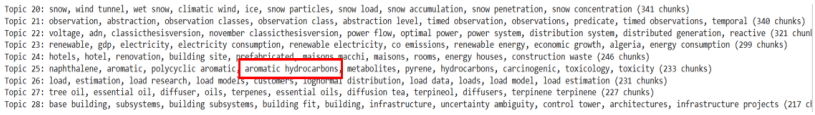

Figure 4. Résultats de l'extraction des mots bi-grammes avec l'approche de clustering des chunks

Document en français

Cependant, lors du test du corpus en français, nous avons rencontré de nombreux résultats incohérents pour les deux approches proposées, comme illustré dans la Figure 4. Par exemple, le sujet 10 dans la Figure 5 comporte plusieurs mots qui n'ont pas pu être interprétés par les experts, tels que « fissap » (« passif » en inversé), qui sont probablement une conséquence du chevauchement de texte utilisé lors de l'étape de fragmentation. Bien que d'autres sujets identifiés aient été considérés comme pertinents par les experts, ce phénomène démontre certaines limites de notre approche, indiquant qu'il reste encore des améliorations à apporter.

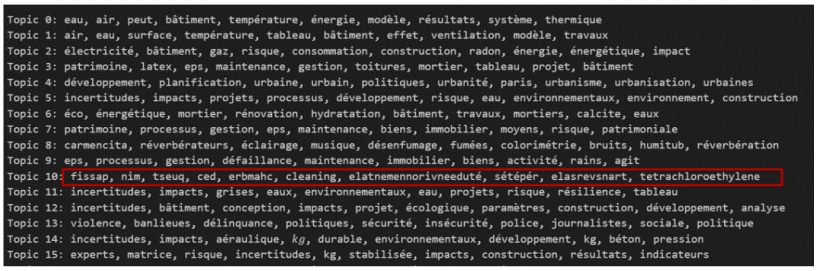

Figure 5. Résultats incohérents de l'extraction avec l'approche basée sur des chunks

5.5 Discussion

À l'issue des 4 réunions de travail, le comité a retenu les résultats de l'approche basée sur les *chunks* pour l'extraction des mots-clés, en mettant temporairement de côté le traitement des documents en français. Lors de la dernière réunion de validation, la méthode *chunk-based* a été désignée comme la solution optimale, en privilégiant l'utilisation de bi-grammes pour garantir une granularité sémantique adaptée aux besoins du projet. En effet, cette approche permet de capturer des unités de sens cohérentes (ex: « qualité de l'air », « changement climatique »), tandis que les bi-grammes évitent la sur-spécialisation des un-grammes isolés.

5 Conclusions et perspectives de recherche

Dans cet article, nous présentons les premiers résultats d'une approche non supervisée associant des grands modèles de langage (LLM) et des techniques de traitement automatique des langues (TAL). Cette combinaison permet d'extraire des mots-clés candidats pertinents, en vue de leur intégration ultérieure au sein d'un référentiel sémantique. L'approche proposée se décline en deux sous-approches de clustering : l'une fondée sur les *chunks* (segments thématiques) et l'autre sur les documents complets. L'expérimentation a été menée sur un corpus multidisciplinaire, bilingue et non étiqueté. Compte tenu de l'importance stratégique des mots-clés candidats et des directives de la politique globale de Gestion des Connaissances au CSTB, une validation humaine des résultats a été privilégiée. Le comité d'évaluation, composé d'experts aux spécialités, profils et ancienneté variés au sein du CSTB, a participé à quatre réunions. Lors de ces sessions, les exigences ont été définies, conduisant à un affinement des techniques de TAL et à une révision itérative des résultats pour en assurer la qualité. Les conclusions du comité soulignent que les résultats les plus performants proviennent de l'approche par *chunks*, offrant une couverture thématique tout en renforçant la cohérence des clusters. Alignés sur ses enjeux opérationnels, ls travaux futurs s'articuleront autour de trois axes :

1. L'optimisation du traitement des documents en français via des modèles linguistiques dédiés (ex. : CamemBERT) ;
2. La résolution des anomalies lexicales résiduelles (ex. : le terme « fissap », probable coquille pour « fissure ») ;
3. L'amélioration de la génération de mots-clés par intégration du *Maximum Marginal Relevance* (MMR), afin d'équilibrer pertinence et diversité. Enfin, l'attribution de libellés thématiques aux clusters (« Acoustique », « Énergétique »…) par le comité positionne ce référentiel comme un outil stratégique pour l'ingénierie multidisciplinaire, consolidant l'accès aux savoirs techniques du CSTB ;
4. Enrichissement du référentiel sémantique du CSTB avec ces mots clés extraits.

Bibliographie

Abilhoa W.D., De Castro L.N. (2014). TKG: A graph-based approach to extract keywords from tweets. *Distributed computing and artificial intelligence, 11th International Conference*, Springer International Publishing, p. 425-432.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459.

