Constitution de corpus thématique : Pour un meilleur suivi du territoire de la Métropole de Montpellier Méditerranée

Rodrique Kafando, Rémy Decoupes, Maguelonne Teisseire, Lucile Sautot, Christiane Weber

▶ **To cite this version:**

Rodrique Kafando, Rémy Decoupes, Maguelonne Teisseire, Lucile Sautot, Christiane Weber. Constitution de corpus thématique : Pour un meilleur suivi du territoire de la Métropole de Montpellier Méditerranée. SAGEO'21 16ème Conférence Internationale de la Géomatique, de l'Analyse Spatiale et des Sciences de l'Information Géographique., May 2021, La Rochelle, France. hal-03243745

HAL Id: hal-03243745
https://hal.science/hal-03243745v1

Submitted on 31 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constitution de corpus thématique : Pour un meilleur suivi du territoire de la Métropole de Montpellier Méditerranée

Rodrique Kafando[1,4], **Rémy Decoupes**[1], **Lucile Sautot**[1,2], **Maguelonne Teisseire**[1], **Christiane Weber**[1,3]

1. *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France*
   prenom.nom@inrae.fr
2. *AgroParisTech, Montpellier, France*
   lucile.sautot@agroparistech.fr
3. *CNRS, France*
   christiane.weber@cnrs.fr
4. *Montpellier Méditerranée Métropole, France*

RÉSUMÉ. *Dans le cadre d'un partenariat avec la métropole de Montpellier (3M) et son service ville intelligente, nous établissons une plateforme permettant l'intégration et l'analyse de données hétérogènes massives pour une observation intelligente du territoire concerné. Dans cet article, nous décrivons un processus de récolte automatique de documents de sources diverses, selon trois thématiques choisies par 3M qui sont l'urbanisation, l'agriculture et l'hydrologie. A partir de termes graines à dire d'experts enrichies grâce à des articles Wikipédia puis Google, un vocabulaire de concepts est automatiquement constitué. Les documents obtenus à partir d'un moteur de recherche sont ensuite évalués à l'aide d'une mesure sémantique spécifique. Nous détaillons les premiers résultats obtenus et présentons les avantages du protocole proposé.*

ABSTRACT. *In collaboration with the metropolis of Montpellier (3M) and its smart city service, we are establishing a platform allowing the integration and analysis of massive heterogeneous data for intelligent observation of the territory concerned. In this article, we describe the process implemented to ensure an automatic collection of documents according to three themes chosen by 3M, which are urbanization, agriculture and hydrology. Results are detailed and the benefits of this scalable solution are highlited.*

MOTS-CLÉS : Observation du territoire, Constitution automatique de corpus
KEYWORDS: Land Planning, automatic corpus constitution

1. Introduction

La Métropole de Montpellier Méditerranée (3M) développe depuis plusieurs années une politique pro-active en faveur de développements numériques innovant par la mise en œuvre d'approches fondées sur l'hétérogénéité des sources, l'extraction de connaissances à partir de données massives dans le but de produire des outils d'aide à la décision. L'intégration de données hétérogènes, en prenant en compte l'écosystème « ville intelligente » de la Métropole (labos, start up, collectifs, etc.), se fonde sur des informations quantitatives (démographiques, économiques, climatiques, etc.) mais aussi sur des informations venant d'autres sources de données, officielles et non officielles. La finalité est de montrer le gain et l'efficacité de la mise en relation de données gérées jusqu'ici en silo et de l'organisation des interactions qui peuvent en découler. C'est dans ce contexte que se positionnent les travaux présentés dans cet article sur un processus de constitution de corpus thématique (ou corpus dédié) à partir du Web sur le territoire concerné. Cette étude rentre dans le cadre de travaux qui visent la mise en relation et l'analyse de données textuelles et hétérogènes. Le but principal est de pouvoir collecter par couple de thématique et territoire, des données textuelles pertinentes qui couvrent une grande diversité de sources. Le challenge et le verrou associé est de ne pas se limiter aux documents purement officiels (souvent disponibles dans l'open data de la métropole). Il devient indispensable de disposer d'une terminologie moins académique que celle utilisée par les différentes collectivités territoriales afin de pouvoir accéder à l'expression de la composante sociétale, indispensable à une véritable perception de la dynamique du territoire dans sa globalité.

