

Master thesis : Generating Topic Models from Corpora Across Languages

Auteur : Thielen, Benoit

Promoteur(s) : Ittoo, Ashwin

Faculté : Faculté des Sciences appliquées

Diplôme : Master en sciences informatiques, à finalité spécialisée en "intelligent systems"

Année académique : 2021-2022

URI/URL : <http://hdl.handle.net/2268.2/13874>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



UNIVERSITY OF LIEGE
FACULTY OF APPLIED SCIENCES

Generating Topic Models from Corpora Across Languages

Author:

Benoit THIELEN

Supervisor:

Prof. Ashwin ITTOO

End of study work carried out with a view to obtaining a Master's degree in
"Computer Science" by Benoit THIELEN.

Academic Year 2021-2022.

Summary

Topic modeling is a learning process aiming to analyze texts to discover their topic composition by associating groups of correlated words. Historically, topic modeling has used unsupervised learning techniques. Bayesian generative models, such as Latent Dirichlet Allocation (LDA), have quickly proven their performance for representing with probabilities the distributions of words across topics and of topics across documents. Recently, new topic models based on LDA have emerged, like the Hierarchical Dirichlet Process (HDP) which self-determines the number of topics in the text and the nested Hierarchical Dirichlet Process (nHDP) which enables a hierarchical representation of the topics.

The performances in topic identification and hierarchical modeling of HDP and nHDP were evaluated in this work, on English and French corpora built from Wikipedia articles. A large number of very coherent and interesting topics were detected in both languages, despite the presence of some less coherent ones. Correlations have been highlighted between the statistics of the corpus and evaluation metrics such as coherence and model perplexity.

Additionally, a more recent approach of learning word embeddings in hyperbolic space, specifically in the Poincaré ball space, has been studied to determine if it could constitute a promising approach to hierarchical topic modeling. Poincaré embeddings of 10 dimensions were trained on hypernymy relations of our English corpus. Our analysis revealed clusters of words which can be linked to topics, unfortunately the 2D representation method we applied did not allow to show hierarchical relations between those clusters.

In conclusion, both HDP and nHDP models have shown good and similar learning performances when trained on French and English corpora, nHDP being also efficient in providing hierarchical representation of the topics. The Poincaré embeddings were successful in learning and representing the hypernymy relations in the Poincaré ball, however suffered from the constraints imposed by the data acquisition methods and required filtering processes.