Rapport final

Identification des thèmes d'un livre par NLP

Pôle Intelligence Artificielle - Projet 2.07

Client : Timothée BOHÉ & Nicolas GAUDEMET – Onepoint

Encadrants : Wassila OUERDANE & Jean-Philippe POLI

Amine CHERIF HAOUAT
Tom LABIAUSSE
Cyrine NABI
Pierre OLLIVIER

2A 2021-2022

Sommaire

I/ Introduction

Le projet d'identification automatisée des thèmes d'un livre fait suite à une demande de Onepoint. Cette entreprise française née en 2002 s'est spécialisée dans la transformation numérique des entreprises et des organisations. Le groupe Onepoint est aujourd'hui présent dans 8 pays répartis sur 4 continents avec un chiffre d'affaire de 300 millions d'euros en 2021. Dans le cadre de notre projet, nos interlocuteurs au sein de l'entreprise sont Timothée BOHÉ, consultant chez Onepoint, et Nicolas GAUDEMET, partenaire de la transformation numérique dans le domaine de la culture.

À l'origine du projet qui nous a été confié se trouvent des maisons d'éditions de livres confrontées à des demandes aussi diverses qu'il existe de types et de thèmes d'ouvrages. Un des enjeux principaux des éditeurs est alors de pouvoir anticiper le nombre d'exemplaires à fabriquer pour chaque livre afin de répondre de manière optimale au marché et ainsi éviter les invendus ou ruptures de stock. Pour ce faire, les maisons d'éditions se basent sur un modèle de prédiction du succès d'un livre nécessitant de connaître les thèmes traités par un ouvrage afin d'être utilisé. Notre mission en tant que groupe projet au sein du pôle IA de CentraleSupélec se situe donc en amont de ce modèle de prédiction et consiste à développer un algorithme permettant, à partir de la lecture d'une 4^{e} de couverture, d'identifier les thèmes traités par un ouvrage.

Afin de mener à bien ce projet, Onepoint a recommandé l'usage de deux méthodes de NLP (*Natural Language Processing*). L'une utilise l'algorithme LDA et l'autre s'appuie sur un modèle de NLP développé en 2018 par Google appelé BERT. Les détails et la pertinence de ces approches seront discutés dans l'état de l'art présenté en section II Onepoint a également fourni deux corpus de textes théoriquement utilisables pour l'entraînement de modèles d'apprentissage non-supervisés. Dans les dernières semaines du projet, Onepoint nous a fait parvenir un troisième corpus du même type. En résumé, les trois corpus prennent la forme de fichiers Excel contenant des titres de livres associés à leur quatrième de couverture et regroupent au total environ 20 000 associations titre-résumé. Les performances du modèle d'identification de thèmes seront évaluées grâce aux métriques classiques de ce domaine détaillées en section III.

II/ État de l'art

L'état de l'art de ce projet est fourni en annexe de ce rapport dans le fichier :
Pole 10 état_de_l'art 2A- Groupe 2.07.pdf

III/ Description du travail réalisé

1) LDA

1.1) Prise en main de LDA et réadaptation aux besoins du projet (septembre)

La toute première étape de notre projet, juste après la phase de documentation, a été de prendre en main l'algorithme LDA. Nous avions donc eu l'occasion de tester un code (dont la documentation se trouvait sur internet) sur une base de données d'articles de journaux BBC. Le code étant bien segmenté, nous avons pu comprendre l'action de chaque bout de code sur la base de données. Ce code a fait office de structure de base pour notre projet, et nous a permis de saisir les différentes parties du code à réadapter (partiellement ou totalement) afin qu'elles fonctionnent sur les bases de données fournies par Onepoint.

Il a donc fallu :

- D'abord, nettoyer la base de données. En effet, elle contenait des résidus de langage html, qui, de plus, se juxtaposaient à certains mots, les rendant inexploitables pour l'algorithme. Beaucoup de titres ne possédaient pas de 4^{e} de couverture, il a donc fallu les retirer de la dataframe.

- Ensuite, il a fallu modifier la partie du code qui permet de « normaliser » la base de données correctement, à savoir bien tokeniser, et surtout bien lemmatiser les données. Par défaut, les modules les plus utilisés sont adaptés à la langue anglaise, beaucoup plus facile à normaliser car peu d'irrégularités touchent les mots. Il a donc fallu chercher un module spécialisé dans la langue française. Nous avons ainsi utilisé *fr_core_news_md*, qui permet une tokenisation et une lemmatisation simultanée de chaque texte.

- Il a aussi fallu chercher une liste de stopwords en français, adaptée à notre projet. Le module initialement utilisé dans le code fournissait aussi les stopwords français, mais la liste fournie était clairement insuffisante. Nous avons donc cherché en parallèle une liste plus complète, que nous importons au cours du code.

- Enfin, nous avons rajouté quelques contraintes lors du processus de tokenisation, qui ne rajoute pas de mots plus petits qu'une certaine longueur par exemple. Nous avons aussi décidé de rajouter le titre dans le sac de mots relatif à chaque texte, car il peut aussi contenir des informations importantes.

1.2) Premiers tests sur les corpus du clients (octobre)

Une fois les différentes corrections apportées, nous avons réalisé les premiers tests sur les corpus fournis par le client. Dans un premier temps (et pour la majorité de notre projet), nous avions deux corpus :

- Le premier est autour de l'histoire et de la politique en France. Il contient environ 1500 résumés utilisables.
- Le second, fourni par leur client Hachette, contient des livres de cuisine, de coloriage, de santé, et quelques romans. Il contient environ 7500 résumés utilisables.
- Un troisième, fourni beaucoup plus tard, n'a pas fait l'objet de tests aussi intensifs que pour les deux premiers corpus.

Les résultats qualitatifs de ces corpus sont les suivants :

- Considéré seul, le corpus 1 n'est pas pertinent. L'algorithme polarise beaucoup les textes en 2 « superthèmes », la politique et l'histoire.
- Le corpus 2 fournit des informations plus variées, où on retrouve des thèmes plus divers.
- La fusion des deux corpus est l'option la plus intéressante, car les 2 « superthèmes » du corpus 1 deviennent alors de simples thèmes.

1.3) Optimisation des paramètres de LDA (novembre-décembre)

Une fois la phase de réadaptation du code et des premiers tests réalisés, il nous a fallu chercher un moyen d'optimiser les hyperparamètres de LDA, à savoir les paramètres α et β des distributions de Dirichlet (cf. documentation), ainsi que le nombre de topics souhaités K. Nous optimisons nos hyperparamètres en utilisant comme fonction de score le *coherence score*, que nous cherchions alors à maximiser. Pour cela, nous avons procédé à 3 méthodes d'optimisation des paramètres :

- La première méthode, la plus classique, à savoir le Grid Search. Nous avons donc fixé des plages de valeurs pour les 3 hyperparamètres et nous avons testé toutes les combinaisons possibles.
- La seconde, un peu plus complexe, est celle de l'optimisation bayésienne. Notre choix a été guidé par le fait que cette méthode était moins chronophage mais sans pour autant perdre de son efficacité.
- La troisième est le Random Search, où des valeurs sont tirées au hasard dans un intervalle fixé.

L'une des difficultés de cette optimisation est la présence d'aléatoire, non seulement dans la reproductibilité des résultats, mais aussi dans le processus de sélection aléatoire dans l'algorithme qui rendait le score de cohérence complètement discontinu, et donc non différenciable.

Le 1er problème est relativement réglable, en fixant un *random_state* (une « issue » de l'aléatoire). Le second problème est ce qui nous a poussés à utiliser les méthodes ci-dessus. En effet, des méthodes se basant sur la différentiabilité de la fonction de score, comme la descente de gradient par exemple, sont inutilisables dans ce genre de situations. Même un grid search avec un pas constant pour les différents paramètres ne fait pas réellement sens car le score de cohérence varie énormément à chaque variation minime des hyperparamètres. L'optimisation bayésienne, quant à elle, permet de choisir des réinitialisations aléatoires au cours de la descente de gradient, ce qui améliore les résultats mais ce qui en revanche, allonge un peu plus les temps de calculs. C'est pour cela que le Random Search est finalement la solution la plus viable.

Le livrable associé au travail réalisé avec la méthode LDA est le notebook ***Topic_Modeling_LDA.ipynb*** contenant une section dédiée à l'optimisation des paramètres, afin qu'elle puisse être réutilisée par le client.

