SORBONNE UNIVERSITES

TP 1

RAPPORT INDEXATION DU CORPUS

UV : **LO17**
Branche : **Génie Informatique**
Auteurs : **Alexandre BALLET - WANG Jiaxin**
Date : **22 octobre 2017**

Table des matières

1. Introduction

Ce rapport déroulera l'ensemble des travaux réalisés dans le cadre du projet d'indexation et recherche d'information. Il comportera les deux parties. Le corpus que nous avons utilisé, est un ensemble de fichiers contenant des articles archivés.

Pour la première grande étape du projet, nous devions préparer ce corpus d'articles en vue de son indexation. Nous devions construire un fichier XML contenant l'ensemble des informations relatives aux documents (identifiant, numéro de bulletin, date, rubrique, titre, texte, images, contacts), qui était facilement indexable. Et puis, dans la deuxième étape, nous devions indexer ce fichier XML que nous avions réalisé dans la première étape, ceci résultant par la création de fichiers inverses. Dans cette partie, nous avons concerné la réalisation des fichiers inverses à partir des mots des titres et des résumés. D'une part, nous souhaitons pouvoir s'affranchir des mots qui ne sont pas porteurs de sens et de ceux qui n'apportent pas d'information. D'autre part nous souhaitons en particulier représenter par un même mot de référence (un lemme) toutes les dérivations d'un même mot.

Nous allons vous montrer et expliquer plus précisement toutes les informations mentionnées ci-dessus.

2. Préparation du Corpus

L'objectif de cette partie est d'extraire les différents types d'informations d'un ensemble de fichiers `.htm` et de créer un nouveau fichier `corpus.xml` qui regroupe toutes ces informations extraites. Nous avons écrit un script unique `file_name.pl` avec le langage Perl, afin de réaliser l'extraction et l'exportation des informations. Nous allons présenter la manière dont nous avons généré le fichier XML final.

2.1 Observation et Analyse des fichiers archivés

Tout d'abord, nous avons observé le modèle général de plusieurs fichiers `.htm`. Nous avons trouvé que la forme d'ensemble est la même.

Ensuite, nous avons vu les codes HTML de plusieus fichiers. Nous avons identifié certaines informations demandées qui étaient uniques et faciles à classifier. Par exemple, le nom du fichier, le numéro du bulletin, la date du bulletin, la rubrique, le titre de l'article et toute l'information de contact. En revanche, nous avons également trouvé certaines informations qui étaient difficiles à classifier. Par exemple, le texte de l'article. Les textes avaient différents nombres de paragraphe. Les images étaient aussi un problème. Il y avait beaucoup de cas pour classifier les images. Certains sites web ne contenaient pas les images. Certaines images étaient intégrées au texte. Il faut également conserver la legende et l'URL correspondantes à l'image.

Nous devons faire attention à tous les principaux points mentionnés au-dessus, lorsque nous dévéloppons la méthode et l'algorithme.

2.2 Méthode de l'extraction des informations

Dans un premier temps, nous avons choisi le mode de lecture de fichiers qui était en ligne par ligne. Et puis, nous avons utilisé le procédé d'identification unique de `la balise`. Les balises correspondantes aux informations demandées sont uniques et sont déclarées de même façon dans les codes HTML pour tous les fichiers. Par conséquent, nous pouvons confirmer l'unicité et l'exhaustivité de chaque information attendue. Nous détaillons au-dessous les unique balises observées pour identifier l'information demandée.

» `le nom du fichier`

En raison du même format `.htm` de tous les fichiers, nous pouvons ouvrir le répertoire « BULLETINS » pour lire la liste des fichiers. Nous avons coupé les informations du format `.htm` et les avons stockées dans une variable tableau `@docs`. Autrement dit, les noms de tous les fichiers sont stockés dans`@docs`. Ensuite, nous avons utilisé la boucle `foreach` pour lire chaque fichier. Par conséquent, nous pouvons directement afficher le nom du fichier par cette boucle.

» `Le numéro / La date du bulletin / Le titre de l'article`

Nous avons trouvé que les informations concernant `<numero>, <date>, <titre>` sont toutes stockées dans la balise `<title></title>`, de plus uniques. Donc, pour extraire le numéro de bulletin, la date de bulletin et le titre de l'article, l'expression rationnelle utilisée est :

```
#On a selectionné trois groupes des informations dans la ligne <title> pour afficher <numero>,<date>,<titre>.
$line =~ qr/\A<title>(?'date'\d{4}\/\d{2}\/\d{2}) &gt; \w{2} \w+ (?'numero'\d+) &gt; (?'title'.*)<\/title>$/;
```

Nous avons seulement extrait le contenu de ces trois parties et ignoré d'autres informations dans cette balise. En fonction de cette façon, nous pouvons directement extraire ces trois informations demandées en

une fois.

