Comparaison de méthodes de modélisation de sujets classiques face à ChatGPT et aux LLM

Samuel Gonçalves

Sous la direction de Fabrice BOISSIER et Marie PUREN

MNSHS - RDI 2025

Le contexte

Les LLM

- Grands modèles de langage
- Beaucoup d'applications

Le contexte

Les LLM

- Grands modèles de langage
- Beaucoup d'applications

La modélisation de sujet

- Résumé de texte
- Sujet : ensemble de mot
- Thème : Nom attribuable à un sujet

Le contexte

Les LLM

- Grands modèles de langage
- Beaucoup d'applications

La modélisation de sujet

- Résumé de texte
- Sujet : ensemble de mot
- Thème : Nom attribuable à un sujet

La subjectivité des résultats

- Bon/mauvais résumé ? Meilleur/pire qu'un autre ?
- Métriques

Introduction au protocole développé

Étapes du protocole

- Pré-traitement
  - Préparation des textes
- Traitement
  - Modélisation de sujets
- Évaluation

Présentation détaillée du travail - Pré-traitement

2 besoins différents

- Pré-traitement léger
  - Textes extérieurs à la base de donnée
  - LLM
- Pré-traitement fort
  - Textes du corpus uniquement
  - Uniquement les notions

- Dégradé de pré-traitement entre les 2 logiques

Présentation détaillée du travail - Méthodes (1)

Méthodes adaptées aux pré-traitements forts

Allocation de Dirichlet Latente (LDA)

- Document : Combinaison de sujets
- Sujet : Ensemble de mots
- Mot : ...

Présentation détaillée du travail - Méthodes (1)

Méthodes adaptées aux pré-traitements forts

Allocation de Dirichlet Latente (LDA)

- Document : Combinaison de sujets
- Sujet : Ensemble de mots
- Mot : ...

CREA

- Matrice d'occurence notion/texte
- Treillis de Galois
- Similarité conceptuelle

Présentation détaillée du travail - Méthodes (2)

Méthodes adaptées aux pré-traitements faibles

Llama2, GPT...

- Prompt aligné aux besoins
  - .json : output = [*sujet*] ; sujet = [*mot*]
- Problèmes liés à ces outils

Présentation détaillée du travail - Méthodes (2)

Méthodes adaptées aux pré-traitements faibles

Llama2, GPT...

- Prompt aligné aux besoins
  - .json : output = [*sujet*] ; sujet = [*mot*]
- Problèmes liés à ces outils

Méthodes basées sur BERT

- Utilisation de *paraphrase-MiniLM-L6-v6*
- Vectorisation de phrases

→ Méthodes classiques ((PCA, Kmeans), (Umap, Hdbscan), ...)

Présentation détaillée du travail - Évaluation

Cohérences

- Cohérence UMass
- Cohérence V

- Évaluation qualitative

Problèmes rencontrés

- Compatibilité pré-traitement / traitement
- Vectorisation de phrases
  - Par ligne
  - Par notion
- Problèmes de LLM
- Subjectivité des métriques

Résultats - Corpus et scénarios

Corpus utilisé: Issu de la validation de CREA [1] → Cours de PHP

test_scenario	Cours 1, 2 et 3 (pour tester les méthodes)
php_courses	Cours de 1 à 19
php_slides	Cours possiblement bruité par OCR
php_texts	Cours sous format texte à l'origine
...	...

→ Utilisation différente selon les méthodes

Résultats - Traitements

- Raw+Babelfy+LDA
- Raw+TreeTagger+Babelfy+LDA
- Raw+RNNTagger+Babelfy+LDA
- Raw+LDA
- Raw+Babelfy+CREA
- Raw+RNNTagger+Babelfy+CREA
- Raw+RNNTagger+LDA
- Raw+TreeTagger+Babelfy+CREA
- Raw+TreeTagger+LDA
- Raw+Babelfy+BERT_(01p, 05p, 1p, 10, 20)
- Raw+BERT_(01p, 05p, 1p, 10, 20)

Résultats - Cours - Cohérence V

	Raw	Babel	Tree	RNN	Tree +Babel	RNN +Babel
LDA	0.465	0.368	0.436	0.445	0.464	0.345
CREA		0.629			**0.631**	0.525
BERT_01p	0.424	0.481				
BERT_05p	0.387	0.47				
BERT_1p	0.385	0.474				
BERT_10	0.437	0.546				
BERT_20	0.425	0.523				

→ **page web dynamique ; changer ; python ; fermant ; gras ; texte gras ; situer ; textuel ; optionnel ; choisir ; déclencher ; ...**

Sujets très longs (tous les mots sont inclus), basés sur les co-occurences.

