École des Mines de Saint-Étienne

Janvier 2022

# Rapport de travail de fin d'études (TFE)

Théo Gady - Élève ICM promo 2018

1 Septembre 2021 / 31 Janvier 2022




## **Sujet de stage :** analyse automatisée de corpus littéraire francophone nativement numérique

**Tuteur de stage :** Julien Velcin
**Co-encadrant :** Enzo Terreau
**Tuteur EMSE :** Benjamin Dalmas

**SOMMAIRE**

**INTRODUCTION**

Ce stage se déroule au sein du laboratoire ERIC (pour Entrepôts, Représentation et Ingénierie des Connaissances), une unité de recherche des Universités Lyon 2 et Lyon 1, ayant pour domaines de spécialité la Science des Données, l'Informatique Décisionnelle ainsi que la gestion et l'analyse de données issues des Humanités Numériques. Le laboratoire traite de sujets de recherche théorique et appliquée ayant pour but de mettre en valeur de grandes bases de données complexes, notamment dans les domaines des lettres, langues, sciences humaines et sociales (LLSHS).

Le projet LIFRANUM (pour Littératures Francophones Numériques) est un projet de recherche multidisciplinaire visant à « l'identification, l'indexation et l'analyse des productions littéraires nativement numériques dans l'aire francophone ». Il est mené par un consortium comprenant le laboratoire MARGE (Université Lyon 3), regroupant des spécialistes de Langues, de Littérature ainsi que des Sciences de l'information et de la communication, le laboratoire ERIC et la Bibliothèque Nationale de France.

L'objectif du projet est multiple. Il vise tout d'abord à constituer un corpus de littérature francophone nativement numérique. À la différence de la littérature au format physique dite « classique », la littérature nativement numérique peut provenir de toutes sortes de supports, que ce soit des sites web, des blogs ou des réseaux sociaux. La difficulté provient alors non seulement de cette variété de supports, posant la problématique de l'indexation de données aussi hétérogènes, mais aussi de la variété de contenu, le tout dans un espace de recherche équivalent au web lui même. La question de ce qui peut ou pas être considéré comme de la littérature francophone nativement numérique surgit alors logiquement. Est-il possible de classer de manière automatique un document comme faisant partie de notre corpus ?

Un autre objectif de ce projet est l'analyse et la structuration du corpus. Comment les écrivains se servent-ils des nouveaux médias et dispositifs numériques ? Comment utilisent-ils les dispositifs numériques pour écrire, diffuser des textes et rejoindre leurs lecteurs ? Comment se servent-ils des dispositifs numériques pour parler de littérature et imaginer de nouvelles formes littéraires ? Toutes ces questions d'ordre littéraire peuvent tirer profit d'analyses de type Data Mining sur les données textuelles composant les oeuvres de notre corpus, ainsi que sur leurs diverses métadonnées comme les auteurs, les commentaires, les images, les abonnés, etc …

Enfin, le but final du projet LIFRANUM est la création d'un outil de micropublication et prise de notes collaboratives, permettant à tout chercheur en LLSHS d'interroger le corpus, d'en récupérer les métadonnées, de le commenter, l'annoter, et de partager ces notes avec les autres chercheurs.

Du point de vue du laboratoire ERIC, l'objectif de ce projet est, entre autre, de progresser sur les questions de la structuration de données textuelles hétérogènes, notamment au travers de lac de données, ainsi que la représentation d'auteurs dans un espace latent, où il devient possible de calculer la similarité entre auteurs ou découvrir des communautés dont la production partage des similitudes. Plus particulièrement, les objectifs de ce stage concernaient l'analyse et l'exploitation d'une base de données de blogs littéraires.

Ce rapport sera articulé de la façon suivante :

- Dans un premier temps nous feront un état des lieux du projet au début du stage, des données disponibles ainsi que des objectifs à court terme.
- Ensuite, nous synthétiserons le travail qui a été fourni lors du stage, notamment au travers des différentes approches qui ont été utilisées.
- Nous présenterons finalement les résultats obtenus lors de ce stage, ainsi que leur apport au projet LIFRANUM.

# 1. ETAT DES LIEUX

## 1.1. LE PROJET LIFRANUM

Lors de mon arrivée, le projet LIFRANUM avait surtout avancé sur l'aspect récupération de données. En effet, un écosystème, disponible dans l'Annexe 1, avait été mis en place afin de récupérer et stocker des productions littéraires numériques d'intérêt. Il avait notamment été réalisé une campagne de crawling grâce à HERITRIX, un outil d'archivage du web utilisé notamment par la BnF. Comme tout robot d'indexation (web crawler en anglais), Heritrix parcours le web à partir d'une liste de pages donnée, desquelles il va récupérer des informations telles que le contenu et notamment les URLs présentes, qu'il va ajouter à sa liste de pages à indexer. Ce processus est répété, jusqu'à ce qu'il n'y ait éventuellement plus de pages à visiter. Afin de ne pas crawler tout le web, on peut indiquer à Heritrix une limite d'exploration, que ce soit sur la distance maximale qu'une page visitée peut avoir vis à vis de la liste de départ ou le nombre total d'URLs obtenues. Dans ce dernier cas, l'ordre de visite des pages dicté par le scheduler est déterminant, plusieurs stratégies pouvant être appliquées : favorisation des sites les plus volumineux, exploration de toutes les pages d'un site avant de passer à un autre, etc …

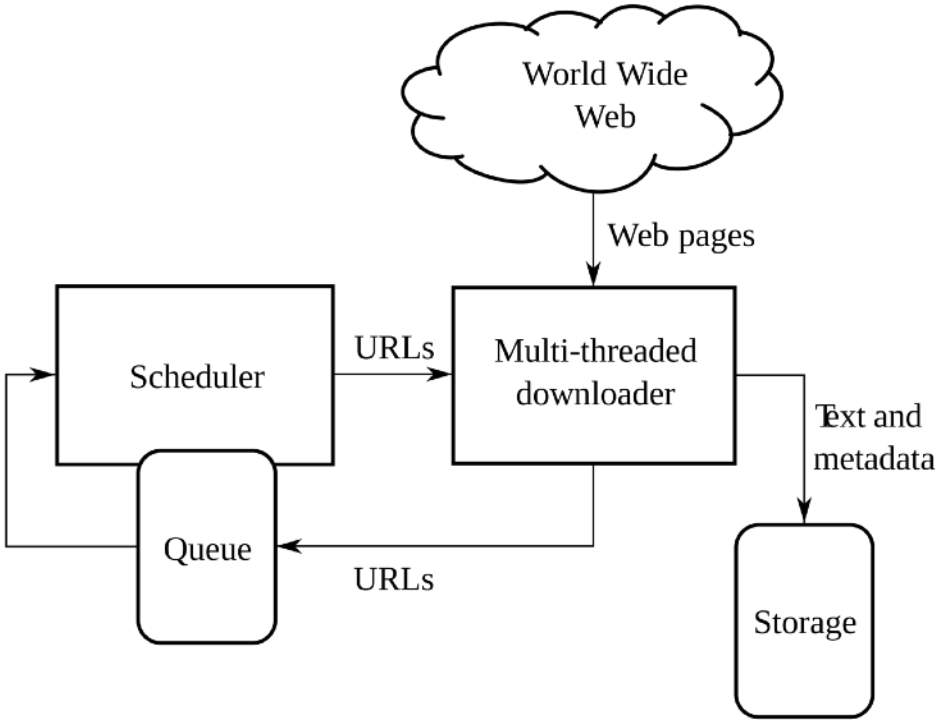



Figure 1-1. Schéma de fonctionnement général d'un web crawler.

Cette liste de départ a été obtenue grâce au laboratoire MARGE qui a pu fournir une liste d'auteurs et de pages web considérées comme étant de la littérature francophone numérique.

Le format de stockage utilisé par Heritrix est le Web ARChive (WARC), capable de prendre en charge des ressources de tout type de format, y compris texte, image, video et audio. Cette base de données peut alors être interrogée au travers de mots clefs ou de métadonnées comme la date par un moteur de recherche, ici Solr. Un exemple de fichier WARC est donné dans l'Annexe 2.

Le problème de ce format, en ce qui concerne notre projet, est qu'il comporte l'entièreté de la page archivée sous format html. Séparer le contenu littéraire du reste des informations de la page est alors une tâche complexe, notamment aux vues de la diversité de structuration des pages web. Il a alors été remarqué qu'une partie importante de ces pages provenaient des plateformes Blogger et Wordpress, deux outils de création de contenu web. Ces plateformes permettant d'avoir une structure commune à toutes les pages crées à travers elles, elles possèdent aussi une API permettant d'obtenir le contenu de leurs pages de manière structurée. C'est donc ce qui a été réalisé, le contenu de tous les blogs issus de Blogger et Wordpress a été obtenu grâce aux API, le résultat étant stocké sous format json.

### 1.2. LES DONNÉES

Comme présenté précédemment, les seules données textuelles facilement exploitables étaient celles issues des API de Blogger et Wordpress. J'ai donc été chargé d'en faire l'analyse, notamment pour Blogger, jugé le plus prometteur car plus diversifié.