Dans la suite de cet article, en Section 2, nous listons un panorama des travaux sur la constitution de corpus thématiques, en particulier ceux qui ont inspiré les bases de notre démarche. Ensuite, la méthodologie que nous proposons pour la constitution d'un corpus propre à notre cas d'étude est détaillée dans la Section 3. Dans la Section 4, nous présentons les premiers résultats quantitatifs suite aux différentes expérimentations réalisées. Enfin, nous dressons un bilan et les perspectives associées en Section 5.

2. État de l'art

La mise en oeuvre des concepts de la ville intelligente devrait permettre aux villes et métropoles d'améliorer l'efficacité de leurs services de gestion (Guéranger, Mathieu-Fritz, 2019), en offrant, notamment, davantage de transparence.

Cette efficacité peut être évaluée en analysant les opinions exprimées sur les différents projets et réalisations en lien avec l'aménagement du territoire (Kergosien *et al.*, 2014). Les ressources disponibles sur le web, et en particulier celles de la presse locale, permettent de suivre les débats que suscitent les projets d'urbanisme. Dans (Kergosien. *et al.*, 2015), les auteurs proposent, notamment, une méthodologie semi-automatisée pour enrichir les scènes issues d'une série temporelle d'images satellites, focalisées sur un projet d'aménagement.

Les genres, ou catégories textuelles, des ressources indexées par les moteurs de recherche sont très variés : blog, publicité, rapport officiels, articles scientifique etc. Plusieurs travaux proposent différentes hiérarchies de genre (Vidulin *et al.*, 2009; Santini, 2011) afin de faciliter la classification de nouvelles ressources. (Madjarov *et al.*, 2019) exploite ces hiérarchies pour renforcer les capacités d'un moteur de recherche à proposer des résultats pertinents selon les genres souhaités. L'évaluation de la pertinence des ressources collectées sur le web est, bien entendu, un enjeu majeur. Une approche classique vise à calculer les co-occurrences de termes entre la requête et les documents rapatriés pour ensuite classer les résultats en fonction du poids du terme dans le document, en utilisant, notamment la mesure TF-IDF(Aizawa, 2003). Cette méthode a cependant des limites. Cette approche par mot clé ne peut retranscrire la proximité sémantique entre deux termes comme, par exemple, vélo et mobilité douce.

Pour remédier ce problème, (Schaeffer. *et al.*, 2020) propose de filtrer les ressources pertinentes en calculant la proximité sémantique, via le modèle de prolongement lexical word2vec (Mikolov *et al.*, 2013), entre des termes qui composent le document et un thésaurus thématique. Dans un contexte de surveillance en épidémiologie animale, (Arsevska. *et al.*, 2016) montre que la construction de requête avec deux mots clés, symptôme et animal, permet de récolter des ressources avec davantage de pertinence. Il est possible, enfin, de combiner ces deux dernières approches, prolongement lexical et requête multi-termes, pour enrichir la sémantique du terme par l'ajout d'un contexte apporté par la phrase, ou le groupe de mots, qui le contient en utilisant le modèle BERT (Devlin *et al.*, 2018). De la même façon, nous proposons la mise en place d'un ensemble de techniques prenant en compte des stratégies d'extraction de concepts (officiels et non officiels) et de constitution automatique de corpus thématiques provenant de sources diversifiées, validées par des mesures de similarité sémantique.

3. Méthodologie

L'approche méthodologique que nous proposons porte sur deux grands points. Le premier est la constitution d'un vocabulaire de concepts thématique. Le second concerne la collecte de documents en utilisant l'ensemble de termes précédemment obtenus.

