PROJETS MOTEURS/SEO

Ce document présente les projets associés au cours sur le fonctionnement des moteurs de recherche et le référencement.

Vous devez rendre votre ou vos projet(s) avant le 31 juillet (inclus). Chaque jour de retard vous pénalise de 1 point.

Le rendu se fait par mail, envoyé à sylvain@ix-labs.org.

Le sujet du mail doit comporter [MTI][2017].

Le corps du mail doit comporter le titre du sujet choisi, les noms (réels) des membres de l'équipe, et les login associés.

Le rendu doit comporter, dans une archive zip ou tgz : votre code, un rapport court (autour de 5 pages max, merci d'éviter le remplissage inutile), des binaires macosX ou linux. Si vous êtes en mode SAAS, vous devez déployer votre projet à un endroit où nous pourrons le tester.

Le langage préféré est python, mais vous pouvez faire ce que vous voulez.

Les documents textes seront en PDF, et les fichiers textes seront encodés en UTF8.

Projet 1 – Cartographie du contenu dupliqué
facile

Il s'agit de réaliser une chaine de traitement qui prend en entrée un ensemble de pages web, qui en extrait le contenu, et qui compare page à page les contenus pour réaliser une matrice d'indication des taux de duplication.

La chaine de traitement doit aussi prendre en entrée le paramètre de taille des shingles qui seront utilisés. L'utilisation de méthodes probabilistes pour accélérer le calcul est un plus.

Les matrices de taux de duplication seront fournis en format CSV, mais aussi sous une forme graphique en pdf.

Projet 2 – Aide à la rédaction
facile

Il s'agit d'implanter un système d'aide à la rédaction qui va prendre en entrée un ensemble de textes, et va en extraire les n-grams les plus signifiants.

Nous conseillons d'utiliser la TF-IDF pour classer les n-grams qui auront été extraits. Nous rappelons aussi qu'il ne faut pas réaliser de lemmatisation pour réussir une extraction de qualité.

Projet 3 – Thématisation
moyen

Il s'agit de réaliser un pipeline qui apprend et reconnaît des thématiques.

1. interaction avec l'utilisateur pour saisir un certain nombre de thématiques
2. soumission de textes dans chaque thématique
3. extraction des caractéristiques (n-grams représentatifs) et validation/suppression des n-grams non signifiants en interaction avec l'utilisateur
4. reconnaissance de la thématique d'un texte à analyser (il est conseillé d'utiliser la distance de jaccard).

Projet 4 – Acquisition automatique de liens thématiques
plus dur

Il s'agit de réaliser une chaine de traitement qui réalise les tâches suivantes :

1. Crawl en continu à un rythme faible
2. Extraction du texte des pages crawlées
3. Lemmatisation, calculs des TF-IDF (et éventuellement co-occurrence)
4. Comparaison à une page cible
5. Lorsque la comparaison (calcul du cosinus de Salton) aboutie à une valeur supérieure à un certain seuil, renvoyer la page pour éventuelle acquisition de lien

Vous aurez besoin pour ce projet d'un dictionnaire de lemmatisation disponible ici : http://www.lri.fr/~syp/dico.zip. N'oubliez pas le traitement des entités nommées.

Projet 6 – Rédaction automatique avec le KG
plus dur

Il s'agit d'utiliser le knowledge graph (ou freebase) pour générer automatiquement du contenu. C'est la diversité des contenus générés et leur apparence de « naturalité » qui donnera une bonne (ou mauvaise) note à ce projet