Livrable 2 - Projet Lifranum

Filippo Sarra, Zoé Noël, Brunel Tchekeli, Ambrine Soltani

23 janvier 2023

Commanditaires : Pantel Alice, Terreau Enzo
Tuteur : Velcin Julien

Table des matières

1 Introduction

1.1 Présentation du projet Lifranum

Dans le cadre du master 2 Humanités Numériques coordonné par Lyon 2, Lyon 3, l'ENS et l'ENSSIB, l'unité d'enseignement "Projet" a été confié aux étudiants et étudiantes. Accompagnés de leur tuteur (professeur du master), les étudiants et étudiantes ont choisi, en fonction de leurs qualifications et de leurs préférences, un projet sur lequel travailler durant quatre mois. Le but était de répondre à un besoin explicité par le commanditaire. Les missions confiées aux étudiants peuvent être variées et correspondent aux fonctionnements et pratiques mises en place par les commanditaires. Il est donc attendu des étudiants une grande capacité d'adaptation et d'organisation.

Le projet Lifranum est porté par le laboratoire en sciences humaines Marge, le laboratoire en sciences informatiques Eric et la bibliothèque nationale de France BnF. Ce projet est financé par l'agence nationale de recherche et a pour but de constituer et analyser un corpus de production littéraire francophone nativement numérique. Lifranum se propose de porter un regard scientifique sur la production littéraire numérique dans le contexte de la francophonie. Face à l'envergure de ces productions, peu d'études ont pour le moment été menées. Le projet se positionne ainsi afin de répondre aux questions suivantes [1] : Comment les auteurs 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 ?

Avant de répondre à ces questions il est essentiel de pouvoir constituer le corpus sur lequel porte l'étude et trouver un moyen d'analyser cet ensemble conséquent de données textuelles. C'est ainsi que le partenariat entre sciences humaines et sciences de l'informatique a été noué. Face à la quantité conséquente de données présentes sur le web, il est nécessaire de pouvoir composer et traiter, au moins en partie, le corpus de façon automatique. Afin de constituer un premier corpus d'analyse, l'équipe de recherche de Lifranum a fait appel aux APIs des hébergeurs de blogs Wordpress et Blogger. Une quantité très conséquente de textes a été extraite, il est donc nécessaire de faire usage des outils d'analyse automatique afin d'extraire des catégories et des thématiques permettant l'étude de cet important corpus.

La pertinence de la pluridisciplinarité de ce projet est à ce stade évident. En effet, l'analyse qualitative manuelle est à ce point impossible vu la quantité de textes extraite, mais aussi à cause du format même des documents extraits, qui

1. https ://projet-lifranum.univ-lyon3.fr/projet

doivent être nettoyés et trier pour permettre l'analyse.

C'est à ce stade que l'équipe du projet a fait appel à des étudiants afin de réaliser les premiers traitement sur ce corpus. Un premier stage a été réalisé entre septembre 2021 et janvier 2022 par l'étudiant en informatique Théo Gady. Il a proposé un traitement automatique des données du corpus. Il a ainsi extrait un ensemble de thématiques représentatives du corpus potentiellement exploitables. Notre équipe a donné suite à ce premier stage à partir d'octobre 2022. Nous nous sommes proposés d'intervenir sur les données extraites de Blogger. Notre travail s'est donc étendu sur quatre mois, en parallèle de nos études respectives.

1.2 Méthodologie

Pour mener à bien ce projet, nous avons opté pour une méthodologie de gestion de projet en cascade (méthode Gantt). Cette méthodologie de gestion de projet consiste en une série d'étapes successives allant de l'exploration des données à l'analyse profonde du corpus, en passant par des réunions d'étape et des livrables. Nous avons également utilisé le dépot `github` afin de pouvoir partager facilement avec nos commanditaires tous types de documents, notamment nos carnets jupyter.

L'exploration nous a permis de prendre connaissance du jeu de donnée afin d'envisager les parties qu'il serait intéressant de garder pour notre analyse. Au vu de la structure du jeu de données, il nous est apparu que seule la balise *content* contenait du texte exploitable. Les autres balises ne contenant que très peu d'informations, elles ont été mises à l'écart. Une fois l'étape exploratoire terminée et le corpus constitué, nous nous sommes résolu à utiliser une méthode combinant l'analyse statistique des données textuelles et l'apprentissage automatique. Ceci afin d'étudier les tendances de thématiques et de style dans notre corpus de textes littéraires. Ce choix de méthode part des hypothèses de recherche suivantes :

- Il pourrait exister des thématiques transversales identifiables dans un corpus.
- Les publications littéraires nativement numérique pourraient contenir des marqueurs thématiques en fonction des tendances marquant la périodicité des publications.
- Les publications littéraires francophones nativement numériques d'un même auteur pourraient se rattacher aux mêmes thématiques.
- De même il pourrait être possible d'identifier un auteur de littérature francophone nativement numérique à partir de ses thématiques de prédilection.
- Les styles d'écriture des auteurs de textes francophones nativement numériques pourraient être identifiables en passant par les thématiques extraites du corpus.

Partant de ces hypothèses, nous avons élaboré notre méthode d'analyse. L'étape

d'analyse du corpus se compose de plusieurs étapes :

1. Pré-traitement : avant de commencer la fouille de texte, il est nécessaire de préparer les documents en les nettoyant, en les structurant et en les normalisant. Cela permet de faciliter l'analyse et d'éliminer les informations inutiles.
2. Tokenisation : cette étape consiste à découper les documents en unités de sens (mots) pour faciliter l'analyse.
3. Nettoyage : cette étape consiste à supprimer les mots inutiles (articles, prépositions, etc.) pour ne conserver que les mots pertinents pour l'analyse.
4. Indexation : cette étape consiste à créer un index des mots pertinents pour faciliter la recherche des informations dans les documents.
5. Analyse : cette étape consiste à utiliser les outils de fouille de texte pour extraire les informations souhaitées (fréquence des mots, co-occurrences, Latent Dirichlet Allocation (LDA) [2] pour l'analyse thématique etc.).
6. Visualisation : cette étape consiste à représenter les informations extraites de manière claire et compréhensible pour faciliter l'interprétation des résultats.
7. Validation : cette étape consiste à vérifier la qualité et la pertinence des résultats obtenus pour s'assurer qu'ils répondent aux besoins du projet.

En suivant ces étapes, la méthode mise en place permet d'extraire, plus ou moins, efficacement des informations à partir de documents numériques et de les rendre facilement compréhensibles pour les utilisateurs.

Quant au livrable final de ce projet, nous avons choisi de présenter nos résultats sous la forme d'un article dans l'optique de participer à un colloque en mars 2023 pour lequel notre intervention a aujourd'hui le titre provisoire suivant : "Apprivoiser LIFRANUM : enjeux et perspectives du traitement d'un corpus de littérature francophone nativement numérique".