Résultats - Cours - Cohérence UMASS

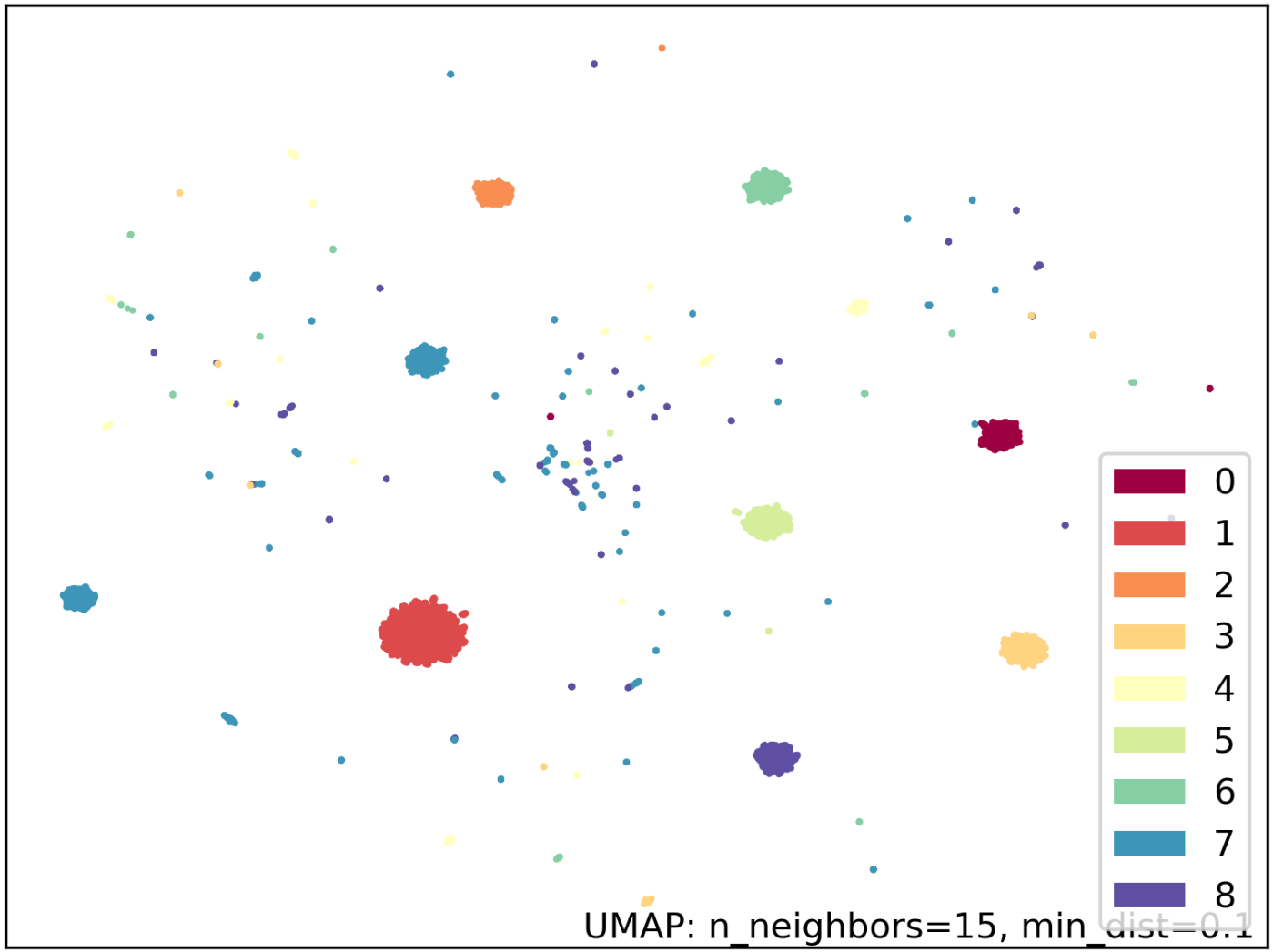
nan : Au moins 1 sujet ne contient qu'un mot (phrase/notion répétée)

	Raw	Babel	Tree	RNN	Tree +Babel	RNN +Babel
LDA	0.947	**0.969**	0.895	0.959	0.943	0.863
CREA		0.011			0.003	0.039
BERT_01p	nan	nan				
BERT_05p	0.014	nan				
BERT_1p	0.004	nan				
BERT_10	nan	nan				
BERT_20	nan	nan				

→ **php ; nom ; pages ; valeurs ; echo ; fonction ; fichier ; type**

Sujets très semblables les uns par rapport aux autres.

Résultats - Cours - Visualisation des résultats de BERT_1p



Sujet 1 : **php**

Futur du projet

- Gestion des LLM
  - Coût des tokens de GPT
  - Résumer le texte... en résumant le texte à l'avance ?
    - ★ Par paragraphe
    - ★ Par vecteurs (CLIP ?)
- Amélioration des méthodes basées sur BERT
  - Marquer les stopwords
  - Modèle de gestion du bruit
- Corpus de textes
  - Impact du bruit
  - Multi-thématique
- Multiplier les métriques

Références I

[1] F. Boissier, "CREA: méthode d'analyse, d'adaptation et de réutilisation des processus à forte intensité de connaissance: cas d'utilisation dans l'enseignement supérieur en informatique," PhD Thesis, Université Panthéon-Sorbonne-Paris I, 2022. [Online]. Available: https://theses.hal.science/tel-03774087/

[2] D. Angelov, "Top2vec: Distributed representations of topics," 2020. [Online]. Available: https://arxiv.org/abs/2008.09470

[3] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com

[5] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[6] F. Hayat, S. Shatnawi, and E. Haig, "Students' experiences and challenges during the covid-19 pandemic: A multi-method exploration," in *European Conference on Technology Enhanced Learning*. Springer, 2024, pp. 152–167.

[7] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," *Programming with TensorFlow: Solution for Edge Computing Applications*, pp. 87–104, 2021.

[8] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018.

Références II

[9] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[10] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[11] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[12] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Shanghai China: ACM, Feb. 2015, pp. 399–408. [Online]. Available: https://dl.acm.org/doi/10.1145/2684822.2685324

[13] H. Rahimi, J. L. Hoover, D. Mimno, H. Naacke, C. Constantin, and B. Amann, "Contextualized topic coherence metrics," 2023. [Online]. Available: https://arxiv.org/abs/2305.14587

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084