



Université de Technologie de Compiègne

Génie Informatique

Rapport de TD

**LO17**

**Steve LAGACHE & Romain PELLERIN**

Chargé de TD : PIERRE MORIZET-MAHOUDEAUX

Printemps 2016 (P16)

Dernière mise à jour : 14 mars 2016

# Table des matières

<b>1</b>	<b>TD1 : Structuration</b>	<b>3</b>
1.1	Méthodologie employée . . . . .	3
1.1.1	Repérer un élément d'intérêt . . . . .	3
1.1.2	Lecture du contenu d'un fichier . . . . .	3
1.1.3	Lecture du contenu de plusieurs fichiers . . . . .	3
1.2	Commandes Unix . . . . .	4
1.2.1	Génération du XML . . . . .	4
1.2.2	Vérification du nombre d'images . . . . .	4
<b>2</b>	<b>TD2 : Indexation</b>	<b>6</b>
2.1	Obtenir les tf . . . . .	6
2.2	Obtenir les idf . . . . .	6
<b>3</b>	<b>Conclusion</b>	<b>7</b>

# 1 TD1 : Structuration

L'objectif de ce TD était de s'initier au “*parsing*” en Perl. Nous avons un peu plus de 300 pages HTML contenant toute un article issu d'un même site scientifique. Il nous fallait extraire des données “uniques” comme le numéro de l'article ou sa rubrique, et des données présentes une (parfois zéro) ou plusieurs fois, comme les paragraphes du texte principal ou les images présentes.

## 1.1 Méthodologie employée

Pour la totalité des éléments à récupérer, nous avons utilisé des *regex* (sauf pour le nom du fichier que nous récupérerions dans \$ARGV).

Pour produire notre XML, nous avons décidé de tout afficher sur la sortie standard. Il s'agira ensuite de rediriger cette sortie standard dans un fichier grâce à l'opérateur > d'Unix.

### 1.1.1 Repérer un élément d'intérêt

Pour chaque élément unique qu'il nous fallait récupérer, nous avons essayé de trouver la structure de balises HTML voisine la plus proche possible qui soit unique. Pour cela, nous avons beaucoup fait appel à la structure `.*` (n'importe quel caractère zéro, une ou plusieurs fois) dans son mode *lazy*. Par exemple, pour le numéro, la structure HTML autour était suffisant simple pour ne pas utiliser “n'importe quel caractère” :

```
$body=~</span class="style95" style="color:inherit">(\d+)</span
></a>;
```

Tandis que nous en avons eu besoin pour la partie contact :

```
$body =~ /Pour en savoir plus, contacts :.*?<p class="style44"><
span class="style85">(.*?)</span>/s;
```

### 1.1.2 Lecture du contenu d'un fichier

Nous avons été capables de récupérer le contenu d'un fichier passé en argument grâce à une boucle sur l'opérateur “diamant”. Fondamentalement, cet opérateur permet de boucler sur chaque ligne de “*l'input*”. Nous avons donc simplement récupéré chaque ligne du fichier et avons concaténé ces lignes à une variable \$body initialisée à une chaîne de caractères vide.

### 1.1.3 Lecture du contenu de plusieurs fichiers

Pour pouvoir récupérer toutes les lignes de chaque fichier tout en dissociant les fichiers, nous avons du opter pour une technique légèrement différente.

1. Nous créons une tableau.
2. Chaque “case” du tableau sera une variable similaire à \$body dont nous avons parlé au-dessus : elle contiendra toutes les lignes **d'un seul fichier**.
3. Grâce à l'opérateur diamant, nous bouclons sur toutes les lignes de *l'input*.
4. Nous sommes capable de savoir à tout instant quel fichier nous sommes en train de lire grâce à la variable \$ARGV.

