Exam 2 : Travaux pratiques à rendre

GjT - Durée 1h
24 Nov 2014

Objectifs

L'objectif de ce travail pratique est de vérifier vos acquis des principales fonctions R utilisées en web content mining

Conseil : Ne cherchez pas la meilleure façon de faire les choses, mais plutôt, la façon la plus simple

Vous pouvez utiliser l'aide disponible sur Internet. Un copier-coller sans la compréhension est facilement répérable

Commentez brièvement votre code

Bonne chance !

Question 1 : Extraction de l'information du web

Récupérer le texte text html d'un des discours devenu célèbres ici
http://www.lemonde.fr/ameriques/article/2009/01/20/le-texte-integral-du-discours-de-barack-obama-en-anglais_1144446_3222.html

Aide :

```
doc.html = htmlTreeParse("",useInternal = TRUE)
# Consultez l'aide de la fonction xpathApply
```

Question 2 : Créer un corpus

Aide : Utilisez la vignette tm pour faire du pre-processing

- Enlever la ponctuation
- Enlever les mots vides
- touts ce qui ressemble à un url
- Les mots vides
- Etc

```
corpus <- Corpus(VectorSource(x))
corpus_bigdata <- tm_map(corpus_bigdata, function)
```

Question 3 : Créer deux tdm

- Créer une tdm avec une pondération tf
- Créer une tdm avec une pondération tfidf
- Compter la fréquence d'apparition de chaque mot

Aide : Le précédent tp

Question 4 : Effectuer une classification hiérarchique des mots du texte et représenter graphiquement les différentes classes

- Bien faire apparaitre les groupes. Option: rect.hclust()

Bon courage