3.1. Constitution des vocabulaires de concepts thématique

Le processus pour la constitution de vocabulaire de concepts thématique est basé sur le principe de construction de corpus proposé dans (Kilgarriff, Grefenstette, 2003; Kilgarriff *et al.*, 2010; Sharoff, 2006). Il consiste à utiliser un ensemble de graines ou une liste de termes d'un domaine caractéristique pour requêter des documents sur le web. Les résultats sont ensuite utilisés pour étendre la liste des termes et donc le corpus, et ainsi de suite. L'approche proposée se décline en deux phases : la constitution de l'ensemble de termes graines noté *TG*, et la constitution du vocabulaire de concepts noté *VC*.

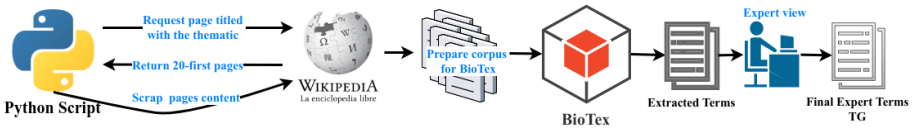

FIGURE 1 – Choix de termes initiaux - TG

3.1.1. Choix des termes graines TG

La proposition de l'ensemble *TG* comprend deux étapes (FIGURE 1). La première consiste à extraire un ensemble de termes à partir d'un mini corpus obtenus de Wikipédia [1] en utilisant l'outil BioTex de (Lossio-Ventura *et al.*, 2014). Ce corpus est construit en utilisant un mot clé thématique (ex : urbanisation), et les vingt premières pages renvoyées par Wikipédia sont utilisées pour former le corpus.

Pour BioTex, nous avons principalement utilisé deux mesures statistiques qui sont C_Value et F-TFIDF-C_M. C_Value donne de l'importance aux termes apparaissant plusieurs fois dans le même document et aux termes composés permettant de valoriser les expressions. F-TFIDF-C_M représente la moyenne harmonique de C_Value et TF-IDF, qui permet de classer les termes en fonction de leur pertinence vis-à-vis du document en prenant en compte l'ensemble du corpus. F-TFIDF-C_M présente l'avantage d'utiliser toutes les valeurs de la distribution et de réduire le bruit tout au long du processus d'extraction. C-Value et F-TFIDF-C_M sont complémentaires, car la première favorise l'extraction des termes composés pertinents et la seconde privilégie les termes discriminants. Discriminants dans le sens ou elle permet de capturer les termes pertinents qui majoritairement ne sont pas reconnus par C_Value et TF-IDF prise individuellement.

La seconde étape consiste à faire choisir un ensemble de termes par des experts du domaine parmi les termes récupérés à partir de Wikipédia. L'ensemble

1. https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

de termes ainsi constitué forme la liste de termes experts que nous appelons ici *TG*.

3.1.2. Constitution du vocabulaire de concepts VC

La constitution du vocabulaire de concepts est basée sur une première étape qui consiste à utiliser la liste *TG* comme une liste de graines permettant de collecter un grand nombre de documents via un moteur de recherche comme par exemple Google. L'objectif est d'obtenir des variantes de termes pour faire évoluer la terminologie de TG, afin de récolter des documents parlant de la thématique mais avec des mots et des termes qui ne seraient pas présents dans Wikipédia. Ensuite, grâce à BioTex, le corpus obtenu est utilisé pour extraire un grand ensemble de termes noté TB (termes BioTex), à partir duquel, nous allons réduire le bruit éventuel pour former le vocabulaire *VC*. Pour éviter de prendre en considération les termes généraux ou les termes qui ne permettent pas de définir la thématique étudiée, nous avons mis en place un filtre basé sur une approche sémantique en utilisant DistilBERT (Sanh *et al.*, 2019), reconnue pour l'évaluation de la Similarité Sémantique Textuelle (STS). La valeur de similarité varie entre -1 pour de très faible similarité et 1 lorsque les termes sont très proches ou identiques.