2. A partir de maintenant nous utiliserons l'acronyme LDA

2 Aspect technique, la mise en application de nos méthodes

2.1 Les données fournies

Les données du projet proviennent de deux hébergeurs de blog : Blogger et Wordpress. Ces données ont été récupérées grâce à la méthode de web scraping [3]. Pour ce faire, les membres du projet Lifranum ont utilisé les APIs de ces deux hébergeurs. Le projet dispose donc de deux grands types de données (celles issues de Wordpress et celles issues de Blogger). Lors de la récupération, les données issues des deux sources ne sont pas organisées de la même façon. En effet, le web scraping produit des fichiers organisés différement en fonction de l'API utilisée. En ce qui concerne Wordpress, toutes les données de tous les blogs sont réunis sous trois fichiers : celui contenant les posts des blogs, celui avec les commentaires et enfin celui avec les informations des blogs (ID, description, url, logo etc...). Pour Blogger, chaque page web à un numéro et chaque numéro a trois ou quatre fichiers : celui avec les posts, celui avec les commentaires et celui avec les informations (et parfois un fichier pages). L'idée est donc similaire puisque nous retrouvons les mêmes informations mais l'organisation est différente. Ceci a joué un rôle important lorsque nous avons traité et récupéré les données. En effet, nous avons dû utiliser deux méthodes différentes pour travailler sur les données de chaque hébergeur. Après avoir examiné les données qui nous ont été fournies, nous avons pensé qu'il serait plus intéressant pour nous de se concentrer uniquement sur les données de Blogger. Nous avons pris cette décisions compte tenu du temps et de nos compétences, et en accord avec les commanditaires. Ceci afin de réaliser un traitement complet et non de faire les choses à moitié pour chaque hébergeur. Nous pensons que l'exemple de ce que nous avons réalisé avec Blogger peut également l'être avec WordPress.

2.2 Nettoyage du corpus

Les commanditaires nous ont fourni un notebook pour récupérer le contenu de la balise *content* à l'aide de l'API de Blogger et aussi des APIs de la librairie BeautifulSoup pour procéder au nettoyage de ces balises. D'abord, nous avons écrit sur un notebook un code permettant de charger une liste contenant toutes les balises *content* [4], ensuite le contenu de cette liste a été chargé sur un fichier `json`. Ce fichier était trop lourd, nous n'avons donc pas pu l'exploiter, même en

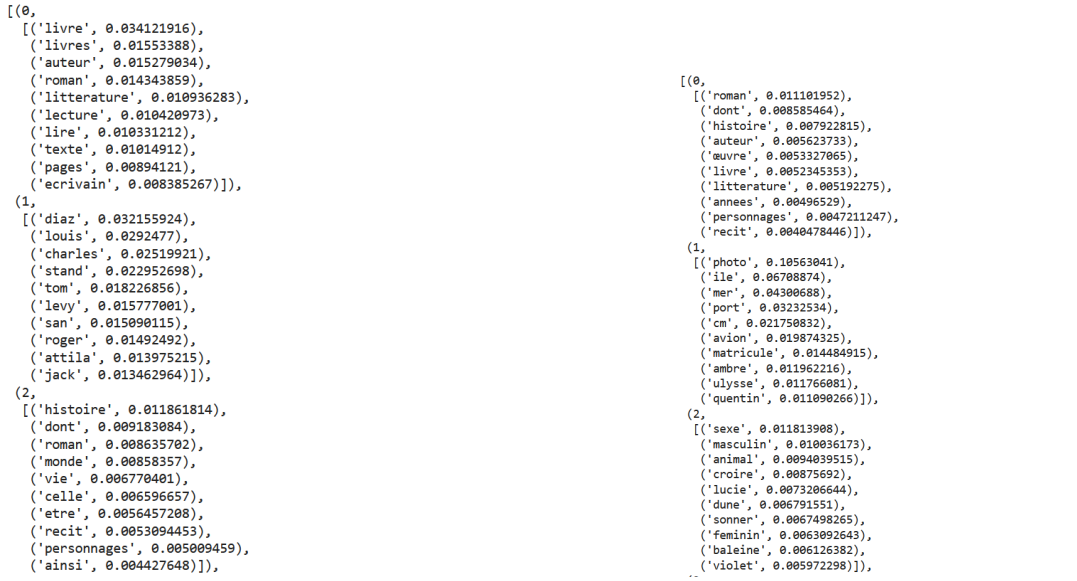
3. Pour plus d'informations, voir le travail de Théo Gady : Théo Gady (2018), *Analyse automatisée de corpus littéraire francophone nativement numérique*. Ecole des Mines Saint-Etienne, 25 pages.

4. La taille de cette appelée liste list_posts est de 187 éléments

découpant le corpus, c'est pourquoi dans le même notebook nous avons décidé de charger la liste des *contents* dans une autre liste, et cette fois nous l'avons nettoyé à l'aide de l'API précédemment fourni. Cette liste propre, contenant le contenu des toutes les balises *content*, est très longue. En effet, chaque post peut avoir plusieurs balises content [5]. Nous avons procédé de même pour charger la liste dans un nouveau fichier, mais encore une fois il était trop lourd à ouvrir et donc inutilisable.

2.3 Traitement automatique

Nous souhaitons préciser que nous avons réalisé trois LDA sur trois machines différentes et qu'à chaque LDA réalisée le résultat était différent. Nous utilisons deux des trois LDA dans nos explications pour des raisons pratiques (difficultés de faire tourner certaines fonctions sur des machines moins puissantes). Nous avons une LDA (cf figure 1a)très intéressante avec des thématiques que nous trouvons pertinentes mais sur laquelle nous n'avons pas réussi à faire la suite du traitement nous permettant l'analyse qualitative. Pour réaliser l'analyse qualitative nous avons donc utilisé une autre LDA (cf figure 1b). Ici, nous nous basons donc sur la première LDA pour les explications qui suivent.



(a) LDA 1 (b) LDA 2

FIGURE 1 – Différence entre les LDA

5. La taille de cette liste appelée list_clean_content est de 110895 éléments

Concernant le traitement de cette liste, nous avons choisi de faire une analyse thématique. Pour ce faire nous avons notamment utilisé les librairies gensim [6], nltk, et sklearn : après avoir chargé le contenu du fichier posts-maior-propre.json dans la liste list-clean-content, nous avons procédé à la tokenisation et enlever les mots outils. Ensuite nous avons lancé une LDA. Le paramétrage de cet outil nous a donné beaucoup de problème en ce qui concerne l'optimisation du modèle. Nous avions notamment deux questions difficiles à gérer : d'un côté établir le bon nombre de thématiques, à savoir la capacité de distinguer et donc de nommer les thématiques sortantes ; et de l'autre de bien filtrer le dictionnaire pour traiter des mots représentatifs, à savoir écarter celles qui étaient utilisées trop souvent ou trop rarement. Après avoir réalisé plusieurs tentatives, nous avons choisi quarante thématiques (suite au calcul d'un score de cohérence [7]), et fixé deux hyper-paramètres de filtrage du dictionnaire : écarter les mots qui comparaissaient dans moins de cinq documents ou dans plus de la moitié de documents.