2) BERTopic

2.1) Pourquoi utiliser BERTopic ?

Les progrès récents réalisés dans le domaine du NLP grâce à l'utilisation du *Deep Learning* nous ont incités à appliquer ce type de méthodes à notre problématique de *Topic Modeling*. Du fait de la popularité grandissante du modèle BERT développé par Google en 2018, nous avons décidé d'opter pour ce dernier. En particulier, l'utilisation de BERT en *Topic Modeling* est rendue possible par la méthode BERTopic développée en 2020 par *Maarten Grootendorst* (*data scientist* titulaire d'un master en *data science* à l'université JADS aux Pays-Bas), reconnue par la communauté Python et disponible en open source. L'implémentation de cette technique de *Topic Modeling* permet notamment d'entraîner un modèle, de visualiser les résultats avec différentes représentations graphiques et de réaliser des classifications sur des nouveaux textes. De plus, des améliorations et nouvelles fonctionnalités continuent d'être ajoutées par *M.Grootendorst*. Il est par exemple possible de calculer un score de cohérence Cv sur l'ensemble des topics identifiés par BERTopic. Cela nous permet ainsi de comparer de manière plus quantitative les résultats obtenus avec ceux de LDA.

2.2) Premiers essais : résultats, avantages & inconvénients de la méthode (novembre-décembre)

BERTopic nécessitant une quantité suffisante de textes afin de saisir au mieux les relations sémantiques, nous avons choisi d'utiliser les deux corpus fusionnés comme données d'entraînement. De plus, les clients souhaitent à terme pouvoir enrichir le modèle en fournissant de nouveaux textes pour l'entraînement. Ainsi, le chargement des données d'entraînement réalisé dans le code a été pensé pour faciliter l'ajout de de nouveaux corpus de textes.

Le corpus fusionné comportant près de 10000 textes, nous n'avons pas pu vérifier la pertinence de chacun d'entre eux. Comme souvent en *Machine Learning*, nous avons identifié certaines données posant problème lors des premiers essais. À la différence de LDA, BERTopic ne nécessite pas de pré-traitement des données susceptible de filtrer certaines anomalies. Ainsi, nous avons remarqué que certains textes contiennent des traces de formatage HTML tandis que d'autres sont presque vides. En remarquant que la grande majorité des textes cohérents contiennent entre 20 et 5000 caractères, nous avons décidé de ne conserver que les textes dans cette plage de taille en définissant deux hyperparamètres *min_size* et *max_size*. Au final, moins d'une dizaine de textes sont écartés du corpus fusionné mais ce filtrage permet de rendre le code plus robuste à l'ajout de nouvelles données.

Disposant du titre du livre correspondant à chaque 4^{e} de couverture, nous avons pensé à concaténer titres et textes afin d'enrichir l'information apportée au modèle pour l'entraînement. Ainsi, nous avons défini un nouvel hyper-paramètre booléen *with_title*.

Enfin, il faut sélectionner un modèle d'*embedding* pour BERT, ce qui correspond plus simplement au type de pré-entraînement de BERT que l'on souhaite utiliser comme base. Pour ce faire, nous avons principalement le choix entre les options *french* ou *multilingual* de BERT en précisant le paramètre *language* lors du chargement d'un modèle. Ce dernier sera alors en état de « comprendre » seulement la langue française ou plus d'une centaine de langues. Les données traitées dans le cadre du projet étant supposées être uniquement des textes écrits en français, nous n'avons a priori pas de raison de vouloir utiliser un modèle multilingue. Cependant, il est par exemple attendu que des anglicismes passés dans le langage français courant doivent apparaître dans les données

d'entraînement. Ainsi, nous avons voulu tester les configurations *french* et *multilingual* au cas où nous décèlerions une différence significative dans les résultats.

Après avoir défini ces différents paramétrages, nous avons donc testé les quatre configurations obtenues en variant les paramètres *with_title* et *language* en calculant pour chacun d'eux le score de cohérence Cv associé comme on peut l'observer en **figure 1**.

Corpus	min_size	max_size	corpus_size	with_title	language	topic_nb	coherence_cv
1+2	20	5000	9080	True	french	156	0.717
1+2	20	5000	9080	False	french	146	0.703
1+2	20	5000	9080	True	multilingual	149	0.721
1+2	20	5000	9080	False	multilingual	152	0.708

Figure 1 : Tests des 4 paramétrages de modèle en fonction des configurations (with_title,language)

On peut tout d'abord observer que le nombre de topics formés par l'algorithme ne varie pas énormément et semble osciller autour de 150 topics. Cela constituera une difficulté pour la suite de notre travail.

De plus, il semblerait que l'interprétation naïve des scores de cohérence nous pousserait à sélectionner un modèle [*with_title=True & language = multilingual*]. Cependant, les scores n'étant pas significativement différents les uns des autres, nous avons réitéré l'expérience de comparaison. Ainsi, nous avons constaté, du fait de la part d'aléatoire dans l'entraînement de BERT, qu'aucun modèle ne se distinguait sensiblement des autres. Plus exactement, nous ne pouvions pas uniquement concentrer toute notre attention sur le score de cohérence. Ainsi, nous nous sommes plutôt fiés à notre jugement humain en observant les topics formés. À titre d'exemple, certains topics obtenus pour le modèle *[with_title=False & language=french]* sont présentés sur la **figure 2**.

```
TOPIC 9 -> plantes jardin potager fleurs aromatiques champignons fraises arbres cultiver légumes
TOPIC 10 -> kitty hello chat chats chaton mon petit chatons autocollants craquer
TOPIC 11 -> jeux jeu questions défis culture catégories boîte cartes geek tester
...
TOPIC 48 -> wars star univers galaxie héros vaisseaux super saga rogue one
TOPIC 49 -> animaux humains nous nature êtres non darwin qui notre anthropocène
...
TOPIC 60 -> bracelets perles couteau bijoux coton réaliser modèles couteaux explications quelques
TOPIC 61 -> musique metal rock groupe jazz musiques albums beatles records disco
TOPIC 62 -> noël anniversaire peppa cadeaux père fête oui jouets cadeau aujourd
...
TOPIC 129 -> yoga stress postures exercices séances respiration relaxation pratiqué gérer programme
TOPIC 130 -> univers bang hasard matière galaxies big étoile physiciens cosmologie scientifique
```

Figure 2 : Exemple de topics obtenus avec le paramétrage de modèle [with_title=False,french]

Nous avons pu constater que les modèles entraînés sur des données avec titres produisent des thèmes très spécifiques. En effet, dans le cas des livres pour enfants, le modèle identifie et sépare différents univers tels que *Star Wars, Hello Kitty, Tortues Ninja, Violetta*… C'est en grande partie une conséquence de la prise en compte des titres qui sont, dans le cas des livres pour enfants, très explicites. Ces derniers contiennent le plus souvent les mots-clefs de l'univers en question et produisent un motif récurrent accaparant l'attention du modèle. Nous avons également pu observer ce phénomène avec d'autres styles d'ouvrages tels que les G*uides du Routard* ou des livres au format « catalogue » : *Les 100 meilleures…* L'inconvénient majeur des résultats obtenus réside dans l'incompréhension développée par le modèle. En effet, ce dernier se concentre alors trop sur les motifs qu'il observe et pas suffisamment sur les relations entre les topics formés.

Une des particularités de BERTopic en tant que méthode de Topic Modeling réside dans l'algorithme de clustering utilisé. En effet, une des particularités qui fait la force d'HDBSCAN consiste à considérer certains points comme étant des outliers. Ainsi, tous les points ne sont pas *in fine* associés à un cluster. Dans le cas de nos deux corpus réunis, cela concerne environ un tiers des textes. Cette proportion n'est pas propre à notre problème et a par exemple déjà été observée en appliquant BERTopic à l'ensemble des textes de *20newsgroup*. Cependant, notre projet doit apporter un supplément d'information dans une architecture de données plus grande et il n'est pas indispensable d'effectuer une classification de thématique pour chaque ouvrage. Ainsi, alors que nous pensions que cette particularité de BERTopic serait un obstacle majeur à sa mise en œuvre, nos clients nous ont tout de même incités à poursuivre sur cette voie.

