NFP 108: projet année 2012/2013

F. Barthélemy

30 novembre 2012

1 Enoncé du projet

Le but du projet est d'extraire un certain nombre d'informations de la page d'accueil du journal *Le Monde* (URL `http://www.lemonde.fr`) au moyen de transducteurs finis.

- extraire les URL des liens
- extraire les URL des images affichées sur la page
- extraire les URL des liens de la page, et pour celle qui sont des URL relatives, les traduire en URL absolues (commençant par http ://).
- extraire les titres des articles dont le texte affiché contient un au moins des mots-clés donnés dans un fichier texte (un mot-clé par ligne).

Chacune des extractions doit donner un automate fini openfst contenant les résultats demandés (et rien qu'eux).

Il faut rendre un ou plusieurs fichiers de grammaire OpenGRM (fichier .grm) accompagnés éventuellement d'explications dans un fichier word ou pdf. Vous pouvez également rendre un ou plusieurs programmes ou un scripts automatisant certaines opérations à réaliser.

2 Indications pour réaliser le projet

Le projet nécessite une connaissance minimale de HTML pour chercher les informations dans le code source de la page du monde. Pour les liens, il s'agit d'explorer les balises `<a...>`, pour les images, les balises `<img...>`. Pour les articles, il faut trouver dans le code des éléments qui les caractérisent et qui les délimitent (pour chercher les mots-clés dans le texte correspondant à un article). Il faut pour cela comprendre le code source de la page du monde.

Pour lire dans un fichier texte les mots-clé à rechercher pour les articles (un mot-clé par ligne) et transformer ce fichier en un automate fini, il faut utiliser la commande `StringFile` d'OpenGRM.

La page du monde a été choisie parce qu'elle utilise un encodage des caractères sur un octet, à savoir l'encodage ISO-8859-1 (ASCII étendu aux caractères accentués d'Europe Occidentale). Un fichier contenant ces caractères est fourni (sur le site web du cours) de même qu'un fichier contenant un automate fini (en mode texte) décrivant l'ensemble des caractères de cet

encodage. La fichier de caractère donne des noms spéciaux à quelques caractères qui ne s'affichent pas comme les autres : l'espace (`<space>`), la tabulation (`<tab>`) les retours à la ligne (`<car_ret>` et `<newline>`).

Pour réaliser le projet, vous pouvez installer les logiciels sur votre machine ou utiliser le serveur dept25.cnam.fr comme lors du TP. De l'extérieur, pour vous connecter, il faut passer par vlad. Pour travailler dans de bonne conditions sous windows, il faut installer un SSH (par exemple PuTTY) et un environnement X (par exemple Xming).

Parmi les tâches à réaliser, il faut transformer la page du monde en un automate fini qui contient une seule longue chaîne : le code source HTML de la page en question. Cette tâche est optionnelle : un automate est fourni sur le site web du cours traduction d'une page du monde, ainsi qu'un script python qui réalise cette traduction. Vous pouvez travailler avec l'un ou l'autre outil ou faire votre propre traduction.

3 Informations pratiques

Le projet sera à rendre via un formulaire web avec accès par login et mot de passe. Il s'agit de vos login et mot de passe du CNAM (les mêmes que sur dept25). Un lien sera mis en ligne sur le site web du cours.

La date limite de remise est le dimanche 13 janvier.

Vous pouvez me contacter par email en cas de besoin.