Projet M2 - cahier des charges

Programmation Python - M2 IDL - 2020-2021

Description du sujet

Le travail comporte deux parties :

1/ la première partie consiste en l'aspiration d'un corpus de textes à partir du Web, en utilisant des librairies python (p.ex. BeautifulSoup, Scrapy, etc.)

2/ la deuxième partir consiste en l'utilisation d'une chaine de traitement de Spacy, en vue de réaliser une des tâches suivantes, au choix :

- évaluation / adaptation de l'analyseur syntaxique en dépendance de Spacy
- classification de textes avec scikit-learn
- anonymisation
- extraction de discours direct, extraction de citation (untel a dit " blablabla ")
- extraction de relations entre entités nommées (p.ex. person → location, etc.)
- autre tâche en fonction de vos envies

A faire en binôme. Vous rendrez une archive zip contenant :

- les codes pythons correctement nommés et commentés. Notamment, pour chaque fonction, une docString doit préciser le but, les entrées et les sorties de la fonction (ou du script).

- dans un répertoire à part, les données sur lesquelles vous avez travaillées, ainsi que des jeux de tests effectués.

Vous accompagnerez ces codes d'un rapport concis expliquant :

- une description du projet

- les packages utilisés, avec les procédures d'installation le cas échéant

- le découpage en modules, les chaines de traitements mises en oeuvres (faites des schémas)

- les choix algorithmiques et les principales structures de données (fichiers, tableaux, etc.)

- une évaluation sommaire des sorties

- les bogues constatés

- les améliorations à apporter et extensions prévues.

Remise des travaux le 5 janvier

O. Kraif