La visualisation du graphique de LDA permet d'apprécier les résultats obtenus. Sachant que le corpus qui fait l'objet de notre travail concerne la littérature numérique francophone il faut remarquer que le mot "poésie" caractérise le quadrant formé entre l'axe négatif PC1 et l'axe positif PC2, alors que le mot "roman" caractérise le quadrant compris entre l'axe négatif PC1 et l'axe négatif PC2. En particulier les mots "poésie", "poème", "poétique", "amour", "Dostoievski", "Baudelaire", et "Gallimard" sont concentrés autour de la thématique seize. Alors que les mots "roman", "livre", "texte", "littérature", "lecture", "écriture", "lire", "écrire", et "auteur" font partie de la thématique sept. Concernant toujours les arts, le mot théâtre caractérise la thématique neuf avec les mots "pièce", "spectacle", "scène", "spectateur", "comédiens", et "musique" sur l'axe négatif PC2 ; en revanche les mots "film" et "cinéma" se situent dans la thématique vingt-quatre sur l'axe positif PC1. En résumant, tout ce qui peut être reconduit à l'activité de lire et écrire, de la poésie, du théâtre, et des romans semble marquer l'axe négatif PC1, et notamment se concentrer entre cet axe et l'axe négatif PC2 (cf figure2). Il est intéressant de souligner que la thématique une, qui contient les mots "être", "vie", "monde", "temps", et qui est aussi la plus grande du graphique, est difficilement déchiffrable. De plus, elle, contient presque entièrement deux autres thématique, à savoir la quatre et la six : la première semblerait renvoyer à la dimension du temps, alors que l'autre à une dimension littéraire, avec des mots comme "histoire", "récit", "roman", "narrateur", "lecteur", "personnage", "auteur", "sens", "être", "œuvre", "monde", "univers", et "vie". Cette analyse permet de détecter des catégories, par exemple les thématiques deux et cinq contiennent des mots concernant le corps humain, les thématiques huit, dix, et douze concernent respectivement la famille,

6. https ://papers.neurips.cc/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf
7. Nous n'avons pas utilisé la perplexité

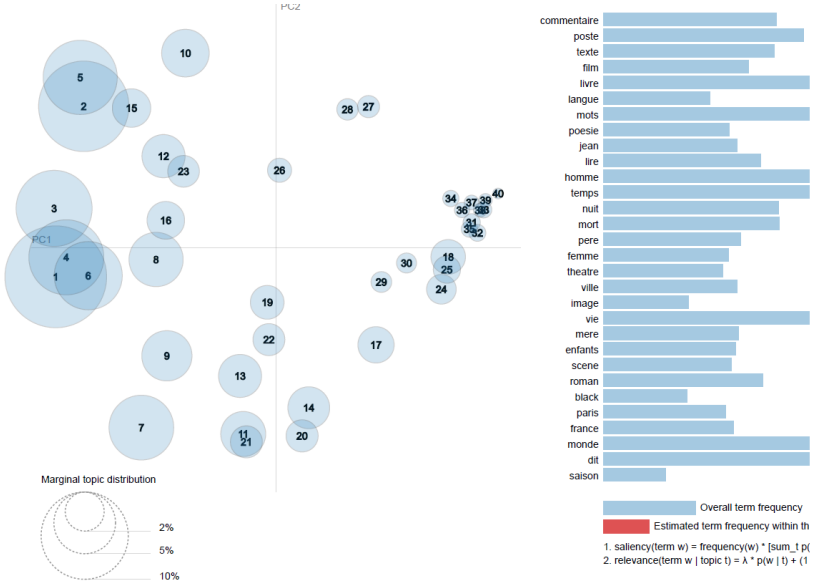


FIGURE 2 – Visualisation LDA 1

la nature, et la mort. Mais elles présentent aussi des sujets d'intérêts généraux comme la politique [8] ou la vie sociale [9].

Enfin, notre analyse a aussi montré une variété linguistique qui devrait mener à une réflexion sur le terme "francophone". En effet, les thématiques trente-une à quarante, qui caractérisent l'axe positif PC1, montrent la présence importante de mots anglais et espagnols.

De plus, à l'aide du plongement lexicale (Word2Vec) nous avons essayé de trouver des confirmations concernant certaines suggestions dérivées de l'analyse précédente. Nous avons choisi de projeter les mots dans un espace à deux dimensions, et ensuite nous avons individués les mots : "être", "monde", "vie","temps", "question", "doute", "réponse", "poésie", "roman", "solitude", "existence", "souffrance", "enfance", "sexualité", "mort". En effet, notre ambition était de voir sur un plan la disposition de ces mots, et de voir s'il était possible de tirer des conclusions. Force est de constater l'existence de deux groupes, à savoir, d'un côté, les mots "souffrance", "question", "vie", "monde", "temps" et "réponse", et de l'autre "roman", "poésie", "enfance", et "mort". Conformément à LDA précédente les mots "vie" et "monde","question" et "doute", et "temps" et "réponse" sont presque superposés et ils tournent autour de la thématique de l'existence. En revanche il est plus difficile d'expliquer la proximité des mots du deuxième groupe qui est plus hétéroclite.

8. thématiques treize et vingt dont il faut rappeler les mots "liberté", "pouvoir", "lutte", "gauche", "gouvernement"

9. thématique vingt-deux

Concernant le plongement lexical nous avons aussi mesuré le cosinus de similarité entre les mots "roman" (côté négative PC1), et "river" (côté positive du même axe), et bien évidemment il nous a donné un résultat conforme à nos attentes car la valeur du cosinus qui varie entre 0 et 1 est de 0.3. Sachant que notre plan n'est qu'en deux dimensions et donc qu'il faut considérer une perte significative d'information, c'est un résultat très faible.

Nous avons essayé de voir l'usage possible de CamemBERT sur le corpus mais au vu de notre niveau et des usages actuels réalisés avec ce dernier nous avons conclu qu'il n'était pas possible de l'exploiter pour le moment.

2.4 Analyse qualitative

Comme précisé en début de partie, les numéros des thématiques mentionnées proviennent de la deuxième LDA. Dès le début du projet, nos commanditaires nous ont fait part de leur demande concernant une analyse plus qualitative des données. L'analyse automatique était un premier pas pour tenter de dégager de grandes catégories, de voir si elles nous semblaient pertinentes et essayer, si possible, de les appliquer à des textes pris dans leur individualité et non plus représentés par quelques mots ressortant d'une thématique. Il n'est pas simple de passer d'un type d'analyse à l'autre, nous nous sommes beaucoup interrogés sur la méthode que nous pouvions adopter. Nous avons pensé à différentes manières de procéder :

1. A partir des thématiques obtenues par LDA, choisir quelques thématiques qui nous semblaient pertinentes et intéressantes et tenter de retrouver avec les mots clés de ces thématiques les posts de blogs qui pouvaient correspondre. Cette méthode ne donne pas de résultat satisfaisant, le corpus est trop important pour chercher manuellement les textes dedans avec pour seule indication des mots clés.
2. Utiliser du code pour essayer de trouver pour chaque document la thématique principale. Utiliser différents documents avec la même thématique principale pour pouvoir les lire et les comparer. En ce qui concerne cette façon de procéder, nous avons eu des résultats intéressants d'un point de vue littéraire mais qui sont à nuancer. En effet, nous avons avec du code python affiché les cinquante premiers documents du corpus (dans l'ordre dans lequel ils apparaissent dans ce dernier), ainsi que la thématique principale rattachée à chacun des cinquante documents. Nous avons obtenu ce résultat (cf figure 3).

