UNIVERSITÉ DE CAEN année 2010-11
U.F.R. de SCIENCES

L2 - option LTAL
Linguistique et Traitements Automatiques des Langues

TP 1-2 (2 séances de TP)

Extraction d'informations de fichiers html, avec traitement de toutes les écritures, en utilisant des expressions régulières

Analyse fonctionnelle :

Le programme à réaliser prend en entrée le nom d'un fichier html (des articles de journaux téléchargés de sites d'organes de presse), en extrait le titre de la fenêtre, fait des statistiques sur les mots, sur les caractères, et sur les groupes de mots contigus.

Les résultats seront sortis sous la forme de 2 fichiers html : résultats statistiques, et source original colorié. L'extraction des groupes de 2 ou plusieurs mots se fera par coloriage pour les mettre en évidence.

Le programme pourra traiter des fichiers dans n'importe quelles langue, écriture, et n'importe quel encoding, et sortira les 2 html en utf-8. Les tests seront à faire au minimum pour le français, l'anglais, l'allemand, le finnois, le grec, le russe, et le chinois (fr, en, de, fi, el, ru, zh).

Vous avez accès à des sites d'organes de presse du monde entier sur :

`http://www.courrierinternational.com/sources_overview`

Traiter toutes les écritures :

Pour pouvoir traiter toutes les écritures, il faut une chaîne de traitement unicode :

- en entrée, détection de l'encoding original
- décoder encoding original --> représentation interne unicode python
- tout traiter en mémoire vive dans cet encoding interne
- en sortie, tout encoder en utf-8.

Les expressions régulières

Traitements possibles :

- délimiter une zone de texte entre 2 balises (<title>, ou <a> par exemple)
- débaliser le source, en remplaçant une balise par un caractère #, pour mémoriser sa place comme macro-ponctuation
- définir un segment négativement (par les caractères qu'il ne contient pas), ou positivement (par les caractères qu'il contient)
- calculer la liste de ces segments
- insérer des balises en remplaçant une sous-chaîne par cette même sous-chaîne entourée des balises à insérer.

Préparation :

Familiarisez-vous avec les sources html en les observant sous emacs ou dans le navigateur (afficher source, ou view source), en téléchargeant des articles provenant de sites variés, de langues variées.

Corpus :

Vous prendrez des documents html dans des langues et des écritures différentes : fr, en, de, fi, el, ru, zh.

Programme à modifier :

Un programme de départ vous est donné, et vous le modifiez pour réaliser les fonctions demandées.

Lecture et écriture de fichiers : le source est chargé entièrement au début en une seule lecture dans une unique chaîne de caractères, tous les traitements sont faits en mémoire vive sur cette chaîne, et les fichiers résultats sont écrits entièrement en une seule écriture.

Travail demandé (par ordre de priorité) **:**

1) Vous ferez une étude statistique sur les longueurs des mots, et vous mesurerez la corrélation entre longueur et effectif des mots, en faisant un nuage de points sur ces 2 séries (sur tableur). Commentez et faites le lien avec l'économie d'effort proposée par Zipf.

2) Vous vérifierez la loi de Zipf sur les effectifs des mots : est-ce que le produit "rang du mot dans la liste classée par effectif décroissant * effectif du mot" est à peu près constant ? Faire le graphe de la fonction : effectif = f(rang) en log-log en sortant un fichier texte lu ensuite par un tableur.

3) Est-ce que les caractères vérifient la loi de Zipf ?

4) Dans un document, combien y a-t-il d'occurrences de mots ? de mots différents ? Quel est le rapport des 2 effectifs pour les différentes écritures ? Est-ce que ce rapport caractérise les écritures ?

5) Compter le nombre d'occurrences des groupes de mots contigus (de 2 à 5 mots) présents au moins 2 fois dans le document, et colorier ceux qui sont présents au moins 2 fois dans le document. On procèdera par fenêtre glissante de 2 à 5 mots sur la liste des occurrences de mots.

Travail à remettre :

Travail en binôme, à remettre en 1 exemplaire par binôme, le vendredi 8 octobre en début de TP.

Chaque binôme envoie un mail à <Julien.Gosme@info.unicaen.fr> avec le fichier du rapport, le programme réalisé, les fichiers placés en entrée et ceux obtenus en sortie.

Le rapport devra comporter :

- une courte introduction présentant le sujet du TP,
- un compte-rendu de vos observations des sources html,
- un compte-rendu sur les statistiques sur les mots et sur les caractères,
- un compte-rendu sur le coloriage des groupes de mots contigus présents au moins 2 fois dans le document,

(à chaque compte-rendu, donnez vos expressions régulières commentées, accompagnées d'exemples tirés de votre corpus, et vos observations sur vos résultats)

- une conclusion rappelant l'ensemble des problèmes rencontrés et les solutions trouvées pour les résoudre

(nombre de pages limité à 5, soin particulier pour la rédaction et l'orthographe).

•