M2 MALIA : cas d'étude

Julien Velcin, Université Lyon 2, Laboratoire ERIC

Février 2024

La commande qui vous est fixée consiste à développer un système de recherche d'information qui permet de naviguer efficacement dans un grand corpus de données textuelles. Il s'agit d'une base d'articles scientifiques publiés en Informatique et indexés dans la célèbre base DBLP (https://dblp.uni-trier.de) et quelques autres sources. En plus du titre, voire du résumé écrit en langage naturel, de l'article, vous disposez d'autres informations intéressantes : auteur(s) de l'article, la date, références bibliographiques, journal ou conférence de publication.

Le jeu de données considéré est tiré du Citation Network Dataset qui rassemble plusieurs millions d'articles : https://www.aminer.org/citation. Vous utiliserez en particulier un extrait de la version 10 qui comprend plus de 3 millions d'articles et 25 millions de citations, extraits le 27 octobre 2017. Cet extrait, fourni sous forme d'export `.csv` comporte environ 64000 articles publiés en 2013 et ne conserve que quelques informations : identifiant, titre, résumé (s'il existe), lieu de publication, année, nombre de citations. Vous pouvez rencontrer des problèmes avec le caractère unicode nul (`'\$x00'`). Pour résoudre ce problème, vous pouvez le remplacer par un caractère vide grâce à la commande `replace`. Le passage à la ligne (caractère `'\n'` peut aussi poser problème, donc éliminez-le systématiquement si vous le rencontrez).