   Nous nous sommes intéressés aux deux thématiques principales (celles qui regroupent le plus grand nombre de documents), c'est à dire la thématique quinze ("être", "dire", "chose", "faut", "réponse", "toujours", "autres", "temps", "jamais", "dont") et la thématique trente-trois ("nuit", "yeux", "corps",

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	33	0.3164	nuit, yeux, corps, jour, ciel, mots, temps, vo...	le premier jour du premier mois de l'année res...
1	1	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
2	2	15	0.6686	etre, dire, chose, faut, reponse, toujours, au...	âmes perduescauses perduesun constat s'impose ...
3	3	33	0.3797	nuit, yeux, corps, jour, ciel, mots, temps, vo...	pour y croîtrej'ai enterré quelque chose de mo...
4	4	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
5	5	33	0.2747	nuit, yeux, corps, jour, ciel, mots, temps, vo...	des sentiments le poursuivent qui lui cuisent ...
6	6	39	0.4677	monde, vie, homme, amour, mort, etre, temps, c...	y a des journées parfaites et fébriles où j'ar...
7	7	15	0.4668	etre, dire, chose, faut, reponse, toujours, au...	je ne suis ni normand ni féministe ni chanteur...
8	8	16	0.3261	terre, espace, lumiere, eau, feu, ciel, ombre,...	je ne sais pas ce qui me peine et je sais à p...
9	9	33	0.4009	nuit, yeux, corps, jour, ciel, mots, temps, vo...	sans les méchants les saints ne le sont guère ...
10	10	15	0.5665	etre, dire, chose, faut, reponse, toujours, au...	si parler n'est pas agir alors les mots ne peu...
11	11	32	0.4658	dit, ans, pere, mere, fille, vie, aime, sais, ...	les questions me perdent et les réponses me fo...
12	12	33	0.3167	nuit, yeux, corps, jour, ciel, mots, temps, vo...	la magie parfois c'est arrêter d'être subtil l...
13	13	15	0.2664	etre, dire, chose, faut, reponse, toujours, au...	tu as toujours cette facilité à aller chercher...
14	14	15	0.2872	etre, dire, chose, faut, reponse, toujours, au...	Toujours pas de nouvelles de mon frère, recue...
15	15	15	0.4321	etre, dire, chose, faut, reponse, toujours, au...	je vous préviens j'ai une terrible envie de ba...
16	16	33	0.8926	nuit, yeux, corps, jour, ciel, mots, temps, vo...	je regarde le ciel et la mer derrière une fenê...
17	17	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
18	18	39	0.2748	monde, vie, homme, amour, mort, etre, temps, c...	DANGERTOUTES LES FAMILLES NUISENTDANGERTOUTES...
19	19	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	

FIGURE 3 – Tableau des correspondances entre documents et thématiques

"jour", "ciel", "mots", "temps", "voix", "vent", "vers"). A priori il n'est pas simple de définir un sujet pour ces thématiques avec ces mots-ci. A la lecture des onze textes du tableau ayant pour thématique majoritaire la numéro quinze, quelque chose de fort nous a interpellé : ils sont tous à connotation négative. Ce sont des textes qui sont pour la plupart écrit à la première personne du singulier, il y a donc une grande importance du "moi" dans ces derniers mais ce "moi" connait des difficultés, plus ou moins personnelles, en fonction du texte. Il ressort aussi de ces textes la présence de la solitude (avec une répétition importante du terme "seul"). Pour ressortir ces différents éléments, nous avons lu ces textes, mais aussi utilisé un outil qui se nomme voyant tools [10] et que nous avons trouvé très intéressant pour faire de l'analyse qualitative de corpus plus petits. Par exemple grâce à cet outil, nous pouvons voir très rapidement à quoi est relier le terme "seul" que nous avons trouvé particulièrement représentatif de ces textes (cf figure 4) et nous constatons qu'il est relier à d'autres mots avec des connotations également négatives comme "cimetière", "mal", "totalement", mais également avec le

10. https ://voyant-tools.org/

terme "vie" qui à notre sens vient appuyer l'idée de solitude comme une constante de la vie. Nous proposons ici d'introduire un début de réflexion concernant la récurrence du la présence du "moi" et de la solitude. Ce qui nous paraît supprenant pour des textes publiés sur des blogs où la notion de communauté ou de partage est fondamentale. Nous avons procédé de

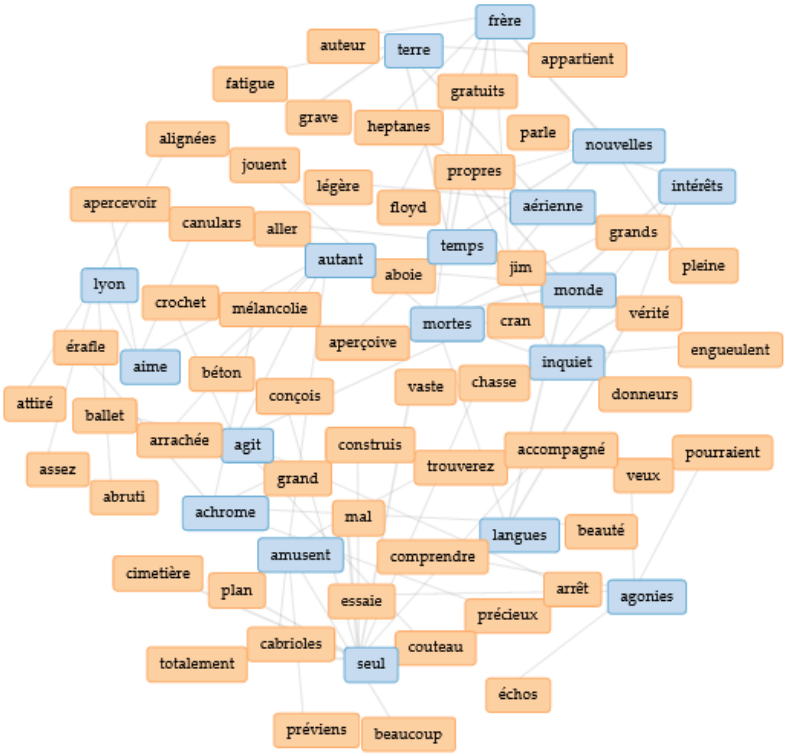


FIGURE 4 – Les termes reliés à "seul"

la même façon pour la thématique trente-trois (total de douze textes étudiés) : lecture et utilisation de voyant tools. A la lecture des textes, on retrouve un "je" assez présent mais qui laisse place à une deuxième et une troisième personne du singulier dans plusieurs textes. Les textes appartenant à la thématique trente-trois sont davantage des textes descriptifs de la vie quotidienne, avec la présence d'éléments tels que la ville, la mer, le temps ou encore les saisons. Avec voyant tools on a pu constater que les sens ("écouter", "voir") sont très reliés à ces éléments du monde qui nous entoure (cf figure 5). Néanmoins nous avons également eu le sentiment, à la lecture de certains textes, d'une certaine similitude avec la thématique quize. On retrouve un effet de chaos dans certains textes, de négatif. Cela

