# Topic Modeling sur des abstracts d'articles scientifiques

Romain Dudoit,
Karidjatou Diaby,
Laïla Djebli

*Université Lumière Lyon 2*

**Abstract**

Ce papier à pour ambition d'évaluer la classification d'abstracts d'articles scientifiques en modélisant L'allocation de Dirichlet latente (LDA) supervisée avec utilisation de glossaires. Nous voulons explorer les différentes méthodes nous permettant d'effectuer une meilleure classification des sujets à l'aide du topic modeling. Nous appuyons notre analyse sur les résumés d'articles scientifiques pour rechercher les méthodes d'analyses qui apportent des nouveautés dans le monde du topic modeling à l'aide d'Abstracts d'articles scientifiques afin d'en améliorer sa performance. Le but étant d'amener notre analyse dans une direction nous permettant de vérifier si la LDA semi-supervisée servirait de classification à nos abstracts d'articles scientifiques.

*Keywords:* LDA, seedLDA , classification, pré-traitement, topic modeling, articles scientifiques, analyse, cohérence, Abstracts

## 1. Introduction

Au fur et à mesure que les informations deviennent disponibles, il devient de plus en plus difficile de découvrir efficacement ce dont nous avons besoin. Nous avons besoin de nouveaux outils pour nous aider à organiser, rechercher et comprendre ces vastes quantités d'informations. C'est dans ce cadre qu'intervient le topic modeling ou la modélisation de sujet. La modélisation de sujet est une technique de fouille de texte qui consiste à découvrir des sujets abstraits ou latents apparaissant dans une collection de documents. En effet, pouvoir détecter des thèmes ou des sujets dans des documents textuels est très utile pour améliorer la classification et l'étiquetage des documents, de même que pour mettre en place des systèmes de recommandation ou encore mettre en évidence des tendances.

## 2. Background

### 2.1. De la LSA à la LDA

Les études sur la modélisation de sujet ont vu leur début dans les années 80. Dans le but de concevoir des programmes de recherche d'information plus précis, une première technique à avoir vu le jour est celle de l'analyse sémantique latente (LSA). Proposée par Deerwester et al [1], cette technique, s'appuyant le procédé d'algèbre linéaire de décomposition en valeurs singulières, a permis de regrouper des documents sous l'hypothèse que les mots sémantiquement proches ont tendance à apparaître ensemble dans les documents. Basée sur cette dernière, l'analyse sémantique latente probabiliste (PLSA) proposée par Hofmann en 2001 [2] a constitué un véritable modèle thématique. Il a toutefois été montré qu'elle souffrait parfois de surapprentissage. Un autre modèle introduit en 2003 par Blei et al [3] sous le nom de Latent DirichLet Allocation (LDA) a justement permis de lever le problème de surapprentissage de la PLSA. La LDA est aujourd'hui considérée comme le modèle de référence pour faire du topic modeling.

### 2.2. Pré-requis à LDA

La première chose à noter est que la LDA, tout comme les modèles thématiques précédemment cités, n'est pas en mesure de déterminer le nombre de topics présents dans un corpus. Le nombre de topics constituera un paramètre d'entrée de l'algorithme. Il est choisit à priori selon la connaissance du domaine d'étude ou peut être approché par une méthode d'essai-erreur, par un processus d'itération sur le nombre de topics associé à des mesures quantitatives (score de cohérence et perplexité en autre).

Ensuite, il faut remarquer qu'une fois que le nombre de thèmes à été identifié, les résultats en sortie de la LDA ne nous disent pas à quoi les thèmes correspondent. Plus exactement pour chaque thème trouvé, nous n'obtenons qu'une distribution de mots, c'est à dire la probabilité que chaque mot du vocabulaire appartienne au thème. Nous devons donc utiliser notre propre interprétation des sujets afin de comprendre de quoi il s'agit et de donner un nom à chaque sujet.

Pour finir, la LDA ne se soucie pas de l'ordre des mots dans le document et utilise la représentation des documents par sac de mots (cf section 2.4).