La mesure sémantique est évaluée à partir 1) des termes extraits TB qui sont les nouveaux termes, et 2) les termes graines ou experts TG. Pour chaque terme de TB, nous calculons sa similarité sémantique avec chacun des termes de l'ensemble TG. Ce qui nous donne un nombre de valeurs de similarité égal au nombre de termes dans l'ensemble TG pour chaque terme t de TB. Pour obtenir la valeur de similarité finale pour un terme, nous faisons la moyenne des TOP@n (n correspond aux n termes les plus pertinents) en tenant compte des plus grandes valeurs de similarités pour garder une certaine exhaustivité vis-à-vis de l'ensemble des termes experts. À la fin du processus, nous obtenons un nouvel ensemble ordonné de termes, constitué des mêmes termes de l'ensemble TB, mais avec des scores de pertinence différents. Les vocabulaires de concepts seront par la suite sélectionnés à partir de ces ensembles de termes pour chaque thématique, en partant de ceux disposant d'un plus grand score. La Figure 2 décrit les différentes étapes pour la construction de vocabulaire de concepts thématique dans son ensemble et le calcul de similarité entre termes est illustré sur la Figure 3.

3.2. Constitution des corpus thématiques

3.2.1. Collecte de documents

Une fois le vocabulaire de concepts thématique construit, nous avons pour chaque thématique, un ensemble de concepts permettant de la décrire au mieux. La phase de collecte (voir Figure 4) consiste à utiliser un module, qui prend en entrée un vocabulaire de concepts *VC* et une emprise spatiale, pour retourner

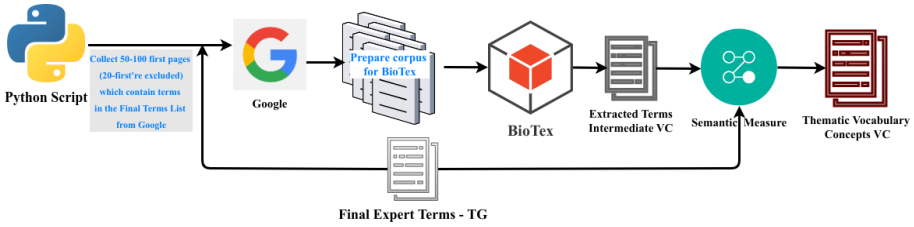

FIGURE 2 – Constitution de vocabulaire de concepts thématique - VC

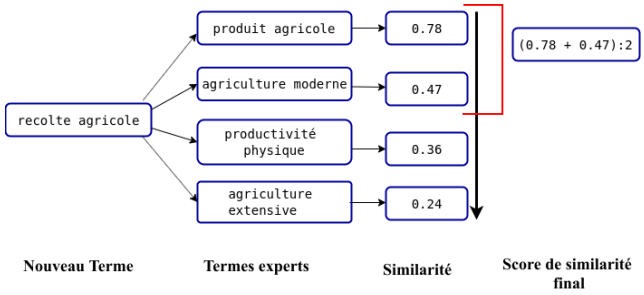

FIGURE 3 – Exemple de calcul de similarité sémantique - score final calculé en faisant la moyenne des 2 plus grandes valeurs

des documents en relation avec la thématique concernée pour former un corpus de taille importante. Nous pouvons ainsi constituer pour chaque thématique, un corpus le plus en cohérence possible avec son contenu. La prochaine étape permet de filtrer les documents les moins pertinents pour chaque corpus thématique. Lorsque le document obtenu est pertinent, une fiche de méta-données contenant les informations sur le document est automatiquement générée avant d'être stockée dans un lac de données (Kafando *et al.*, 2020).