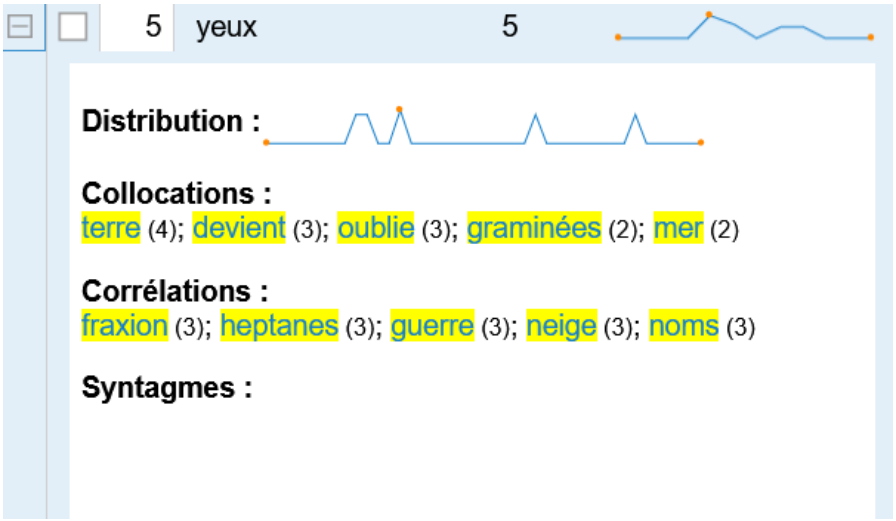


FIGURE 5 – Les termes reliés aux "yeux"

n'est pas si étonnant que cela, car lorsque l'on regarde la LDA, ce sont des catégories assez proches sur le premier axe, soit le plus important en terme d'informations. Après avoir étudié ces textes plus en profondeur nous avons voulu savoir sur quels blogs ils apparaissent et comment ils sont présentés sur ces derniers. Nous nous sommes rendu compte à ce moment que tous les textes que nous avions étudié ont été écrits par un seul et même auteur. C'est en cela qu'il est important de nuancer les résultats, car on peut se demander si ce ne sont pas les auteurs (ou un groupement d'auteurs bien défini) et leur style propre, qui sont à l'origine de certaines catégories que l'on retrouve dans la LDA ? [11]

3. Prendre des textes de manière aléatoire, afin de les lire et de voir s'il est possible de leur attribuer une des thématiques qui est ressorti de notre LDA. Nous avons choisi trente cinq textes au total, dont ceux qui ont été étudiés dans les thématiques quinze et trente-trois. En ajoutant ces textes nous avons fait ressortir une autre thématique qui est celle de la littérature (avec les mots "éditions", "livre", "poème", "poésie"). Nous avons donc lu les textes dans lesquels apparaissent ces différents termes. Ces textes sont en fait des critiques de littérature. Cette méthode confirme un peu plus notre idée selon laquelle les auteurs forment les thématiques puisque en prenant totalement au hasard les textes dans le corpus, sur les douze nouveaux textes ajoutés huit sont issus du même blog.

11. Hypothèse que nous avions émise en introduction

3 Aspect critique

3.1 Analyse critique du traitement du corpus

Concernant les résultats de nos analyses nous avons détecté des thématiques telles que la famille, le temps, la politique et les différents styles littéraires. Nous nous posons tout de même une question : pouvons nous tirer des conclusions générales ? Autrement dit, est-il possible d'affirmer que la littérature numérique se caractérise par des sujets de façon particulière ? Dans ce sens, nos résultats ne permettent pas de donner une réponse définitive. C'est pourquoi nous allons indiquer dans la partie *Propositions* des méthodes plus performantes pour obtenir des résultats plus concluants. De plus, dans cette partie nous croyons qu'il serait bien de réfléchir sur les modalités de visualisation des résultats.

Il est maintenant nécessaire de commenter notre travail réalisé sur l'analyse automatique de données. A cet égard, force est de constater que notre protocole d'analyse automatique de la littérature numérique présente certaines limites. Premièrement, il y a le problème de la maîtrise des outils, étant donné que nous ne sommes pas des informaticiens mais des littéraires, nous avons paramétré nos LDA à l'aide des lignes de codes "guides" [12] de la page de gensim dédiée aux APIs de cet outil. Cependant, nous ne maîtrisons pas parfaitement cet outil qui a été le support de notre analyse quantitative. Par ailleurs, nous nous sommes surtout affairés à réaliser une analyse qualitative à la main. Ceci afin de questionner et contextualiser les résultats de l'analyse quantitative précédente. De plus, nous avons aussi testé une TF-IDF (notamment sur bi-grams et tri-grams) et un moteur de recherche pour obtenir d'autres résultats. Cela reste cependant difficile à exploiter. Nous ne maîtrisons pas encore totalement ces outils mais nos résultats sont disponibles dans les annexes. Peut-être cela est-il dû à notre formation littéraire et nos compétences limitées en informatique qui ne nous permettent pas de comprendre entièrement ces résultats.

La structuration du jeux de données nous a également posée problème. En effet, les fichiers blog-posts contenant du texte exploitable sont imbriqués dans des dossiers par *ID* qui sont eux-mêmes imbriqués dans un dossier. Il a été difficile de sélectionner les textes exploitables automatiquement avec du code. De plus nous n'avons aucun moyen de vérifier dans tout le corpus généré si toutes les parties contenant du texte exploitable ont bien été prises en compte. L'autre difficulté concerne les auteurs que nous n'avons pas pu associer aux posts. Il aurait été intéressant de pouvoir regrouper les posts par auteur puis de les associer plus tard aux marqueurs stylistiques ou aux marqueurs thématiques. Nous pensons que cela aurait pu donner une autre piste de réflexion à notre étude.

