

Méthodes Empiriques et Langages de Script

TP 5

G. A. Musillo
musillo4@etu.unige.ch

October 24, 2006

1 Exercice 1: attachement des syntagmes prépositionnels

Une solution empirique au problème de l'attachement des PPs a été formulée par COLLINS et BROOKS, et vous a été présentée en détails par Paola.

On vous demande de compléter le code que Paola a commenté durant le cours (<http://www.lat1.unige.ch/mels/cours5-ppexemple.ppt>). On vous demande également de tester le programme complété et de mesurer l'exactitude de la méthode de COLLINS et BROOKS.

Pour mener cette évaluation, vous utiliserez le fichier

```
http://www.lat1.unige.ch/mels/experimental-quads.rand
```

que vous devez partitionner (9/10 du fichier pour l'entraînement et 1/10 pour les tests). Afin de réaliser le partitionnement, vous pouvez utiliser la commande `split`.

2 Exercice 2: un devineur de langues

On peut mesurer la distance qui sépare deux langues L_1 et L_2 (qui partagent le même alphabet) de la manière suivante:

$$dist(L_1, L_2) = \sum_{x \in \mathcal{A}} |h_{L_1}(x) - h_{L_2}(x)|$$

(\mathcal{A} est l'alphabet commun aux deux langues et

$h()$ est l'entropie d'un caractère)

Codez et utilisez cette distance pour deviner la langue source d'un texte. Afin de réaliser ce programme, il vous faut:

- représenter une langue par un texte d'au moins 10000 lignes.

- partitionner ce texte en un fichier d'entraînement (9/10 du texte) et un fichier de test (1/10 du texte); le partitionnement doit être aléatoire, vous devez donc coder un programme capable de sélectionner aléatoirement les lignes d'un fichier.
- évaluez l'entropie $h()$ des caractères des fichiers d'entraînement et des fichiers de test.
- soumettre chacune des entropies des fichiers de test au programme et les comparer à chacune des entropies des fichiers d'entraînement (la langue du fichier d'entraînement qui minimise la distance est supposée être la langue source du fichier de test).
- reporter les langues devinées.

Vous trouverez à l'adresse <http://www.gutenberg.org/catalog/> de nombreux textes en de nombreuses langues. On vous demande de traiter au moins l'anglais, l'italien, le français et l'allemand.