Fiche de proposition d'un projet tutoré

Proposé par : Stéphane GORIA

Date : 9 septembre 2014

Type : Exploratoire

Description du projet

Développer un logiciel qui permettent de déduire le sens de certains textes à partir des termes qui s'y trouvent et dans une certaine mesure de regrouper les textes ayant des termes ou expressions en commun.

Objectifs à atteindre

Développer un analyseur de textes (une dizaine à une vingtaine) au format pdf, doc et odt. Dans un premier temps une liste de mots d'une langue identifiée manuellement devra être reconnue et mémorisée par le logiciel. Ce seront les mots les plus courts et les plus fréquents de la langue du texte (par exemple : a, et, le, la l, les, de, du, un, en, donc). Il s'agira de repérer les séquences de caractères entre deux « blancs » les plus fréquentes. On pourra d'abord lister les séquences de 1 à 4 caractères, puis rechercher les suites de 2 séquences les plus répétées (par exemple : d un, de la, dans un, mais si, n est). Ces suites et séquences formeront une petite base stockant les « ***séquences fréquentes*** ». Ensuite, une deuxième analyse recherchera les autres séquences que l'on retrouve au moins 2 fois dans les textes. Ce seront des séquences de 3 ou 4 caractères, mais qui ne sont pas dans liste des « ***séquences fréquentes*** » ainsi que les séquences entre deux blancs de 5 caractères ou plus. Elles formeront les « ***séquences intéressantes*** » Toutes les « séquences intéressantes » seront alors recherchées dans le texte pour identifier si on ne les trouve pas précédées ou suivi au moins 2 fois par une même « séquence fréquente ». Elles donneront les la liste des « ***longues séquences intéressantes*** ». Cette troisième liste pourra être étendue en recherche si deux séquences intéressantes ne sont pas séparées à deux reprises au moins par une séquence fréquente ou un blanc. Elles formeront ainsi la liste des « ***très longues séquences intéressantes*** ». Cette dernière liste pourra, avec le même raisonnement, être complétée par des séquences encore plus longues.

À la fin de l'analyse d'un texte, nous disposerons de différentes séquences qui sont répétées dans le texte. À la manière d'un nuage de tags leur fréquence représentera le contenu du texte. Mais, afin compenser les effets de longueur des séquences, un **poids** sera associé aux différentes ***séquences***

intéressantes qui correspondra à leur nombre de caractères multiplié par leur fréquence. La taille d'affichage des séquences du nuage sera fonction du **poids des séquences**.

Enfin, un affichage de nuages plusieurs textes sera proposés à l'utilisateur. Lorsque l'on cliquera sur une séquence présentée dans un nuage, toutes les séquences identiques des autres textes seront colorées. Un réseau de nuages pourra aussi être dessiné. La force des liens du nuage (épaisseur) correspondra à la somme des séquences communes (à deux nuages liés) en fonction du **poids le plus faibles** de chacune de ces séquences dans l'un ou l'autre texte.

Outils à utiliser

Le développement est à faire de préférence pour tourner sur PC avec Windows.

Le choix du langage et des outils est laissé aux étudiants.