Les données disponibles au travers de l'API de Blogger sont organisées hiérarchiquement de la manière suivante :


Figure 1-2. Structure hiérarchique des données sur Blogger avec exemples d'attributs.

Ainsi, un blog peut contenir plusieurs posts qui eux même peuvent contenir plusieurs commentaires. Concernant les attributs, chaque objet possède un identifiant unique, une date de publication, de révision, et une URL. Les blogs ont un nom ainsi qu'une description tandis que les posts possèdent un auteur, un titre et un contenu. Idem pour les commentaires dont on peut en plus savoir s'il s'agit de la réponse à un autre commentaire, mais qui ne possèdent pas d'URL. La liste de l'ensemble des attributs est fournie dans l'Annexe 3.

Afin d'obtenir une base de données unique et homogène, on enrichira la base de données des posts ainsi que que celle des commentaires avec les données des blogs, grâce à une jointure sur l'id du blog. Ces deux bases de données seront ensuite homogénéisées en ajoutant les attributs manquants puis additionnées.

On trouvera dans l'Annexe 4 un exemple de données issues d'un post et d'un commentaire.

Les commentaires étant difficiles à analyser en tant que contenu textuel (fautes d'orthographe, abréviations, smiley, …) et ne faisant pas partie de notre corpus littéraire, nous nous intéresserons principalement aux posts. On ne prendra pas non plus en compte les posts ayant une langue majoritaire autre que le français.

| | Auteurs | Posts | Commentaires | Blogs |
|---|---|---|---|---|
| **Données totales** | 187 | 110 895 | 242 945 | 187 |
| **Données utilisées** | 180 | 80 528 | 0 | 186 |

Figure 1-3. Informations générales sur la base de données utilisée.

On obtient alors avec une base de données de plus de 80 000 posts, écrits par 180 auteurs et publiés sur 186 blogs. En effet, si un auteur peut posséder plusieurs blogs, un blog peut recevoir les publications de plusieurs auteurs.

Afin de se faire une première idée de l'homogénéité des données, on représente la distribution du nombre de posts en fonction des auteurs. On observe alors aisément que celle-ci est déséquilibrée, les cinq auteurs les plus prolifiques représentant près de 25% du nombre total de posts avec plus de 3 000 publications chacun, lorsque la moyenne se trouve autours de 206 par auteur.

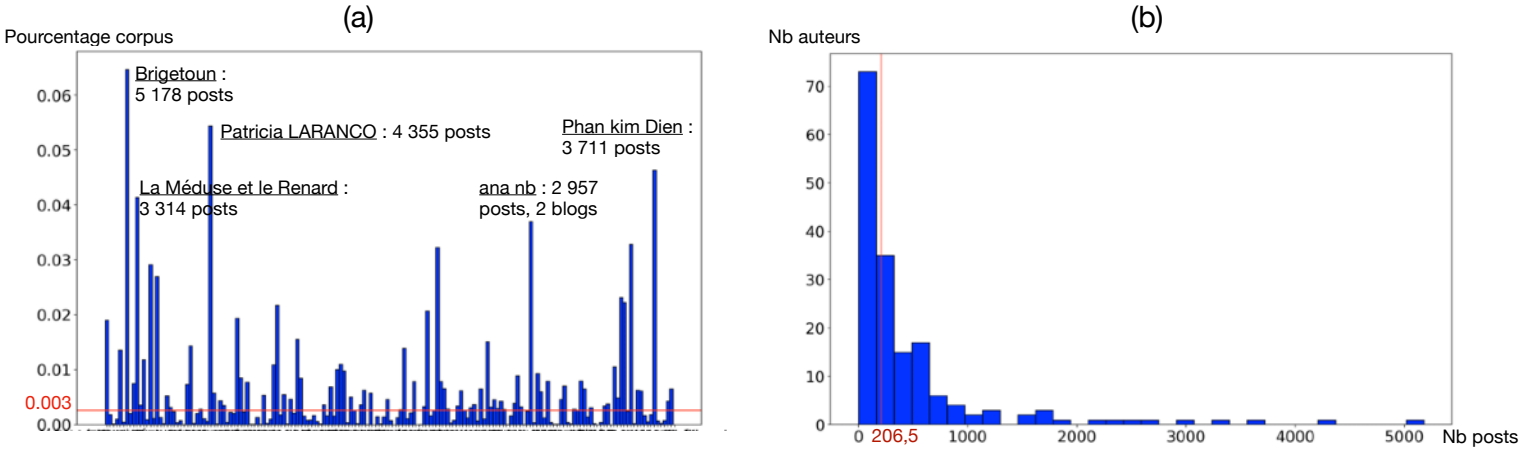



Figure 1-4. (a) Auteurs les plus représentés du corpus. (b) Distribution des auteurs en fonction du nombre de posts.

Une autre variable d'intérêt pour l'analyse de l'homogénéité de nos textes est leur longueur. En effet, si tous les blogs ont été manuellement confirmés comme étant producteurs de littérature française nativement numérique, ils n'en restent pas moins très divers dans leurs formats. Ainsi, un blog spécialisé dans la poésie ou les haïkus produira des textes beaucoup plus courts qu'un autre publiant des nouvelles ou des anthologies. Cet effet doit notamment être pris en compte lors de la construction de vocabulaires, afin qu'un genre littéraire ne soit pas sur-représenté.

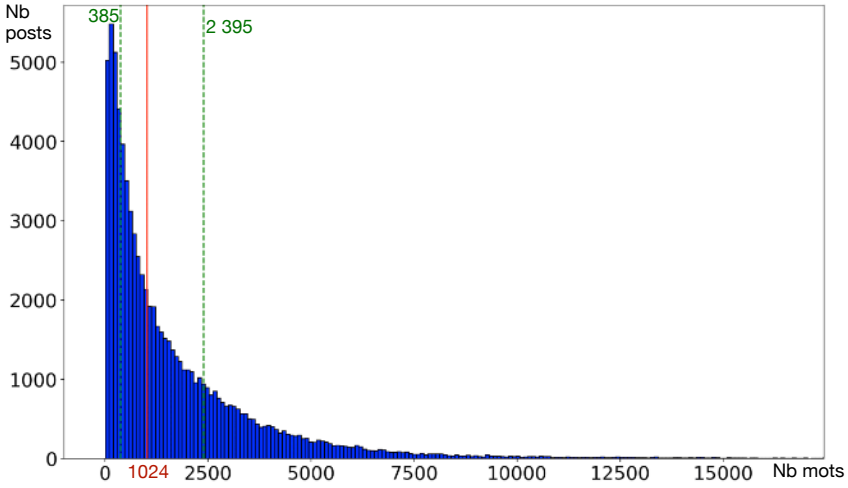



Figure 1-5. Distribution des posts en fonction du nombre de mots.

On remarque que non seulement cette distribution est relativement étendue, avec le troisième quartile valant plus de sept fois le premier, mais aussi qu'elle comporte une majorité de textes « courts ».

### 1.3. OBJECTIFS

Après cette première étape de familiarisation avec les données et plusieurs échanges avec les experts du laboratoires MARGE, plusieurs pistes d'intérêt ont été définies :

- Tout d'abord la description du corpus. Si tous les blogs ont été sélectionnés manuellement, l'entièreté de leur contenu n'a pas été lu et encore moins analysé. L'idée est alors de considérer ce corpus comme un échantillon, sûrement non représentatif, de ce qu'est la littérature française numérique. Son analyse permettrait alors de se faire une image plus précise de ce qu'elle est et ainsi avancer vers un possible début de définition. Plus concrètement, on se posera la question du contenu des posts, de quoi les gens parlent, et du format, comment ils en parlent. On se demandera aussi s'il est possible de regrouper les posts, les blogs voir les auteurs dans des groupes cohérents, ce qui revient à une tache de clustering non supervisée. Une autre piste se trouve dans l'exploitation des métadonnées, comme la localisation du blog ou les informations personnelles de l'auteur. Celles-ci étant non standardisées et non

obligatoirement rendues publiques, ces données sont à la fois très hétérogènes et incomplètes, mais pourraient possiblement offrir une grille de lecture plus fine à notre analyse, en prenant en compte des critères socio-culturels comme le sexe, l'âge ou l'appartenance géographique.

- Analyser les données sous le prisme d'un réseau social. En effet, si écrire un post est un acte individuel, l'action de le rendre visible par tous fait que Blogger peut néanmoins être considéré comme un réseaux social. L'espace commentaire en est une concrétisation, divers internautes, auteurs de blogs ou pas, anonymes ou pas, y donnant leurs opinions et se répondant les uns les autres. Il serait alors intéressant d'essayer de construire ce réseau social, qui n'est pas directement disponible, afin de pouvoir par exemple effectuer de la détection de communauté et repérer les auteurs qui gravitent autours des mêmes groupes sociaux, ou encore de définir l'influence d'un auteur.