» `Rubrique / Contact`

Nous avons trouvé d'abord le contenu de la rubrique afin de confirmer la position de cet attribut. Et puis, nous avons confirmé l'unicité de la position. La balise observée et l'expression régulière utilisée est :

```
elsif ($line =~ /<p class="style96"><span class="style42">/) {
  $line =~ qr/<p class="style96"><span class="style42">(?'rubrique'.*)<br>/;
```

Nous pouvons également extraire l'information `<contact>` par cette même façon. La balise observée et l'expression rationnelle utilisée est :

```
($line =~ qr/<p class="style44"><span class="style85">(?'name'.*)<a href="mailto:.*">(?'mail'.*)<\/a><\/span>.*$/)
```

» `Texte`

Comme nous vous l'avons présenté, en raison du mode de lecture de fichier par ligne, nous avons résumé deux conditions pour la description du texte. Au début, le style HTML pour presenter le texte est le même, c'est `<p class=« style96 »><span class=« style95 »>`. Ensuite, le premier cas est la balise fermante qui s'apparente à </p></td>. Ça signifie qu'il n'y a qu'un paragraphe et toute la ligne lue contient le texte de l'article. Dans le deuxième cas, la ligne termine par
. Ça signifie qu'il y a plusieurs paragraphes au texte. Ces deux conditions sont montrés au-dessous :

```
$line =~ qr/<p class="style96"><span class="style95">(?'texte'.*)<\/span><\/p><\/td>/;
$line =~ qr/<p class="style96"><span class="style95">(?'texte'.*)<br \/>/;
```

» `UrlImage / LegendeImage`

De la même manière ci-dessus, nous devons trouver la déclaration unique de la balise pour confirmer que les informations demandées correspondent à <image>, <urlImage> et <legendeImage>. Ce que nous devons être conscients, c'est que la conformité de l'URL ou la légende avec l'image. Il peut y avoir plusieurs photos, et eds photos sans légende. La balise observée et l'expression régulière utilisée pour l'URL est :

```
$line =~ qr/<\/span><div.*><img src="(?'url'.+)" border.*<br \/>/;
```

La regex utilisée dans nos scripts d'extraction pour la légende :

```
($line =~ qr/<span class="style21"><strong>(?'legende'.*)<\/strong>.*$/)
($line =~ qr/<span class="style21"><strong>(?'legende'.*)<br \/>$/)
```

2.3 Exactitude d'extraction

Nous avons utilisé la façon suivante pour vérifier l'exactitude de l'extraction. Nous avons ajouté une nouvelle variable dans chaque partie demandée pour compter le nombre de traitement. Et puis, nous avons également noté le nombre de fichiers traités. Si ces deux résultats obtenus sont égaux, cela montre que nous avons bien extrait l'information dans tous les fichiers.

Par ailleurs, nous avons aussi testé le nombre d'images dans le terminal.

```
[alex on Alex-MBP in ~/Documents/Cours UTC/GI05/LO17/LO17_git/TD1] cat corpus.xml | grep "<image>" -c
154
[alex on Alex-MBP in ~/Documents/Cours UTC/GI05/LO17/LO17_git/TD1] cd BULLETINS/
[alex on Alex-MBP in ~/Documents/Cours UTC/GI05/LO17/LO17_git/TD1/BULLETINS] cat *.htm | grep '</span><div.*><img src=' *.htm | wc -l
     155
```

Nous constatons que notre script nous récupère une image en moins par rapport au nombre d'images du corpus.

3. Indexation du Corpus

A partir du fichier `corpus.xml` obtenu en partie précédente, nous allons réaliser une indexation sur ce fichier. L'objectif final de l'indexation sera un ensemble de fichiers inverses, chaque fichier correspondant à une balise ou un ensemble de balises (date ou numéro du bulletin, rubrique, titre, titre+résumé, légende des images).

3.1 Création d'une stop-list

La Stop-List, c'est une liste de mots qui ne fait pas partis de l'indexation et qui stocke les mots `non significatifs`. Nous allons supprimer ces mots dans notre fichier `corpus.xml` pour générer un nouveau fichier XML. Afin d'obtenir la stop-list, nous allons calculer le coefficient $tf \times idf$.

