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

L2 : option LTAL

Linguistique et Traitements Automatiques des Langues

TP 5-6

Évaluation d'un modèle de calcul de termes générique endogène

Termes et concepts :

terme : groupe de mots faisant sens ensemble pour un lecteur

générique : le même programme prend en entrée des textes de langues variées

endogène : les seules ressources utilisées par l'algorithme sont présentes dans le matériau analysé (modèle calculatoire sans ressources a priori)

Fonctions attendues :

Le programme à réaliser prend en entrée un document html, et ressort les termes répétés de ce texte classés par « surface » décroissante (effectif multiplié par longueur en nombre de syllabes).

Hypothèses de propriétés linguistiques :

- un terme a une des formes suivantes : P+ P+v+P+ P+v+P+v+P+ (v=vide, P=plein)
- un mot vide est monosyllabique ou bisyllabique
- un mot vide est présent dans tout document d'une langue donnée
- une frontière vp ou Pv est caractérisée par le fait que le mot vide est plus court et plus fréquent que son voisin mot plein (effectifs dans tous les documents d'une langue donnée)
- un mot vide dans un contexte est vide dans tout contexte
- un mot plein a 3 syllabes ou plus

Ces propriétés seront exploitées pour étiqueter vide – plein.

Métrique de longueur des mots :

longueur = nombre de syllabes = nombre de noyaux vocaliques

noyau vocalique = suite de voyelles

pour éviter de lister toutes les voyelles accentuées de toutes les langues traitées, on désaccentue les lettres en utilisant la décomposition unicode : è = e + `

cas particulier du "y", voyelle ou consonne selon le contexte :

valeur par défaut : consonne (rayon, loyalty, beyond, citoyen), toujours entre 2 voyelles

voyelle en début ou fin de mot, ou seule (usually, by, y)

voyelle s'il y a une consonne avant (carrying, myopie), ou après (employment, pays), ou les 2 (symbol, hymne)

donnée : listes des voyelles non accentuées (aeiou)

Protocole expérimental :

- constituer un corpus : au moins 3 langues, au moins 3 documents thématiquement différents par langue
- évaluer les hypothèses sur la collection de documents d'une même langue
- implémenter l'étiquetage vide – plein
- implémenter le calcul de terme
- calculer les termes de chaque document, en calculant les mots vides à partir de tous les documents d'une même langue
- évaluer les termes calculés
- évaluer les différents comportements du modèle :
  - selon les langues ?
  - selon la taille des textes analysés ?
  - selon la taille de la collection de textes d'une même langue ?
  - selon le nombre de documents d'une même langue ?

Sorties :

un fichier html avec les termes répétés classés par « surface » décroissante (effectif multiplié par longueur en nombre de syllabes)

Travail à remettre :

Date de remise : le **mardi 30 novembre 2010,** 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.

N'oubliez pas de commenter vos résultats, et évaluez-les qualitativement.

•