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

Samuel Gonçalves

Sous la direction de Fabrice BOISSIER et Marie PUREN

MNSHS - RDI 2025

Question de recherche et objectifs

Question de recherche

Dans quelle **mesure** les LLM sont-ils **pertinents** dans le résumé de **textes**, en comparaison des méthodes de modélisation de sujets ?

- Comment mesurer ça ?
- Comment définir la pertinence dans ce contexte ?
- Quel genre de textes ? Quelle base de données ?

Intégration de la méthode **CREA** aux comparaisons préexistantes.

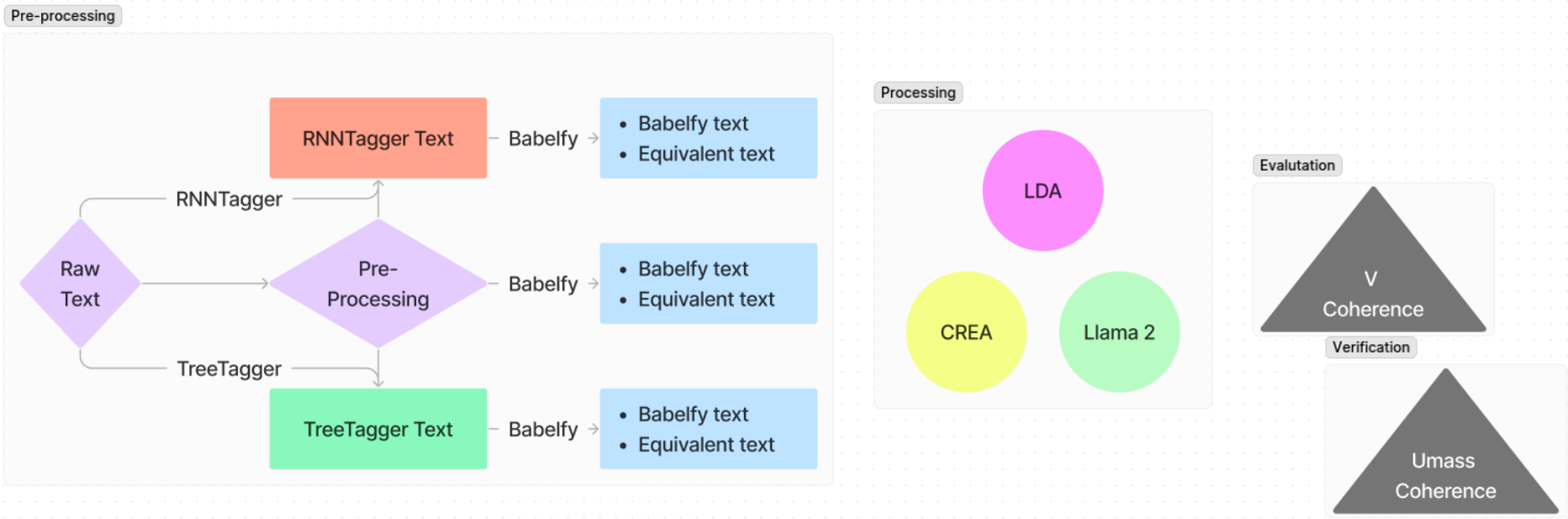
État du travail le 23/02/2024 (LT1)

- État de l'art des **LLM**
  - ChatGPT
  - Llama 2
- État de l'art en **Topic Modeling**
  - LDA
- Choix des outils et bibliothèques utilisées
  - RNNTagger, TreeTagger, Babelfy
  - openai, replicate, gensim, nltk

État du travail le 15/04/2024 (LT2)