3.1.1 Choix de l'unité documentaire

Le calcul de ce coefficient s'appuie sur la fréquence d'un mot dans un document et sur le nombre de documents qui contiennent un mot donné. L'unité documentaire doit donc être clairement définie. Nous avons choisi `<un document = un article>` . Ça signifie que nous allons calculer la fréquence des mots qui est basée sur les titres et textes de chaque article.

Supposons que nous ayons choisi `<un document = un bulletin>`, alors, la fréquence d'un mot qui apparaît dans un document va augmenter. Ce qui signifie que $tf_{i,j}$ sera plus grand, puisque nous prenons en compte un ensemble d'articles dans un document. Par ailleurs, le nombre de document traité(N) va réduire , en comparant avec d'autre choix. Puisque c'est possible qu'il y a plusieurs articles dans un bulletin. Et puis, si N réduit, la fréquence d'un mot qui apparaît dans les différents documents va relativement augmenter. Par conséquent, quand nous calculons idf_i, en fonction de la formule $idf_i = log_{10} \frac{\mathbb{N}}{df_i}$, le résultat de idf_i va réduire. Ce qui signifie que l'importance du mot sera réduite. Par contre, en raison de la réduction du nombre de document traité, les mots obtenus seront relativement réduits, puisque certains mots répétitifs seront fusionnés. Alors, l'algorithme fonctionnera plus vite. Si dans d'autre cas, le nombre de document / bulletin soit trop grand, nous pouvons choisir cette façon de traitement. Mais pour notre projet, le choix de `<un document = un article>` est plus adaptable.

3.1.2 Calcul du $tf_{i,j}$

D'abord, nous avons lu les balises<fichier> du corpus.xml, et puis nous avons obtenu la liste des mots contenus dans les balises textes et titres. Ensuite, nous avons utilisé la boucle `foreach $mot (@mots)` pour compter la frequence de chaque mot qui apparaît dans chaque texte et titre $tf_{i,j}$ et compter le nombre de document dans laquelle ce mot apparaît df_i.

```
foreach $mot (keys %ind_mots) {
    print OUTPUT_FILE $fichier . " " . $mot . " " . $ind_mots{"$mot"} . "\n";
    $files_word_count[$i] = $fichier . " " . $mot . " " . $ind_mots{"$mot"};
                            #le nom du fichier ,ce mot,le nombre de fois dans ce fichier(tf)
    $i = $i + 1;
}
```

3.1.3 Calcul du idf_i

Grâce au calcul de $(tf_{i,j})$ précédent, nous avons calculé le (df_i). Ensuite, nous pouvons appliquer la formule pour calculer le idf_i :

$$idf_i = log_{10}\frac{\mathbb{N}}{df_i}$$

Les résultats obtenus avec les mots correspondants sont affichés dans `idf.txt`.

```
foreach $mot (keys %df) {
            $idf = log($N / $df{"$mot"})/log(10);
            print OUTPUT_FILE2 $mot . " " . $idf . "\n";
}
```

3.1.4 Calcul du $tf_{i,j} \times idf_i$

À partir du script précédent, nous avons pris le $tf_{i,j}$ et le idf_i correspondant au mot. Afin de calculer le coefficient $tf_{i,j} \times idf_i$, nous avons repris la troisième colonne $tf_{i,j}$ de chaque ligne dans le fichier `tfij.txt` et utilisons le résultat du idf_i du chaque mot. Et puis, nous avons effectué la multiplication et affiché le résultat final dans le fichier `tf_idf.txt`.

```
foreach $file_word_count (@files_word_count) {
            @sp = split(/ /, $file_word_count);
            $res = $sp[2] * log($N / $df{"$sp[1]"})/log(10); #( tf*idf )
            print OUTPUT_FILE3 $sp[0] . " " . $sp[1] . " " . $res . "\n";
}
```

3.1.5 Détermination de la stoplist

L'objectif du calcul du coefficient $tf_{i,j} \times idf_i$, c'est pour obtenir la stop-list qui stocke les mots non significatifs.