Après avoir présenté à nos clients les premiers résultats obtenus avec BERTopic, ces derniers nous ont demandé de réfléchir à un moyen permettant de réduire le nombre de topics formés par l'algorithme. En effet, en tentant d'optimiser la qualité du clustering, BERTopic forme environ 150 topics. Cependant, nos clients souhaitent obtenir un nombre plus restreint de topics, qui refléteraient les grandes catégories de livres qui existent (histoire, sciences, santé, cuisine, livres pour enfants, romans policiers, guides touristiques…) : pas plus d'une trentaine de thèmes. Ainsi, il devient crucial que le modèle sache par exemple faire le rapprochement entre les topics *Star Wars* et *Tortues Ninja* et distinguer correctement les topics *Hello Kitty* et cuisine. Pour ces raisons, nous avons décidé de ne pas ajouter les titres aux données d'entraînement puisque nous avions constaté que cela affectait la capacité de compréhension du modèle.

Enfin, en comparant visuellement les topics obtenus pour les deux modèles avec données non-titrées, nous avons décidé de privilégier le modèle *french* qui nous a semblé plus cohérent, bien que la différence soit difficilement perceptible.

Finalement, compte tenu des attentes de nos clients et des premiers observations des résultats obtenus avec BERTopic, nous avons décidé de retenir la configuration suivante :

- *min_size = 20 & max_size = 5000,*
- *with_title = False,*
- *language = french.*

2.3) Réduction des topics : mise en œuvre & résultats (décembre/janvier)

2.3.1 - Réduction du nombre de topics proposée par BERTopic

BERTopic peut être vu comme une projection des données d'entraînement dans un espace vectoriel réel multi-dimensionnel permettant de définir une notion de similarité entre des textes via le concept de distance mathématique. En regroupant les documents semblables au sein de clusters qu'on nomme dans notre cas topics, l'algorithme parvient à calculer une représentation vectorielle de chaque topic. De la même manière qu'avec les documents, il est possible de définir une notion de similarité entre ces topics via des calculs de distance.

Ainsi, il devient possible de réduire le nombre de topics identifiés en fusionnant certains d'entre eux. Cette fonctionnalité est d'ailleurs déjà implémentée dans BERTopic et on peut alors simplement réduire n'importe quel modèle à n_0 topics en un modèle à $n_1<n_0$ topics. Cela peut même être réalisé directement lors de l'entraînement du modèle sur les données via le paramètre *nr_topics* comme on peut l'observer en **figure 3.**

```
[ ]  BERT_model = BERTopic(language="french", verbose=True, nr_topics=None)
```

Figure 3 : Instanciation d'un modèle BERTopic

Ce paramètre *nr_topics* peut prendre les valeurs *None*, *'auto'* ou bien n'importe quelle valeur entière n_0 non nulle. Dans ce dernier cas, après entraînement du modèle sur les données, le *Topic Modeling* obtenu contiendra n_0 topics. La réduction, si elle est nécessaire, est effectuée par des fusions successives entre topics considérés comme proches via leur représentation TF-IDF. Dans le cas où le paramètre *nr_topics* est spécifié à *'auto'*, une réduction automatique du nombre de topics est réalisée en fixant une sorte de distance minimale de dissimilarité lors de l'exécution d'HDBSCAN, forçant certains points à former un même cluster. Cette réduction « automatique » ne permet cependant pas de contrôler précisément le nombre de topics formés. Enfin, affecter la valeur *None* à *nr_topics* n'impose aucune contrainte sur le nombre de topcis.

Il est également possible de réduire le nombre de topics d'un modèle après entraînement avec la fonction *reduce_topics* recevant en argument un *nr_topics* qui se doit cette fois-ci d'être un nombre entier non nul. La réduction est ensuite réalisée par fusions successives.

Après avoir testé la réduction *'auto'* à plusieurs reprises, nous avons pu observer que le modèle produit rassemblait toujours environ 90 topics. Ce nombre est encore bien trop élevé par rapport à nos attentes (une trentaine de thèmes). Ainsi, la réduction *'auto'* proposée par BERTopic ne nous permet pas d'atteindre l'objectif fixé et il nous a fallu réfléchir à une autre méthode.

2.3.2 – Réduction de topics par valeur imposée

La réduction de n_0 topics en n_1 topics proposée par BERTopic requiert évidemment de préciser n_1. Déterminer la meilleure valeur de n_1 compte tenu des données et de nos attentes a été pour nous un problème majeur. Nous avons tout d'abord pensé pouvoir attaquer le problème par recherche exhaustive. Autrement dit, construire un modèle pour chaque valeur de n_1 à tester puis calculer un score de cohérence pour chacun d'entre eux. Deux obstacles majeurs se présentent alors. Premièrement, le temps nécessaire sur Google Colab pour le calcul d'un score de cohérence est d'environ 5 minutes. Cela peut sembler raisonnable mais tester 30 valeurs de n_0 nécessiterait près de 3 heures. De plus, l'exploitation des ressources calculatoires est limitée sur Google Colab et il n'est donc pas certain que nous puissions mettre en œuvre cette méthode. Par ailleurs, bien que le score de cohérence puisse donner une indication sur la qualité d'un *Topic Modeling*, ce dernier n'est pas parfait et il serait dangereux d'accorder une telle responsabilité de choix d'un modèle à cette métrique. En utilisant un jugement humain comme métrique, nous pourrions tout aussi bien sélectionner une valeur de n_1 différente. Pour toutes ces raisons, nous avons préféré réfléchir à une autre méthode de sélection de n_1 moins brutale et exploitant mieux la technicité de BERTopic.

Rappelons que l'entraînement de BERTopic permet de produire un modèle contenant un nombre de topics supposé être optimal pour garantir la pureté et la séparabilité de ces derniers. Nous ne savons pas nous-même à l'avance de combien de topics va être composé le modèle bien que l'utilisation de BERTopic avec les données du corpus aboutit toujours à l'identification d'environ $n_0 = 150$ topics. En observant les mots prépondérants de chacun de ces clusters, nous pouvons alors réaliser nous-mêmes des regroupements et ainsi aboutir à une estimation du nombre de topics qu'il serait bon d'imposer. Par conséquent, nous avons dû réfléchir à un moyen d'automatiser ce regroupement de topics.

Parmi les différentes fonctionnalités implémentées dans BERTopic et permettant de visualiser les résultats, l'*Intertopic Distance Map* a particulièrement retenu notre attention. La **figure 4** ci-dessous en donne un exemple.

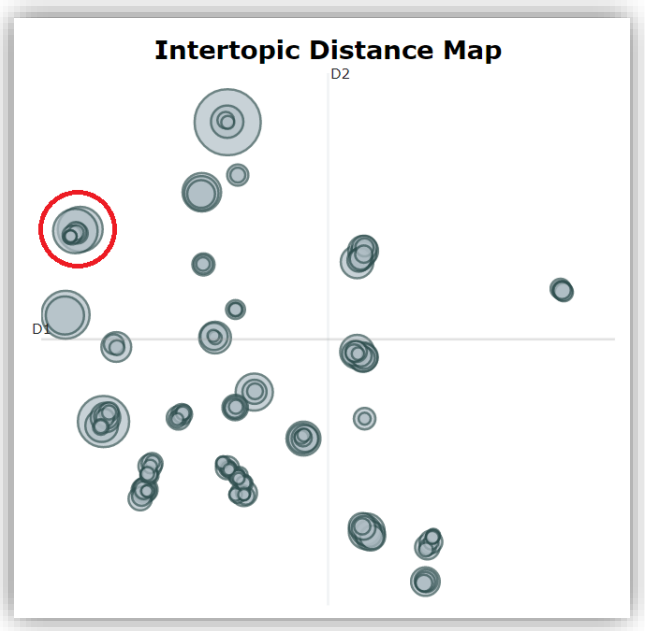

Figure 4 : Intertopic Distance Map obtenue pour le paramétrage de modèle choisi

C'est carte n'est autre qu'une projection en deux dimensions des différents topics, représentés par les disques gris, d'un modèle BERTopic. L'*Intertopic Distance Map* étant de plus interactive, il est possible de savoir précisément quel topic correspond à quel disque. Nous avons alors pu remarquer qu'il se formait des clusters de topics plutôt bien identifiables. À titre d'exemple, l'ensemble des topics inclus dans le cercle rouge contient des topics présentés en **figure 5** ci-dessous. Comme on peut l'observer, le rassemblement effectué est très pertinent.

```
TOPIC 3 -> chocolat recettes cake gâteaux gâteau pain crème desserts cakes caramel
TOPIC 35 -> recettes ingrédients plats cuisine cuisiner recette quotidien super poulet apéro
TOPIC 51 -> recettes terrine carrément tomates légumes cuisine chou tartelettes merguez charcuteries
TOPIC 52 -> riz recettes sushis cuisine rouleaux boeuf asiatique wok nems porc
TOPIC 58 -> recettes cuisine couscous poulet aux tajine agneau classiques annoncer au
TOPIC 117 -> recettes robot plats cuiseur chef techniques cuisine atelier chefs cookeo
TOPIC 131 -> burger hamburgers burgers fernand hamburger big fermier terroir veau recettes
```