### 2.3. Principes de la LDA

*2.3.1. Principes*

L'idée de base de la LDA est que chaque document peut être décrit par une distribution (un mélange probabiliste) de sujets latents et chaque sujet peut être décrit par un mélange de mots. Pour découvrir ces thèmes, la LDA utilise des distributions de Dirichlet et un modèle génératif probabiliste, c'est à dire un modèle qui peut générer de manière aléatoire de nouvelles données à partir des données observables (les documents dans notre cas) et qui peut évaluer la probabilité qu'un nouveau document ait été généré à partir des documents observés.

*2.3.2. Hyperparamètres*

Dans la LDA, seul les mots sont observés, la distribution des topics par document et la distribution des mots par topic sont cachées.

Les modélisations réalisés par une LDA dépendent de 3 hyperparamètres (variables déterminées à priori) :

- $K$ : le nombre de topic à choisir
- $\alpha$ : vecteur de dimension $K$ correspondant au paramètre de la loi de Dirichlet qui contrôle le paramètre modélisant la distribution multinomiale des sujets par document (noté $\theta$). Lorsque $\alpha$ est au plus bas, il est possible de limiter chaque document à un seul sujet et lorsqu'il est au plus haut, on force les documents à partager uniformément les mêmes sujets.
- $\beta$ : paramètre de la loi de Dirichlet qui contrôle le paramètre modélisant la distribution multinomiale des mots par sujet (noté $\varphi$ ou $\eta$). Un faible $\beta$ signifie que chaque topic contiendra peu de mots et inversement avec une valeur élevée.

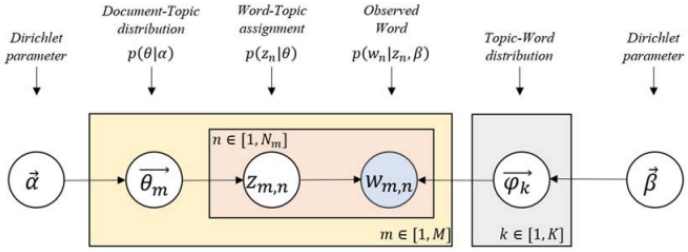


Figure 1: Représentation de la LDA sous forme de modèle graphique

Le but de la LDA est d'estimer $\theta$ et $\varphi$. Néanmoins l'estimation de ces paramètres selon un corpus d'apprentissage donné est complexe. Ces estimations peuvent se faire suivant différentes méthodes, ce qui explique des implémentations de la LDA qui varient selon la méthode privilégiée. Parmi ces méthodes, certaines utilisent l'inférence variationnelle, d'autre utilisent l'échantillonnage de Gibbs ou encore l'espérance-maximisation (EM).

Pour finir, l'algorithme de la LDA est un algorithme itératif. Un topic est premièrement assigné aléatoirement à chaque mot de chaque document. A partir de là, on peut calculer la distribution des mots par topics et la distribution des topics par documents. Ces dernières sont utilisées à l'itération suivante afin de réassigner à chacun des mots un topic plus pertinent. Le procédé est répété jusqu'à ce que les assignations se stabilisent.

*2.3.3. Les mesures de la cohérence*

Afin d'effectuer un modèle, plusieurs préparations se font en amont. Les mesures de cohérence de sujet font parties de ces préparations, ils attribuent un score à un seul sujet en mesurant le degré de similitude sémantique entre les mots les mieux notés du sujet. Ces mesures aident à distinguer les sujets qui sont des sujets sémantiquement interprétables et les sujets qui sont des artefacts d'inférence statistique. Cette mesure va nous permettre de mesurer le nombre de Topics à réaliser pour obtenir des résultats optimaux. Parmi ces mesures on peut trouver :

- La mesure C_v : elle est basée sur une fenêtre glissante, une segmentation en un seul ensemble des premiers mots et une mesure de confirmation indirecte qui utilise des informations mutuelles ponctuelles normalisées (NPMI) et la similitude des cosinus.
- La mesure C_p : elle fonctionne sur une fenêtre glissante, une segmentation précédente des premiers mots et la mesure de confirmation de la cohérence de Fitelson
- La mesure C_uci : elle est utilisée sur une fenêtre glissante et l'information mutuelle ponctuelle (PMI) de toutes les paires de mots des premiers mots donnés
- La mesure C_umass : elle est faite à partir du nombre de cooccurrences de document, une segmentation précédente et une probabilité conditionnelle logarithmique comme mesure de confirmation
- La mesure C_npmi : elle est une version améliorée de la cohérence C_uci mais utilise les informations mutuelles ponctuelles normalisées (NPMI)
- La mesure C_a : elle est basé sur une fenêtre de contexte, une comparaison par paires des mots principaux ainsi qu'une mesure de confirmation indirecte qui utilise des informations mutuelles ponctuelles normalisées (NPMI) et la similitude des cosinus.