D'abord, nous allons utiliser le fichier `tf_idf.txt` qui représentent le niveau de signification du chaque mot. Mais en raison de l'extraction de tous les mots dans l'ensemble du document, nous avons trouvé les mots répétitifs correspondants aux coefficient $tf_{i,j} \times idf_i$ différents . Par conséquent, nous avons calculé la somme de ces coefficients différents$tf_{i,j} \times idf_i$ du mot correspondant et obtenu la moyenne du coefficient $tf_{i,j} \times idf_i$ du mot correspondant. Et puis, nous avons affiché le résultat dans le fichier `result_moy_tf_idf.txt`. Le but de cette étape est pour obtenir plus précisément l'importance du mot dans l'ensemble du document. Nous pouvons utiliser d'autre façon comme le calcul de la médian du coefficient $tf_{i,j} \times idf_i$ pour générer le mot correspondant au du coefficient $tf_{i,j} \times idf_i$ unique.

```
foreach my $mot (sort { ($coeffs{$a} / $occurence_mot{$a}) <=> ($coeffs{$b} / $occurence_mot{$b}) } keys %coeffs) {
                    # la sortie en fonction de l'ordre croissant.
        $moy = $coeffs{"$mot"} / $occurence_mot{"$mot"};
        # moyenne= la somme du coefficient tf_idf / la frequence du même mot
        @moy_occurence_mot[$i] = $moy;
        $i = $i + 1;
        print OUTPUT_FILE $mot . " " . $moy . "\n";
}
```

Alors, à partir du fichier sorti `result_moy_tf_idf.txt` ci-dessus, nous avons obtenu d'abord la liste croissante du coefficient $tf_{i,j} \times idf_i$ avec les mots. Ensuite, nous avons analysé cette liste et trouvé une seuil minimum (voir maximum) de deux façons différentes en fonction du coefficient tf×idf. Dans un premier temps, et de manière subjective, nous avons choisi la seuil minimum en 0.71. Et puis, nous avons pris les mots dont les coefficients $tf_{i,j} \times idf_i$ sont inférieurs que 0.71, afin de créer la stop-liste avec ces mots qui sont supprimés du

corpus. Dans un second temps, nous avons utilisé une librairie perl de statistiques afin de récupérer le deuxième quantile, ce qui nous donne un seuil minimum de 2.21. Ce seuil est largement supérieur au seuil précédent, ce qui nous donne une stop-list bien plus importante.