Figure 5 : Exemple de clustering représentatif du thème « cuisine »

De manière générale, nous avons alors constaté que les regroupements présentés étaient très fiables. En particulier, nous avons pu identifier un îlot cuisine, un îlot politique/économie, un îlot dessin/coloriage… L'*Intertopic Distance Map* correspond donc parfaitement à ce que nous recherchions. Cependant, cette fonctionnalité de BERTopic permet seulement de visualiser ce regroupement sans pouvoir extraire d'information autre que par l'observation de la carte. Pour remédier à cela, nous avons dû aller explorer le code source de BERTopic disponible sur le lien Github de *Maarten Grotendorst*. Nous avons alors adapté le code nous permettant de reproduire les éléments d'intérêt de l'*Intertopic Distance Map*. Cela consiste principalement en l'utilisation d'UMAP pour projeter les topic embeddings en deux dimensions. La **figure 6** présente la carte des topics obtenus, chaque point bleu représentant un des n_0 topics identifiés par BERTopic lors de la phase d'entraînement :

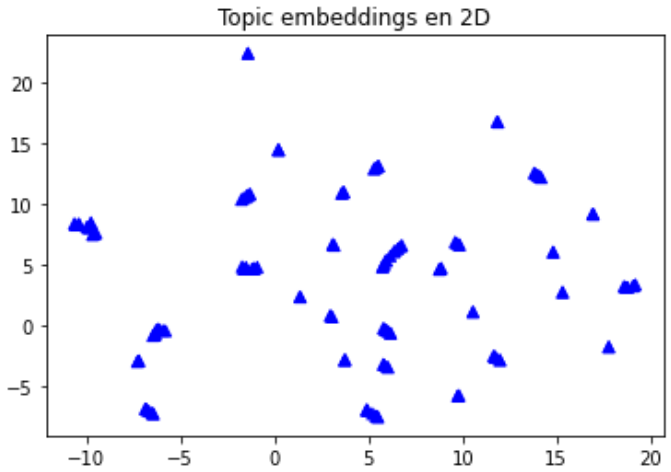

Figure 6 : Visualisation des topic embeddings réduits en 2D

Nous avons ensuite réalisé un clustering sur ces topics en 2D. Les premiers essais furent réalisés à l'aide d'HDBSCAN supposé être bien adapté pour l'identification de clusters denses indépendamment de leurs formes. Cependant, les résultats furent peu concluants, probablement du fait de la faible quantité de points au total et par cluster observable sur la **figure 6**. Les clusters semblant plutôt distants les uns des autres, nous avons pensé à tester un simple clustering hierarchique. Son exécution sur un faible nombre de points comme le nôtre étant quasi instantanée, nous avons pu tester en quelques secondes une plage d'une vingtaine de valeurs de *threshold* et sélectionner une de celles fournissant un score de silhouette maximal. La **figure 7** présente les résultats obtenus par clustering hiérarchique sur les données de la **figure 6**. Nous avons pu constater que cette méthode de clustering correspond très bien à nos attentes sur les différents *Topic Modeling* produits par BERTopic que nous avons pu obtenir. On constate alors que le clustering hiérarchique avec *fine-tuning* de l'hyper-paramètre *threshold* correspond parfaitement à nos attentes. Nous avons donc choisi cette méthode.

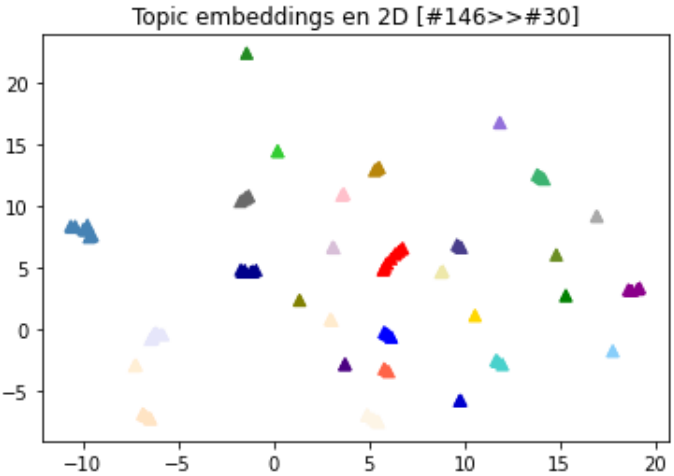

Figure 7 : Visualisation du clustering réalisé sur les topic embeddings réduits en 2D

Finalement, nous avons réussi à identifier n_1=30 clusters de topics que nous appelons big topics. La **figure 8** présente quelques-uns des big topics obtenus.

```
# BIG TOPIC 6 = [3, 35, 51, 52, 58, 117, 131]
TOPIC 3 -> chocolat recettes cake gâteaux gâteau pain crème desserts cakes caramel
TOPIC 35 -> recettes ingrédients plats cuisine cuisiner recette quotidien super poulet apéro
TOPIC 51 -> recettes terrine carrément tomates légumes cuisine chou tartelettes merguez charcuteries
TOPIC 52 -> riz recettes sushis cuisine rouleaux boeuf asiatique wok nems porc
TOPIC 58 -> recettes cuisine couscous poulet aux tajine agneau classiques annoncer au
TOPIC 117 -> recettes robot plats cuiseur chef techniques cuisine atelier chefs cookeo
TOPIC 131 -> burger hamburgers burgers fernand hamburger big fermier terroir veau recettes

# BIG TOPIC 15 = [13, 36, 75, 80, 133]
TOPIC 13 -> plantes jardin potager fleurs champignons aromatiques arbres légumes sol cultiver
TOPIC 36 -> vins vin cave appellations dégustation mets hachette 000 whisky aveugle
TOPIC 75 -> médecine santé homéopathie maladie symptômes maux médecines maladies acupuncture hygiène
TOPIC 80 -> voitures 911 automobile modèles voiture cv marque porsche icône bond
TOPIC 133 -> huiles essentielles élixirs utiliser remèdes he utilisation soulager thérapeutique propriétés

# BIG TOPIC 20 = [32, 134]
TOPIC 32 -> exercices programme corps votre entraînement vous muscles vos silhouette pilates
TOPIC 134 -> yoga postures stress respiration exercices séances relaxation pratiqué corps méditation

# BIG TOPIC 21 = [22, 31, 64]
TOPIC 22 -> bible église la christianisme religion dieu histoire du textes jésus
TOPIC 31 -> empire rome romain histoire civilisation civilisations de la par ont
TOPIC 64 -> islam islamique islamistes arabe arabes empire la orient musulmane musulmans

# BIG TOPIC 23 = [25, 40]
TOPIC 25 -> politique macron emmanuel démocratie la république président droite pouvoir il
TOPIC 40 -> france française république la histoire guerre révolution français siècle ont
```

Figure 8 : Exemple de big topics obtenus par clustering des topic embeddings sur le corpus 1+2

On peut alors remarquer que certains d'entre eux répondent très bien à nos attentes comme par exemple les big topics « cuisine » (6), « corps/gymnastique » (20), « histoire/religion » (21), « histoire/politique/France » (23). Ces derniers semblent à la fois spécifiques et homogènes, tandis que d'autres sont plus discutables comme le big topic 15 qui semble concerner les plantes et les produits dérivés mais contient étonnament le topic 80 traitant de l'automobile. Nous avons également pu constater que les big topics concernant des univers pour enfants semblaient de moins bonne qualité comme on peut l'observer en **figure 9**.

```
# BIG TOPIC 3 = [10, 16, 30, 45, 53, 68, 90, 111]
TOPIC 10 -> galaxie jedi will empire tom ils luke seigneur dark wan
TOPIC 16 -> île océan capitaine mer cruz bateau pirate mais bord trésor
TOPIC 30 -> royaume elsa reine kristoff elles anna pouvoirs magique cristal malice
TOPIC 45 -> mistral ladybug ranch celesteville babar va cheval il forêt amis
TOPIC 53 -> dragons skylanders captain skylands bruce kaos armée rocket planète amis
TOPIC 68 -> masha michka concours zara elle équestre orage tommi mais compétition
TOPIC 90 -> lycée elles club billie isa collège mal pat pension retrouvent
TOPIC 111 -> spider man peter héros parker super robot hiro miles parviendra
```

Figure 9 : Exemple de big topic de livres pour enfants

En effet, les univers décrits par les topics sont déjà très spécifiques et il peut alors sembler peu pertinent de vouloir les regrouper ensemble. Il est de toute manière difficile d'identifier un niveau intermédiaire de classification entre les topics obtenus et une simple catégorie « fictions enfant/ado ».