3.2.2. Évaluation automatique des documents par mesure sémantique

Dans le but de ne considérer que les documents qui sont pertinents pour chaque thématique, nous avons introduit une évaluation par mesure de similarité. Identique à la précédente, l'évaluation des documents se fait entre un document collecté par rapport à l'ensemble des termes de son vocabulaire étendu de concepts thématique VEC obtenu avec WordNet (Miller, 1998). L'objectif principal de cette étape est de disposer 1) d'une couverture plus large sur les termes en introduisant leur synonyme grâce à WordNet, et 2) d'obtenir des termes moins experts et/ou administratifs qui sont importants pour l'évaluation des documents moins administratifs tels que les blogs, les annonces, etc. Cette technique permet de prendre en compte ces synonymes lors de l'évaluation des documents. Comme décrite précédemment, la valeur de similarité avec

DistilBERT varie entre -1 et 1. Un document de la thématique urbanisation sera évalué avec le vocabulaire étendu de concepts de la thématique urbanisation. Plus la valeur de similarité est forte, plus le document sera considéré comme pertinent vis-à-vis de la thématique, et vice-versa.

Le processus pour l'évaluation des documents est illustré sur la FIGURE 4. Afin d'éliminer les documents qui ne sont pas pertinents, nous définissons lors des expérimentations, une valeur seuil à partir de laquelle, nous décidons de garder le document ou de l'exclure du corpus.

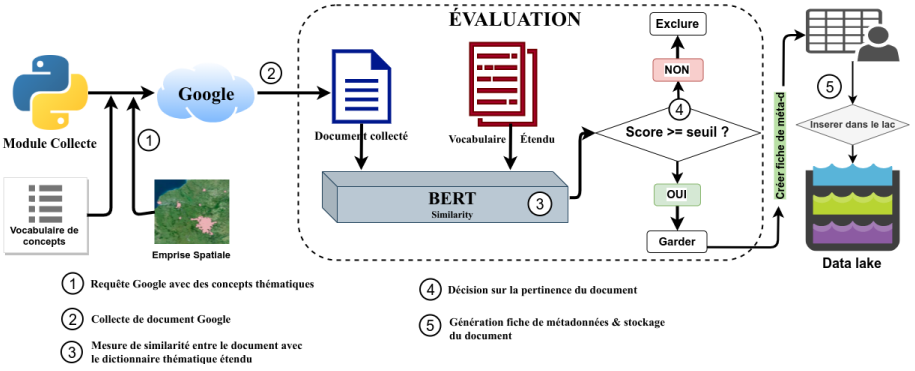

FIGURE 4 – Collecte et évaluation des documents

Les différentes étapes, illustrées sur la FIGURE 5, constituent le protocole de collecte proposé.

4. Expérimentation

Nous cherchons des documents traitant de la Métropole de Montpellier et relatifs à l'une des trois thématiques qui sont l'Agriculture, l'Urbanisation et l'Hydrologie. Les codes sources de la solution mise en place seront disponibles et accessibles dans notre *répertoire partagé* [2].

4.1. Vocabulaire de concepts thématique

La première étape est le choix de termes graines *TG* décrit dans la section 3.1. La seconde étape consiste à soumettre cet ensemble de termes à un groupe d'experts dont le rôle est de sélectionner un sous-ensemble (liste finale de TG) de termes permettant de décrire au mieux la thématique. Pour chaque requête, nous utilisons le libellé de la thématique comme mot clé principal de

