Sujet de TER

Titre

Les metromap de l'information

Sujet

Digérer les masses importantes d'informations diffusées sur les plate-forme en ligne, qu'il s'agisse de réseaux sociaux, blogs, articles de presse, est difficile pour un simple être humain. Il faut alors recourir à des techniques d'analyse de données textuelles afin de calculer une forme de résumé de l'information. Parmi les techniques existant dans la littérature, la modélisation thématique (Blei et al., 2003) est très employée. Elle est intégrée dans la plupart des modules de fouille de données disponibles, en particulier en Python. Une autre manière de résumer l'information est de recourir à des "cartes de métro" (metromap) de l'information (Shahaf et al., 2012). Ces deux techniques n'ont, à ma connaissance, pas encore été combinées pour résoudre le problème du résumé d'une large collection de documents.

Objectif

- lire et comprendre l'état de l'art
    - acquérir une connaissance minimale du (vaste) champ de recherche d'information : introduction et principalement les chapitres 1, 2, 4, 6 et 7 du livre de Manning et al. 2008
    - comprendre les techniques qui se trouvent derrière le résumé à base de metromap : deux articles de Shahaf et al. 2012
- réimplémenter l'affiche d'une metromap à partir d'une collection de documents
    - vous pouvez vous aider du rapport réalisé par D. Talary-Brown en 2007
    - vous veillerez à ce que le code de la visualisation soit bien séparé de celui qui extrait la carte à partir des données textuelles
    - privilégiez un langage Javascript avec des librairies simples à installer et à utiliser
    - je peux fournir un sous-ensemble d'une large collection d'articles de presse publiés par le New-York Times, mais vous pouvez aussi travailler sur un jeu de données de votre choix.
- réfléchir à la manière d'intégrer des thématiques dans ce cadre
    - cette partie peut rester purement théorique, c'est-à-dire sans proposition d'un algorithme ni d'une implémentation

Références

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Manning C. D., Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. https://nlp.stanford.edu/IR-book/
- Shahaf, D., Guestrin, C., & Horvitz, E. (2012). Trains of thought: Generating information maps. In Proceedings of the 21st international conference on World Wide Web (pp. 899-908). ACM.
- Shahaf, D., Guestrin, C., & Horvitz, E. (2012). Metro maps of science. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1122-1130). ACM.
- Talary-Brown, D 2017, Mind the Gap: Newsfeed Visualisation with Metro Maps. Department of Computer Science Technical Report Series, Department of Computer Science, University of Bath, Bath, U. K.

Contact

julien.velcin@univ-lyon2.fr