3.1.6 Sortie du nouveau fichier XML filtré

À l'aide du script `newcreeFiltre.pl`, qui permet d'éliminer un mot ou de remplacer un mot par un autre mot, nous avons créé un fichier `filtre_corpus.pl`. Dans ce fichier-là, nous avons lu le fichier `stop_list.txt` et stocké ces mots non significatifs dans une variable tableau. En même temps, nous avons lu le fichier `corpus.xml` pour extraire tous les mots dans le texte et le titre. Et puis, nous avons filtré les mots non significatifs dans le fichier `corpus.xml`, en fonction du script `newcreeFiltre.pl`. Enfin, nous avons sorti un nouveau corpus sans ces mots, donc avec seulement des informations importantes dans un nouveau fichier `corpus_filtre.xml`. Donc, nous avons obtenu un nouveau fichier XML filtré :


```
<?xml version="1.0" encoding="utf-8"?>
<corpus>
<bulletin>
<fichier>67068.htm</fichier>
<numero>258</numero>
<date>21/06/2011</date>
<titre>Physique bel chercheur innove</titre>
<rubrique>Focus</rubrique>
<contact>Institut Langevin 'Ondes et Images' - Mathias Fink - email : http://www.institut-langevin.espci.fr</contact>
<texte>avril Médaille Innovation Valérie Pécresse ministre Enseignement Supérieur origine création souhaite ingénieurs travaillant établissements publics universités écoles développent innovations distinction
attribuée réputés économiste roboticien physicien Directeur créé janvier Ecole Physique Chimie Industrielles Ville remarquable innovent abouti création quatre start personnes variés médecine objets tactiles
transformer idées physique fondamentale produits innovants emblée parvenir fallait trouver terrain propice rencontre prix Physique découvrir coeur Montagne Geneviève plein Alors physicien physique solide
souhaite innovation concept belle physique connaisseurs intéresse quatre aéronautique Dassault Snecma prêts financer époque milieu enclin travailler industrie industrielle retrouve isolé lui propose
rencontrer Pierre directeur Jacques directeur lui Je ai rencontré était Ils vous Alors venez ajoute ma carrière Vous chercheur vous dès physicien crée Acoustique Innover particulièrement vu vous donne feu
vert moment idée proposée originale Autre particularité vraiment Ne cherchez vous trouverez département physique département chimie département biologie Ici vous dès départ vous chercheur vous pointe humour
lui fameux miroirs enregistrer champ incident puis réémettre chronologie inversée collaborateurs rendus revivre passée focaliser source efficace Ils travailler abord puis micro aurait pu limiter publier
publications réputées Physical Review compris vite delà belle physique avait potentielles brûler tumeur détruire électronique voir mur envoyer informations précis pièce nous avons certain fini déboucher
création start concept découvert était créer images parvenus fabriquer corps humain auxquels obtenir images corps humain avons pu découvrir corps humain était parcouru jamais personne avait observé
enthousiasme clé réussite aurait effets âge pratiquent delà âgé laisser croire enthousiasme reste rapprochement laboratoire laboratoire Physique créant directeur Arnaud directeurs Mixte UPMC Diderot centaine
personnes coeur métier physique menées concepts fondamentaux imagerie multi optique photo originales filtre inverse contrôle front passant création entreprises mettant oeuvre technologies domotique etc
aventure poursuit clé réussite impérativement préserver conclut</texte>
<images>
</images>
</bulletin>
<bulletin>
<fichier>67071.htm</fichier>
<numero>258</numero>
<date>21/06/2011</date>
<titre>environnement publie rapport</titre>
<rubrique>Actualité Innovation</rubrique>
<texte>nationale Environnement février initiative Valérie Pécresse ministre Enseignement supérieur environnement système complexe recouvre champ étendu agroécologie aménagement territoires passant sciences
climat Biologie écologie physico chimie appliquées géographie sciences économiques sociales Toutes disciplines sciences environnementales introduction Roger Genet Président directeur général nécessité
coordonner caractérise principaux opérateurs environnement nationale Environnement membres fondateurs Alimentation Eau Climat constituent quatre enjeux véritables défis relever décennies venir puisqu agira
alimentation parvenir nourrir milliards individus horizon Concernant eau accès plan mondial qualité nécessairement quantité Autre défi taille humanité face changements climatiques biodiversité populations
planète devront respecter impératif qualité environnementale territoires organisé groupes thématiques réunissant experts Agroécologie Alimentation Biodiversité Biologie Climat Evolution Eau Chimie durable
marines Risques environnementaux naturels naturelles environnementale nouvel élan initié Président maintenant nos ingénieurs techniciens identifier concevoir valider déployer solutions permettront relever
défis transition écologique croissante conclut introduction rapport publié nationale Environnement CEA Cirad INRA IRD Conférence Présidents Universités National</texte>
<images>
</images>
</bulletin>
<bulletin>
<fichier>67383.htm</fichier>
<numero>259</numero>
<date>22/07/2011</date>
<titre>médaille innovation récompense roboticien cesse innover</titre>
<rubrique>Focus</rubrique>
<contact>Lirmm - François Pierrot - email : francois.pierrot@lirmm.fr</contact>
<texte>avril décerné Médaille Innovation Parmi directeur Informatique Robotique Microélectronique Montpellier mondialement roboticien participé conception parallèle rapide brevet entreprise leader systèmes
Autant contrairement gens innovation chercheur lui sait innovation puisqu vit pratique quotidien parle innover évidemment idées indispensable souvent début ma carrière je ai cessé emblée carrière
professionnelle recruté me soutient vingt liberté mon lui innovation contraire nécessité disposer recruté Montpellier ville dispose potentiel exceptionnel laboratoire réputé mondialement laboratoire
exceptionnel renferme talents lesquels je travaille quotidien nous avons pu transformer nos idées produits États Unis Japon créer entreprise Robotics enthousiasme innovation problème théorique général sert
fil conducteur thèse problème industriel rencontre émerger prototypes théorique proche besoin industriel prototypes nous allons pouvoir démonstrations auprès potentiellement intéressés chercher aller loin
aboutir solutions puissent sens éventuellement économique marché appliquant conçu parallèle rapide marché mondial clé innovation quatre chaînes parallèle agissent seul organe déplacer vite boîte déplacer vite
composants cherche assembler poursuit privée nous transformer nos connaissances avantages entreprises chercheur quatre créé filiale Montpellier évoque Celle travailler établissement vivre collaborer collègues
créatifs trouver puis idées originales Telle recette innover selon chercheur Montpellier</texte>
<images>
</images>
```


3.2 Création des lemmes

Dans cette partie, nous devons trouver différentes dérivations d'un même mot pour l'indexation dans le fichier XML filtré et les représenter par un mot de référence (un lemme).

Nous avons utilisé d'abord `segmente.pl` afin d'obtenir la liste de tous les mots du corpus. Ensuite, à l'aide du script `successeurs.pl`, nous pouvons calculer les successeurs. Et puis, à l'aide du script `filtronc.pl`, nous pouvons générer les lemmes en fonction des différents successeurs obtenus précédemment.

