Exercices

Les fichiers à utiliser sont disponibles sur http://ressources.labex-efl.org/formation-regexp.php

1) Traitement de texte (OpenOffice)

A) Recherche de motifs

*Sur le fichier **retombees_atmospheriques.doc***

- Chercher les années ex : 2001
- Chercher les éléments chimiques ex : Al, Fe

*Sur le fichier **YvesDelaporte_hal00349638_v1.odt***

- Chercher les mots entre guillemets fr. ex : « un » (attention aux car. d'espacement)
- Chercher les appels aux figures dans le texte ex : (fig. 2)
- Chercher les références à la biblio dans le texte ex : (Chauveau 1991)

*Sur le fichier **Chassot_resubmission3.odt***

- Chercher les sigles ex : IFREMER
- Chercher les URL ex : http://www.seaaroundus.org/

*Sur le fichier **2Chianea.doc***

- Répertorier les noms des rois
- Rechercher les titres
- Débusquer les paragraphes vides
- Repérer les appels de notes

B) Substitutions

*Sur le fichier **Chassot_resubmission3.odt***

- Remplacer les parenthèses par des crochets droits dans les citations d'URL ex : (http://www.seaaroundus.org/) devient [http://www.seaaroundus.org/]
- Inverser les séquences noms d'auteur, initiales dans la biblio ex : Agnew, D.J. devient D.J., Agnew

*Sur le fichier **article_transcranial.odt***

- Modifier les entrées biblio pour avoir auteurs, année, titre

  ex : K. Hynynen, FA. Jolesz, “Demonstration of potential noninvasive ultrasound brain therapy through an intact skull.” Ultrasound in Medicine and Biology, Vol. 24, pp 275-283, February 1998
  devient
  K. Hynynen, FA. Jolesz, 1998, “Demonstration of potential noninvasive ultrasound brain therapy through an intact skull.” Ultrasound in Medicine and Biology, Vol. 24, pp 275-283

*Sur le fichier **2Chianea.doc***

- Corriger un défaut de typographie ex : une espace avant une parenthèse fermante ou un point
- Remplacer la pagination des références comme suit :
  pp 456-458 => p. 456-458
  pp 32-58 => p. 32-58
- Remplacer tous les guillemets anglais entourant des chaînes de caractères en guillemets français

3) Éditeurs de texte

A) Recherche de motifs

Corpus : Flexique http://www.llf.cnrs.fr/flexique-fr.php

Sur le fichier ***nlexique.20130910.csv***

1. Chercher tous les mots qui commencent par "accro"
2. Chercher tous les mots qui contiennent un tiret (remarque sur les caractères accentués)
3. Chercher tous les mots qui riment en "ette"

B) Substitutions et groupes de capture

Corpus : Est républicain (fichiers dans l'archive)

Sur le ***fichier 2002-01-02.xml*** *(et 2002-01-03.xml)*

1. Transformer le prénom de la balise <name> en prénom complet (l. 11)
    - Tester d'abord votre expression régulière en ligne : http://rubular.com/r/0aA6D2ohfk
2. Transformer le paragraphe (l. 73 à 78) en deux paragraphes (segmentation sur le point)
3. Ajouter une classe qui aura pour valeur "premier" au premier paragraphe de la balise (l. 81)
4. Transformer tous les mois par les trois premières lettres du mot dans les fichiers 2002-01-02.xml et 2002-01-03.xml
5. Ajouter un étiquetage morpho-syntaxique pour la phrase "Péguy s'est moqué de nous !" (l. 836) de la forme mot/TAG
6. Transformer les numéros de téléphone du <head>L'EST REPUBLICAIN (l. 1924 à 1933) pour aboutir à une forme : +33... , en prenant soin de transformer les séparateurs (.) en espaces.
7. Transformer les paragraphes du <head> BIBLIOTHÈQUE MUNICIPALE (l. 1954) de façon à obtenir la forme : Information : Adultes | Ouverture : 14h | Fermeture : 18h

4) Outils de recherche sur corpus

A) CQP **http://ressources.labex-efl.org/cqp-berder/**

Corpus : FrenchTreebank

1. Rechercher toutes les interjections
2. Rechercher tous les noms dont le lemme finit en "or"

Corpus : C-Oral-Rom

1. Rechercher tous les verbes présents
2. Rechercher les séquences : adverbe suivi d'un adjectif qualificatif

Corpus : Opensubtitle_fr_2013

1. Rechercher les séquences N de N
2. Rechercher les séquences truc de N

B) TXM (installé sur les ordinateurs de la salle)

Corpus : Germinal *http://www.gutenberg.org/cache/epub/5711/pg5711.txt*

1. Chercher les occurrences de mot + Bonnemort
2. Chercher les occurrences de ADJ NOM

C) Frantext **http://frantext.fr/**

1. Identifier les apparitions de Perceval sous ses graphies Perceval ou Percevax
2. Même chose avec Arthur (Artur, Arturus, Artures, Arurum)
3. Rechercher tous les mots qui finissent en -ant : pour ce faire, aller dans l'onglet "mots du corpus" et choisir l'option "expression régulière"
4. Afficher tous les mots qui contiennent deux fois la lettre "z"