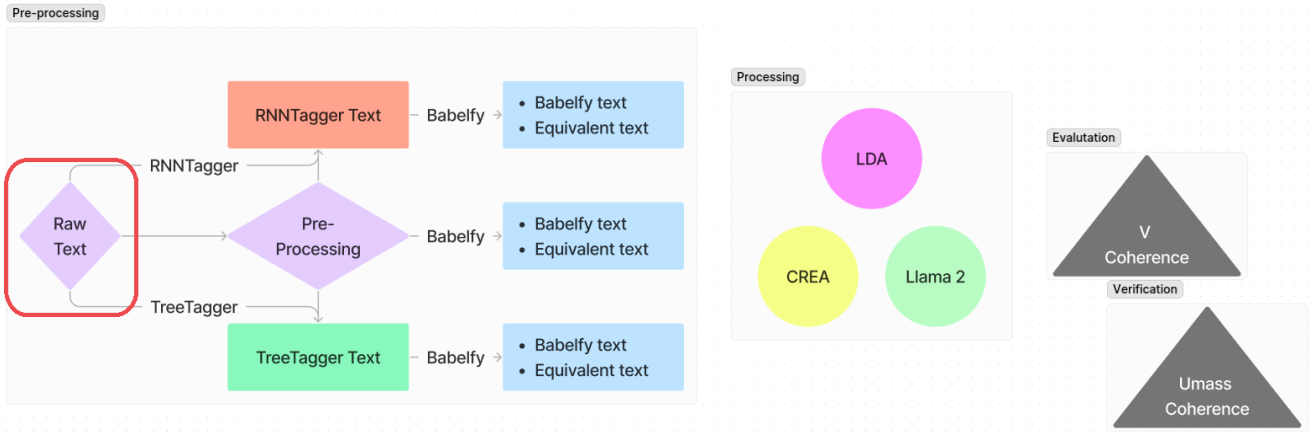
- Implémentation...
- Évaluation des résultats
  - Cohérences
    - ★ Cohérence V
    - ★ Cohérence UMASS