Les méthodes consisteront au calcul d'un graphe avec en abscisse le nombre de Topics à choisir. Selon la courbe on sélectionnera le point le plus pertinent, c'est à dire le moins grand nombre de Topics avant que le score de cohérence ne se mette à baisser.

### 2.4. Représentation des textes et pré-traitements

*2.4.1. Représentation des textes*

Les algorithmes d'apprentissage automatique nécessitent des entrées bien définies et ne peuvent pas travailler directement avec des textes bruts. Une manière de représenter un texte est de le transformer par une série de termes (tokens) qui le constituent.

Pour en extraire des caractéristiques, un moyen simple et populaire est d'utiliser le modèle de sac de mots. Ce

dernier décrit sous forme d'une liste non ordonnée l'occurrence des termes dans un texte. Cela implique un vocabulaire de mots connus et une mesure de la présence de ces mots. L'inconvénient de cette représentation est que toute information sur l'ordre des mots dans le document est supprimée.

Partant de ce modèle, on peut représenter numériquement un corpus par une matrice appelée matrice document-terme (chaque ligne représentant un document et chaque colonne un mot) qui va renseigner la fréquence d'apparition de l'ensemble des mots dans chaque document. Cette matrice constitue un paramètre d'entrée de la LDA.

Notons qu'il est courant d'appliquer sur cette matrice une certaine pondération qui tient compte d'un niveau local (fréquence d'un terme dans un document) et d'un niveau global (distribution d'un terme dans l'ensemble du corpus) permettant ainsi d'évaluer l'importance d'un terme dans un document relativement à un document. Il existe différentes fonctions de pondération dont la fonction TF-IDF (Term Frequency-Inverse Document Frequency) et la fonction Okapi BM25 qui figurent parmi les plus connues. Selon l'algorithme utilisé, on choisira ou non d'appliquer une pondération.

*2.4.2. Pré-traitements*

Avant même de s'intéresser à l'étude des méthodes de modélisation de sujet et avant même la construction de la matrice document-terme, il est très important d'effectuer un ensemble de pré-traitements sur les données brutes car la qualité de la sortie de l'algorithme sera déterminée en grande partie par la qualité de l'entrée. Un des premiers pré-traitements à appliquer est celui de la suppression de la ponctuation. En effet, comme expliqué précédemment, l'ordre d'apparition des mots n'étant pas pris en compte dans les topic models, la ponctuation n'apportent aucune information.

Ensuite afin d'évaluer correctement l'occurrence d'un mot dans un corpus et d'éviter par la même occasion la construction vocabulaire trop conséquent, il convient de transformer en bas de casse (minuscule) l'ensemble des chaînes de caractères du corpus.

Un autre pré-traitement important et des plus basiques est celui de la suppression des mots vides, mots essentiellement grammaticaux tels que les déterminants, les prépositions, les conjonctions et les adverbes qui apportent peu d'information dans la compréhension d'un sujet. Sans l'utilisation d'une mesure de pondération adéquate (pour sélectionner les mots clés), ces derniers de part leur occurrence élevée détériore l'interprétation des sujets en sortie. Selon les topics models utilisés, certains n'acceptent pas directement de matrice document-terme pondérée (ce qui est le cas de la LDA). C'est pourquoi afin de supprimer ces mots vides, on peut faire appel à une des listes prédéfinie qui contiendront les mots à exclure du modèle. Différentes bibliothèques comme NLTK, SpaCy, Gensim ou Scikit-learn ont leur propre liste de mots vides qu'il est possible d'allonger selon le corpus d'étude.

