TER L3 : Comparaison de titres de documents

Encadrant : Michel Leclère

Contact : leclere@lirmm.fr / 04 67 41 86 16

Contexte

L'ABES (agence bibliographique de l'enseignement supérieur) gère le catalogue des métadonnées documentaires de toutes les bibliothèques de l'enseignement supérieur français. Ce catalogue contient des notices documentaires décrivant un document : titre, auteur, date, mots-clés, éditeur…

Pour les auteurs, on n'indique pas directement leur nom dans la notice documentaire ; on référence une notice dite d'autorité qui contient des informations sur l'auteur : nom, date de naissance, sexe…

L'objectif est de s'assurer que les documents référencent les « bonnes » autorités et non des autorités « homonymes ». Pour cela, une méthodologie de regroupement des documents en fonction de critères de similarité a été mise en place ; l'idée étant que deux documents écrits par le même auteur ont des caractéristiques communes : des co-auteurs communs, des dates de publications proches, des mots-clés similaires…

Objectifs

Le but du TER est de contribuer à cette méthodologie en proposant un critère exploitant les titres des documents et permettant d'évaluer la possibilité que deux documents soient de(s) même auteur (désambiguïsation d'auteurs) et/ou d'évaluer la proximité de leurs champs thématiques.

Sujet

Pour une telle tâche, on pourra mettre en œuvre une méthode consistant à présenter les titres sous forme vectorielle afin de mesurer la proximité entre titres. Des pondérations statistiques pourront être associées afin de ne pas rapprocher des titres partageant des mots « vides de sens » (de, la, introduction…) ou trop fréquents (France, histoire…). A l'inverse l'utilisation d'une distance d'édition entre mots ou l'utilisation d'outils de traitement linguistique de types lemmatiseur seront mis en œuvre pour tenir compte de variations morphologiques d'un même terme ou de la présence de coquilles. On pourra par ailleurs envisager d'exploiter un thesaurus pour rapprocher des mots sémantiquement proches. L'ensemble du processus à mettre en place sera rigoureusement évalué sur les données de l'ABES un fichier de 12 millions de titres de documents.