M1 Informatique

Projet Recherche Documentaire

Cyril Lepinette
Romain Grelier

1 Introduction

Dans le cadre de l'UE "Recherche Documentaire" nous avons implémenté un programme de recherche de document en évaluant la pertinence de ceux-ci en fonction d'une requête envoyée par l'utilisateur. Dans ce document nous décrivons le pipeline utilisé pour l'indexation des documents et le traitement des reqêtes.

2 Pipeline

2.1 Lecture des fichiers

Les textes constituant le corpus sont dans plusieurs fichiers qui contiennent les textes des documents aussi plusieurs autres données, comme le titre, des notes ... Ces données doivent donc être récupérée à l'aide d'un lecteur de fichier.

Les fichiers utilisent un language de balise qui vont nous servir à la lecture de celui-ci, les balises qui vont nous intérésser sont les balises "DOCNO" et "TEXT". Le lecteur est capable de gérer la plupart de ces balises, lors de la lecture, les données sont regroupées par document et stocké dans un objet python "Document".

2.2 Suppression des caractères spéciaux

Une fois les fichiers lu et stockés, nous observons des caractères qui ne seront pas forcement utile comme la ponctuation, dans notre implémentation nous avons choisit de les supprimer. La suppression de ces caractères se fait à l'aide d'une expression régulière pour la ponctuation et les chiffres.

2.3 Stemming

Le Stemming permet de trouver la racine des mots pour des recherches plus éfficace, nous avons utiliser une implémentation existance appelée Porter Stemming (voir source).

2.4 Index inversé

L'index inversé est l'ensemble des mots présents dans le corpus qui réfère aux textes qui les contiennent après être passé par les étapes de suppresion des caractères spéciaux et de stemming. L'index des implémenté sous forme d'un dictionnaire python dont la clé est le mot et la valeur est une list contenant en première position le nombre d'occurance du mot dans tout le corpus puis des tuples donnant l'index du document et le nombre d'occurance dans ceui-ci.

2.5 Pertinence des documents

2.6 Interface graphique

Afin de fournir une interface pouvant fonctionner sur un maximum de plateforme ainsi permettant d'être développée rapidement, c'est une interface web qui a été choisit. C'est à l'aide du framework web Flask que nous avons implémenté l'interface, elle peut être lancée facilement, et se trouve dans le navigateur web de l'utilisateur.

Un champ de recherche permet de taper une requête et en réponse, le programme renvoit la liste des documents les plus pertinent trouvés.

Références

[1] The Porter stemming algorithm
http://snowball.tartarus.org/algorithms/porter/stemmer.html