12. https ://radimrehurek.com/gensim/models/ldamodel.html

Quid du nettoyage ? Le nettoyage du corpus a nécessité beaucoup de temps. Le traitement des corpus textuels peut être affecté par les fautes d'orthographe. Cela peut rendre difficile la reconnaissance automatique des mots et des phrases. Dans notre corpus, il apparaît des mots qui sont collés les uns aux autres mais aussi des mots mals orthographiés. Pour corriger toutes les fautes d'orthographe (même de manière automatique), il aurait fallu connaître toutes les fautes du corpus, ce qui est pour le moins impossible à faire.

Nous avons également rencontré des problèmes lors de la suppression des mots outils (stop words). Malgré tout le soin apporté au code et avec tous les mots outils ajouté à la liste des stop words, certains mots outils reviennent sans cesse dans les thématiques (exemple avec les sigles "https", "org", "www" etc.) Nous nous posons donc la question de la place de ces mots dans les thématiques qui ressortent.

De plus, comme nous l'avons signalé auparavant les résultats obtenus changent sensiblement d'une machine à une autre, c'est pourquoi nous insistons encore une fois sur l'importance d'avoir réalisé trois LDA. Nous avons cherché à mettre en lumière les convergences et les différences entre les LDA. Dans ce sens, nous rappelons que les résultats obtenus doivent être interprétés avec précaution. Comme nous l'indiquerons par la suite, il faudra combiner les différentes méthodes d'analyse. Ceci afin d'obtenir une compréhension complète des textes. Aussi, les algorithmes de classification automatique tels que LDA peuvent varier d'une machine à l'autre en raison des versions des logiciels, de la puissance de calcul et des données utilisées pour l'analyse. Il est donc important de tenir compte de ces facteurs pour obtenir des résultats fiables et comparables.

Enfin, dans le cadre de notre projet, nous avons dû réaliser des tâches qui ne correspondent pas à priori à nos profils. Bien que nous ayons porté une grande attention à la redéfinition des missions en début de projet, nous n'avions pas directement perçu la place que prendrait l'analyse automatique des données. Ceci a bien entendu été intéressant puisque nous avons pu diversifier nos compétences et nous avons aussi pu nous intéresser à divers modes de traitement automatique des données textuelles. Cependant, nous aurions voulu nous concentrer davantage sur l'analyse qualitative (une des raisons pour lesquelles nous trouvons ce projet particulièrement intéressant). Mais nous nous sommes heurtés à cette étape de l'analyse automatique des données. Nous avons dû passer par une étape de traitement des données très longue et nous n'avons pu consacrer qu'un temps réduit à l'analyse qualitative.

3.2 Analyse critique du projet

Nous avons été confrontés à un véritable problème avec les données, tant en ce qui concerne la quantité que la qualité de ces dernières. En effet nous avons du

faire face à des données très nombreuses (110895 posts de blogs uniquement pour Blogger), et très hétérogènes. Cela nous a posé différents problèmes.

Premièrement, traiter une telle quantité de données n'a pas toujours été simple car nos machines ne sont pas adaptées à ce type de traitement. Cela a parfois entrainé de long temps d'attentes pour obtenir des résultats. Ensuite, nous lorsque nous avons voulu regarder de plus près les différents posts que l'on pouvait trouver sur les blogs, nous nous sommes rendus compte que nombre d'entre eux étaient vides ou contenaient seulement quelques mots. Là encore, c'est un problème lorsque l'on fait du traitement automatique de texte : les textes plus courts ne vont pas avoir le même poids que des textes plus longs. Pour autant, ces textes courts (poèmes ou haiku par exemple) peuvent être très intéressants à analyser. Nous pouvons aussi nous demander si ces champs de textes vides ou courts doivent être pris en compte dans un projet portant sur la littérature numérique.

Nous nous sommes alors posé la question de l'ambivalence de ce corpus. En effet, il est très intéressant puisqu'il rassemble tous les blogs, donc tous les genres et tous les styles existants en ligne. Cela permettrait d'avoir un aperçu général et précis de la littérature francophone numérique. Cependant, d'un point de vue purement littéraire, on peut se demander s'il est intéressant et pertinent de comparer des textes si différents. De quelle manière pourrait-on justifier cette analyse ? Qu'est ce que l'on espère pouvoir démontrer avec ce corpus ? Comment déterminer ce qui relève de la littérature et ce qui n'en relève pas ? Ce sont quelques unes des questions que nous nous sommes posés après avoir réalisé les différents traitements exposés dans la première partie de notre travail.

Ces réflexions nous ont également amené à nous questionner sur la méthode de collecte de ce corpus. Nos connaissances sur le fonctionnement des APIs et la collecte par web scraping sont très restreintes. Pour autant, nous avons pensé qu'il serait intéressant de se demander, en premier lieu, quel type de texte souhaite-t-on réellement récolter ? Quel est l'intérêt littéraire de travailler avec un corpus aussi imposant et homogène ? Par extension, est-ce qu'il ne faudrait pas faire des sous-catégories **lors** de la collecte (par genre, par auteur, par année de publication...) et les analyser par la suite à plus grande échelle. Nous ne prétendons pas apporter des réponses à ces différentes questions : ce sont des questions que nous nous sommes posées lors de notre travail sur ce corpus. Nous n'avons pas essayé de le scinder, ne disposant pas des métadonnées nécessaires pour le faire. Nous ne pouvons donc pas garantir que cela apporterait de meilleurs résultats. Inversement, est-ce qu'en choisissant de récolter les données via cette méthode, ne passerions-nous pas à côté d'élements qui pourraient correspondre à littérature numérique francophone, mais qui ne ressortent pas lors de la récolte ? Les textes les plus courts seront quasi systématiquement nettoyés lors d'un traitement comme celui que nous avons réalisé. Pour autant un texte de moins de dix tokens pourrait être une maxime,

un aphorisme ou encore un haiku.

Ainsi, nous avons prolongé ces questionnements en nous interrogeant sur le rôle des acteurs impliqués dans la récolte des données. En effet, un projet comme LIFRANUM amène nombre de profils différents à collaborer ensemble. N'ayant que peu de connaissances relatives à l'organisation du projet nous ne pouvons que spéculer. Comment a été décidée et conceptualisée la récolte de ces données issues des blogs ? Les API sont certes très pratiques pour récupérer un maximum de données, pour autant la qualité de ces dernières peut être questionnée. Cette méthode de récolte correspond-t-elle aux ambitions qualitatives, littéraires voir sociologiques du projet ? Comment analyser le profil des auteurs quand on ne dispose pas, ou presque pas, de données sur ces derniers ? Bien que les méthodes de récolte soient efficaces et reconnues dans le domaine de l'informatique, elles ne correspondent peut être pas tout à fait au format qui peut être analysé plus finement par des profils littéraires. Quelle est la part de collaborations entre les personnes chargées de récolter les données et celles chargées de les analyser ? Les uns s'adaptent ils aux attentes méthodologiques des autres ? Nous savons par ailleurs que le projet s'intéresse aussi aux productions littéraires issues des réseaux sociaux et des sites internets [13], comment s'est adaptée la récolte des données à chacun des formats explorés par le projet ?

