Annotation avec un modèle de sujet LDA

Laura Darenne, Inalco

11 mars 2025

1 Introduction

Le topic modeling est une technique qui permet d'identifier les thèmes sous-jacents dans une collection de documents pour la classification de documents, l'analyse de sentiments, la recommandation de contenu.

Le Latent Dirichlet Allocation (LDA) [1] est un algorithme utilisé pour le topic modeling qui repose sur une approche probabiliste pour modéliser les documents. LDA suppose que chaque document est un mélange de plusieurs sujets, et que chaque sujet est une distribution de mots. Mathématiquement, cela signifie que chaque document peut être représenté comme une distribution sur un nombre fixe de sujets, et chaque sujet comme une distribution sur un vocabulaire fixe de mots.

2 Implémentation de LDA

Gensim [2] sera utilisé pour implémenter cette méthode. Les principales variables que l'ont retrouve sont :

- `corpus` est une liste de documents ;
- `docs` est une liste de documents tokenisés, lemmatisés (pour l'anglais) et filtrés (pour les stopwords et la ponctuation) ;
- `id2word` est un dictionnaire ayant pour clé les vecteurs et en valeur des mots, nécessaire pour l'entraînement du modèle ;
- `dictionary` est le dictionnaire, un objet spécifique à Gensim (il a sa propre classe) de n unique tokens.
- `model` est le modèle LDA entraînés [3].

Pour l'analyse des documents, Gensim permet le filtrage des sujets selon leur poids dans le document avec :

```
get_document_topics(corpus, minimum_probability=conf.minimum_probability
```

Par exemple, pour un document composé de 55% du sujet "santé", 25% du sujet "technologie", et 20% du sujet "éducation", un minimum de 30 % filtrerait les sujet "technologie" et "éducation".

Pour l'analyse des sujets, Gensim

1. Blei, D. and Ng A. and Jordan, M. (2003). Latent Dirichlet Allocation.
2. Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora.
3. Dont vous pouvez retrouver les spécificités ici : `https://radimrehurek.com/gensim/models/ldamodel.html`

3 Évaluation

Cette section se base en partie sur l'État de l'art réalisé par Elizabeth Savatier dans son mémoire « Suivi de la propagation de sujets dans les médias ».

Il est nécessaire de déterminer si les groupes (*clusters*) obtenus peuvent être considérés comme des « bons sujets ». Plusieurs données sont à prendre en compte :

- **Cohérence** : Les mots qui composent le sujet doivent être cohérents entre eux et former un ensemble logique. Elle est évaluable de manière quantitative avec le score de cohérence. C'est une mesure de la qualité des sujets générés par le modèle, évaluant la similarité sémantique entre les mots les plus probables dans chaque sujet. Une cohérence élevée indique que les mots d'un sujet sont fortement liés sémantiquement, ce qui suggère que le sujet est bien défini et interprétable ;
- **Qualité des groupes** : Les éléments d'un groupe doivent être les plus proches possibles les uns des autres. Les groupes doivent être éloignés les uns des autres. [4] Parmi les nombreuses mesures existantes, trois seront présentés plus loin ;
- **Stabilité** : Les sujets doivent être stables lorsque le modèle est exécuté plusieurs fois sur le même corpus de données. Cela indique que le modèle est robuste et fiable.
- **Interprétabilité et Granularité** : Les sujets doivent être facilement interprétables par un humain. Les mots qui composent le sujet doivent permettre de comprendre de quoi il s'agit. Les sujets ne doivent être ni trop généraux ni trop spécifiques.

3.1 Évaluation de l'entraînement

Les mesures conventionnelles d'exactitude ne sont pas applicables aux résultats des modèles LDA, étant donné l'absence de sujets de comparaison. Par conséquent, il est important de surveiller l'apprentissage du modèle à l'aide des mesures de perplexité, cohérence et convergence. Ces mesures sont calculées par la même bibliothèque que celle utilisée pour entraîner le modèle [5].

La perplexité est une mesure utilisée pour évaluer la capacité d'un modèle à prédire une distribution de probabilité. Une valeur plus faible de perplexité indique un meilleur ajustement du modèle aux données.

$$perplexité = 2^{-\frac{bornein férieuredelavraisemblancelogarithmique}{nombretotaldemotsdanslecorpus}}$$

La borne inférieure de la log-vraisemblance du corpus, toujours inférieure ou égale à la log-vraisemblance du corpus, fournit une estimation inférieure de la qualité du modèle de sujet sur le corpus donné, représentant ainsi le scénario le moins favorable. Cette mesure est calculée pour chaque document du corpus et ensuite additionnée sur tous les documents. Elle est particulièrement utile pour comparer différents modèles de sujets et pour déterminer le nombre optimal de sujets à utiliser dans un modèle.