- Une fois que nous avons récupéré tous les différents documents dans chaque case du tableau, nous commençons à boucler sur ce tableau de la même manière que nous le faisons pour un seul fichier.

```
@htmls;
while (<>) {
    $fichier = $ARGV;
    $fichier=~s/.*\\///g;
    if (!defined(@htmls{$fichier})) {
        $htmls{$fichier} = $_;
    }
    else {
        $htmls{$fichier} .= $_;
    }
}

print "<corpus>\n";
while (($fichier,$html) = each(%htmls)) {
    print "<bulletin>\n";
    ...
    print "</bulletin>\n";
}
print "</corpus>\n";
```

## 1.2 Commandes Unix

### 1.2.1 Génération du XML

Voici la commande Unix utilisée pour récupérer la sortie standard de notre programme et la rediriger dans un fichier XML. Il faut également préciser que nous avons rendu notre programme exécutable grâce à `chmod`. De plus, nous utilisons le script `convert.pl` pour convertir les entités HTML en caractère Unicode.

```
./td1.pl BULLETINS/*.htm | perl convert.pl > output.xml
```

### 1.2.2 Vérification du nombre d'images

Nous avons également utilisé quelques commandes Unix supplémentaires dans le but de vérifier que nous récupérerions bien le nombre exactes d'images présentes dans les articles. Par exemple, nous avons utilisé `grep` dans son mode *regex*.

```
./td1.pl BULLETINS/*.htm | perl convert.pl > output.xml && { grep "<image>" output.xml } | wc -l && echo "Images parsées en Perl"
&& { grep -aoE "streaming.+?\.jpg" BULLETINS/*.htm && grep -aoE "<img.*?www\.bulletins-electroniques\.com\/Resources_fm\/actualites.*?\.jpg" BULLETINS/*.htm } | wc -l && echo "Images dans les fichiers d'origine";
```

```
~/g/u/L/T/TD1 >>> rm output.xml -f ; ./td1.pl BULLETINS/BULLETINS/*.htm | perl convert.pl > output.xml && { grep "<image>" output.xml } | wc -l && echo "Images parsées en Perl" && { grep -aoE "streaming.+?.jpg" BULLETINS/BULLETINS/*.htm && grep -aoE "<img.*?www\\.bulletins-electroniques\\.com/Resources_fm/actualites.*?.jpg" BULLETINS/BULLETINS/*.htm } | wc -l && echo "Images dans les fichiers d'origine";  
removed 'output.xml'  
155  
Images parsées en Perl  
155  
Images dans les fichiers d'origine  
~/g/u/L/T/TD1 >>> master *
```

FIGURE 1.1 – Résultat de la commande

## 2 TD2 : Indexation

Le but de ce second TD est d'utiliser le fichier XML que nous avons produit lors du TD1 afin de créer des index (des fichiers inverses). Pour cela, nous avons à notre disposition plusieurs scripts Perl.

1. **segmente\_TT.pl** : permet, à partir d'un fichier XML, d'extraire tous les mots individuellement contenus entre les balises `<titre>` ou `<texte>`.
2. **newcreeFiltre.pl** : permet de ...

### 2.1 Obtenir les tf

Grâce au premier script, nous sommes capable d'obtenir la liste des mots contenus dans les textes et titres de tous les documents HTML. Puis, à l'aide des commandes Unix telles que **sort** et **uniq** (avec l'argument `-c`), nous obtenons pour chaque mot son **tf**.

```
cat ../TD1/output.xml | ./segmente_TT.pl -f | sort | uniq -c | sed -e 's/^\\s*//' > tf.txt
```

La commande **sed** nous permet de supprimer les espaces inutiles en début de chaîne.

### 2.2 Obtenir les idf

Ensuite, pour obtenir les **idf**, nous avons deux possibilités : réutiliser le fichier **tf.txt** directement ou repartir du fichier XML produit lors du TD1. Voici comment faire à partir du fichier **tf.txt** :

```
cat tf.txt | cut -d' ' -f2 | cut -f1 | sort | uniq -c | sed -e 's/^\\s*//' > idf.txt
```

La commande **cut** nous permet de sélectionner uniquement certaines “colonnes” de notre fichier source.

## 3 Conclusion