Toutes ces interrogations nous amènent à questionner les relations entre acteurs des différentes disciplines impliqués dans le projet. Les techniciens ont ils une compréhension fine des données qui intéressent les sciences humaines ? Les humanistes ont ils une connaissance des capacités techniques de l'informatique et de la forme que prennent les données une fois récoltées ? Dans le cadre de notre travail, il nous a semblé évident que chacun et chacune est porté par une même volonté et un même objectif. Pour autant, il semble que les frontières disciplinaires cloisonnent les champs de compétences et empêchent de porter un regard holistique sur les corpus. Ce regard englobant (d'ailleurs porté par les humanités numériques) est nécessaires afin de ne pas s'arrêter à une simple collaboration pluridisciplinaire, et d'entrer dans une réelle réflexion transdisciplinaire inhérente à notre objet d'étude.

Ces questions illustrent la nécessité de mettre en place une méthode commune se situant à la jonction des sciences humaines et sociales (plus particulièrement la littérature dans ce cas) et des sciences informatiques. La mise en place d'une méthode mixte permettrait de faciliter le travail, mais aussi les échanges au sein de l'équipe. La méthode mixte doit être pensée dans le cadre de la mise en place du protocole de recherche et intégrer les attentes, les besoins et les contraintes des différentes disciplines intégrées au projet. La réflexion autour de la méthode mixte permet de mettre au clair les spécificités relatives aux différentes disciplines

13. https ://projet-lifranum.univ-lyon3.fr/projet

concernées. Il est par exemple question de mettre en place un vocabulaire commun, de comprendre les possibilités techniques et théoriques de chacun, ou encore, de déterminer les outils ou les traitements futurs en fonction des contraintes et attentes. Ceci permettrai, à terme, de trouver des réponses aux différentes questions que pose ce projet. Il nous semble primordial de pouvoir intégrer activement chacun et chacune dans l'ensemble des réflexions relatives aux différentes étapes du projet. Ainsi, les littéraires doivent pouvoir prendre part (au moins sous l'angle réflexif) à la collecte des données et à leur traitement automatique, comprendre la forme que prennent ces manipulations de données et ce qui en découle. De même, les informaticiens doivent pouvoir prendre part aux réflexions sur les analyses qualitatives prévues sur le corpus. Ils pourront ainsi identifier l'importance d'avoir des métadonnées exploitables pour chacun des blogs ou encore définir des paramètres de récolte pertinents (longueur, langue, format etc.). Sans cette approche mixte, il semble très difficile de communiquer et de se fixer des objectifs communs.

4 Nos propositions

Après avoir abordé les différentes critiques que nous avons pu soulever dans le cadre de notre travail, nous nous proposons d'avancer quelques pistes pour la poursuite de ce projet. Les connaissances que nous avons pu acquérir au contact de ce corpus nous ont permis de réfléchir à de nouvelles pistes, que celles-ci soient d'ordre méthodologique ou technique. Nous allons donc préciser nos recommandations, tout d'abord sur le plan méthodologique en explicitant les avantages d'une méthodologie dite "mixte". Puis, nous mettrons en avant la nécessité de repenser le protocole de recherche afin de rendre plus aisé le travail pratique autour de ce corpus. Enfin, nous aborderons des propositions plus organisationnelles afin de faciliter les échanges et la compréhension au sein des équipes de recherches mixtes et pluridisciplinaires.

Comme évoqué plus tôt, il est selon nous nécessaire de mettre en place une méthode mixte. Celle-ci permettrait de correspondre aux attentes des humanistes tout en exploitant le plein potentiel des méthodes informatiques. Nous pouvons proposer une application de la méthodologie mixte. Afin de s'assurer de la qualité des données récoltées, il serait possible de procéder en amont des APIs en triant plus finement les textes composants notre corpus. Nous avons par exemple pensé à un système permettant à la fois d'avoir des textes d'auteurs (autoproclamés), des textes exploitables pour l'analyse thématique et des métadonnées nous permettant par exemple de travailler sur la géolocalisation des auteurs (pertinent dans le cadre de la francophonie), leur genre, ou leur âge. Pour cela, nous procéderions comme suit :

1. Mettre en place un recensement des auteurs sur le principe de la participation volontaire en mettant en avant les différents niveaux de qualifications dans l'écriture (d'amateur à professionnel). ‘
2. Le recensement permettrait de récolter des informations pouvant être pertinentes lors de l'exploitation des données comme l'identité de genre, l'âge, la localisation, l'autodéclaration du ou des genres d'écriture, l'autodéclaration du ou des styles d'écriture, l'autodéclaration des thématiques des textes et enfin le lien du ou des blogs de l'auteur.
3. Se servir de la section "lien du blog" pour constituer une liste des blogs pour le web scraping.
4. Réaliser le web scraping.
5. En se concentrant sur une seule plateforme, il est possible d'utiliser les APIs (ici Wordpress ou Blogger).
6. Les données obtenues grâce à ce format de récolte pourront être exploitées dans le cadre d'une LDA par exemple.
7. Il est possible de comparer les résultats de la LDA avec les catégories déclarées par les auteurs. Ceci pourrait aussi nous aider à déterminer les catégories.

Cette méthode de récolte serait certes plus longue à mettre en place, certainement plus coûteuse et les textes récoltés seraient sans doute moins nombreux. Cependant cette approche nous permettrait d'obtenir des données de plus grande qualité et de s'assurer de deux choses :

1. Les textes analysés sont bien de la "littérature" (autodéclaré). Cela répond au problème des textes vides, et des textes qui ne sont pas de la littérature.
2. Les métadonnées seront présentes pour chacun des blogs récoltés et il sera possible d'aller plus loin dans l'analyse (localisation des auteurs, les questions de genre, interactions entre auteurs etc.).

Concernant la partie pratique et le traitement automatique du corpus, après avoir constaté les limites de nos outils, nous proposons de déployer des méthodes permettant d'optimiser la détection des thématiques. Pour ce faire, il faudrait d'abord procéder à une lemmatisation du corpus pour assurer un nettoyage des textes plus complet. En effet le notebook qui a pour objet la TF-IDF montre clairement dans les bi-grams et dans les tri-grams les limites de notre nettoyage. De plus il s'agit seulement d'une petite partie du corpus, il n'affiche donc pas tous les problèmes. Malheureusement, nous n'avons pas réussi à opérer dans ce sens à cause de nos machines. C'est pourquoi il devrait être envisagé de travailler sur la plateforme Colab par exemple, pour palier à ce problème. Nous sommes arrivés à cette conclusion trop tard, sans avoir le temps de nous former à ce type d'outil. Notre analyse nous amène aussi à considérer la nécessité d'exploiter des outils

plus modernes de topic-modeling comme BERTopics, dans sa version française camemBERT. Dans le cadre de notre projet, il avait été indiqué comme traitement envisageable. Nous précisons cependant que cette possibilité ne peut être prise en considération sans une formation sur le fonctionnement de ce modèle. En effet, force est de constater que pour utiliser BERTopics les données doivent être regroupées : une partie demande une entraînement alors qu'une autre fait l'objet d'une validation.

