Projet de Fouille de Données

par Jordan Baudin, Corentin LeGuen et Geoffrey Spaur

19 février 2018

Contents

1 Présentation

Ce projet a pour but d'extraire des connaissances à partir d'un annuaire de connaissances sous forme ontologique, qui pourra être utilisé ultérieurement afin d'annoter des connaissances, afin de contribuer à l'indexation et la recherche de documents.

2 Extraction des données

Nous avions à traiter des données de type PubMedArticle, extraits depuis le site : PubMed.gov. De ces articles, nous avons extraits les titres et les abstracts, que nous avons associé par couple, grâce à du code Java produit par nous-même. On obtient alors un fichier txt dans lesquel chaque paire de lignes correspondent à un article, avec son titre préfixé par "T." et son abstract préfixé par "A.".

3 Traitement des données

De ces fichiers d'où sont extraits les titres et abstracts, nous avons effectué une recherche de token unique, grâce à TreeTagger, nous donnant un nouveau fichier qui taggera chacun des termes selon le système de tags de TreeTagger, qui est configuré pour traiter des termes en anglais et qui les simplifiera en retirant la conjugaison appliquée aux termes, ainsi que le pluriel. Ainsi, des termes seront annotés comme inconnus par TreeTagger, parfois, ils seront simplement copié et encore d'autres fois modifié au niveau orthographique ou de la casse.

4 Conclusion

TODO