Protocole développé



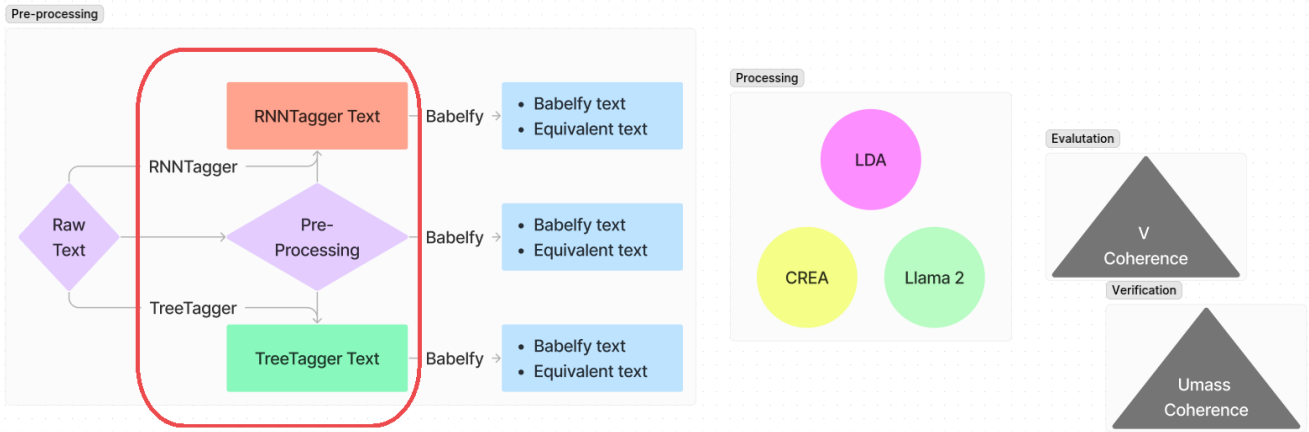
- Scénarios

Textes d'entrées



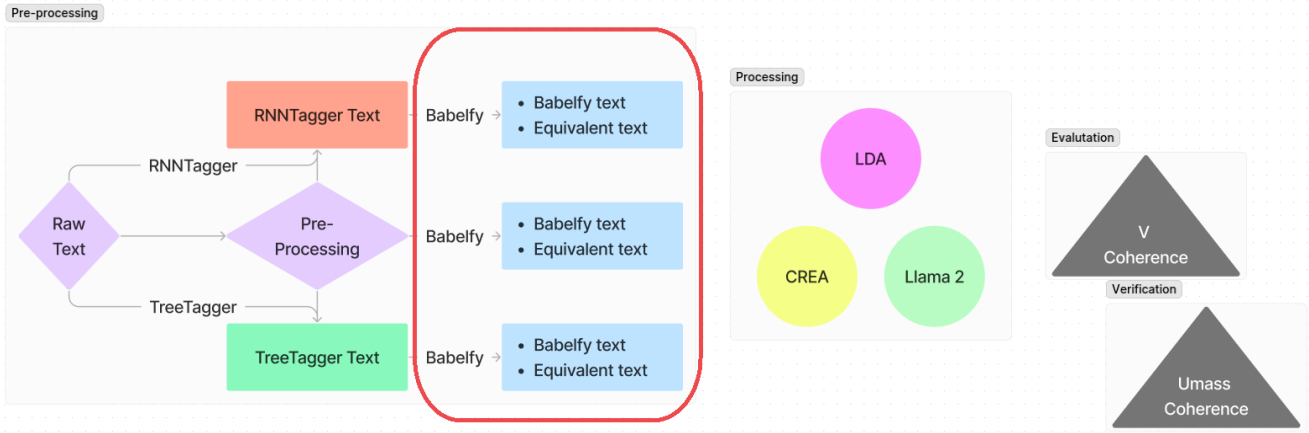
- En français
- Sujet
  - Cours de PHP
- Format
  - Slides
  - Textes
  - Code

RNNTagger
TreeTagger
NLTK



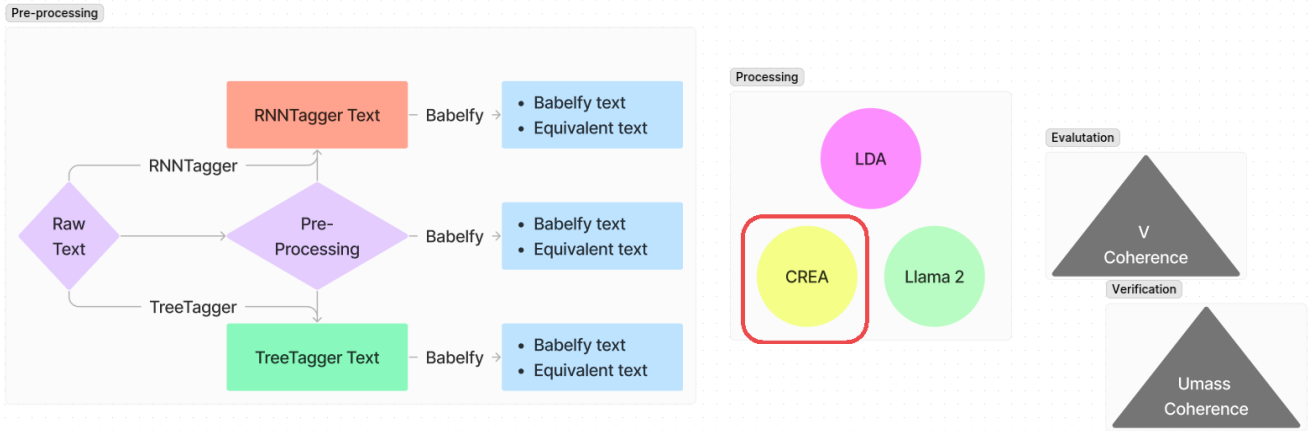
- Tokenisation / Pré-traitement
  - NLTK
- Lemmatisation
  - TreeTagger (masculin singulier, indicatif) - stopclasses
  - RNNTagger (deep learning) - stopclasses / erreurs

