LA36GM11 – Programmation
Projet

1 Objectifs

L'objectif du projet est d'écrire un programme Perl qui identifie la langue d'un texte contenu dans un fichier. Le programme devra être capable d'identifier au minimum les langues suivantes : Allemand, Anglais, Français, Italien et Néerlandais. Vous êtes toutefois libres d'ajouter des langues supplémentaires.

Pour cela, vous utiliserez les deux méthodes suivantes : (i) une méthode basée sur les mots et (ii) une méthode basée sur les suffixes. Le programme devra afficher la langue identifiée par chacune des deux méthodes, puis combiner ces informations afin de donner la décision finale, de la manière suivante :

```
Bonjour, quel est le nom du fichier à analyser?
texte.txt
La langue du texte d'après l'analyse des mots est : français.
La langue du texte d'après l'analyse des suffixes est : anglais.
Combinaison des résultats...
La langue du texte est : français.
```

1.1 Méthode basée sur les mots

Une langue peut être caractérisée par un ensemble de mots très fréquents. Il s'agit par exemple des articles ou des prépositions. En fonction de leur nombre d'occurrences dans la langue, il est possible de donner un poids à ces mots, les mots les plus fréquents ayant un poids plus important. Le poids (ou fréquence) peut être estimé en divisant le nombre d'occurrences du mot dans un texte de la langue cible par le nombre total de mots du texte. On obtient alors un poids compris entre 0 et 1. Ce poids peut également être exprimé sous forme de pourcentage en multipliant le chiffre précédent par 100.

La Table 1 détaille les poids obtenus après analyse de textes issus du Projet Gutenberg. [1]

Allemand		**Anglais**		**Français**		**Italien**		**Néerlandais**	
und	3,82	the	8,66	de	3,82	di	3,03	de	5,48
er	3,57	of	5,73	le	2,48	e	3,03	en	3,04
die	2,46	in	2,59	et	2,34	la	2,00	van	2,78
der	2,16	and	2,46	la	2,33	il	1,95	het	2,60
in	1,61	a	2,26	à	2,00	che	1,94	in	2,07
war	1,46	to	1,99	les	1,88	a	1,24	een	1,88
zu	1,32	is	1,81	l'	1,61	un	1,13	den	1,25
sich	1,30	as	0,87	il	1,41	non	1,11	ik	1,23
nicht	1,17	it	0,81	un	1,13	in	1,09	die	1,23
es	1,14	with	0,80	d'	1,05	del	1,04	te	1,14

TABLE 1 – Scores des 10 mots les plus fréquents de chaque langue obtenus à partir des données d'apprentissage.

1. Les textes qui ont été utilisés sont les suivants : *Clelia* de Giuseppe Garibaldi (italien), *De Val van Antwerpen* de Jozef Muls (néerlandais), *Le Docteur Ox* de Jules Verne (français), *Der Schwimmer* de John Henry Mackay (allemand) et *A History Of Greek Art* de F. B. Tarbell (anglais).

La langue d'un texte inconnu pourra alors être déterminée en calculant un score pour chaque langue afin de déterminer la langue de score maximal. Pour donner un score à une langue, il suffit d'additionner les poids des mots du texte dans cette langue.

Prenons l'exemple de la séquence « in die ». Les scores pour les différentes langues sont les suivants :

- allemand : 1,61 + 2,46 = 4,07
- anglais : 2,59 + 0 = 2,59
- français : 0 + 0 = 0
- italien : 1,09 + 0 = 1,09
- néerlandais : 2,07 + 1,23 = 3,30

La langue de la séquence « in die » est donc l'allemand. Si aucun score ne dépasse 0, alors la langue du texte est inconnue.

Pour plus de détails sur cette méthode, voir [McN05] (accessible depuis le portail documentaire de l'Université de Strasbourg dans l'ENT).

1.2 Méthode basée sur les suffixes

Une langue peut également être identifiée par les suffixes (ou terminaisons) des mots. On considèrera les suffixes de longueur 1, 2, 3 et 4. Par exemple, pour le mot "industrialisation", le suffixe de longueur 1 est "n", celui de longueur 2 est "on", celui de longueur 3 est "ion" et celui de longueur 4 est "tion". On donnera des poids à ces suffixes en fonction de leur fréquence d'occurrence, comme pour les mots.

2 Tâches

Le projet se décompose en un certain nombre de sous-tâches :

1. Collecte de textes pour chacune des langues à identifier.
2. Extraction des mots et suffixes avec leur fréquence.
3. Développement du système de pondération utilisant les tables de fréquence.

3 Modalités

Le projet devra être réalisé par groupe de 2 étudiants. Les éléments suivants devront être remis :

1. Programmes en Perl **commentés**
2. Données utilisées : corpus, lexiques
3. Rapport de projet de 3 à 5 pages indiquant de manière claire les contributions de chaque membre du groupe. Le rapport décrira ce que fait le programme et présentera une évaluation du système et l'analyse des erreurs.

Votre travail sera jugé en fonction du bon fonctionnement et de la clarté de votre programme, de sa présentation (indentation et commentaires), des choix opérés pour la définition de la solution et de la qualité finale de votre rapport.

La composition des groupes devra impérativement être communiquée au plus tard le 15 novembre 2011.

Le projet complet devra être déposé au plus tard le 8 janvier 2012 sur la plateforme Moodle.

Références

[McN05] Paul McNamee. Language identification : a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3) :94–101, 2005.