- Aider à l'obtention de nouvelles données. Une des problématiques majeures de ce projet est l'acquisition d'un grand nombre de nouvelles données qui soient facilement utilisables. En effet, la campagne de crawling avec HERITRIX n'a pas permis d'obtenir de données textuelles « propres », tandis que la méthode d'identification de nouveaux blogs littéraires « à la main » est longue et peu envisageable dans l'optique éventuelle de l'utilisation de méthodes de Deep Learning, nécessitant de grandes quantités de données. Toujours dans cette éventualité, il y a aussi la question de la labellisation des données. Est-il possible de se mettre d'accord sur des catégories objectives qui permettaient d'étiqueter l'ensemble de notre corpus afin d'entrainer un classifieur ?

## 2. TRAVAIL FOURNI
### 2.1. ANALYSE THÉMATIQUE

La première approche utilisée pour traiter ces données textuelles a été la mise en place d'un modèle thématique dit Latent Dirichlet Allocation (LDA). Ce modèle se base sur deux hypothèses : chaque document est un mélange de thématiques, chaque thématique est un mélange de mot. Le but est alors l'apprentissage de ces distributions grâce aux données d'entrainement. Cette approche est dite probabiliste et générative car les thématiques ne sont pas définies à l'avance mais découvertes au travers de l'apprentissage. Plus concrètement, le modèle va parcourir l'ensemble des mots de l'ensemble des documents et considérer que des mots souvent présents dans les mêmes textes feront partie d'une même thématique. Une fois ces thématiques crées, il suffit de calculer la proportion de mots de chaque thématique dans un document pour connaître ses thématiques. Une description plus détaillée du modèle est fournie dans l'Annexe 5.

Ce modèle, basé sur la co-occurence de mots dans un même texte, nécessite certaines phases de préparation des données. En effet, chaque document doit passer d'un texte html à un vecteur à V dimensions, où V représente la taille du vocabulaire considéré.

« Ce ciel est beau … »
« Ah, ce sont de beaux chiens. » →

| Ce | ciel | est | beau | Ah | ce | sont | de | beaux | chiens |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

Figure 2-1. Vectorisation de phrases sous format Bag-of-worlds.

Un premier travail est donc de normaliser le texte en retirant la ponctuation et les majuscules, séparant les mots les uns des autres et éventuellement en effectuant une lemmatisation ou racinisation des termes, c'est à dire ne garder respectivement que leur lemme (forme infinitive masculin singulier) ou leur racine (forme sans déclinaison).

[« ce », « ciel », « est », « beau »]
[« ah », « ce », « est », « beau », « chien »] →

| ce | ciel | est | beau | ah | chien |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 | 1 |

Figure 2-2. Vectorisation de phrases, lemmatisation.

Une autre étape intéressante pour améliorer le modèle est de retirer les mots outils, qui ne sont pas porteur d'une thématique au sens intuitif du therme. Pour cela, nous avons utilisé la liste de 507 mots outils français de la librairie « fr_core_news_sm » de Spacy.

[«ciel », « beau »]
[« beau », « chien »] →

| ciel | beau | chien |
|---|---|---|
| 1 | 1 | 0 |
| 0 | 1 | 1 |

Figure 2-3. Vectorisation de phrases, suppression des mots outils.

En plus de ces problématiques classiques lors de la création d'un modèle LDA, nous avons dû faire face à certaines difficultés propres à notre jeux de données.

Tout d'abord, un problématique vis à vis de la la taille et la composition de certains textes. En effet, nous avons remarqué que les documents les plus longs étaient enfaite des compilations de textes, qui n'ont pas forcement de caractéristiques thématiques communes. Cela fausse alors notre apprentissage, le modèle considérant le document comme un texte unique cohérent avec peu de thématiques, et ce malgré la variation de sujets effectivement abordés dans les différents sous-textes. Il a donc été décidé de mettre à profil le fait d'avoir le format html des différentes

pages pour séparer les documents de plus de 500 caractères au niveau des indentations. Celles-ci étant disponibles au travers des balises html et permettant de créer des sous textes ayant un minimum de cohérence et de ne pas couper au milieu d'une phrase ou d'un paragraphe.

Une seconde considération est venue de certaines formes d'écritures particulières, qui comportent une importante répétition d'un mot ou d'un groupe de mots. Cette pratique a pour effet de totalement déséquilibrer le modèle, qui considère alors ce groupement de mots comme une thématique à part entière aux vues de son importante fréquence d'apparition. Afin de limiter cet effet, nous avons instauré une limitation de 10 occurrences de mot par texte maximum.

Une ultime question dans la paramétrisation du modèle est le choix du nombre de thématiques. Plusieurs méthodes automatiques existent notamment par minimisation de la perplexité du modèle, au sens de la théorie de l'information, ou maximisation de sa cohérence. Nous avons néanmoins choisis arbitrairement une valeur de 30 thématiques, ces deux indicateurs ne donnant pas de résultats particulièrement intéressants lors de l'analyse des résultats.

Après entraînement du modèle on obtient, pour chaque texte, la probabilité d'appartenance à chacune des trente thématiques, et pour chacune des thématiques, la distribution de probabilité sur l'ensemble des mots. Vient ensuite la phase d'interprétation, qui consiste dans notre cas à labelliser chacune des thématiques.

On utilise pour cela différents outils. On peut tout d'abord visualiser les mots ayant la densité de probabilité la plus importante pour chaque thématique. Mais on peut aussi extraire les mots les plus représentatifs d'une thématique. Pour cela, on associe chaque texte à sa thématique principale, celle de probabilité la plus élevée. Puis, pour chaque mot w de chaque thématique k, on associe le score suivant : $\lambda \,.\, log[f_k(w)] + (1-\lambda) \,.\, log[f_k(w)/f_{tot}(w)]$ où $f_k(w)$ représente la fréquence d'apparition de w dans le sous corpus des textes associés à la thématique k, $f_{tot}(w)$ étant la fréquence d'apparition de w dans tout le corpus. $\lambda$ est un paramètre permettant de régler l'importance donnée à l'apparition d'un mot dans une thématique ($\lambda = 1$) ou à sa spécificité envers une thématique ($\lambda = 0$). On choisira $\lambda = 0.5$ pour repartir les deux aspects de manière équitable. On peut effectuer le même travail après apprentissage des bi-gramme, c'est à dire après avoir regroupé les mots étant souvent utilisés à la suite. On obtient alors ce genre de résultats :

(a) (b) (c)


Figure 2-4.(a) Mots les plus représentatifs de la thématique labellisée « Musique ». (b) Bi-grammes les plus représentatifs de la thématique labellisée « Pétition en ligne ». (c) Bi-grammes les plus représentatifs d'une thématiques non labellisée.

On remarque que certaines thématiques sont alors facilement interprétables de par les mots qui les représentent, à l'image des deux premières. D'autres nécessitent de continuer le travail de recherche, en extrayant par exemple les textes les plus représentatifs de la thématique. On associe à chaque document i de chaque thématique k, le score suivant : $\theta_i(k) \,.\, N_i$ où $\theta_i(k)$ représente la probabilité du document i d'appartenir à k et $N_i$ le nombre de mots contenus dans ce document. La prise en compte de $N_i$ a pour but de contre balancer le fait que les petits textes contiennent moins de thématiques et obtiennent donc des valeurs de $\theta_i(k)$ plus élevées. Cela a permis de réaliser par exemple que la thématique non labelisée de la Figure 2-4.(c), ayant dans ses vingt textes les plus représentatifs un unique écrivain, était liée directement à cet auteur nommé Phan Kim Dien et dont le vocabulaire très particulier en a fait une thématique à part.

Certaines thématiques ont nécessité l'aide du laboratoire MARGE pour être labellisées, notamment grâce à leurs connaissances du monde littéraire. Un graph représentant l'ensemble de nos trente thèmes est fournie dans l'Annexe 7. Les valeurs de ce graph ont été obtenues en effectuant une mesure de similarité cosinus sur la distribution des mots entre chaque thématique. Un seuil minimal de 0.1 a été choisi pour la visualisation des arêtes, afin de ne pas obtenir un graph complet.