Nous avons également pensé à changer la dimension de réduction des topics embeddings. En effet, le choix de la dimension 2 fut principalement guidé par l'*Intertopic Distance Map*. Nous avons alors testé les dimensions 3, 4 et 5 mais n'avons pas observé de différences significatives dans les big topics formés. En particulier, l'intrusion du topic 80 parmi le big topic « plante/produits phyto » est toujours présente pour ces dimensions. Ainsi, nous avons décidé de conserver la réduction en dimension 2 qui facilite la visualisation et l'interprétation éventuelle des clusterings.

Enfin, nous avons cherché à homogénéiser chaque big topic. Par cela, on entend identifier les mots les plus représentatifs de chaque big topic à la manière de ce qui est réalisé par BERTopic sur les topics à l'aide de la méthode c-TF-iDF. Cependant, nous avons estimé que les représentations des big

topics obtenus par méthode c-TF (comptage simple des mots les plus fréquents hors stop-words) étaient plus cohérents. En particulier, nous avons observé que les mots prépondérants qui en résultent sont plus aptes à décrire un thème. Par exemple, on préfère obtenir « recette/alimentation/cuisine/gâteaux » (avec c-TF) que « nutritionniste/glycémie/vitamines/nutriments » (avec c-TF-IDF). Les résultats obtenus sont présentés en **figure 10**.

```
#================================================= WORDS of BIG TOPICS ==================================================#
 BIG TOPIC 0 -> kitty petit hello chat amis livre monde vie aujourd histoire
 BIG TOPIC 1 -> recettes alimentation gluten régime santé aliments conseils livre sucre poids
 BIG TOPIC 2 -> livre pages histoire coloriages grand stickers disney format petits film
 BIG TOPIC 3 -> lecture texte court petits images mieux aller simple coucher endormiront
 BIG TOPIC 4 -> pokémon livre sacha amis activités kai jeux dresseur petit chevaux
 BIG TOPIC 5 -> adresses guide sites quartier restaurants cartes balades carnet voyage pratique
 BIG TOPIC 6 -> recettes chocolat cuisine gâteaux cake ingrédients plats gâteau pain desserts
 BIG TOPIC 7 -> amis jeune monde mission île vie sauver temps royaume aide
 BIG TOPIC 8 -> livre super jeux héros stickers grâce personnages coloriages univers préférés
 BIG TOPIC 9 -> stickers princesses mode tenues grâce habiller livre accessoires robes autocollants
 BIG TOPIC 10 -> jeux questions jeu défis culture cartes 100 mots boîte jour
 BIG TOPIC 11 -> vie livre méditation zen monde vivre temps bonheur jour sommeil
 BIG TOPIC 12 -> petit lapin blanc bébé koala aujourd jour franklin allistair maman
 BIG TOPIC 13 -> guerre histoire politique siècle france monde juifs russie ans livre
 BIG TOPIC 14 -> recettes cocktails salade légumes cuisine pâtes salades spaghettis italienne produits
 BIG TOPIC 15 -> vins plantes vin livre ouvrage jardin monde guide conseils santé
 BIG TOPIC 16 -> mois famille blagues année toto agenda fous page rires jour
 BIG TOPIC 17 -> enfant livre colorier petits couleur couleurs coloriages enfants plaisir grâce
 BIG TOPIC 18 -> routard cartes adresses plans infos découvrir jour souvent introuvables remises
 BIG TOPIC 19 -> langue mots expressions 000 ouvrage examen permis française exemples vie
 BIG TOPIC 20 -> exercices corps yoga programme livre entraînement forme séances respiration conseils
 BIG TOPIC 21 -> histoire monde siècle livre empire ouvrage temps vie politique aujourd
 BIG TOPIC 22 -> vie enfant philosophie ouvrage livre monde questions parents pensée psychanalyse
 BIG TOPIC 23 -> politique france histoire république monde française macron pouvoir français ouvrage
 BIG TOPIC 24 -> vie ans jeune fille père mère nouvelle pourtant mort enquête
 BIG TOPIC 25 -> art histoire vie monde siècle exposition artiste ouvrage oeuvre musée
 BIG TOPIC 26 -> monde santé politique ouvrage économie économique vie écologique numérique politiques
 BIG TOPIC 27 -> droit leçon droits partie titre international chapitre action justice principes
 BIG TOPIC 28 -> violetta journal vie studio musique nouvelle père fille jeune ans
 BIG TOPIC 29 -> travail monde ouvrage histoire éducation école europe vie foot scolaire
```

Figure 10 : Exemple d'homogénéisation des big topics avec c-TF

Remarquons tout de même que le thème « automobile » se retrouve alors totalement éclipsé au sein du big topic 15. En effet, le topic « automobile » (80) ne rassemble qu'une vingtaine de textes, ce qui est peu comparé aux 185 textes du big topic 15. Cela n'est pas résolu en utilisant c-TF-IDF, le problème résidant de toute manière dans l'association du topic 80 avec les topics « plantes/produits phyto ».

Un dernier point crucial dans le regroupement par topics concerne le nombre de textes capturés par chaque big topic. On constate grâce à la **figure 11** qu'environ deux tiers des big topics contiennent au moins une centaine de textes, ce qui est plutôt satisaisant.

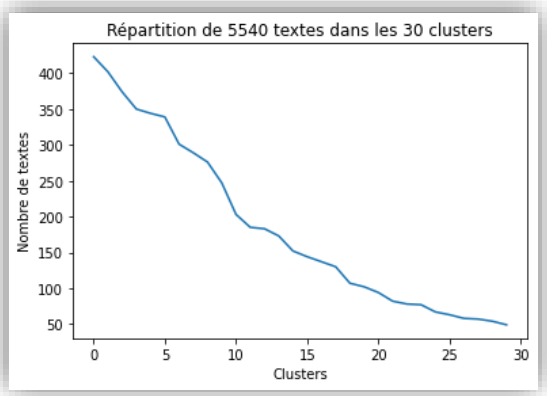

Figure 11 : Répartition des textes dans les 30 big topics identifiés sur le corpus 1+2

