Rapport projet Analyse de Données Textuelles

MARIJON Pierre, PIVERT Jérôme, PICARD DRUET David

24 novembre 2015

Problématique

Notre problématique était d'extraire les termes présents dans un ensemble de mails, provenant de la liste de diffusion de la SFBI sur la période d'un an. Ces termes devaient ensuite être associées à des mots clefs. **Pas satisfait de la formulation, à ajuster. Il faut plus préciser sur les données**

L'idée était de pouvoir obtenir, à partir des mails, les termes associés aux différents thèmes de la bioinformatique. Il fallait également éviter d'utiliser les librairies déjà existantes sur le sujet, en dehors de NLTK.

Enfin, il fallait pouvoir évaluer la qualité et la quantité de données récupérées, ainsi que les limites de notre méthode d'extraction de terminologie.

Conception du programme

Workflow

— Parsing des mails : récupération du corps et du sujet de chaque mail.
— Tokenisation
— Filtration
— Racinisation/Lemmatisation

Association thème et termes

3 méthodes : - thème = tout les termes qui y sont associés, en vrac. - thème = chaque terme associé, avec une valeur de pondération dépendant du nombre de fois que le terme a été trouvé. - thème = terme trouvé dans TOUT les mails avec ce thème.

Utilisation

Le programme a été conçu pour être utilisé en console, en ligne de commande.

Usage :
./parse_email.py (–input=<repository>) (–output=<file>) [options]

Options :
–help, -h Show help message.
–input, -i=<repository> The directory containing all data.
–output, -o=<file> The file with results.
–stopword_fr=<file> French stop words file. (default value include)
–stopword_en=<file> English stop words file.(default value include)
–debug Activate debug mode

Évaluation des résultats obtenus

Améliorations, alternatives

Conclusion