### 2.2. ANALYSE STYLISTIQUE

Une seconde approche pour l'analyse du corpus a été l'étude du style des différents textes. Cette approche a notamment été motivée par l'article publié par un doctorant du laboratoire ERIC et montrant l'intérêt de l'étude du style dans la représentation d'auteurs et de documents[1]. On considère pour cela que le style d'un texte peut être représenté par certains marqueurs calculables numériquement. La question du choix de ces marqueurs est centrale, car ils doivent être assez nombreux et variés pour capturer la pluralité des styles possibles, tout en restant au maximum indépendants d'autres considérations littéraires comme le thème. Cette hypothèse d'indépendance entre thématique et stylistique est vivement critiquable, le genre du haïku par exemple présente à la fois une stylistique et une thématique extrêmement cadrée, et la question de la possibilité de séparer ces deux composantes reste une question ouverte. L'article propose néanmoins une liste de marqueurs stylistiques issus de la littérature scientifique actuelle, tous aussi indépendants que possible de la thématique, et repartis en huit catégories :

- Lettres : fréquence d'apparition de chaque lettre dans le texte.
- Nombres : fréquence d'apparition de chaque nombre dans le texte.
- Structure : taille moyenne des mots utilisés, nombre de syllabes, …
- Ponctuation : fréquence d'apparition de chaque signe de ponctuation dans le texte.
- Mots outils : fréquence d'apparition de chaque mot outil (et, est, ayant, …) dans le texte.
- POS Tag : fréquence d'apparition de chaque catégorie de l'étiquetage morpho-syntaxique (PRO, ADJ, NAM, …) dans le texte.
- Ner : fréquence d'apparition de chaque catégorie de la reconnaissance d'entités nommées (personne, ville, pays, entreprise, …) dans le texte.
- Indexes : indexes de complexité et de lisibilité du texte (indexe de Simpsons, test de lisibilité Flesch-Kincaid, …).

Cette première liste de marqueurs stylistiques, issue de la littérature anglophone, a été adaptée au français et affinée, notamment dans la variété des conjugaisons possibles dans les catégories morpho-syntaxiques, sur les conseils des experts du laboratoire MARGE. Il en résulte une liste de 775 marqueurs stylistiques, dont la liste complète est donnée dans l'Annexe 6.

On peut alors utiliser ces marqueurs stylistiques pour faire de la clusterisation sur notre jeux de données et regrouper les auteurs ayant des styles similaires. Mais on peut aussi s'en servir comme outil pour repérer des différences stylistiques entre communautés. Reprenons par exemple notre clusterisation des documents basée sur leur thématique principale. En moyennant les valeurs stylistiques des documents appartenant à une même thématique, on obtient un vecteur de style pour chaque thème. En le comparant à la moyenne effectuée sur l'ensemble des documents, et grâce à une analyse statistique type test de Student, on peut extraire quelles valeurs stylistiques sont particulièrement significatives par rapport à la distribution normale.

On remarque sur la Figure 2-5.(a) que le cluster labellisé « pkd » qui correspond à l'auteur Phan Kim Dien possède des valeurs stylistiques significativement extrêmes (représentées par des cercles) sur pratiquement tous les marqueurs, rendant compte encore une fois de sa particularité d'écriture. Quand au cluster « langue orale » présent sur la seconde figure, il possède une valeur élevée pour les mots « me », « ça » et pour les mots outils en général (func_w_freq). Et ce au contraire du cluster « tournois nouvelles » qui comporte des compilations de nouvelles et qui présente des valeurs significativement faibles pour ces marqueurs.

[1] : E. Terreau, A. Gourru, J. Velcin, Writing Style Author Embedding Evaluation, EMNLP (Eval4NLP) 2021.

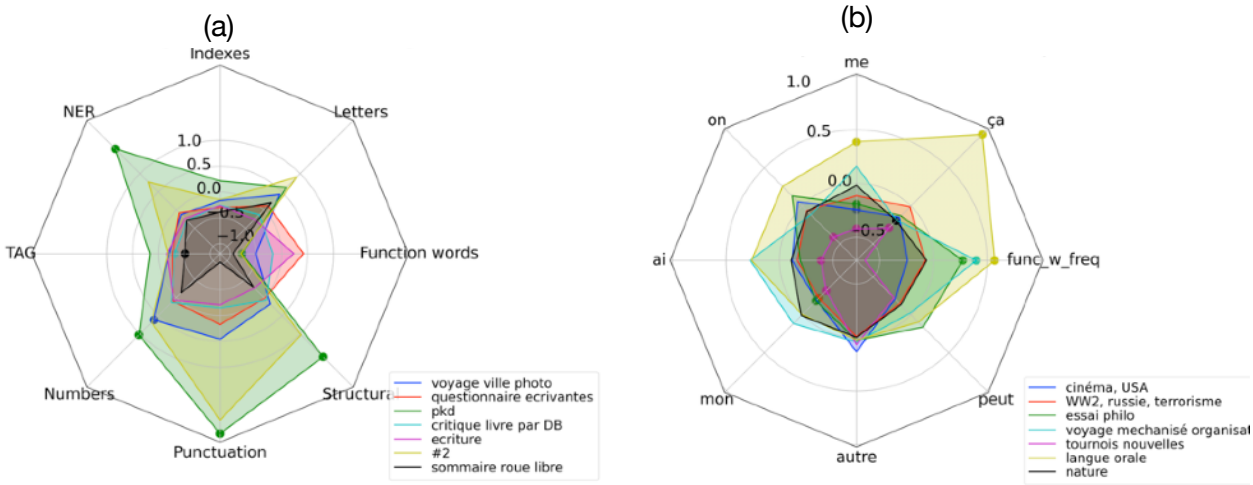



Figure 2-5. (a) Comparaison stylistique agrégée de sept thématiques différentes. (b) Comparaison stylistique non agrégée de sept autres thématiques.

Ce nouvel outil est plus globalement utilisable dans n'importe quel contexte nécessitant l'extraction et l'analyse de marqueurs stylistiques dans une base de données contenant des textes regroupés par catégorie, par auteur classiquement.

### 2.3. ANALYSE SOCIALE

Comme mentionné précédemment, une des pistes avancées pour enrichir l'analyse du corpus était de créer un réseaux social des auteurs. Ce réseaux serait classiquement représenté sous forme de graph, dont les noeuds seraient les auteurs et les arêtes les liens sociaux entre eux. Ayant récupéré les commentaires des différents posts, une première idée a été de considérer qu'un auteur était connecté à un autre si et seulement si il avait commenté un post de ce second auteur. Malheureusement cette première approche fut infructueuse aux vues du nombre relativement réduit d'auteurs laissant des commentaires sur d'autres blogs.

La seconde idée fut alors de remarquer que certains blogs fournissaient, sur leur page, la liste des profils de tous leurs abonnés. En cliquant sur un profil, on a alors accès à l'ensemble des blogs auxquels le profil est abonné. De même, ayant accès au profil de tous les auteurs des posts et des commentaires de notre base de données, il est possible d'obtenir, pour chacun, la liste de ses abonnements. On pourra donc créer un graph social des blogs, en considérant que deux blogs sont reliés si et seulement si ils possèdent un abonné en commun, ce nombre d'abonnés étant le poids du lien entre les deux blogs.

Ces données sur les abonnés n'étant pas disponibles au travers de l'API de Blogger, il a été

```
liste_blogs
bdd_abonnés

Pour blog dans liste_blogs :
        récupération Liste_abonnés
        Pour abonné dans Liste_abonnés :
                récupération Liste_abonnements
                récupération pseudo
                récupération id_abonné
                Si id_abonné n'est pas dans bdd_abonnés :
                        ajout de (id_abonné, pseudo, Liste_abonnements) dans bdd_abonnés
```

Figure 2-6. Algorithme de scrapping des listes d'abonnements.

nécessaire d'implémenter un robot de scrapping sur le modèle suivant :
À ce point, il est nécessaire d'expliciter que si ces données ne sont pas disponibles au travers de l'API de Blogger c'est que certaines d'entre elles ne sont pas sensées pouvoir être récoltées automatiquement. On en a d'ailleurs la confirmation en consultant la page https://

www.blogger.com/robots.txt qui définie ce qu'il est possible ou pas de récolter sur les sites Blogger. Nous précisons donc que cette campagne de scrapping a été réalisée uniquement dans un but scientifique et que les données récoltées n'ont pas vocation à être rendues publiques ou à être utilisées de quelque autre façon que ce soit.

Le nombre de blogs rendant disponible la liste de leurs abonnés étant relativement limité, nous avons élargie notre scrapping au niveau supérieur, en réitérant le même procédé sur la liste de tous les abonnements des abonnés à nos blogs de départ.

On obtient alors un réseaux de 38 127 blogs, dont 112 faisant partie de nos blogs de départ. On peut alors s'intéresser au sous graph comportant nos 112 blogs littéraires.