D'autres pré-traitements plus complexes dits linguistiques peuvent être utilisés afin de ramener chaque mot à une forme invariable qui peut être sa racine (racinisation) ou sa forme canonique (lemmatisation). La lemmatisation vise à transformer les formes que peut prendre un mot (singulier ou pluriel, masculin ou féminin, conjugaison pour les verbes) par son lemme, mot qui correspond à la plus petite unité lexicale faisant sens et pouvant constituer une entrée dans un dictionnaire. Elle a pour avantage de regrouper les différentes réalisations d'une même forme tout en augmentant le nombre d'occurrence. Elle a pour inconvénient de nécessiter d'un dictionnaire complexe.

La racinisation (stemming) en anglais) est une opération plus simple à mettre en place mais plus brutale. Elle va augmenter les occurrences d'une racine mais contrairement à lemmatisation, la racinisation d'un mot va la plupart du temps entraîner une perte de sens et ne va pas nécessairement retourner un mot qui existe dans la langue. A noter qu'il existe différents 'stemmers' (Porter, Snowball, Lancaster, Krovetz) dont les résultats diffèrent. Les bibliothèques Python précédemment citées sont aujourd'hui suffisamment puissantes et disposent suffisamment de ressources pour faire des lemmatisation efficaces et dans des temps relativement courts, ce qui fait que la lemmatisation est la plus part du temps privilégiée par rapport à la racinisation.

Enfin on pourra également citer, l'étiquetage morpho-syntaxique (part-of-speech tagging en anglais) qui permet de sélectionner des mots selon leur appartenance grammaticale. En effet, selon le corpus étudié, il peut être par exemple être intéressant de ne conserver que les noms communs.

## 3. État de l'art

### 3.1. Restriction du domaine d'étude

Les méthodes de développement de la *L'allocation de Dirichlet latente* (LDA) obtiennent des informations essentielles à partir de données en utilisant l'inférence bayésienne. On retrouve plusieurs articles notamment pour répondre à des erreurs de généralisation dans la réduction de dimension ou encore pour résoudre des problèmes structurels de mesure du texte et des co-variables.

La LDA est également utilisé pour effectuer du topic modeling. Plusieurs articles font part de cela notamment pour analyser les sentiments, la satisfaction ou connaître les avis des gens sur un sujet. Ce fut dernièrement le cas sur le réseau social Twitter concernant la pandémie du COVID 19, dans un article publié en septembre dernier par plusieurs co-auteurs : Jia Xue, Junxiang Chen et Chen Chen [4]. Une étude vise à comprendre le discours et les réactions psychologiques des utilisateurs de Twitter au COVID-19. Des techniques d'apprentissage automatique pour analyser environ 1,9 million de Tweets liés au coronavirus collectés du 23 janvier au 7 mars 2020.

La LDA est également utilisé pour effectuer de la segmentation d'image. Dans un article du Journal of Marine Science [5] and Engineering parut en février 2021, Xi Yu, Bing Ouyang et José C. Principe utilisent une allocation de Dirichlet latente pour effectuer une segmentation d'image. Ici la lda sert principalement pour générer de manière itérative des pseudo étiquettes à l'aide d'une cohérence spatiale sur l'espace des caractéristiques. La méthode proposée est évaluée sur l'ensemble de données d'images qui permet ainsi d'améliorer les performances de segmentation d'image contre des échantillons faiblement étiquetés et obtenir de meilleurs résultats par rapport à d'autres approches semi-supervisées.

*L'allocation de Dirichlet latente* est aussi utilisée pour extraire et valider des sujets d'intérêt dans un ensemble de données. Dans l'article Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation écrit par plusieurs co-auteurs (Ian Sutherland, Youngseok Sim, Seul Ki Lee, Jaemun Byun et Kiattipoom Kiatkawsin) [6] la LDA est utilisé pour faire du Topic modeling afin d'exploiter de grandes données textuelles non structurées en vue d'étudier la satisfaction des clients et le comportement des consommateurs.

A travers ces nombreux exemple nous vous montrons que *l'allocation de dirichlet latente* peut être utilisé pour traités un grand nombre de sujet tant mathématique que textuelle. Nous allons nous servir de cela afin de faire du *Topic modeling* pour tenter de détecter des liens entre les titres d'articles et le contenu de ces derniers, dans le but d'estimer la pertinence de ces derniers.