Cette modélisation prendra beaucoup de temps. Elle demandera probablement un effort pour trouver un vocabulaire capable de gérer les formes stylistiques (poésie, roman, récit etc.) de ces productions, et les registres linguistiques présents : cela pourrait en effet présenter un problème compte tenu du fait que nous avons détecté des mots en français, en anglais et en espagnol.

Aussi, le développement du plongement lexical peut être amélioré à l'aide de l'intelligence artificielle, surtout en ce qui concerne la visualisation des données. Il serait donc intéressant d'entraîner un modèle pour une visualisation des mots sur Embedding projector [14]. En effet cela permettrait de faire des classifications sur les mots du corpus.

Concernant la visualisation des données, sans utilisation de l'intelligence artificielle, il serait aussi utile de considérer les utilisations de Gephy pour afficher les liens des auteurs, et les représenter sur une carte géographique. Toutefois, tout travaille de visualisation nécesiterai une restructuration de la récolte des données initiales. Il serait essentiel de disposer de plus d'information, comme nous l'avons indiqué auparavant (localisation par exemple).

Nous avons également souligné l'importance de mettre en place un vocabulaire commun. Pour ce faire, nous avons commencé à travailler sur l'élaboration d'un glossaire qui sera disponible à la fin de ce travail. Notre rapport s'adressant aussi bien à des informaticiens qu'à des littéraires, et dans un soucis de mettre à l'honneur les valeurs de transdisciplinarité propre aux humanités numériques, il nous semble primordial de s'assurer que chacun et chacune soit en mesure de comprendre le vocabulaire employé dans le cadre de ce projet. Ceci afin de comprendre les réalisations et les applications de nos manipulations.

Dans tous la cas, ce paragraphe ne prétend pas relater toutes les pistes possibles, mais seulement d'amener une réflexion sur le développement de ce projet. Bien évidemment ces suggestions sont issues de notre travail et de ce que nos capacités nous ont permis de comprendre dans le domaine de l'apprentissage non-supervisé.

14. https ://projector.tensorflow.org/

5 Conclusion

Rappelons les différentes tâches que nous avions identifié lors du rendu de la première étape du projet :

1. Découvrir le corpus et le nettoyer en identifiant les balises et les points pertinents pour notre travail.
2. Réaliser nos premiers essais et premiers traitements en utilisant les librairies python nous permettant de mettre en lumière des pistes de catégorisations.
3. Nommer et classer, en d'autres mots, réaliser notre propre typologie adaptée à notre corpus. Sur cette partie, les commanditaires attendaient de nous diverses propositions. Les marqueurs peuvent être thématiques, linguistiques, ou autres. C'était à nous de trouver des catégories parlantes et cohérentes dans le cadre de l'étude de la littérature nativement numérique.
4. Une étape d'analyse se découpant en deux sous-étapes :

   - Analyse macro avec des outils permettant une analyse automatique afin de réaliser des analyses statistiques ou du datamining par exemple.

   - Analyse micro où nous mettrons à profit nos connaissances en SHS. Nous pourrons ainsi proposer des analyses plus fines, concernant les interactions entre auteurs, dresser des profils, identifier des spécificités en fonction des zones géographiques, analyser les discours (côté SHS) ou réaliser des analyses de textes, des analyses de langage ou de style d'écriture, une identification des références ou inspirations (côté littérature).

Les étapes 1, 2 et 4 (en partie) ont été réalisées et sont exposées dans ce document. L'étape 3, quant à elle n'a pas été menée à bout. En effet les analyses quantitatives et qualitatives que nous avons réalisé ne nous ont pas permis de produire une rééelle typologie comme nous le pensions au début de ce projet. Nous avons mis en lumière des thématiques, qui, à notre sens, ne sont pas toujours des plus pertinentes. Nous n'avons pas trouvé de marqueurs linguistiques ou temporels propre à ce corpus. Avec plus de temps, et plus de connaissances [15] peut être aurions nous trouvé davantage de catégories à mettre en avant. Comme nous l'avons mentionné dès le début de ce projet, nous n'avons pas travaillé sur les auteurs, ni sur les commentaires de ce corpus. Il pourrait être très intéressant de se pencher sur cette question. Cela pourrait peut être aider à réaliser la catégorisation.

Pour finir, Les Humanités Numériques occupent une place centrale dans le projet Lifranum. Il est essentiel de se le rappeler. Ce projet vise à promouvoir l'utilisation des technologies numériques pour l'analyse et la compréhension des œuvres littéraires nativement numérique. Les humanités numériques permettent

15. Que ce soit sur le corpus en lui-même ou sur différentes méthodes d'analyses que nous aurions pu mettre en place.

de développer des outils et des méthodes transdisciplinaires pour l'analyse de textes numérisés, la visualisation de données et la création de corpus de textes. Les résultats obtenus par les Humanités Numériques sont utilisés pour améliorer la compréhension des œuvres littéraires et pour éclairer les débats sur l'histoire et la culture littéraire . Il est donc primordial de conserver une approche au croisement des disciplines impliquées plutôt que de proposer un agrégat d'interventions successives.

Rappelons enfin, que notre proposition d'analyse n'est que le résultat de notre approche. Nos choix ont été influencés par nos hypothèses mais aussi par nos encrages disciplinaires. Il y a aurait autant de possibilités d'angles qu'il y a "d'Humanités". Il serait possible d'adopter un angle d'analyse sociologique en étudiant les interactions entre auteurs, un angle plus anthropologique en se concentrant sur les questions de communautés rassemblées sous les thèmes du genre ou du post-colonialisme par exemple ; ou encore un angle historique pour voir comment les formes traditionnelles de la littérature sont mises en forme dans l'espace numérique. Ajoutons à cela le fait que nous devons intégrer les questions relatives à l'intelligence artificielle qui permettront de réaliser des recherches sous de nombreux angles en conservant le socle commun des Humanités Numériques.

6 Glossaire

LDA Dans le domaine du traitement automatique des langues, l'allocation de Dirichlet latente (de l'anglais Latent Dirichlet Allocation) ou LDA est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données [16]

Plongement lexical Méthode d'apprentissage d'une représentation de mots utilisée notamment en traitement automatique des langues. Le terme devrait plutôt être rendu par vectorisation de mots pour correspondre plus proprement à cette méthode [17]

Score de cohérence Niveau d'interprétabilité pour les humains des thématiques proposées par la machine [18]

TF-IDF Méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document [19]

7 Annexes

Voir fichier compressé (ci-joint).

16. https ://fr.wikipedia.org/wiki/Allocation_de_Dirichlet_latente
17. https ://fr.wikipedia.org/wiki/Plongement_lexical
18. https ://www.baeldung.com/cs/topic-modeling-coherence-score
19. https ://fr.wikipedia.org/wiki/TF-IDF

Table des figures