2. https://github.com/aidmoit

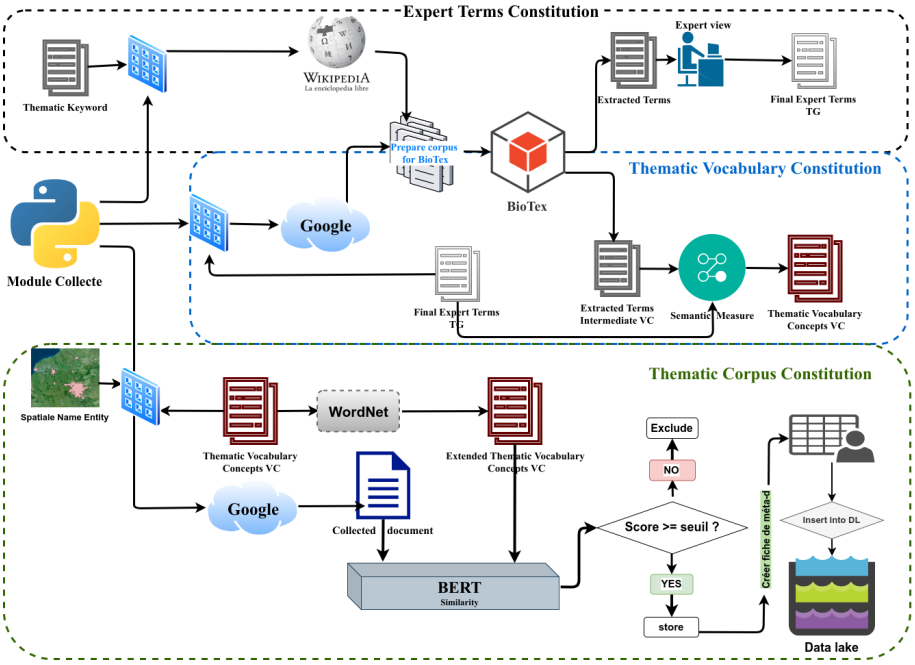

FIGURE 5 – Les étapes du protocole de collecte

la recherche sur Wikipédia, soit 'urbanisation', 'agriculture', ou 'hydrologie'. Les corpus ainsi constitués de chaque thématique sont ensuite utilisés pour l'extraction des termes parmi lesquels seront issus les termes graines.

TABLE 1 – Statistiques sur la constitution de la liste de termes graines TG

Termes requêtes	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille Corpus (Nb docs)	20	20	20
Nombre de termes retenu	100	100	100

TABLE 2 – Extrait de termes experts par thématique

Urbanisation TG-1	**Agriculture TG-2**	**Hydrologie TG-3**
Politiques urbaines	Récolte agricole	Hydrogéologie
Aménagement territoriale	Agriculture	Eau pluviale
Aménagement foncier	Aménagement agricole	Mesure hydrologique
Aménagement touristique	Production agricole	Cycle de l'eau
Architectonique	Exploitation agricole	Réseau hydrologique

Les tableaux 1 et 2 présentent respectivement les statistiques et les TOP@5 de la liste de termes graines TG de chaque thématique.

Une fois les termes graines obtenus, nous procédons à la constitution des vocabulaires de concepts *VC*. Pour chaque thématique, la liste TG est utilisée pour servir de mots clés aux requêtes pour constituer de corpus plus importants, sur le web, sans distinction de site. Ceci nous permet d'avoir une forte diversité des données tant dans leurs contenus que dans leurs sources. Nous avons récolté respectivement 105, 120 et 117 documents pour former les corpus Urbanisation, Agriculture et Hydrologie à partir desquels seront extraits les ensembles de termes que nous nommons termes Biotex (TB). Afin de pouvoir constituer les vocabulaires de concepts thématiques finaux, nous avons utilisé la mesure sémantique proposée dans le Section 3.1.2 qui nous a permis de limiter le bruit, en donnant des poids très faibles aux termes dont la sémantique est éloignée de la thématique.

Dans le tableau 3, nous présentons les TOP@10 des termes de l'ensemble TB avant et après l'application de la mesure sémantique pour la thématique Agriculture. Tout d'abord, nous constatons que l'ordre de pertinence des termes

Table 3 – TOP@10 des termes de la thématique Agriculture avant et après la mesure sémantique (MS)