2.3.3 – Retour sur la réduction de topics via BERTopic

D'une certaines manière, en recherchant n_1 pour réaliser la réduction du nombre de topics, nous sommes parvenus à réaliser la réduction elle-même. Comme nous l'avons vu, les big topics obtenus ne sont pas encore parfaits. De plus, réaliser une prédiction de big topic pour un nouveau texte nécessite d'abord de prédire son topic via le modèle BERTopic puis de lui associer le big topic correspondant. Cela augmente la complexité de prise en main du code pour l'étape de prédiction et nécessite de stocker les correspondances topic-bigtopic en plus du fichier modèle BERTopic. En résumé, cette méthode n'est pas optimale. Nous avons donc pensé utiliser la réduction automatique de topics proposée par BERTopic avec la valeur de n_1 calculée permettant d'aboutir à un modèle réduit BERTopic clef en main pour lequel il est également possible de calculer un score de cohérence. En réalisant cette réduction, nous obtenons alors un score de cohérence d'environ 0,53. Comme attendu, le score de cohérence chute pour n_1 comparé à la valeur optimale n_0. Les topics obtenus sont présentés en **figure 12**.

```
#================================================== NEW TOPICS ==================================================#
NEW TOPIC 0 -> routard cartes adresses plans infos découvrir jour souvent introuvables remises
NEW TOPIC 1 -> vie ans fille petite jeune mère jamais jour pourtant père
NEW TOPIC 2 -> stickers livre pokémon jeux super activités retrouve décorer grâce affaires
NEW TOPIC 3 -> livre activités stickers jeux pages autocollants petits puzzle grand pièces
NEW TOPIC 4 -> adresses sites quartier boutiques restaurants pratique plan guide carnet quartiers
NEW TOPIC 5 -> coloriages livre pages coloriage colorier cahier couleurs grand format couleur
NEW TOPIC 6 -> histoire lecture disney texte petits album images lire court film
NEW TOPIC 7 -> stickers princesses mode tenues livre grâce habiller accessoires robes autocollants
NEW TOPIC 8 -> recettes chocolat gâteaux cake gâteau pain crème desserts base dessert
NEW TOPIC 9 -> colorier crayons couleurs couleur plaisir coloriages art créer retrouvez motifs
NEW TOPIC 10 -> recettes cuisine plats légumes ingrédients pâtes cocktails temps recette cuisiner
NEW TOPIC 11 -> enquête crime détective police ans mort meurtre alex femme vie
NEW TOPIC 12 -> langue mots expressions 000 française vie exemples prononciation comprendre sens
NEW TOPIC 13 -> enfant tableaux couleur calme cartes coloriages colorier enfants concentration application
NEW TOPIC 14 -> recettes alimentation gluten santé régime aliments conseils livre sucre poids
NEW TOPIC 15 -> art exposition musée artiste oeuvre histoire siècle vie oeuvres ouvrage
NEW TOPIC 16 -> galaxie amis empire jeune guerre jedi mission royaume will sauver
NEW TOPIC 17 -> enfant parents enfants vie conseils guide livre questions développement bébé
NEW TOPIC 18 -> recettes salade cuisine légumes produits poulet classiques burger salades 100
NEW TOPIC 19 -> kitty hello chat chats chaton petit livre petits amis activités
NEW TOPIC 20 -> wars star univers couleurs super héros personnages grâce livre coloriages
NEW TOPIC 21 -> jeux questions jeu défis culture cartes 100 boîte mots jour
NEW TOPIC 22 -> cartes routard découvrir adresses partie culturelles respect visites introuvables photos
NEW TOPIC 23 -> plantes jardin potager fleurs livre conseils légumes ouvrage nature champignons
NEW TOPIC 24 -> guide voyage adresses cartes evasion auteurs infos balades pratiques étapes
NEW TOPIC 25 -> île mer océan capitaine jeune bateau bord mission cruz amis
NEW TOPIC 26 -> monde économie climatique économique politique écologique vie ouvrage aujourd crise
NEW TOPIC 27 -> vie littérature livre histoire monde siècle grands thibaudet ans travers
NEW TOPIC 28 -> philosophie pensée ouvrage bergson grands temps science philosophique monde philosophes
NEW TOPIC 29 -> histoire temps monde ouvrage bible livre église vie textes religion
```

Figure 12 : Réduction à 30 topics obtenue avec BERTopic et réprésentation par c-TF

On ne constate alors pas de différence très marquée par rapport aux big topics de la **figure 10**. Cependant, dans le cas de la réduction BERTopic, plus de la moitié de l'ensemble des données n'est pas classifié et considéré comme outlier. Comme on peut le voir en **figure 13**, les 30 topics identifiés regroupent alors peu de textes. Graphiquement, on peut voir que la distribution représentée est « en dessous » de celle de la **figure 11**.

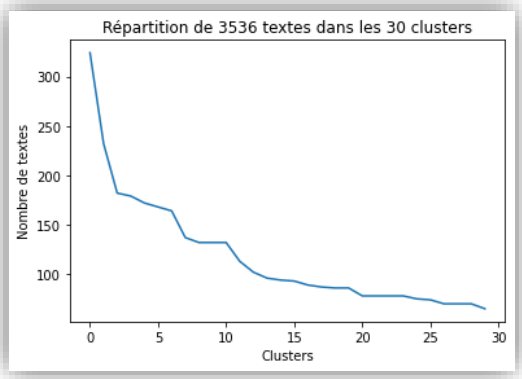

Figure T13 : Répartition des textes dans les 30 topics réduits par BERTopic sur le corpus 1+2

La réduction de topics via BERTopic présente cependant l'avantage de livrer un modèle prêt à l'emploi sans nécessiter de stocker des correspondances entre topics et big topics. Pour toutes ces raisons, nous ne pouvons pas réellement conseiller une des deux méthodes. Le choix doit surtout dépendre de l'usage que l'on souhaite en faire.

2.4) Apport d'un nouveau corpus (janvier)

2.4.1 Topic Modeling sur le corpus 3

Durant les dernières semaines du projet, le 3 janvier 2022 exactement, Onepoint nous a fait parvenir un nouveau corpus de 4es de couvertures en provenance de l'éditeur Editis. Ce corpus, dont nous ferons référence en tant que corpus 3, rassemble environ 13000 textes exploitables. Contrairement aux corpus 1 et 2 précédents, certains des textes de ce nouveau corpus sont écrits en anglais. Nous avons donc décidé d'utiliser un paramétrage *multilingual* plutôt que *french* pour BERTopic. Le Topic Modeling qui en résulte rassemble les textes dans environ 130 topics. Après application de notre méthode de regroupement, nous parvenons à identifier 24 big topics dont certains sont présentés en **figure 14**.

```
# BIG TOPIC 7 = [52, 63, 68, 78, 99, 130]
TOPIC 52 -> données excel vous office 2016 outlook powerpoint graphiques word gestion
TOPIC 63 -> android iphone tablette smartphone votre applications galaxy google fonctionnalités vous
TOPIC 68 -> prévisions votre mois astro ascendant vous décan vos conseils 2018
TOPIC 78 -> windows vous ordinateur 10 pc applications votre web ipad mac
TOPIC 99 -> windows 10 ordinateur microsoft internet démarrer fichiers votre vos applications
TOPIC 130 -> synopsis coming soon powerpoint excel word 2019 office diapositives déménagent

# BIG TOPIC 18 = [24, 40, 92, 105, 118, 121, 128, 133, 136]
TOPIC 24 -> what to your and the see await discoveries skip advice
TOPIC 40 -> and planet lonely to your the travel for of you
TOPIC 92 -> your what to skip trusted companion the passport advice await
TOPIC 105 -> with phrases language your dictionary phrasebook travel for the get
TOPIC 118 -> ireland and lonely planet your the it to travel dublin
TOPIC 121 -> road trips to the planet lonely your britain freedom great
TOPIC 128 -> national parks and park canyon to usa lonely planet your
TOPIC 133 -> and planet the to lonely your travel for of guide
TOPIC 136 -> top of best and the world year planet travel lonely

# BIG TOPIC 22 = [12, 45, 85, 109, 122, 125]
TOPIC 12 -> recettes cuisine plats légumes desserts vos aux repas ou salade
TOPIC 45 -> chocolat gâteaux desserts recettes cakes fruits tarte gâteau cake crème
TOPIC 85 -> gluten recettes fruits aliments noix alimentation légumes gourmandes ou avocat
TOPIC 109 -> fromage cuisine recettes fromages eat food bruccio to aux italienne
TOPIC 122 -> huile huiles essentielles coco recettes votre vous olive vos santé
TOPIC 125 -> réussissez leçon exercice cuisine recettes macarons recette paupiettes bavarois aux

# BIG TOPIC 23 = [0, 5, 10, 11, 16, 18, 22, 25, 26, 41, 47, 53, 59, 111, 120]
TOPIC 0 -> de du la le et des un art ses les
TOPIC 5 -> politique la france de en le du et qui des
TOPIC 10 -> femmes elles sexualité hommes les ou leur femme des et
TOPIC 11 -> football joueurs sport psg club coupe france équipe du champions
TOPIC 16 -> islam musulmans islamique arabe terrorisme orient syrie musulmane la coran
TOPIC 18 -> guerre hitler juifs allemands la allemagne en mondiale de nazisme
TOPIC 22 -> rock musique musicale du chansons beatles années de chanson le
TOPIC 25 -> guerre algérie la français de mondiale les et des du
TOPIC 26 -> elle roman amour son ans vie une écrivain la de
TOPIC 41 -> russie soviétique russe russes moscou la ruskin poutine staline pays
TOPIC 47 -> aviation pilote pilotes avion vol aéronautique bord guerre formation avions
TOPIC 53 -> elle guerre sa mère juifs camp louise son 1939 roman
TOPIC 59 -> afrique il de en khmers le du guerre la au
TOPIC 111 -> petrescu ai petipeux justin léa pekkala le dgse benjamin ne
TOPIC 120 -> nègre noirs vdm blanc baldwin racisme américains je suis black
```