La cohérence est une mesure de la qualité des sujets générés par le modèle LDA, évaluant la similarité sémantique entre les mots les plus probables dans chaque sujet. Une cohérence élevée indique que les mots d'un sujet sont fortement liés sémantiquement, ce qui suggère que le sujet est bien défini et interprétable. [6].

La convergence évalue si le modèle LDA a atteint un état stable où les paramètres du modèle ne changent plus de manière significative. Une faible convergence indique que le modèle

4. Berry, M. J. and Linoff, G. S. (1997). Data mining techniques : For Marketing, Sales, and Customer Support. John Wiley & Sons.

5. `https://radimrehurek.com/gensim/models/callbacks.html`

6. Gensim implémente la pipeline de Röder, M. and Both, A. and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures.

apprend peu ou a terminé son apprentissage. Par exemple, les résultats ci-dessous montrent que les distances sont presque égales à 1 pour chaque sujet à la première époque, car le modèle apprend beaucoup. À la cinquième époque, les distances sont beaucoup plus petites, indiquant que le modèle se stabilise.

```
    Epoch 1
[1., 1., 0.99497487 0.99497487 0.99497487, 1., 1., 1., 1., 1.,
1., 1., 1., 0.98989899, 0.99497487, 1., 0.99497487, 0.99497487,
0.99497487, 1.]
    Epoch 5
[0.51923077, 0.26470588, 0., 0., 0.26470588, 0., 0., 0., 0., 0.,
0., 1., 0.88317757, 0., 0.51923077, 0.13366337, 0., 0.51923077,
0.13366337, 0.]
```

La convergence est la somme des distances de chaque sujet entre deux modèles. Une mesure possible pour calculer cette distance est la distance de Jaccard, dont la valeur se situe entre 0 (haute similarité) et 1 (basse similarité). Pour deux modèles epoch n et epoch $n-1$, nous avons la somme des distances de Jaccard tel que :

$$distancejaccard = 1 - \frac{|sujetdumodèlen \cap sujetdumodèlen - 1|}{|sujetdumodèlen \cup sujetdumodèlen - 1|}$$

3.2 Évaluation quantitative des groupes

Évaluer la qualité des clusters, c'est mesurer la cohésion interne des clusters, c'est-à-dire la proximité des mots au sein d'un même sujet, et la séparation entre les clusters, c'est-à-dire la distance entre les différents clusters. Trois indices, implémentés par Scikit-Learn sont utilisés dans le fichier python :

- **Indice de Davies Bouldin** qui varie entre 0 (meilleure classification) et $+\infty$ (pire classification) ;
- **Indice de Calinski-Harabasz** qui varie entre 0 (pire classification) et $+\infty$ (meilleure classification) ;
- **Le coefficient de silhouette** qui varie entre -1 (pire classification) et 1 (meilleure classification).

Il est important de noter que ces mesures prennent toutes en entrée une représentation TF-IDF des documents et un sujet par document.

3.3 Évaluation qualitative

Les évaluations qualitatives pour évaluer des groupes sont en général des tâches permettant aux humains d'évaluer la cohérence des groupes. Une tâche assez répandue consiste à attribuer une note entre 1 et 3 à chaque groupe selon un critère fixé en amont (généralement la cohérence).

Deux autres tâches populaires, la détection d'un mot intrus (*word intrusion*) et la détection d'un sujet intrus (*topic intrusion*), permettent d'évaluer les sujets (qui contiennent un mot intrus) et l'affectation des sujets aux documents (qui contiennent un sujet intrus). Si l'être humain parvient à détecter le mot intrus dans le sujet ou le sujet intrus dans la liste des sujets affectés à un document, alors les sujets et l'attribution de ces derniers aux documents sont cohérents.

Comme nous souhaitons savoir si notre approche permet de créer des groupes représentant chacun un sujet, pour chaque groupe, nous nous demanderons si ses éléments discutent d'un même sujet. Pour ce faire, nous observerons les visualisations en deux dimensions des groupes d'articles, à l'aide d'un outil de visualisation tel que pyLDAvis. Le titre étant en général repré-

sentatif du sujet dont l'article discute, les documents seront représentés par leur titre. Couplé avec l'observation des groupes d'articles, nous observerons aussi les mots représentatifs des sujets dégagés par notre modèle de sujet. Il est important de noter que la visualisation en deux dimensions n'est pas très représentative des groupes lorsque les éléments de ces derniers ont plusieurs dimensions.