Babelfy



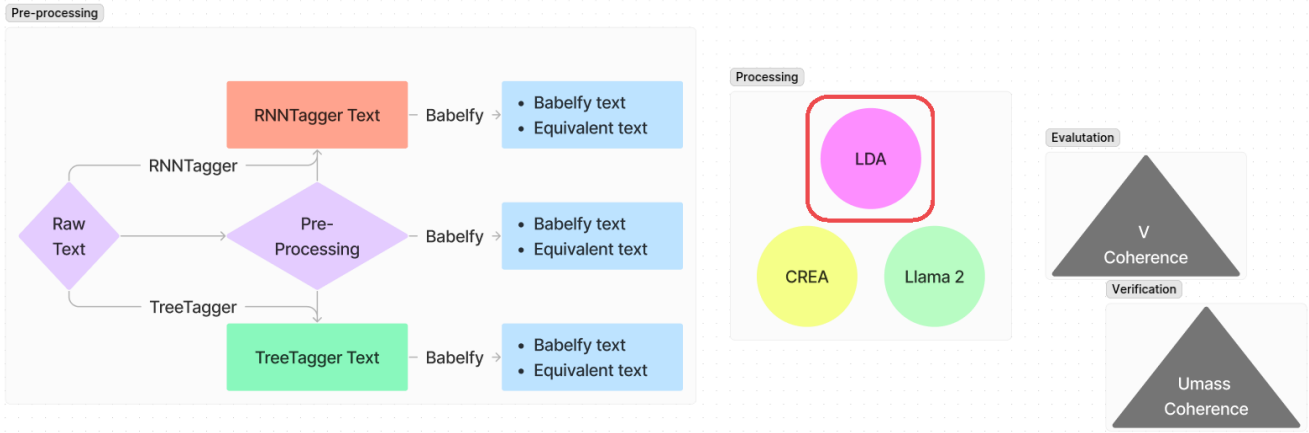
- Notions / Scores
- Identifiants / Dictionnaire

CREA



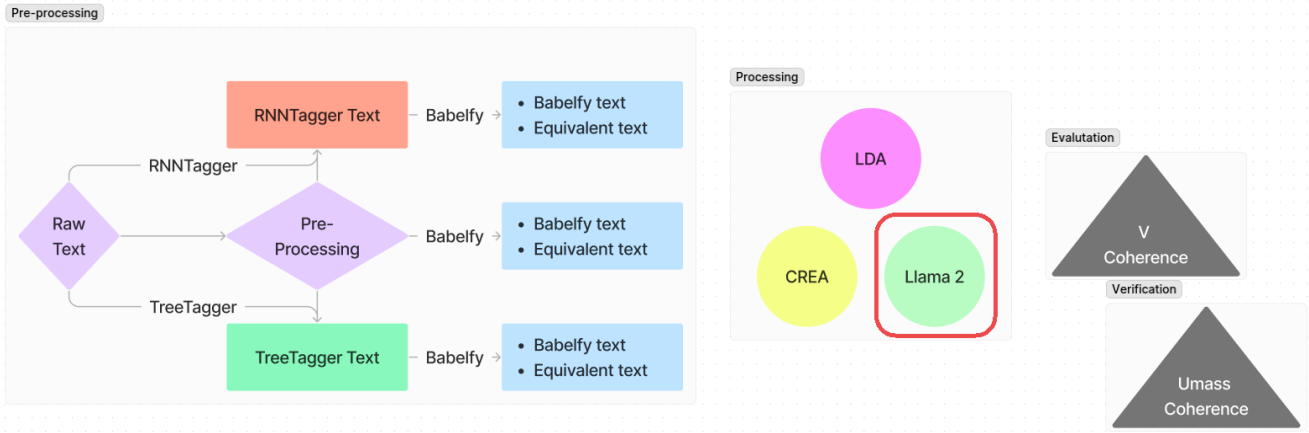
- Occurrences / document
- Double-normalisation
- Concept formel
- Treillis de Galois
- Similarité conceptuelle
- Classification hiérarchique $\rightarrow$ Clustering

LDA



- Distribution de probabilité de Dirichlet des mots / sujets / documents
- Jusqu'à stabilité : itération sur les mots / sujets / documents