### 3.2. Le prétraitement

A partir de la méthode LDA, nous recherchons qu'elles sont les moyens d'analyser les titres de manière pertinente, pour cela différents articles abordent le sujet :

Certains articles illustrent les différents syntagmes sur les prétraitements de *Topics modeling* (notamment Amaury Delamaire et al., 2019 [7] ) avec comme objectif d'améliorer la qualité du l'analyse du modèle LDA. En démontrant qu'une analyse syntaxique, c'est-à-dire une lemmatisation ainsi que la sélection de syntagmes, peuvent compléter les méthodes statistiques.

Par la suite, des auteurs (Fiona Martin et Mark Johnson, 2019 [8]) vont étudiés des méthodes qui vont montrer que les textes lemmatisés réduits aux noms permettent d'obtenir des résultats plus pertinents que des textes lemmatisés.

D'autres auteurs (Qingsheng Wan et al., 2013 [8])) vont directement s'intéresser à l'analyse de texte court, c'est-à-dire en intégrant la modélisation de sujet avec une courte intégration de texte lors de l'inférence de sujets. L'agrégation sera alors fondée sur l'affinité d'actualité. Ces textes courts sont générés à la suite de textes long, et vont regrouper l'essentiel du texte. Cette méthode va alors permettre d'obtenir davantage de pertinences que le fait d'utiliser le topic modeling sur les longs textes.

On peut notamment trouver des études concernant différents prétraitements et leurs performances [9] [10] [11] . Il faut savoir que le la perplexité du modèle et les informations mutuelles entres elles sont sensibles au volume de données et à la taille du vocabulaire. C'est une des choses que l'on abordera dans notre article. Par exemple on comprend que dans certains cas, la méthode de suppression de mots vides peut diminuer la performance des analyses.

### 3.3. Topic Modeling sur tout le texte ou seulement sur l'abstract

La cohérence d'un sujet est basée sur la l'hypothèse selon laquelle les mots ayant une signification similaire ont tendance à se co-produire dans un contexte similaire. La longueur des mots dans un document et la taille du vocabulaire ont des effets sur la cohérence du sujet. Les longs textes sont moins affectées par des termes incorrects ou parasites faisant partie des distributions sujet-mot, ce qui fait que les sujets sont plus cohérents et mieux classés. La différence entre un résumé et un long texte est que les données sont plus apparentes dans un petit document, avec des différences allant jusqu'à 90 pourcent de sujets de haute qualité pour les données en texte intégral, contre 50 pourcent de sujets de haute qualité pour les résumés. L'utilisation de LDA pour découvrir des sujets latents à partir de données textuelles est très courant dans le domaine de recherche. L'utilisation des résumés peut être pratique pour plusieurs raisons. Notamment un accès plus facile aux données abstraites et une rapidité du calcul du temps qui permettrait de réduire les étapes de prétraitement nécessaire lorsque nous travaillons sur un tout un article. Ces étapes de prétraitement sont longues et comprennent plusieurs taches : la recherche sur les référentiels des éditeurs, la conversion de PDF en texte brut (directement ou par la reconnaissance optique de caractères), une phase de nettoyage passe-partout accrue notamment via un logiciel spécifique (OCR)... Cependant, une justification plus scientifique est nécessaire pour aider à choisir un résumé ou un texte intégral pour faire du Topic Modeling avec la LDA. C'est pourquoi nous allons nous intéresser à l'analyse d'abstracts avec la LDA. Notre but étant de perfectionner l'analyse et ainsi produire des résultats cohérents et pertinents.

## 4. Données et expérimentations

### 4.1. Description des données choisies et axe d'analyse

Les données sur lesquelles nous avons décidé de travailler sont des abstracts d'articles scientifiques. Ces derniers sont disponibles sur la plateforme Kaggle (Topic Modeling for Research Articles). Ces articles en question proviennent de différents domaines de recherche dont l'informatique, la physique, les mathématiques, les statistiques ou la biologie quantitative. Nous disposons d'un fichier

d'entraînement contenant plus de 20 000 abstracts associés à un ou plusieurs domaine de recherche. Notre objectif est de voir si l'utilisation d'un topic model sur des abstracts peut nous permettre d'affecter correctement un topic (parmi ceux précédemment cités) à un article en se basant à partir d'un certain seuil sur les probabilités de distribution des topics obtenus dans les documents. Les titres n'ont pas été utilisés pour faire nos analyses. Nous supposons donc ici que le nombre de topics à savoir les domaines de recherches dans lesquels les articles s'inscrivent est connu à l'avance et égale ici à 5.