Figure 14 : Exemple de big topics obtenus par clustering des topic embeddings sur le corpus 3

Certains regroupements comme les big topics 7 et 22 sont très pertinents puisqu'on peut aisément identifier les thèmes « informatique/bureautique » et « cuisine/alimentation ». Le big topic 18 pouvant s'interpréter comme le thème « voyage/tourisme » semble contenir de nombreux textes en anglais et démontre la capacité de compréhension de plusieurs langues par ce modèle. Enfin le big topic 23 est beaucoup plus discutable. En effet, de nombreux topics semblent être proches du thème « guerre/conflit » mais certains comme le 11 (« sport/football ») ainsi que le 22 (« musique/pop rock ») font plutôt figure d'anomalie au sein de ce big topic. De nouveau, les résultats obtenus par la méthode de regroupement des topics sont encore nettement améliorables.

Nous avons également utilisé la fonctionnalité de réduction incluse dans BERTopic afin de réduire le modèle à 24 topics. Comme nous avons déjà pu l'observer avec le corpus 1+2, le nombre de textes considérés comme outliers augmente alors drastiquement et atteint alors plus de la moitié du corpus 3 (7473/12928). La **figure 15** donne un aperçu des topics obtenus.

```
#================================================ WORDS of BIG TOPICS ================================================#
 BIG TOPIC 0 -> vie malko ans jeune mer père homme monde jour grand
 BIG TOPIC 1 -> santé livre vie corps yoga sommeil alimentation aliments conseils quotidien
 BIG TOPIC 2 -> vie jeune ans monde fille jamais jour découvre petit seule
 BIG TOPIC 3 -> ans vie âge livre toilettes jamais blagues petit book humour
 BIG TOPIC 4 -> intelligence monde artificielle robots numérique robot machine réseaux humanité entreprises
 BIG TOPIC 5 -> livre plantes and ouvrage jardin techniques the jardins réaliser créer
 BIG TOPIC 6 -> ouvrage livre cahier exercices histoire monde questions propose français travail
 BIG TOPIC 7 -> livre windows prévisions mois ordinateur internet applications conseils êtes année
 BIG TOPIC 8 -> the lune and espace étoiles temps univers mission spatiale monde
 BIG TOPIC 9 -> jeune monde vie nouvelle fille mort temps homme japon enfin
 BIG TOPIC 10 -> enfants and the livre jeux autocollants pages illustrations with couleurs
 BIG TOPIC 11 -> vie livre monde histoire temps amour ouvrage enfant the siècle
 BIG TOPIC 12 -> and the guide sites découvrir ville voyage itinéraires planet for
 BIG TOPIC 13 -> roi louis france histoire napoléon grand ans guerre homme jeune
 BIG TOPIC 14 -> vie ans jeune femme mère homme jamais jour fille famille
 BIG TOPIC 15 -> mots phrases guide restaurant and décoder pays information expressions mini
 BIG TOPIC 16 -> and the mots langue livre with français children vocabulaire enfants
 BIG TOPIC 17 -> the petit lapin forêt carmélito loup poules poulailler ours carottes
 BIG TOPIC 18 -> and the planet your lonely travel for you with all
 BIG TOPIC 19 -> question conseils règle exercices application révision comprendre corrigés anglais enfant
 BIG TOPIC 20 -> siècle histoire vie postales cartes moto ouvrage travers paris ville
 BIG TOPIC 21 -> prénom livre avez album portant histoire année cadeau cinéma années
 BIG TOPIC 22 -> recettes cuisine livre plats desserts légumes fruits chocolat découvrez the
 BIG TOPIC 23 -> livre histoire vie monde france guerre ans années politique siècle
```

Figure 15 : Réduction à 24 topics obtenue avec BERTopic sur le corpus 3 et représentation par c-TF

2.4.2 Test sur le supercorpus 1+2+3

Forts des résultats obtenus sur les corpus 1+2 et 3, nous avons décidé d'appliquer BERTopic à l'ensemble de nos données, c'est-à-dire le corpus 1+2+3. Ce regroupement massif de données présente cependant le risque de réduire le proportion finale de textes abordant un thème spécifique comparée à cette même proportion considérée dans un des trois corpus. Une trop faible représentativité de certains thèmes pourrait alors pousser le modèle à considérer comme outliers des textes qu'il prenait bien en compte lorsque les corpus 1+2 et 3 lui étaient présentés séparément. Après avoir entraîné un modèle BERTopic sur le corpus 1+2+3 et formé 44 big topics, nous obtenons les résultats présentés en **figure 16**.

```
#================================================== WORDS of BIG TOPICS ==================================================#
 BIG TOPIC 0 -> enfant vie livre femmes amour enfants parents ouvrage hommes histoire
 BIG TOPIC 1 -> vie livre bonheur monde japon émotions vivre ouvrage and japonais
 BIG TOPIC 2 -> jour chien monde vie forêt amis petite grand petit poules
 BIG TOPIC 3 -> petit lapin livre humour blagues jour rire blanc petite loup
 BIG TOPIC 4 -> yoga exercices corps vie livre programme conseils cahier pratique méditation
 BIG TOPIC 5 -> vie homme ans monde jeune famille père années mère hommes
 BIG TOPIC 6 -> jeux and jeu the enfants wipe questions pages cahier clean
 BIG TOPIC 7 -> enfant tableaux enfants couleur colorier calme coloriages petits cartes concentration
 BIG TOPIC 8 -> and the planet lonely your travel for you with what
 BIG TOPIC 9 -> adresses sites guide quartier restaurants découvrir plan pratique carnet ville
 BIG TOPIC 10 -> histoire guerre monde france livre politique vie ans aujourd années
 BIG TOPIC 11 -> histoire roi siècle vie louis france monde empire ans ville
 BIG TOPIC 12 -> vie noël père mia jeune monde petit grand amis nouvelle
 BIG TOPIC 13 -> vins vin monde guide cave france livre dégustation whisky 000
 BIG TOPIC 14 -> année mois famille conseils prévisions avez album mode 2018 enfant
 BIG TOPIC 15 -> petits gommettes colorier crayons couleur pochette confortablement soucis installez disney
 BIG TOPIC 16 -> recettes cuisine salade légumes pâtes cocktails soupe soupes plats ingrédients
 BIG TOPIC 17 -> île mer jeune vie monde grand ans océan petite jour
 BIG TOPIC 18 -> autocollants tenues accessoires stickers the habiller mode princesses robes habille
 BIG TOPIC 19 -> art histoire vie livre années siècle oeuvre musique ouvrage monde
 BIG TOPIC 20 -> petits images mieux texte aller coucher yeux lecture jolies endormiront
 BIG TOPIC 21 -> and the prénom livre animaux with histoire informations portant planet
 BIG TOPIC 22 -> philosophie vie monde pensée histoire temps science ouvrage livre univers
 BIG TOPIC 23 -> santé recettes livre vie alimentation ouvrage quotidien aliments conseils maladies
 BIG TOPIC 24 -> langue mots français cahier exercices anglais française ouvrage and the
 BIG TOPIC 25 -> monde livre ouvrage droit plantes histoire vie jardin aujourd france
 BIG TOPIC 26 -> jeune monde vie nouvelle amis enfin mort héros temps combat
 BIG TOPIC 27 -> techniques livre réaliser and modèles projets explications the technique astuces
 BIG TOPIC 28 -> chat chats kitty hello livre petit chaton chien chevaux monde
 BIG TOPIC 29 -> recettes cuisine chocolat livre desserts plats gâteaux légumes ingrédients crème
 BIG TOPIC 30 -> the terre espace ciel and planète lune ans spatiale mission
 BIG TOPIC 31 -> expressions mots guide prononciation conversation phrases and vie langue situations
 BIG TOPIC 32 -> livre enfants petits illustrations the and pages découvrir jeunes matières
 BIG TOPIC 33 -> pokémon livre sacha jeux activités kai dresseur jeu retrouve héros
 BIG TOPIC 34 -> stickers décorer affaires pages planche relief super coloriages indispensable trentaine
 BIG TOPIC 35 -> règle question conseils application exercices révision comprendre anglais corrigés enfant
 BIG TOPIC 36 -> histoire album livre disney film soir histoires enfant lecture aventures
 BIG TOPIC 37 -> livre coloriages pages couleurs colorier stickers grâce jeux activités personnages
 BIG TOPIC 38 -> livre windows applications internet ordinateur web maîtriser utiliser musique tablette
 BIG TOPIC 39 -> champignons espèces reconnaître guide insectes livre petit identifier habitat nature
 BIG TOPIC 40 -> routard cartes adresses plans jour infos souvent introuvables remises ailleurs
 BIG TOPIC 41 -> vie ans jeune femme mère fille jamais jour famille père
 BIG TOPIC 42 -> and the guide voyage new planet découvrir ville lonely sites
 BIG TOPIC 43 -> examen ouvrage formation code bac questions permis préparation jour conseils
```