Llama 2 et autres LLM

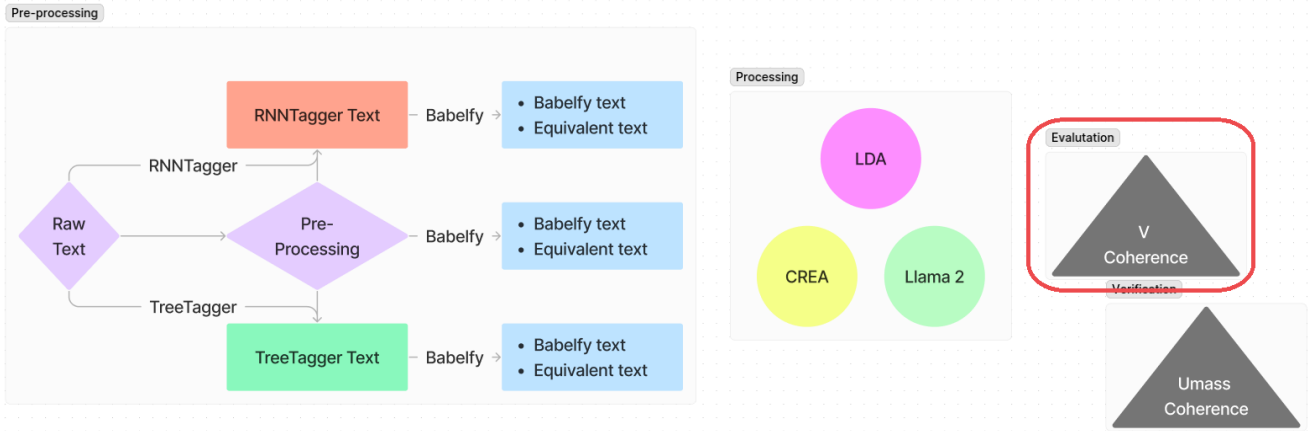


Problème de taille !

- Création du prompt
  - Format
  - Objectif
  - Texte d'entrée
  - ... ?

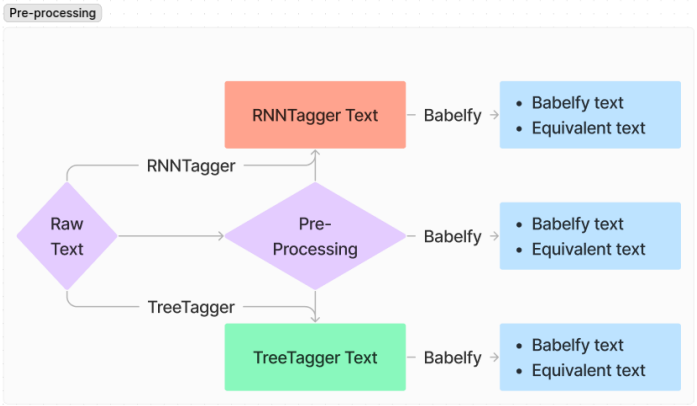
- Génération des sujets à partir du prompt

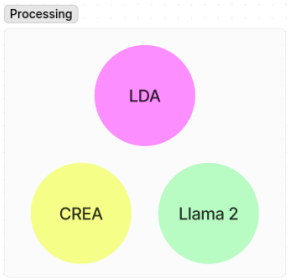
Cohérence V

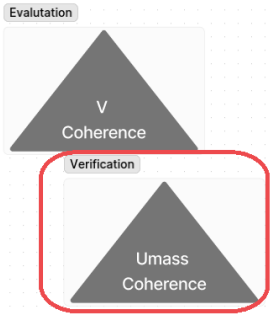


- Entre 0 et 1
- Cohérence principale (état de l'art) $\rightarrow$ Utilisée dans les résultats
- Représentation: $c_v = (P_{sw(110)}, S^{one}_{set}, m_{cos(nlr,1)}, \sigma_a)$
  - $P_{sw(110)}$ - Fenêtre coulissante booléenne de taille 110
  - S^{one}_{set} - Pour chaque mot, traitement avec tous les sous-ensembles possibles (disjoints ou non !)
  - $m_{cos(nlr,1)}$ - Mesure de similarité cosinus - appliqué sur des vecteurs contextuels représentant les mots
  - σ_a - Agrégation par la moyenne