Thématique Agriculture	
TOP@10 avant la *MS*	**TOP@10 après la *MS***
dégradation des terres	production agricole
agriculture urbaine	exploitation agricole
etats membres	agricole production
changement climatique	oeuvre agricole
google scholar	entreprise agricole
occupation du sol	service agricole
field sizes	techniques de production agricole
lutte contre la désertification	production agricole perdu
production agricole	agricole perdu
matière de dégradation des terres	gestion de la production agricole

a changé entre les deux étapes. Dans le tableau 3, première colonne, de la thématique *Agriculture*, nous remarquons la présence de terme comme *'google scholar'*, qui n'a pas un sens sémantique proche à l'agriculture, mais qui occupe un rang important dans les classements de BioTex (c'est-à-dire dans l'ensemble TB). Après avoir appliqué le calcul de similarité sémantique entre ces termes et les termes experts (*conf* Table 2), nous constatons dans la seconde colonne, l'apparition de nouveaux termes plus pertinents dans les TOP@10. Cela signifie que les nouveaux termes (ex: production agricole au 9ème rang) présents dans la deuxième étape occupaient un rang de pertinence moins fort avant la mesure de la similarité. Cette remarque est aussi valable pour les thématiques Urbanisation et Hydrologie.

4.2. Constitution des corpus thématiques

4.2.1. Vocabulaire de concepts thématique final

Le choix du vocabulaire de concepts thématique final est effectué en fixant un seuil minimal spécifique à chaque thématique. Pour ce faire, nous regardons donc à partir de quelle valeur du score, les termes commencent à être éloignés de la thématique. Pour la suite de notre étude, nous avons considéré les TOP@1000 termes de chaque thématique. Le tableau 4 indique le seuil correspondant pour chaque thématique.

TABLE 4 – Statistiques sur la constitution du vocabulaire de concepts thématique VC

Thématique	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille Corpus (Nb documents)	105	120	117
Taille du VC	TOP@1000	TOP@1000	TOP@1000
Valeur du seuil de similarité	0.80	0.75	0.77

Une fois le seuil fixé, la liste de termes est étendue en utilisant WordNet. Les synonymes obtenus permettent de sortir du cadre de langage officiel ou académique, et d'utiliser des expressions moins techniques, qui permettront par exemple d'évaluer des documents en relation avec des réseaux sociaux, des blogs ou encore des annonces telles que les offres d'emplois. En exemple, certains synonymes obtenus avec le terme *"zone d'aménagement concerté"* de la thématique Urbanisation sont : *district d'aménagement concerté, domaine d'aménagement concerté, aire d'aménagement concerté, quartier d'aménagement concerté, etc.*.

4.2.2. Évaluation automatique des documents par mesure sémantique

Cette évaluation est faite par calcul de similarité sémantique entre chaque document collecté avec le vocabulaire étendu de concepts thématique de la thématique concernée.

Pour chaque thématique, nous avons considéré les TOP@1000 termes de son vocabulaire étendu de concepts thématique pour l'évaluation des documents la concernant. Ce choix est fixé suite aux différentes expérimentations qui montrent une très faible variation, voir nulle au delà des TOP@1000. À la fin du processus, pour chaque thématique, un score est affecté à chaque document, représentant sa proximité sémantique avec le vocabulaire étendu de concepts thématique précédemment obtenu. Nous avons défini une valeur seuil de 0.5 comme valeur minimale. Un document dont la valeur de la proximité sémantique est inférieur à 0.5 ne sera donc pas considéré dans le corpus, et ce pour l'ensemble des thématiques. Cette valeur est fixée après une vérification des résultats obtenus. Le tableau 5 indique les caractéristiques pour chaque corpus thématique.

TABLE 5 – Statistiques sur les corpus constitués par thématique

Thématique	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille du VC utilisé	TOP@500	TOP@500	TOP@500
Taille du VEC utilisé	TOP@1000	TOP@1000	TOP@1000
Seuil de similarité	0.5	0.5	0.5
Taille Corpus	867	1.400	1380