Comme nous avons fait avant, nous avons créé un fichier pour faire un nouveau filtrage en base des lemmes des mots. À l'aide du fichier `newcreeFiltre.pl` et du fichier `corpus_filtre.xml`, nous pouvons générer un nouveau fichier `corpus_filtre_lemme.xml` dont les mots sont remplacés par leur lemme :

```
<?xml version="1.0" encoding="utf-8"?>
<corpus>
<bulletin>
<fichier>67068.htm</fichier>
<numero>258</numero>
<date>21/06/2011</date>
<titre>physi bel cherch innov</titre>
<rubrique>Focus</rubrique>
<contact>Institut Langevin 'Ondes et Images' – Mathias Fink – email : http://www.institut-langevin.espci.fr</contact>
<texte>avril médaille innov valérie pécresse ministre enseignement supérieur origin création souhait ingénieurs travail établissements publi universités écoles développent innov distinction attribuée réputés économiste robot physi direct créé janvier ecole physi chimi industrielle vill remarquable innov abouti création quatr start personn variés médecine obje tactiles trans idées physi fondamenta produ innovant emblée parven fall trouver terrain propice rencontr prix physi découvrir coeur montagne geneviève plein alors physi physi solid souhait innov concept belle physi connaiss intéresse quatr aéronautique dassault snecma prêts financ époque milieu enclin travail industri industrielle retrouv isolé lui propose rencontr pierre direct jacques direct lui je ai rencontré était ils vous alors venez ajout ma carrière vous cherch vous dès physi crée acoustique innov particulièrement vu vous donn feu vert moment idée proposée origin autre particularité vrai ne cherch vous trouverez département physi département chimi département biolog ici vous dès départ vous cherch vous point humour lui fameu miroirs enregistre champ inciden puis réémettre chro inversée collaborat rendus revivre passée focaliser sourc efficac ils travail abord puis micro aurait pu limit publi publi réputées physi review compris vite delà belle physi avait potentiel brûler tumeur détruire électronique voir mur envoy informati précis pièce nous avons certain fini débouncher création start concept découvert était créer imag parven fabriqu corps humain auxquel obten imag corps humain avons pu découvrir corps humain était parcour jamais personn avait observé enthousias clé réussite aurait effets âge pratique delà âgé laisse croire enthousias reste rapproche laboratoire laboratoire physi créant direct arnaud direct mixte upmc diderot cent personn coeur métier physi menées concept fondamenta imag multi optique photo origin filtr invers contrôle front passant création entrepris mettant oeuvre technolog domotique etc aventure poursui clé réussite impérativement préserver conclu</texte>
<images>
</images>
</bulletin>
<bulletin>
<fichier>67071.htm</fichier>
<numero>258</numero>
<date>21/06/2011</date>
<titre>environnement publi rapport</titre>
<rubrique>Actualité Innovation</rubrique>
<texte>nationa environnement février initi valérie pécresse ministre enseignement supérieur environnement système complexe recouvre champ étendu agroécologie aménagement territoire passant science climat biolog écologie physi chimi appliquées géographie science économiques soci toutes discipline science environnement introdu roger genet président direct général nécessité coordonn caractérise princip opérateurs environnement nationa environnement membre fondateur aliment eau climat constitu quatr enjeu véritables défis relev décennies venir puis agira aliment parven nourri milliard individu horizon concern eau accès plan mondia qualité nécessairement quantité autre défi taille humanité face changement climat biodiversité population planète devront respect impératif qualité environnement territoire organisé groupe thématiques réunissant exper agroécologie aliment biodiversité biolog climat evolution eau chimi durable marines risque environnement nature naturelle environnement nouve élan initié président mainten nos ingénieurs techn identifi concevoir valid déployer solution permett relev défis trans écologique croissan conclu introdu rapport publié nationa environnement cea cirad inra ird conférence présidents universités nationa</texte>
<images>
</images>
</bulletin>
<bulletin>
<fichier>67383.htm</fichier>
<numero>259</numero>
<date>22/07/2011</date>
<titre>médaille innov récompense robot cesse innov</titre>
<rubrique>Focus</rubrique>
<contact>Lirmm – François Pierrot – email : francois.pierrot@lirmm.fr</contact>
<texte>avril décerné médaille innov parmi direct informati robot microélectronique montpellier mondia robot participé concept parallèle rapide brevet entrepris leader systèmes autant contraire gens innov cherch lui sait innov puis vit pratique quotidien parle innov évidemment idées indispensable souvent début ma carrière je ai cessé emblée carrière professionnel recruté me soutien vingt liberté mon lui innov contraire nécessité disposer recruté montpellier vill dispose potentiel exceptionnel laboratoire réputé mondia laboratoire exceptionnel renferm talents lesquel je travail quotidien nous avons pu trans nos idées produ etat unis japon créer entrepris robot enthousias innov problème théorique général sert fil conduct thèse problème industri rencontr émerger prototyp théorique proche besoin industri prototyp nous allon pouvoir démonstrations auprès potentiel intéressés cherch aller loin abouti solution puis sens éventuellement économique marché appliqu conçu parallèle rapide marché mondia clé innov quatr chaînes parallèle agiss seul organ déplacer vite boîte déplacer vite composant cherch assembl poursui privée nous trans nos connaiss avantage entrepris cherch quatr créé filiale montpellier évoque celle travail établissement vivre collabor collègues créatifs trouver puis idées origin telle recette innov selon cherch montpellier</texte>
<images>
</images>
</bulletin>
```

