NLP
Travail pratique 2
Date de remise : 13 novembre 2015, 23h55.

Jonathan Gingras

8 novembre 2015

Nom : Jonathan Gingras
Matricule : 111 004 940
Numéro du cours : IFT-7022

1 a&b) Préalales (installations et fichiers nécéssaires)

Premièrement le choix de l'engin de recherche est *ElasticSearch*. Il est donc nécéssaire, avant de rouler le projet, de démarrer une instance de *ElasticSearch* roulant sur le port 9200 (le port par défaut). La version utilisée est 2.0.0. Il est possible qu'une autre version fonctionne, toutefois aucune autre n'a été testée.

Également, le projet est implémenté en Ruby (version 2.2.3 de l'interpréteur sous ma machine). Certaines dépendances sont nécéssaires pour rouler le projet. Voici la procédure pour les installer :

- Avoir `bundler` :
  `$ gem install bundler`
- Avoir `imagemagick` : disponible via homebrew sous Mac ou sur environ tous les bon package managers dans le monde GNU/Linux (Le projet ne fonctionne pas sous Windows)
- `cd` vers la racine du projet
- `$ bundle install --path vendor/bundle`

Les fichiers fournis pour le travail se trouvent dans les fichiers `corpus.txt`, `requests.txt` et `pertinence.txt`. Ils proviennent directement du site du cours.

2 c) Indexation

Avant de commencer, il est primordiale qu'une instance de *ElasticSearch* roule sur le port 9200. L'indexation est faite en utilisant `index.rb`. Il ne s'agit que de rouler :

```
$ ruby index.rb
```

Il est toutefois assez long d'indexer les fichiers, il faut prévoir au moins 10 minutes pour les 4 configurations (4 indexes différents correspondants aux 4 configurations de la prochaine question).

Les détails quant à l'implémentation des appels `REST` vers `ElasticSearch` se retrouvent dans `elastic_search.rb`.

3 d) Expérimentation sur l'évaluation des différents facteurs

3.1 La normalisation de mots par stemming

3.2 Le retrait de mots outils (stop words)

3.3 La pondération des mots (ex. tf*idf)

4 e) Estimation de la précision, rappel et F-mesure pour les configurations testées