(a) (b)

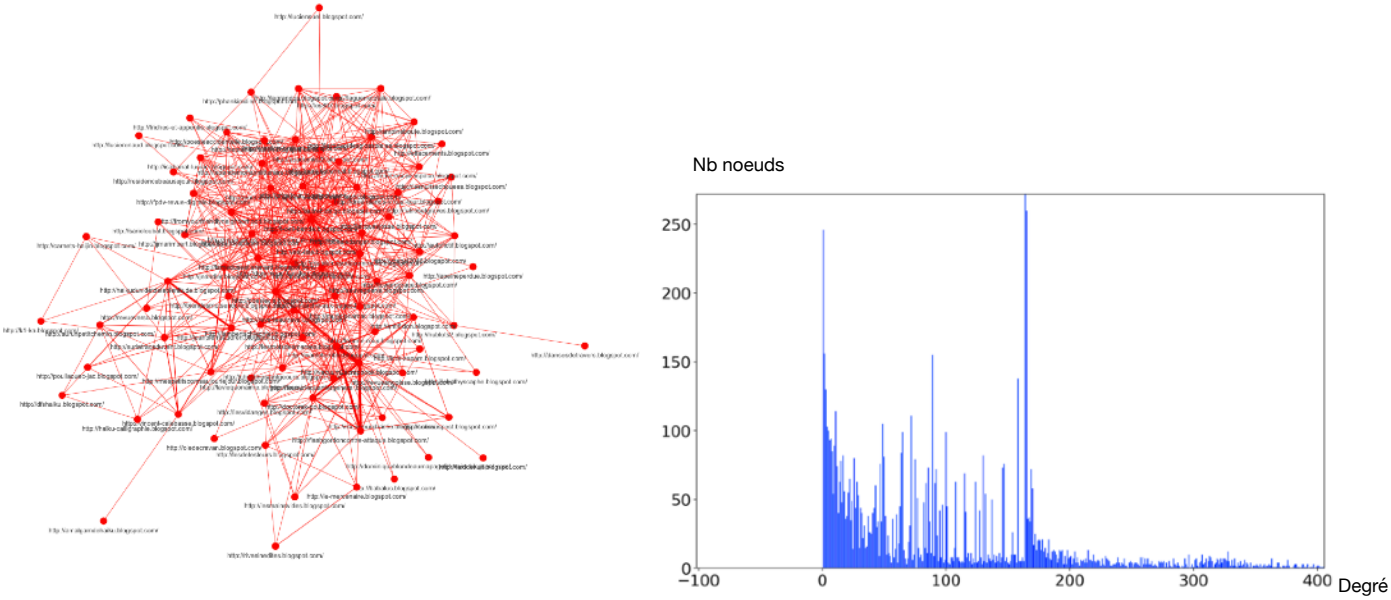



Figure 2-7. (a) Sous-graph des 112 blogs littéraires.
(b) Distribution du degré des noeuds au sein du graph entier.

Ce sous graph est connexe mais il est largement déséquilibré par les biais introduits lors de sa création. En effet, n'ayant pas accès à l'ensemble des données nécessaires à sa construction, une grande partie des liens est manquante. Les noeuds se divisent donc clairement en deux ensembles : ceux dont le blog a pu être scrappé et qui ont donc un grand nombre de liens, et ceux qui n'ont pas été scrappés et qui ont donc été découverts au travers des abonnés d'autres blogs, ils n'ont logiquement que peu de liens. On observe bien ces deux ensembles dans la distribution du degré des noeuds, c'est à dire leur nombre de voisins, avec une première distribution centrée sur 0 et une seconde sur 180.

## 3. EXPLOITATION DES MODÈLES

### 3.1. DESCRIPTION DU CORPUS

Les trois approches décrites précédemment, thématique, stylistique et sociale, sont autant d'outils nous permettant de décrire notre corpus, notamment par l'extraction de cluster, c'est à dire des regroupements de textes, de blogs ou d'auteurs partageant des caractéristiques communes.

On peut tout d'abord utiliser le modèle thématique, en associant à chaque texte un vecteur thématique de taille 30, notre nombre de topics. Dans ce nouvel espace à 30 dimensions, on peut alors entrainer un modèle de clustering simple, dans notre cas un K-means à 40 clusters. Le résultat est visualisé en deux dimensions grâce à l'algorithme t-SNE, une méthode classique pour la visualisation de données de grandes dimensions.

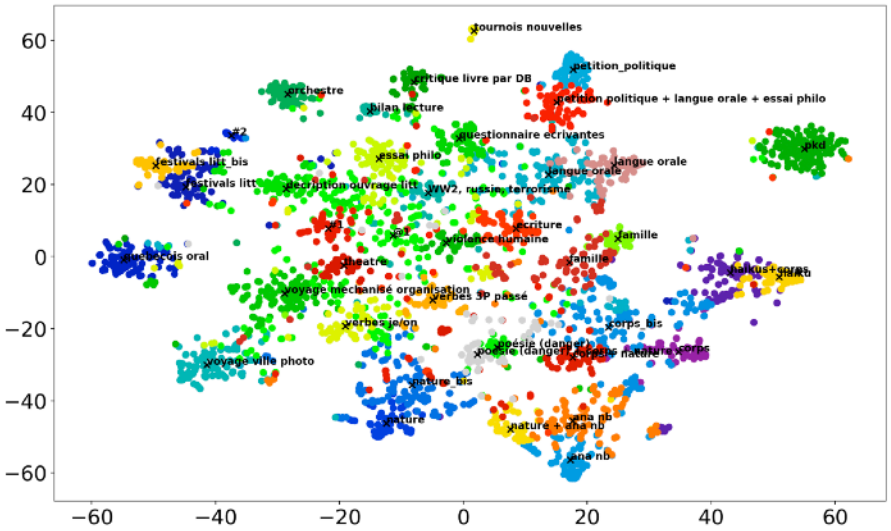



Figure 3-1. Visualisation des 40 clusters obtenus dans l'espace des thématiques ainsi que leurs centroïdes.

La plupart des clusters étant définis par une ou deux thématiques, leur analyse et définition est assez simple, puisqu'elle reprend l'analyse de la thématique en question.

Sur le même principe, on peut associer chaque texte à un vecteur stylistique de taille 775, notre nombre de marqueurs stylistiques. En répétant le même procédé, avec un K-means à 10 clusters, on obtient le résultat suivant :

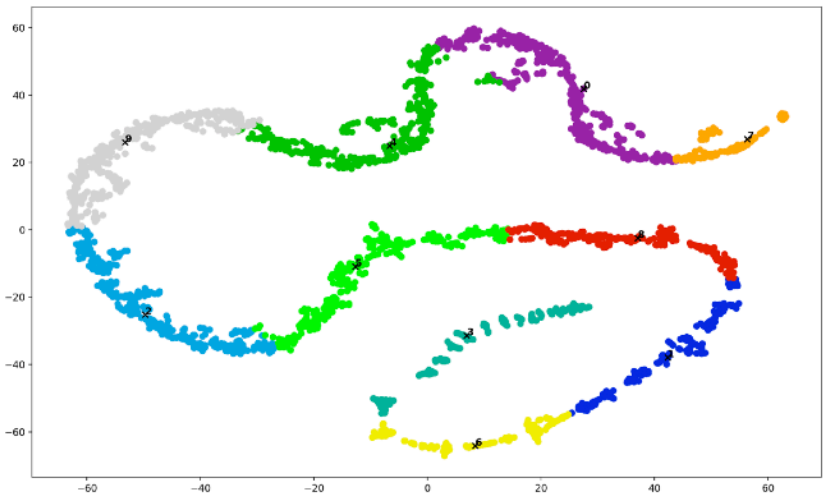



Figure 3-2. Visualisation des 10 clusters obtenus dans l'espace des marqueurs stylistiques ainsi que leurs centroïdes.

Cette fois-ci les clusters sont plus difficilement interprétables car faisant appel à des données plus techniques. Néanmoins, grâce aux coordonnées des centroïdes dans l'espace stylistique, une analyse des mots, bi-grammes et textes les plus représentatifs de chaque cluster ainsi que l'apport littéraire du laboratoire MARGE, il semble possible de mettre en place une labélisation des différents clusters, comme cela a été fait avec les thématiques.

Enfin, concernant l'étude du graph social, nous avons déjà présenté le fait qu'il était incomplet de par sa construction, ce qui en rend l'étude particulièrement complexe. Par exemple, une méthode classique pour la détection de communautés dans un réseaux social est celle de Girvan-Newman, qui se base sur la suppression des arêtes d'indice d'intermédiarité la plus élevée. Cette approche est ici relativement décevante, la plupart des clusters obtenus n'étant composés que d'un individu. D'autres méthodes, basées sur la modularité du réseaux, sont plus efficaces. C'est le cas de l'algorithme multilevel utilisé lors de la clusterisation présentée ci dessous, et qui présente cinq clusters.

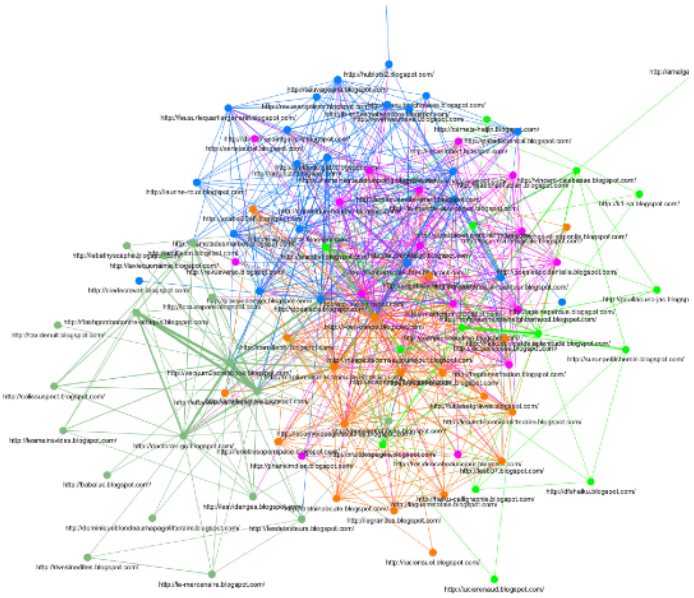


Figure 3-3. Sous graph des 112 blogs littéraires, 5 clusters.