3.3 Création des fichiers inverses

Nous devons créer les fichiers inverses à partir de notre nouveau fichier XML filtré `corpus_filtre_lemme.xml`. La commande suivante (composée de pipes) permet de créer un fichier texte qui contient dans une première colonnes les rubriques, et ensuite les noms des fichiers et les numéros d'articles associés à cette rubrique.

```
[alex on balletal-hz-4 in ~/Documents/Cours UTC/GI05/LO17/LO17_git/TD2] cat corpus_filtre2.xml | perl index.pl rubrique
création d'index pour la balise <rubrique>...
A lire 68283.htm 263 69821.htm 269 70429.htm 271 70753.htm 272 71366.htm 274 71621.htm 275 72401.htm 278 72940.htm 280 74457.htm 286 74752.htm 287
76213.htm 291
Evénement 67392.htm 259 67561.htm 260 67803.htm 261 67804.htm 261 67945.htm 262 67946.htm 262 68889.htm 266 69186.htm 267 69542.htm 268 69543.htm
268 69820.htm 269 70168.htm 270 70169.htm 270 70170.htm 270 70428.htm 271 70749.htm 272 70751.htm 272 70752.htm 272 71363.htm 274 71620.htm
275 71845.htm 276 72120.htm 277 72121.htm 277 72122.htm 277 72399.htm 278 72400.htm 278 72937.htm 280 72938.htm 280 72939.htm 280 73190.htm 28
1 73438.htm 282 73689.htm 283 73690.htm 283 73691.htm 283 73881.htm 284 73882.htm 284 73883.htm 284 73884.htm 284 74174.htm 285 74175.htm 285
74176.htm 285 74456.htm 286 75071.htm 288 75464.htm 289 75465.htm 289 75466.htm 289 75795.htm 290 75796.htm 290 75797.htm 290 76212.htm 291 7651
4.htm 292 76515.htm 292 76516.htm 292
Actualité Innovation 67071.htm 258 67385.htm 259 67386.htm259 67387.htm 259 67556.htm 260 67557.htm 260 67796.htm 261 67797.htm 261 67940.htm 262 67941.htm
262 67942.htm 262 68277.htm 263 68278.htm 263 68384.htm 264 68385.htm 264 68386.htm 264 68639.htm 265 68640.htm 265 68641.htm 265 68642.htm
265 68883.htm 266 68884.htm 266 69180.htm 267 69181.htm 267 69536.htm 268 69537.htm 268 69814.htm 269 69815.htm 269 69816.htm 269 70163.htm 27
0 70164.htm 270 70165.htm 270 70423.htm 271 70424.htm 271 70743.htm 272 70744.htm 272 70917.htm 273 70918.htm 273 70919.htm 273 70920.htm 273
71360.htm 274 71615.htm 275 71616.htm 275 71617.htm 275 71838.htm 276 71839.htm 276 71840.htm 276 71841.htm 276 72632.htm 279 72633.htm 279 7293
4.htm 280 72935.htm 280 73186.htm 281 73187.htm 281 73686.htm 283 73687.htm 283 73880.htm 284 74170.htm 285 74171.htm 285 74451.htm 286 74452
.htm 286 74453.htm 286 74454.htm 286 74747.htm 287 75791.htm 290 76208.htm 291
Au coeur des régions 67798.htm 261 67943.htm 262 68279.htm263 68387.htm 264 68643.htm 265 68885.htm 266 69182.htm 267 69817.htm 269 70166.htm 270 70745.htm
272 70921.htm 273 71361.htm 274 71842.htm 276 72117.htm 277 72118.htm 277 72395.htm 278 73434.htm 282 73435.htm 282 73878.htm 284 74455.htm
286 74748.htm 287 75066.htm 288 75067.htm 288 75068.htm 288 75461.htm 289 75792.htm 290 76511.htm 292 76512.htm 292
Horizon Enseignement 67944.htm 262 69185.htm 267 70426.htm271 70748.htm 272 71362.htm 274 72398.htm 278 72636.htm 279 72637.htm 279 72936.htm 280
Du côté des Pôles 67388.htm 259 67799.htm 261 72396.htm278 72634.htm 279 73879.htm 284 74172.htm 285 74749.htm 287 75069.htm 288 75462.htm 289 75793.htm
290 76209.htm 291
Horizon enseignement 73437.htm 282
Horizon formation 71619.htm 275