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP,

LE, ICA, t-SNE). Computer Science Review, 40, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Ahadh A., Binish G.V., Srinivasan R. (2021). Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process Safety and Environmental Protection*, 155 (Nov. 2021), 455-465. https://doi.org/10.1016/j.psep.2021.09.022.

Bisht R.K. (2022). A Comparative Evaluation of Different Keyword Extraction Techniques. *International Journal of Information Retrieval Research (IJIRR)*, 12(1), 1-17.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A. et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Carbonell J., Goldstein J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335-336. ACM, Melbourne, Australia.

Ciancone M., Kerboua I., Schaeffer M., Siblini W. (2024). MTEB-French: Resources for French Sentence *Embedding* Evaluation and Analysis. *arXiv*, arXiv:2405.20468. Disponible sur : https://arxiv.org/abs/2405.20468.

Church Kw. Word2Vec. *Natural Language Engineering*. 2017; 23(1):155-162.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, 2-7 juin 2019.

Delamaire, A., Beigbeder, M., & Juganaru-Mathieu, M. (2019, May). Exploitation de syntagmes dans la découverte de thèmes. Actes de la conférence CORIA (Conférence en Recherche d'Information et Applications).

Grootendorst M. (2022). BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling. *arXiv*, arXiv:2203.05794. Disponible sur : https://arxiv.org/abs/2203.05794.

Hasan H.M., Sanyal F., Chaki D. (2018). A novel approach to extract important keywords from documents applying latent semantic analysis. *2018 10th International Conference on Knowledge and Smart Technology (KST)*, pp. 117-122. IEEE, Chiang Mai, Thaïlande.

Khelifa LN., Lammari N., Akoka J., Bouabana-Tebibel T. (2012), *Building Contextualized Topic Maps*. 19th IBIMA (International Business Information Management Association) conference on Innovation Vision 2020: Sustainable growth, Entrepreneurship, Real Estate and Economic Development, Nov 2012, Barcelone, Spain. ⟨hal-01126211⟩

Laihonen H., Kork A.A., Sinervo L.M. (2023). Advancing public sector knowledge management: towards an understanding of knowledge formation in public administration. *Knowledge Management Research & Practice*, 22(3), 223–233. https://doi.org/10.1080/14778238.2023.2187719.

Leland McInnes, John Healy, and James Melville (2018). “UMAP: Uniform manifold approximation and projection for dimension reduction”. In: arXiv preprint arXiv:1802.03426

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852.

Maharjan P. (2020). Knowledge Management Enablers for Knowledge Creation Combination in Nepalese Hospitality Industry. *Journal of Balkumari College*, 9(1), 25-33. https://doi.org/10.3126/jbkc.v9i1.30064.

Malzer C., Baum M. (2020). A hybrid approach to hierarchical density-based cluster selection. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 223-228. IEEE, Karlsruhe, Allemagne.

McInnes L., Healy J., Melville J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, arXiv:1802.03426. Disponible sur : https://arxiv.org/abs/1802.03426.

Mihalcea R., Tarau P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411.

Morse R. (2000). Management in the 21st Century Knowledge Management Systems: Using Technology to Enhance Organizational Learning. *Proceedings of the 2000 Information Resources Management Association International Conf. on Challenges of Information Technology Management in the 21st Century*, pp. 426–429. IGI Global, Anchorage, USA.

Narazaki R.Y., Silveira Chaves M., Drebes Pedron C. (2020). A project knowledge management framework grounded in design science research. *Knowledge and Process Management*, 27(3), 197-210. https://doi.org/10.1002/kpm.1627.

Priti J. (2012). An Empirical Study of Knowledge Management in University Libraries in SADC Countries. In : Hou H.T., *New Research on Knowledge Management Applications and Lesson Learned*. IntechOpen. https://doi.org/10.5772/36309.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Tsoumakas, Grigorios & Katakis, Ioannis. (2007). Multi-label classification: An overview. IJDWM. 3. 1-13.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Wan X., Xiao J. (2008). Single Document Keyphrase Extraction Using Neighbourhood Knowledge. *Proceedings of the 23th AAAI Conference on Artificial Intelligence*, pp. 855-860. American Association for Artificial Intelligence, Chicago, Illinois, 13-17 juillet.

Wenhui W., Furu W., Li D., Hangbo B., Nan Y., Ming Z. (2020). MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Proceedings of the 34th International Conf. on Neural Information Processing Systems (NIPS '20)*, Article 485, pp. 5776–5788. Curran Associates Inc., Red Hook, NY, USA.

Xu D.D., Wu S.B. (2014). An improved TFIDF algorithm in text classification. *Applied Mechanics and Materials*, 651 (Sep. 2014), 2258-2261.

Yao-Sheng L. (2007). The effects of knowledge management strategy and organization structure on innovation. *International Journal of Management*, 24(1), 53-60.

Zhou J., Jia Y., Qiu Y., Lin L. (2023). The potential of applying ChatGPT to extract keywords of medical literature in plastic surgery. *Aesthetic Surgery Journal*, 43(9), NP720-NP723.