Gestion Sémantique des Contenus : TAL **M2 ICONE**
TD2 - Modèles de Langue **gael.lejeune@paris-sorbonne.fr**

Exercice 1: Modèles de langue et diagnostic

Dans les exercicesprécédents, nous avons répéré que l'effectif des mots d'une langue suivait certaines lois de distribution. Nous allons maintenant exploiter ces propriétés pour calculer des "modèles de langue". Ces modèles de langue permettent d'identifier la langue d'un texte inconnu.

Récupérez le corpus de documents multilingue à l'adresse suivante : `https://lejeuneg.users.greyc.fr/ressources/corpus_multi_europarl.zip` Décompressez cette archive dansun dossier dédié. Il y a un sous-dossier pour chaque langue (22 langues en tout) et pour chaque langue un corpus d'apprentissage ("appr") et un corpus de test ("test").

Pour chaque langue du corpus, utilisez l'ensemble des textes figurant dans le dossier « appr » pour calculer un modèle de langue. Ce modèle de langue sera constitué des n mots les plus fréquents identifiés dans ce sous-corpus. Sauvegardez tous ces modèles dans un unique fichier au format Json.

Exercice 2: Appliquer les modèles de langue

Nous allons maintenant regarder si ces modèles fonctionnent bien. Pour chaque langue nous allons nous intéresser aux fichiers contenus dans le dossier "test". Nous allons comparer les 10 mots les plus fréquents de chacun de ces fichiers avec nos modèles de langue. Nous allons ainsi pouvoir établir un diagnostic de langue pour chaque texte.

Exercice 3: Evaluer le diagnostic

Nous devons évaluer l'efficacité de notre système : pour chaque langue d'une part et pour l'ensemble des langues d'autre part. Nous allons considérer que notre but est de trouver la bonne classe (i.e. la bonne langue) pour chaque document. Nous utiliserons le Rappel qui permet de vérifier que l'on a bien classé tous les documents d'une même langue ainsi que la Précision qui évalue combien de documents sont classés dans la bonne langue. Il s'agit donc de calculer le nombre de Vrais Positifs (VP), Faux Positifs (FP) et Faux Négatifs (FN).

Pour chaque document traité à l'exercice précédent nous allons comptabiliser nos VP, nos FP et nos FN. Nous pourrons ainsi pour chaque langue calculer :

- le rappel : VP/(VP+FN)
- la précision : VP/(VP+FP)
- la F_1-mesure (moyenne harmonique du rappel et de la précision) : (2*Rappel*Précision) / (Précision+Rappel)

Ensuite, vous comparerez les résultats selon le nombre n de mots fréquents pris en considération et vous identifierez les erreurs les plus fréquentes : couples de langues et documents.

Exercice 4: Et en caractères ?

Reprenez les 3 exercices précédents en utilisant non plus desmots mais des n-grammes de caractères. Testez pour des valeurs de n de 1 à 4.