Figure 16 : Réduction à 44 topics obtenue avec BERTopic sur le corpus 1+2+3 et représentation par c-TF

Malheureusement, comme nous l'avions pressenti, la proportion d'outliers considérés par le modèle augmente fortement et atteint quasiment la moitié du corpus total (10770/22008). Cependant, le regroupement des corpus permet au final d'obtenir des big topics de tailles plus conséquentes que tous les précédents observés jusque là comme on peut le voir en **figure 17**.

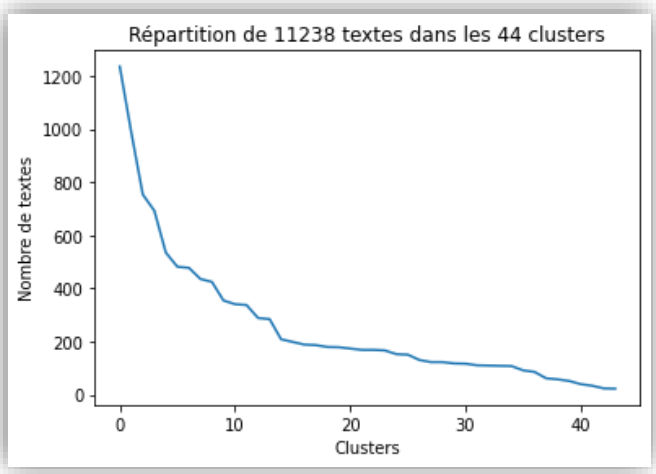

Figure 17 : Répartition des textes dans les 44 big topics identifiés sur le corpus 1+2+3

2.5) Livrables associés à BERTopic

Le principal livrable associé au travail réalisé avec la méthode BERTopic est bien sûr le notebook Python ***Topic_Modeling_BERTopic.ipynb*** permettant de réaliser toutes les opérations décrites précedemment ainsi que la prédiction du topic d'un nouveau texte étant donné un *Topic Modeling* BERTopic. Nous fournissons également quatre modèles *BERTopic* pouvant être utilisés via le notebook pour réaliser ces prédictions.

Deux de ces modèles ont été entrainé sur le corpus 1+2, l'un ayant subi une réduction du nombre de topics via la fonctionnalité *reduce_topics* de BERTopic (annoté *reduced_t30*) et l'autre n'ayant subi aucune réduction (annoté *tNone*) mais disposant de fichiers *.txt* annexes établissant les correspondances entre documents et big topics (*Doc2bBigtopic*) et entre topics et big topics (*Topic2Bigtopic)*. Nous avons également produit, grâce à une fonctionnalité implémentée dans le notebook, un fichier Excel *Data2Topics_#1+2.xls* listant les correspondances entre big topic et textes du corpus 1+2.

Les deux autres modèles restants ont respectivement été entrainé sur les corpus 3 et 1+2+3 et sans réduction BERTopic. Pour chacun d'eux, nous avons également réalisé des regroupements en big topics accessibles via des fichiers du type *Doc2bBigtopic* et *Topic2bBigtopic.*

IV/ Répartition des tâches

La figure 18 ci-dessous présente l'organisation du travail au sein du groupe.

	Amine	Tom	Cyrine	Pierre
Recherches LDA	x	x	x	
Recherches BERT		x		x
Recherches métriques			x	
Prise en main de LDA	x	x		
Prise en main BERT		x		x
Prise en main des métriques			x	
Fine-Tuning LDA	x		x	
Utilisation de BERTopic		x		x

Figure 18 : Répartition des tâches principales du projet

IV/ Conclusion et perspectives

La méthode LDA a été une très bonne introduction au traitement de langage naturel. En effet, nous avons pu y découvrir les bases du prétraitement de données textuelles, afin de normaliser les différentes écritures d'un mot, ainsi que la manière dont on représente informatiquement un mot dans une approche statistique de Topic Modeling. Les premiers résultats étaient étonnants à observer car nous y avions retrouvé tes topics très parlants dès les premiers essais. Seulement nous avons aussi rapidement compris les limites d'un modèle statistique pour du NLP. En effet, la régularisation des mots n'est pas parfaite, les mots rares (qui sont censés être critiques pour déterminer le topic d'un texte) ont très peu de poids, et considérer les textes comme des sacs de mots fait parfois perdre le sens particulier de chaque texte. De plus, LDA n'a aucun moyen de juger de la pertinence de mots qui ne se trouvaient pas dans les corpus d'entraînement. Enfin, LDA est surtout performant pour des catégorisations assez "grossières" : le modèle va plutôt chercher à déterminer des thèmes globaux du texte plutôt que ses sujets d'intérêt à proprement parler. Ainsi, des catégories comme "cuisine", "art", "santé", "coloriage" ou encore "société" sont reconnus.

Le travail réalisé avec l'algorithme BERTopic a tout d'abord permis de confirmer les réels bénéfices liés à l'utilisation d'une représentation contextualisée du langage (BERT) sur ce projet de NLP. Ainsi, nous avons pu constater que les topics formés par la méthode BERTopic décrivaient des ouvrages très spécifiques notamment lorsqu'il est question d'univers pour enfants/adolescents (*Star Wars, Hello Kitty…*) comme c'est le cas de très nombreux textes du corpus 2. Cela a cependant porté atteinte à la qualité des *Topic Modeling* obtenus. En effet, les thèmes sous-représentés au sein des différentes combinaisons de corpus testés sont moins bien appréhendés par l'algorithme et souvent éclipsés au profit d'autres plus abondants. De plus, nous avons également été confrontés à un nombre élevé de topics identifiés par l'algorithme par rapport aux attendus des clients. Ainsi, nous avons développé une méthode de regroupement des topics similaires basée sur les résultats de l'algorithme BERTopic. De manière générale, les big topics obtenus démontrent la pertinence de cette approche. En effet, les thèmes globaux liés à la cuisine, l'histoire ou le tourisme sont reconnus de manière plutôt constante à travers les différents modèles développés. Toutefois, il persiste dans ces big topics ce qui pourrait être considéré comme des anomalies pour le jugement humain. Certains regroupements de topics semblent également moins pertinents. Ainsi, le travail réalisé pose des bases solides mais pouvant encore être perfectionnées. En particulier, il pourrait être bénéfique de se plonger plus en détail dans le paramétrage d'UMAP ou d'HDBSCAN quitte à devoir réimplémenter des fonctionnalités essentielles de BERTopic. De plus, la mise en place réelle de l'algorithme de *Topic Modeling* pour enrichir un modèle de prédiction de ventes devrait permettre de pouvoir mieux estimer le niveau de détail attendu pour les topics et donc mieux adapter leur nombre.

Ce projet nous aura permis de monter en compétence dans différents domaines, notamment l'utilisation de LDA et BERT mais aussi, de manière plus surprenante mais bien réelle, dans l'art de gérer le partage des fichiers avec d'autres personnes et les problèmes d'authentification liés à l'utilisation de stockage sur un Drive Google.