5. Conclusion

Le protocole que nous avons élaboré dans cette étude vise à proposer une démarche générique permettant de constituer 1) des vocabulaires thématiques et 2) de constituer des corpus thématiques spécifiques (l'agriculture, l'urbanisation et l'hydrologie pour la métropole de Montpellier) pour permettre une meilleure observation du territoire. Elle s'appuie principalement sur le principe de faire évoluer une terminologie allant de la constitution de graines à la formation d'un vocabulaire de concepts thématique bien spécifique. Quant à la constitution des corpus, elle se fait en utilisant chaque vocabulaire thématique comme ensemble de mots clés de recherche et une spécification de l'emprise spatiale d'intérêt. Chacune des étapes est soumise à une évaluation sémantique, dans le but de limiter le bruit aussi bien dans le vocabulaire de concepts que dans les corpus thématiques. Nos travaux futurs porteront sur l'intégration de l'ensemble de ces documents au sein du lac de données de la métropole de Montpellier et leur mise en lien selon les dimensions spatiale, temporelle et thématique. Notre contribution s'inscrit en sciences de l'environnement avec le développement de méthodes permettant de gérer une grande masse de données pour assurer le suivi d'un territoire donné et d'en extraire une meilleure connaissance. Ces travaux ont été soutenus par l'Agence nationale française de la recherche dans le cadre du programme Investissements d'avenir #DigitAg, référencé ANR-16-CONV-0004.

Bibliographie

Aizawa A. (2003, 01). An information theoretic perspective of tf-idf measures. *Information Processing Management*, vol. 39, p. 45-65.

Arsevska. E., Roche. M., Hendrikx. P., Chavernac. D., Falala. S., Lancelot. R. *et al.* (2016). Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. *International Journal of Agricultural and Environmental Information Systems*.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guéranger D., Mathieu-Fritz A. (2019, novembre). The smart city at work. *Reseaux*, vol. No 218, n° 6, p. 41–75.

Kafando R., Decoupes R., Sautot L., Teisseire M. (2020). Spatial Data Lake for Smart Cities: From Design to Implementation. *AGILE: GIScience Series*, vol. 1, p. 1-15.

Kergosien. E., Alatrista-Salas. H., Gaio. M., Güttler. F., Roche. M., Teisseire. M. (2015). When textual information becomes spatial information compatible with satellite images. In *Proceedings of the 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management - volume 1: Kdir, (ic3k 2015)*, p. 301-306. SciTePress.

Kergosien E., Laval B., Roche M., Teisseire M. (2014). Are opinions expressed in land-use planning documents? *International Journal of Geographical Information Science*, vol. 28, n° 4, p. 739-762.

Kilgarriff A., Grefenstette G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, vol. 29, n° 3, p. 333–347.

Kilgarriff A., Reddy S., Pomikálek J., Avinesh P. (2010). A corpus factory for many languages. In *Lrec*.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M. (2014). Biotex: A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the 2014 international conference on posters demonstrations track - volume 1272*, p. 157–160. Aachen, DEU, CEUR-WS.org.

Madjarov G., Vidulin V., Dimitrovski I., Kocev D. (2019). Web genre classification with methods for structured output prediction. *Information Sciences*, vol. 503, p. 551 - 573.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, vol. 26, p. 3111–3119. Curran Associates, Inc.

Miller G. A. (1998). *Wordnet: An electronic lexical database*. MIT press.

Sanh V., Debut L., Chaumond J., Wolf T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santini M. (2011). Cross-testing a genre classification model for the web. In A. Mehler, S. Sharoff, M. Santini (Eds.), *Genres on the web: Computational models and empirical studies*, p. 87–128. Dordrecht, Springer Netherlands.

Schaeffer. C., Interdonato. R., Roche. M. (2020). Construction d'un corpus sur la problématique de la sécurité alimentaire guidée par un lexique et des approches de fouilles de textes, in : Toth 2020 - terminologie et ontologie : théories et applications. roche christophe (ed.). chambéry : Presses universitaires savoie mont blanc.

Sharoff S. (2006). Creating general-purpose corpora using automated search engine queries. *WaCky*, p. 63–98.

Vidulin V., Lustrek M., Gams M. (2009, 01). Multi-label approaches to web genre identification. *JLCL*, vol. 24, p. 97-114.