Après une rapide analyse des blogs composants chacune de ces communautés, on peut définir celle en vert kaki comme les auteurs Québécois et celle en vert clair comme les auteurs de haïkus. Les trois autres n'étant pas facilement identifiables, elles nécessiteraient elles aussi l'analyse des experts du laboratoire MARGE.

On peut ajouter à ces trois visualisations l'apport des différentes métadonnées obtenues au travers de l'API et de la campagne de scrapping. Ainsi, on peut par exemple se représenter la distribution de certains auteurs ou d'une certaine région dans l'espace des thématiques.

(a) (b)

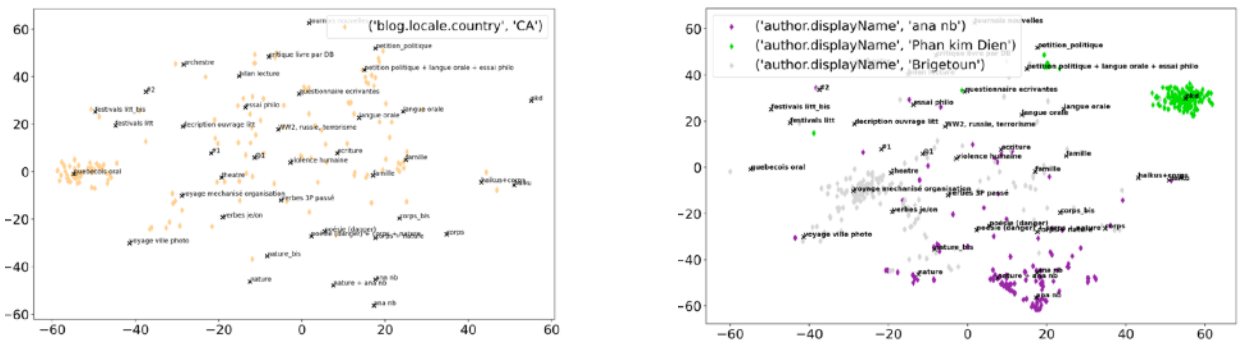



Figure 3-4. (a) Visualisation des textes ayant donné pour localisation le Canada. (b) Visualisation des textes des trois auteurs les plus prolifiques.

### 3.2. CONSTRUCTION D'UNE PIPELINE

Nous l'avons déjà mentionné, l'un des objectifs forts du projet LIFRANUM est la construction et l'enrichissement d'un corpus de littérature francophone numérique. L'idée de developper des outils permettant, au minima, de semi automatiser cet enrichissement, surgit alors de manière naturelle. Cette augmentation de notre base de donnée permettrait aussi d'enrichir l'apprentissage de nos modèles qui sont, par construction, limités à nos données actuelles.

Il s'agirait alors d'un problème de classification binaire : étant donné un nouveau texte, fait il oui ou non partie de la littérature francophone ? La difficulté réside alors dans le fait qu'il est nécessaire d'avoir de nombreux exemples de chacune des deux classes pour pouvoir construire ce genre de classifieur. Or, si nous possédons bien un corpus entier de blogs étiquetés comme faisant partie de la catégorie « littéraire », nous n'avons pas de corpus de blogs « non-littéraires », ce qui rend la piste du classifieur bien moins pertinente.

Nous nous sommes donc orienté vers un système de scoring, plus un nouveau document aura un score fort, plus sa probabilité de faire partie de notre corpus littéraire sera élevée. Il suffira alors à un expert d'analyser les blogs les mieux classés avant de décider si oui ou non ils devraient être intégrés à la base de données LIFRANUM. À noter que l'analyse des documents ayant les scores les plus faibles ainsi que le rejet de certains textes pour le corpus nous permettra d'obtenir des exemples négatifs.

L'idée principale de notre outil est la suivante : plus un blog est proche de notre corpus, du point de vue thématique, stylistique, du contenu et du réseau social, plus sa probabilité d'être un candidat valable semble élevée. On construit alors quatre scores, correspondant à chacune de ces quatre dimensions d'analyse, pour chaque nouveau blog, la compilation de ces quatre scores donnera alors le score final.

Concernant le score relatif aux thématiques, il faut tout d'abord noter que, au cours d'échanges avec le laboratoire MARGE, certaines thématiques ont été définies comme n'étant pas littéraires. Par exemple, la thématique « Pétition en ligne » qui se retrouve fortement dans des posts politiques invitant à signer des pétitions en ligne, n'est que très peu présente dans d'autres types de textes. Sa présence rend donc peu probable le fait d'avoir découvert un texte littéraire. On notera que, si tous les blogs de notre corpus ont été définis comme contenant de la littérature, l'ensemble de leurs posts ne sont pas de la littérature, d'où l'émergence de ces thématiques non littéraires. On peut alors donner à chacune de nos trente thématiques un « indice de littérature » compris entre 1 et -1. Avec un indice de 1 la thématique est considérée comme littéraire, -1 non-littéraire et 0 neutre. Soit $u$ ce vecteur des trente indices et $v$ la distribution thématique d'un texte. On donne alors à ce texte le score thématique de $S_t = u^T . v$

Pour le style, on commence par entrainer un One-class-SVM sur l'ensemble de nos données, dans l'espace des indices stylistiques. L'objectif du One-class-SVM, qui se base sur le classifieur SVM, est de détecter les nouvelles données anormales au sens statistique par rapport à la distribution d'apprentissage. Ce classifieur construit ainsi une frontière autours des données d'entrainement et, pour chaque nouveaux point à classifier, calcul la distance de ce point à la frontière. Si cette distance est positive, le point est du même coté de la frontière que les données d'apprentissage, il est donc considéré comme « normal », si la distance est négative, il est rejeté. On utilisera ici cette distance algébrique comme valeur pour le score stylistique $S_s$ de chaque nouveau texte.

Une méthode plus classique pour placer des textes dans des espaces est de faire appel à des modèles d'embedding de mot. Ces méthodes, basées sur de l'apprentissage par réseau de neurones, permettent de placer les mots dans un espace de dimensions réduites devant la taille du vocabulaire tout en faisant en sorte que des mots étant utilisés dans des contextes similaires soient proches. Il existe pour cela plusieurs implémentations, une des plus en vogue étant nommée BERT, qui a la particularité d'avoir été pré-entrainée sur de gigantesques bases de données du type Wikipedia. On peut alors définir la position de phrases ou de textes dans l'espace d'embedding comme étant la moyenne des positions des mots les composant. Ayant positionné nos textes du corpus dans ce nouvel espace, on peut appliquer le même processus

que précédemment, avec un One-class-SVM chargé d'évaluer la distance des nouveaux textes à nos textes d'apprentissage. On obtient alors un score de contenu $S_c$.

Concernant le score $S_r$ en lien avec l'aspect social, plusieurs pistes de définition sont encore à l'étude. Il pourrait s'agir d'un simple calcul de plus court chemin ou du nombre de liens entre le nouveaux blogs et ceux de notre corpus. Ou alors, et à l'image du score précédent, il serait imaginable de plonger nos blogs dans un espace d'embedding conservant les propriétés de voisinage de notre graph. $S_r$ serait alors encore une fois calculé grâce à un One-class-SVM comme étant la distance algébrique à la distribution d'apprentissage.

En compilant les différents indices de l'ensemble des posts ainsi que du blog, on obtient notre score final $S_f$. Cette compilation pourrait consister, par exemple, en une moyenne pondérée.

On fourni dans l'Annexe 8 un schéma résumant la construction de notre pipeline de scoring.

**CONCLUSION ET PERSPECTIVES**

L'objectif de ce stage était de fournir des outils d'analyse pour le corpus de blogs littéraires du projet LIFRANUM. Ces outils devaient notamment être capable de faire émerger des regroupements de textes, de blogs ou d'auteurs afin d'aider à la future structuration du corpus. L'objectif était aussi d'explorer de nouvelles pistes comme la création d'un graph social ou la prédiction de nouveaux blogs littéraires.

