

Recherche et extraction d'informations dans les textes

Proposition de mini projet

Nom: Issa Hammoud

Nom: Salim Tabarani

Objectif: classification des oeuvres par genre.

Corpus: théâtre classique.

Métrique: accuracy

Procédure:

Ce corpus est formé des fichiers xml, chaque fichier contient des informations concernant l'oeuvre. Le but est de classer ces oeuvres par genre. Pour cela, il suffit d'extraire la balise <genre> qui contient l'annotation, puis on fait du prétraitement sur le contenu de la balise <text> pour extraire les textes.

Enfin, on obtient des textes avec leurs annotations correspondantes qui seront prêts pour être utilisés dans les modèles des réseaux neuronaux.