### 4.2. Expérimentations

#### *4.2.1. Pré-traitements*

Comme nous l'avons détaillé plus haut, la qualité des pré-traitements réalisés est très importante. Si ils ne sont pas toujours garants d'une bonne interprétabilité des topics obtenus en sortie suite l'application d'un topic model, ils sont toutefois indispensables pour espérer obtenir des résultats pertinents.

Les abstracts que nous avions à notre disposition ont nécessité un certain nombre de pré-traitements allant des plus basiques (passage des abstracts en chaîne de caractères minuscules, suppression de la ponctuation, suppression des caractères issus de l'obtention des données textuelles (balises html, retour à la ligne), suppression des chiffres et des éventuels caractères spéciaux) aux plus complexes (lemmatisation, pos-tagging, n-grams).

Pour les traitements les plus basiques nous avons usé d'expressions régulières et utilisé la libraire *spaCy*. Avec cette dernière nous avons également procédé à la suppression des mots vides, à la lemmatisation et au pos-tagging en ne sélectionnant que les noms communs et les noms propres (afin d'interpréter au mieux les topics). En plus des mots vides de base présent dans *spaCy*, nous avons après analyse rajouté les mots les plus fréquemment observés dans les abstracts tel que 'model', 'result', 'problem', 'method', 'approach', 'paper', 'effect', 'application' ...

Nous avons également à l'aide de la librairie *Gensim* procéder à la construction des bigrams et des trigrams. La détermination de ces derniers est assez complexe et repose sur le calcul d'un score qui dépend de l'occurrence des n-grams et d'un seuil prédéfini.

#### *4.2.2. Modèle LDA*

Une fois le vocabulaire obtenu, nous avons réalisé une LDA avec *Gensim* en gardant les paramètres par défaut. Nous avons constaté pour un nombre de topic égale à 5, c'est à dire pour le nombre de domaines de recherche sur les lesquels les abstracts portent, que les sujets n'étaient pas très différents les uns des autres. De plus nous avons cherché à mesurer le score de cohérence *cv* de différents models LDA en faisant varier le nombre de topic entre 1 et 9. Le score de cohérence maximal (score avant que la courbe de cohérence ne se mettent à décroître) à été trouvé pour 4 topics, soit a peu près 0.28. Cette faible valeur bien

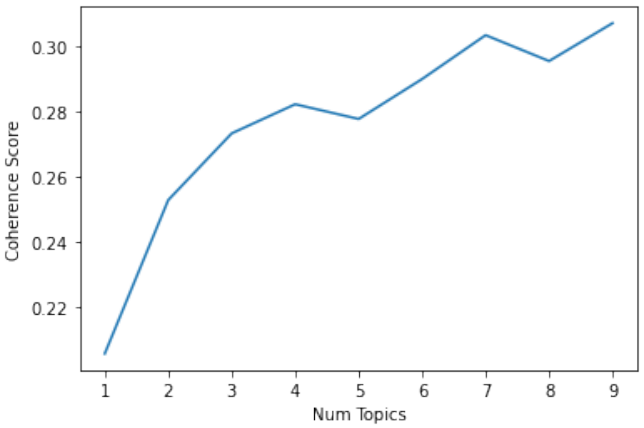


FIGURE 2: Scores de cohérences

que la mesure de cohérence ne soit pas une mesure absolue permettant de déterminer précisément si un topic est interprétable ou non, nous confirme que les topics obtenus ne sont pas très distincts. Nous avons cependant souhaité garder un nombre de topic égale à 5 afin de pouvoir comparer les distributions des topics trouvés aux attributions des domaines de recherche qui ont été faites.