En direct des labos 75463.htm 289
En direct des laboratoires 67389.htm 259 67390.htm 259 67391.htm 259 67558.htm 260 67800.htm 261 67801.htm 261 67802.htm 261 68280.htm 263 68388.htm 264
68389.htm 264 68390.htm 264 68391.htm 264 68644.htm 265 68645.htm 265 68646.htm 265 68886.htm 266 68887.htm 266 69183.htm 267 69184.htm 267 6953
9.htm 268 69540.htm 268 69819.htm 269 70167.htm 270 70746.htm 272 70747.htm 272 70922.htm 273 71618.htm 275 71843.htm 276 72119.htm 277 72635
.htm 279 73188.htm 281 73189.htm 281 73436.htm 282 73680.htm 283 74173.htm 285 74750.htm 287 75070.htm 288 75794.htm 290 76210.htm 291
Focus 67068.htm 258 67383.htm 259 67553.htm 260 67554.htm 260 67555.htm 260 67794.htm 261 67795.htm 261 67937.htm 262 67938.htm 262 67939.htm 262
68273.htm 263 68274.htm 263 68275.htm 263 68276.htm 263 68383.htm 264 68638.htm 265 68881.htm 266 68882.htm 266 69177.htm 267 69178.htm 267 6917
9.htm 267 69533.htm 268 69534.htm 268 69535.htm 268 69811.htm 269 69812.htm 269 69813.htm 269 70161.htm 270 70162.htm 270 70420.htm 271 70421
.htm 271 70422.htm 271 70914.htm 273 70915.htm 273 70916.htm 273 71357.htm 274 71358.htm 274 71359.htm 274 71612.htm 275 71614.htm 275 71835.htm
276 71836.htm 276 71837.htm 276 72113.htm 277 72114.htm 277 72115.htm 277 72116.htm 277 72392.htm 278 72393.htm 278 72394.htm 278 72629.htm
279 72630.htm 279 72631.htm 279 72932.htm 280 72933.htm 280 73182.htm 281 73183.htm 281 73184.htm 281 73185.htm 281 73430.htm 282 73431.htm 28
2 73432.htm 282 73683.htm 283 73684.htm 283 73685.htm 283 73875.htm 284 73876.htm 284 73877.htm 284 74167.htm 285 74168.htm 285 74169.htm 285
74449.htm 286 74450.htm 286 74744.htm 287 74745.htm 287 74746.htm 287 75063.htm 288 75064.htm 288 75065.htm 288 75457.htm 289 75458.htm 289 7545
9.htm 289 75788.htm 290 75789.htm 290 75790.htm 290 76206.htm 291 76207.htm 291 76507.htm 292 76508.htm 292
En direct des Laboratoires 72397.htm 278
Horizons Enseignement 68281.htm 263 68392.htm 264 68393.htm264 68888.htm 266 69541.htm 268 70923.htm 273 74751.htm 287 76513.htm 292
Horizons Formation Enseignement 76211.htm 291
Du côté des Pôles 69818.htm 269
Actualité-Innovation 70425.htm 271
Actualités Innovations 69538.htm 268 75460.htm 289 76509.htm292 76510.htm 292
```

4. Conclusion

A travers ces 2 TDs, Nous avons essentiellement maîtrisé la façon d'indexation et recherche d'information. Nous avons résumé les basiques étapes de génération d'index ci-dessous :

» `Génération du XML`

Selon le corpus fourni, nous pouvons d'abord observer et analyser les informations demandées, et puis conformément aux exigences, nous pouvons générer un fichier XML.

» `Génération de la stop-list`

Afin de pouvoir trouver le mot-clé plus précisément et plus rapidement pendant l'indexation, nous devons supprimer les mots qui n'ont pas de sens.

— `Calcul du coefficient` $tf_{i,j} \times idf_i$

Afin de calculer l'importance de chaque mot.

— `Définition du seuil minimum`

Nous pouvons selectionner les mots non significatifs dans pour générer la stop-list.

— `Suppression des mots de stop-list dans le corpus XML.`

» `Génération du fichier inverse`

Nous pouvons savoir les mots que nous voulons chercher avec leurs sources indiquées.