Nous avons ainsi produit :

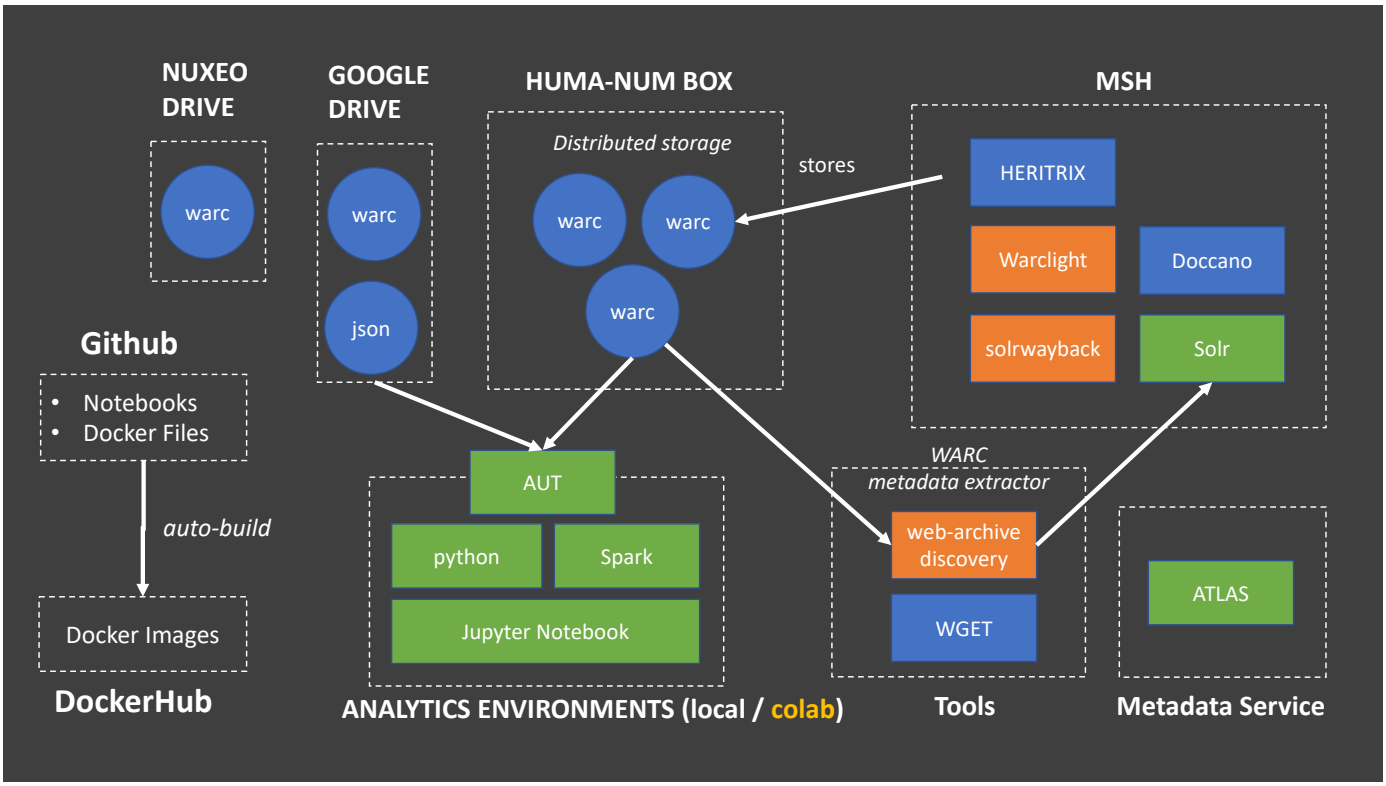
- Trois outils permettant de clusteriser le corpus selon une approche thématique, stylistique et sociale.
- Une méthode pour labelliser ces clusters, à partir de l'analyse des mots, n-grammes et textes les plus représentatifs. Méthode qui a déjà permis de labelliser les clusters thématiques.
- Un modèle de pipeline visant à détecter des blogs proches de notre corpus de départ et qui pourraient potentiellement venir s'y ajouter.

De nombreuses pistes restent encore à explorer :
- Analyser et labelliser les clusters stylistiques et sociaux, à l'aide notamment du laboratoire MARGE.
- Effectuer une analyse détaillée de notre pipeline, du réglage des ses nombreux hyper-paramètres (dimension des espaces d'embedding, construction des SVM, choix de la méthode de compilation des scores, …), à la quantification des performances des différents blocs (quelle est l'influence de chacun dans le score final ?) pour ensuite l'utiliser concrètement sur notre corpus de départ.
- Obtenir des données labellisées afin de faire de l'apprentissage supervisé. Les labels de nos différents clusters pourraient venir alimenter cette réflexion, donnant plusieurs pistes de comment structurer le corpus et montrant la diversité d'approches et d'analyses qu'il est possible d'effectuer.

Du point de vue personnel, ce stage aura été pour moi une grande source d'enrichissement. Un enrichissement scientifique tout d'abord, travailler au seins du laboratoire ERIC au contact d'experts dans les domaines des sciences des données et des humanités numériques m'aura permis de me familiariser rapidement avec des modèles complexes, issus de la littérature scientifique contemporaine. Ce fut aussi une belle introduction au monde de la recherche, à ses codes et à ses méthodologies. Enfin, ce fut aussi l'occasion d'évoluer dans un contexte hautement pluridisciplinaire, notamment au contact du laboratoire MARGE, avec tout ce que cela implique du point de vue de la communication, de la vulgarisation et plus généralement de la gestion de projet.

**ANNEXES**

## Annexe 1 : Ecosystem de LIFRANUM



**Annexe 2 : Exemple de block « réponse » d'un fichier WARC**

WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Warcinfo-ID: <urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>
WARC-Date: 2006-09-19T17:20:24Z
WARC-Block-Digest: sha1:UZY6ND6CCHXETFVJD2MSS7ZENMWF7KQ2
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-IP-Address: 207.241.233.58
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: application/http;msgtype=response
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 1902

HTTP/1.1 200 OK
Date: Tue, 19 Sep 2006 17:18:40 GMT
Server: Apache/2.0.54 (Ubuntu)
Last-Modified: Mon, 16 Jun 2003 22:28:51 GMT
ETag: "3e45-67e-2ed02ec0"
Accept-Ranges: bytes
Content-Length: 1662
Connection: close
Content-Type: image/jpeg

[image/jpeg binary data here]

**Annexe 3 : Description complète des attributs sur blogger**

## Blogs Ressource :

```
{
 "kind": "blogger#blog",
 "id": value,
 "name": value,
 "description": value,
 "published": value,
 "updated": value,
 "url": value,
 "selfLink": value,
 "posts": {
  "totalItems": value,
  "selfLink": value
 },
 "pages": {
  "totalItems": value,
  "selfLink": value
 },
 "locale": {
  "language": value,
  "country": value,
  "variant": value
 }
}
```

## Post Ressource :

```
{
 "kind": "blogger#post",
 "id": string,
 "blog": {
  "id": string
 },
 "published": datetime,
 "updated": datetime,
 "url": string,
 "selfLink": string,
 "title": string,
 "titleLink": string,
 "content": string,
 "images": [
  {
   "url": string
  }
 ],
 "customMetaData": string,
 "author": {
  "id": string,
  "displayName": string,
  "url": string,
  "image": {
   "url": string
  }
 },
 "replies": {
  "totalItems": long,
  "selfLink": string,
  "items": [
   comments Resource
  ]
 },
 "labels": [
  string
 ],
 "location": {
  "name": string,
  "lat": double,
  "lng": double,
  "span": string
 },
 "status": string
}
```

## Comment Ressource :

```
{
 "kind": "blogger#comment",
 "status": string,
 "id": string,
 "inReplyTo": {
  "id": string
 },
 "post": {
  "id": string
 },
 "blog": {
  "id": string
 },
 "published": datetime,
 "updated": datetime,
 "selfLink": string,
 "content": string,
 "author": {
  "id": string,
  "displayName": string,
  "url": string,
  "image": {
   "url": string
  }
 }
}
```

**Annexe 4 : Exemple de données issues d'un post et d'un commentaire**

| | Data 1 | Data 2 |
|---|---|---|
| **kind** | blogger#post | blogger#comment |
| **published** | 2008-11-25T16:00:00+01:00 | 2008-11-25T17:53:33+01:00 |
| **updated** | 2008-12-23T01:01:02+01:00 | 2008-11-25T17:53:33+01:00 |
| **url** | http://versminuit.blogspot.com/2008/11/rien-lire.html | Nan |
| **title** | Rien à lire | Nan |
| **content** | Il n'y a rien à lire sur ce blog, c'est frustrant, j'en conviens. Sachez que j'y travaille : je cherche des photos pour faire plus gai et des textes littéraires pour faire sérieux. Il y a tout de même quelques liens à droite qui vaillent qu'on clique dessus. Rendez-vous prochainement, vers minuit. | Rien à lire, mais... tout à espérer?! <BR/>Bonne chance à ce blog, Franck (entre nous, quel beau titre!) |
| **labels** | ['blog'] | Nan |
| **author.displayName** | fg | M agali |
| **author.url** | https://www.blogger.com/profile/17264828114690065245 | https://www.blogger.com/profile/07996811380904857505 |
| **blog.name** | Vers minuit | Vers minuit |
| **blog.description** | Franck Garot, littérature | Franck Garot, littérature |
| **blog.locale.language** | fr | fr |
| **blog.locale.country** | Nan | Nan |

**Annexe 5 : Modèle LDA**

On défini les variables suivantes :

