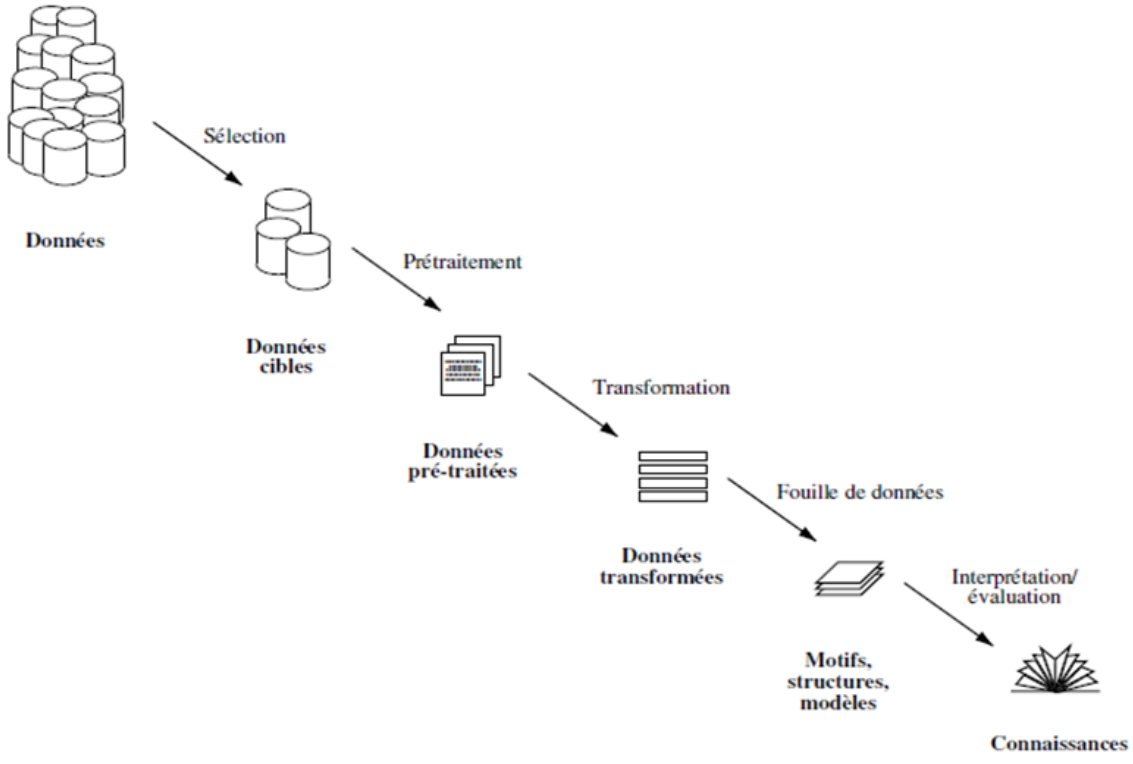
M2 GIL - Fouille de Textes
Indexation de documents XML issus de MEDLINE

26 octobre 2017
Lina Soualmia

1 Préambule

L'objectif des séances de TP est de développer une application efficace reposant sur une chaîne de traitements permettant de traiter des corpus de textes en entrée et d'en déduire des connaissances formalisées.

Les étapes intermédiaires de cette chaîne de traitement reposeront sur des outils réutilisables que vous développerez et qui permettront d'indexer efficacement les textes.



Contexte Général

On s'intéresse ici à l'extraction de connaissances à partir de données en vue de les formaliser sous format ontologique. L'ontologie peut être réutilisée comme base de connaissances et servir de support dans les processus d'annotation et de recherche d'information. Le format des données concernées est non structuré (i.e. texte) et le type d'ontologie est une ontologie de domaine formelle (i.e représentée à l'aide d'un langage offrant des mécanismes de raisonnement).

Démarche proposée

L'application de la méthode retenue se fera sur un corpus de textes en entrée (au format XML, résumés d'articles scientifiques en anglais) à partir duquel on souhaite obtenir :

1. une ontologie représentative au format standard OWL (hiérarchies de concepts, relations, hiérarchies de relations, domaines et co-domaines ...etc)(La consistance de l'ontologie devra notamment être être vérifiée)
2. les documents du corpus en entrée seront dans une seconde étape représentés au format RDF et annotés avec l'ontologie obtenue en 1.

TP 1 : Constitution des corpus de textes

L'objectif de ce premier TP est de préparer les données pour les traitements futurs et de réaliser les pré-traitements.
Les fichiers XML devant être traités sont des extractions au format XML MEDLINE, de la banque de données des articles scientifiques en santé de la NLM (National Library of Medecine : http://www.ncbi.nlm.nih.gov/pubmed).
Les requêtes devant être lancées par chaque groupe sont fournies en annexe (chaque binôme choisira 1 ensemble de mots clés, sur les 5 proposés, tous les mots clés devront être traités, permettant ainsi de constituer 5 corpus de textes).
Les fichiers XML résultant des requêtes sont de taille conséquente. Il est recommandé de prévoir une limite supérieure du nombre d'articles traités (par exemple au plus 100 000 articles pour les premiers tests).

Pour chaque corpus de documents (c-à-d. pour chaque fichier XML correspondant à chaque requête), écrire un programme qui permet d'en extraire :

1. Chaque titre et résumé stockés dans des documents au format texte. Chaque fichier aura pour nom le PMID de l'article.
2. Chaque titre, résumé et MeSH terms stockés dans des documents au format texte. Chaque fichier aura pour nom PMID_indexed.

Format fichier 1 : nom = PMID ;

- ✓ ligne 1 : T. suivi du titre de l'article
- ✓ ligne 2 : A. suivi du texte de l'abstract

Format fichier 2 : nom = PMID_indexed ;

- ✓ ligne 1 : “T.” suivi du titre de l’article
- ✓ ligne 2 : “A.” suivi du texte de l’abstract
- ✓ ligne xx : “I.” suivi par tous les MeSH terms séparés par des “|”.

Annexes :

Format des requêtes PubMed :

http ://www.ncbi.nlm.nih.gov/pubmed ?term=[terme]
Liste des éléments XML de MEDLINE :http://www.nlm.nih.gov/bsd/licensee/elements_alphabeti
Description des XML Element de MEDLINE http://www.nlm.nih.gov/bsd/licensee/elements_desc

Liste 1 :

myocardial infarction ; post-operative complications ; breast neoplasm ; hypertension ; lung neoplasm ;

Liste 2 :

coronary disease ; HIV infection ; pain ; obesity ; asthma ;

Liste 3 :

rare diseases ; liver neoplasm ; thrombosis ; nosocomial infections ; hypertension ;

Liste 4 :

arthritis rhumatoïd ; inflammation ; tuberculosis pulmonary ; alcoholism ; nosocomial infections ;

Liste 5 :

nosocomial infections ; hypertension ; rare diseases ; asthma ; diabetes mellitus ;

Liste 6 :

myocardial infarction ; hypertension ; colon neoplasm ; nosocomial infections ; back ;

Liste 7 :

thrombosis ; Alzheimer disease ; individualized medicine ; colon neoplasms ; tuberculosis pulmonary ;

Liste 8 :

ebolavirus ; nosocomial infections ; rare diseases ; post-operative complications ; individualized medicine ;

Liste 9 : (monôme)

post-operative complications ; rare diseases.

Stop words

http ://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/