#### *4.2.3. Modèle seeded LDA ou guided LDA*

Les résultats n'étant pas satisfaisant nous avons longuement réfléchit à moyen d'obtenir des topics qui diffèrent un peu plus les uns des autres, bien que la frontière qui distingue ces derniers soit parfois difficile à observer en raison d'un vocabulaire très technique, d'où la complexité de notre problème d'étude. En regardant les différentes variantes de modèle LDA existantes et constant que la LDA 'classique' ne fonctionne pas très bien sur les textes courts tel que les abstracts, nous avons finalement opté pour un modèle LDA semi-supervisé nommé *guided LDA* ou *seeded LDA* qui est un modèle de LDA qui peut prendre en paramètre des listes de mots. Ces listes de mots sont créées de tel sorte à ce qu'on puisse les associées à des topics.

C'est ici que nous avons eu l'idée d'utiliser des mots clés contenus dans des glossaires afin de mieux faire converger la LDA et obtenir ainsi des topics plus distincts. Nous avons ainsi choisir d'extraire les mots de différents glossaires anglais Wikipédia sur les thèmes suivants :

- Physique
- Computer-Science et IA
- Mathématiques
- Biologie
- Probabilités et Statistiques

En exécutant l'algorithme avec les listes de mots que nous avons choisit pour chaque topic, nous avons obtenu en sortie des topics plus clairs, notamment pour les mathématiques, les statistiques, la physique et l'informatique. En revanche nous n'avons pas réussi à déterminer un topic relatif à la biologie quantitative malgré nos listes de mots. En effet, nous avons systématiquement obtenus deux topics plus ou moins équivalents liés au domaine de la physique

et de ce fait nous n'avons pas réellement pu comparer les distributions des topics de chaque document au domaine de recherche pour lesquels les articles ont été associés dans notre jeu de données.

### 4.3. Conclusion

Le sujet de notre article nous a permis de comprendre les enjeux de la LDA. Tout au long de notre analyse nous avons été face à différents problèmes (topics peu différents, scores de cohérences bas...). Cette recherche nous a donc amenés à pousser notre réflexion afin d'améliorer notre modèle. Il nous a fallut nous décider quant à la manière d'apporter notre contribution. C'est pourquoi nous avons choisit la LDA semi-supervisée. C'est l'approche sur laquelle nous pouvions apporter une réelle contribution. Avec celle-ci et grâce à l'utilisation de glossaires, nous avons pu améliorer et trouver des topics assez homogènes pour en faire une classification.

Cependant, des améliorations peuvent s'ajouter à notre analyse, et différentes pistes de recherches pourraient éventuellement venir compléter notre cheminement. Dans un premier temps nous avons pensé à tester le dictionnaire, voir si le fait d'en fournir un plus chargé nous permettrait d'obtenir une meilleur classification. Nous avons aussi pensé à la modification des paramètres de toutes les fonctions utilisées, comme alpha, bêta, ou encore seed-confidence. Toutes ses pistes conduiront potentiellement à d'autres recherches qui viendront compléter les découvertes existantes, afin de pouvoir évoluer dans le domaine du topic modeling et de la classification.

## Références


[1] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, page 17, 1990.

[2] Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. page 20.

[3] David M Blei. Latent Dirichlet Allocation. page 30.

[4] Chen Chen Chengda Zheng Sijia Li Tingshao Zhu Jia Xue, Junxiang Chen. Public discourse and sentiment during the COVID 19 pandemic : Using Latent Dirichlet Allocation for topic modeling on Twitter. 2020.

[5] Journal of Marine Science and Engineering. 2019.

[6] Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. 2020.

[7] Amaury Delamaire, Michel Beigbeder, and Mihaela Juganaru-Mathieu. Exploitation de syntagmes dans la découverte de thèmes. 2019.

[8] Fiona Martin and Mark Johnson. More Efficient Topic Modelling Through a Noun Only Approach. 2015.

[9] Pooja Kherwa and Poonam Bansal. Topic Modeling : A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, 0(0) :159623, July 2018.

[10] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1), September 2016.

[11] Concept-based Topic Model Improvement.

[12] Shaheen Syed and Marco Spruit. Full-Text or Abstract ? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 2017.

[12]