$M$ : le nombre de documents
$N_i$ : le nombre de mots dans le document i
$K$ : le nombre de thématiques
$V$ : taille du vocabulaire sur l'ensemble des documents
$\alpha$ : paramètre de Dirichlet à prioris de la ditribution des documents sur les topics
$\beta$ : paramètre de Dirichlet à prioris de la ditribution des topics sur les mots
$\theta_i$ : la distribution sur les topics du document i
$\varphi_k$ : la distribution sur les mots du topic k
$z_{ij}$ : la thématique du j-ème mot du i-ème document
$w_{ij}$ : le j-ème mot du i-ème document

On considère alors que le corpus a été construit de la manière suivante :

1) Pour chacun des M documents :
   $\theta_i \sim Dir(\alpha)$ avec $i \in \{1,...,M\}$ où $Dir(\alpha)$ représente une distribution de Dirichlet de paramètre $\alpha < 1$
2) Pour chacun des K topics :
   $\varphi_k \sim Dir(\beta)$ avec $k \in \{1,...,K\}$ et $\beta < 1$
3) Pour chacune des positions de mot (i, j) avec $i \in \{1,...,M\}$ et $j \in \{1,...,N_i\}$ :
   - On choisit un topic $z_{ij} \sim Multinomiale(\theta_i)$
   - On choisit un mot $w_{ij} \sim Multinomiale(\varphi_{z_{ij}})$

On justifie la supposition que $\alpha$ et $\beta < 1$ par l'hypothèse qu'un texte ne contient que certaines thématiques et une thématique ne contient que certains mots.

On peut représenter ce modèle de dépendance entre variables par le schéma ci dessous. W étant la seule distribution grisée car c'est la seule que l'on connait effectivement, ce sont les mots de notre corpus de documents.

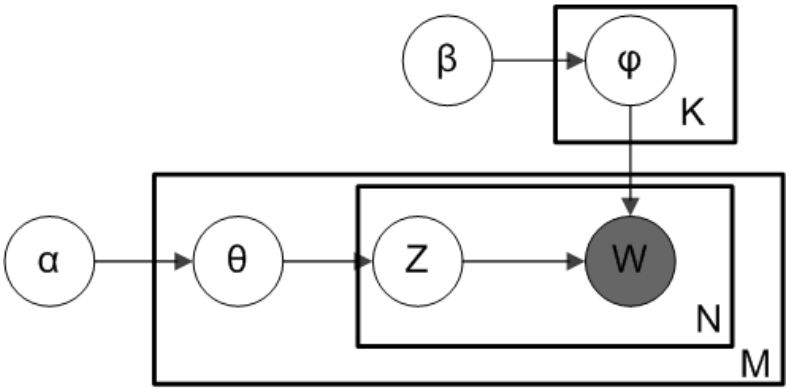


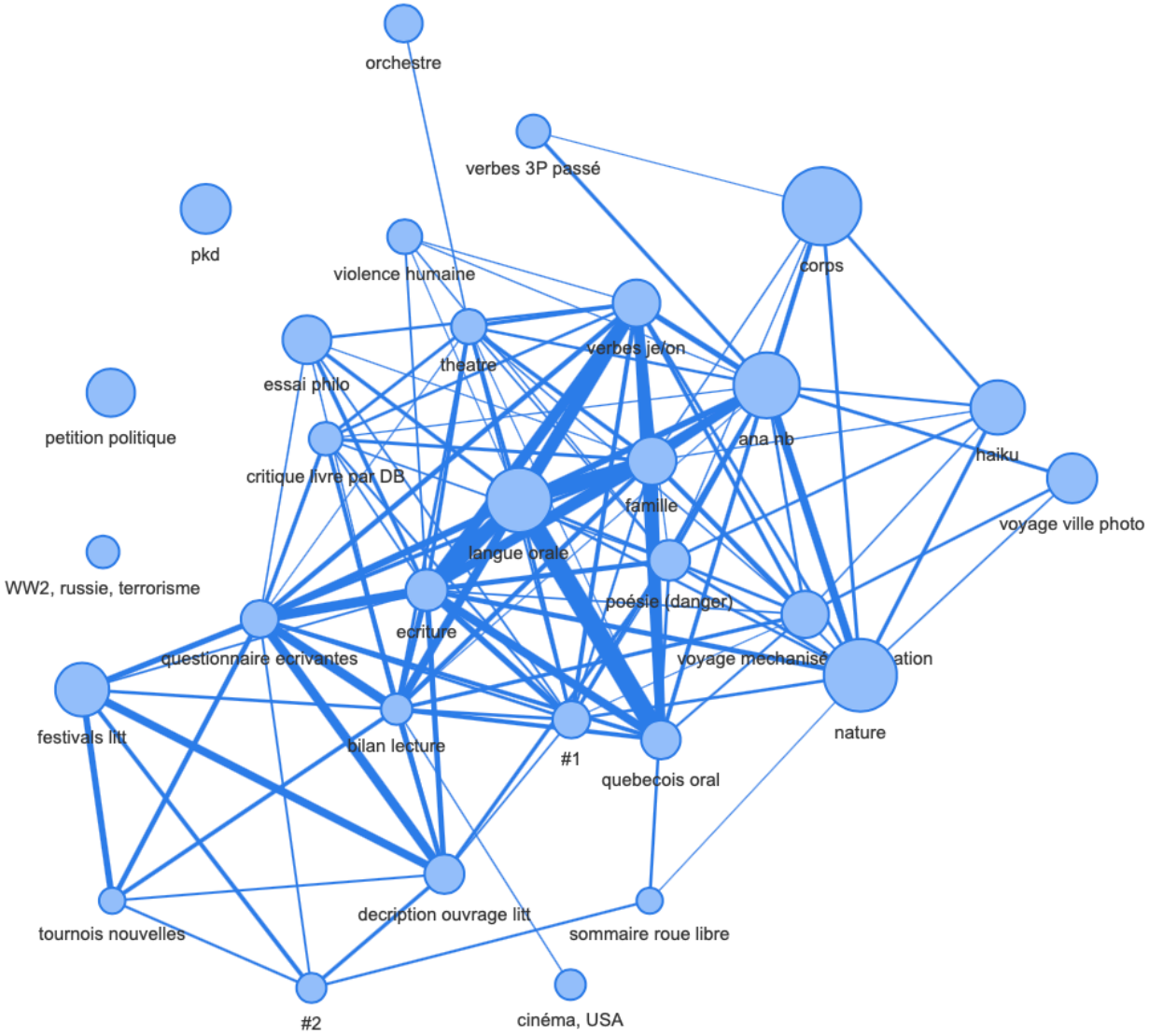

Schéma du modèle LDA.

C'est alors à partir de W que les autres distributions sont apprises par inférence statistique. Différentes techniques sont utilisées, par échantillonnage (méthodes de type MCMC) ou par optimisation. Celle utilisée par la librairie Gensim est une technique d'optimisation issue des méthodes bayésiennes variationnelles, adaptée spécifiquement pour être utilisable sur de grands jeux de données.

**Annexe 6 : Liste des marqueurs stylistiques (extrait)**

| structure | lettres | nombres | mots outils | indices | TAG | ponctuation | NER |
|---|---|---|---|---|---|---|---|
| avg_w_len | a | 0 | dis | hapax | ADJ\| | ! | LOC |
| tot_short_w | b | 1 | à | yules_K | ADV\| | " | MISC |
| tot_digit | c | 2 | â | shannon_entr | CCONJ\| | £ | ORG |
| tot_upper | d | 3 | abord | simposons_ind | DET\| | € | PER |
| func_w_freq | e | 4 | afin | flesh_ease | INTJ\| | # | |
| avg_s_len | f | 5 | ah | flesh_cincade | NOUN\| | $ | |
| syllable_count | g | 6 | ai | gunnin_fox | PART\| | % | |
| avg_w_freqc | h | 7 | aie | | PRON\| | & | |
| | i | 8 | ainsi | | PUNCT\| | ( | |
| | j | 9 | ait | | SCONJ\| | ) | |
| | k | | allaient | | SPACE\| | * | |
| | l | | allons | | SYM\| | + | |
| | m | | alors | | X\| | , | |
| | n | | anterieur | | ADJ\|Gender=Fem\|Number=Plur | - | |
| | o | | anterieure | | ADJ\|Gender=Fem\|Number=Sing | . | |
| | p | | anterieures | | ADJ\|Gender=Fem\|NumType=Ord\|Number=Plur | / | |
| | q | | apres | | ADJ\|Gender=Fem\|NumType=Ord\|Number=Sing | : | |
| | r | | après | | ADJ\|Gender=Masc | ; | |
| | s | | as | | ADJ\|Gender=Masc\|Number=Plur | < | |
| | t | | assez | | ADJ\|Gender=Masc\|Number=Sing | 0 | |
| | u | | attendu | | ADJ\|Gender=Masc\|NumType=Ord\|Number=Plur | > | |
| | v | | au | | ADJ\|Gender=Masc\|NumType=Ord\|Number=Sing | ? | |

**Annexe 7 : Graph des thématiques**



**Annexe 8 : Schéma pipeline**