Cohérence UMASS







- Entre $-\infty$ et 0
- Cohérence secondaire (évaluation des tendances)
- Représentation: $c_{Umass} = (P_{bd}, S^{one}_{pre}, m_{lc}, \sigma_a)$
  - P_{bd} - Document booléen
  - S^{one}_{pre} - Pour chaque mot, traitement de sa paire avec tous les précédents
  - m_{lc} - Log-probabilité conditionnelle
  - σ_a - Agrégation par la moyenne

Résultats (I)

- Pour LDA :
  - Inefficaces (0.35) : Babelfy / RNNTagger+Babelfy
    → Dont RNNTagger+Babelfy dernier selon Umass
  - Efficaces (0.45) : RNNTagger / TreeTagger / TreeTagger+Babelfy / Simple
    → Dont RNNTagger premier selon Umass

Divergence entre les cohérences Umass et V

- Pour CREA (uniquement via Babelfy) :
  - Babelfy+RNNTagger moins sensible que les deux autres (cas du fichier CJA.txt) mais moins performants en cas normal (autres cas)

Résultats (II)

Divergence GIGANTESQUE entre les cohérences Umass et V

- LDA vs CREA (uniquement via Babelfy) :
  - Avec CJA.txt - Environ 0.4 pour LDA et 0.4 pour CREA
  - Sans CJA.txt - Environ 0.4 pour LDA et 0.6 pour CREA

→ Comment expliquer cette divergence ?

- Intrinsèque à CREA (résultat précédent)
- Sujets recouvrants / non-recouvrants
  - Très centré sur php
  - Document booléen vs Fenêtre booléenne
- Erreur d'implémentation

Travail futur

- Résoudre le conflit "Cohérence V / Cohérence UMASS"
- Résoudre la limitation de taille des prompts
- Inclure ChatGPT à l'analyse
- Inclure d'autres pré-traitements
- Appliquer le protocole à un corpus de textes moins centré sur un thème/mot précis (php)

Références I

[1] R. Belohlavek, "Introduction to formal concept analysis," *Palacky University, Department of Computer Science, Olomouc*, vol. 47, 2008.

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

[3] 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

[4] 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/

[5] E. Ekinci and S. İlhan Omurca, "Concept-lda: Incorporating babelfy into lda for aspect extraction," *Journal of Information Science*, vol. 46, no. 3, pp. 406–418, 2020.

[6] B. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster analysis*, 5th ed. Wiley, 2011.

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

[8] A. Jaffal, "Aide à l'utilisation et à l'exploitation de l'analyse de concepts formels pour des non-spécialistes de l'analyse des données," Ph.D. dissertation, Université Panthéon-Sorbonne-Paris I, 2019.

Références II

[9] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178–203, Apr. 2011. [Online]. Available: https://doi.org/10.1007/s10791-010-9141-9

[10] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, and T. Häussler, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," in *Computational methods for communication science*. Routledge, 2021, pp. 13–38. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003082606-2/applying-lda-topic-modeling-communication-research-toward-valid-reliable-methodology-danie C3%A4ussler-schmid-petri-adam

[11] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.

[12] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 231–244, 2014.

[13] 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.

Références III

[14] 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

[15] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *New methods in language processing*, 1994, p. 154.

[16] ——, "Improvements in part-of-speech tagging with an application to german," in *In Proceedings of the ACL SIGDAT-Workshop*, 1995, pp. 47–50.

[17] ——, "Deep learning-based morphological taggers and lemmatizers for annotating historical texts," in *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 2019, pp. 133–137.

[18] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Ordered Sets*, ser. NATO Advanced Study Institutes Series, I. Rival, Ed. Springer Netherlands, 1982, vol. 83, pp. 445–470. [Online]. Available: http://dx.doi.org/10.1007/978-94-009-7798-3_15