INALCO - Licence LMFA - TNM L3, Algorithmique et structures de données

Caroline Koudoro-Parfait

20 Mars 2023

Nom de l'étudiant :
n° d'étudiant :

Projet

1 Modalités :

Le projet :

- est individuel,
- sera à terme, constitué :
  - d'un programme python commenté dans ses grandes lignes,
  - d'un fichier PDF de 1 à 3 pages répondant aux questions ci-dessous,
- il doit être envoyé par email le 27 mars 2023 au plus tard, à caroline.parfait@outlook.fr.
- en cas de problèmes pour la remise du fichier en parler avant le 27 mars 2023.

2 But du projet

La problématique pour le projet à programmer consiste à réaliser un outil d'évaluation et de comparaison des résultats produits par différents traducteurs automatiques pour un même texte source.

3 Constitution du corpus

Le corpus est constitué comme suit :

- Chanson, poème, texte court en langue française (50 lignes environ) ;

- Traduction officielle du texte choisi dans la langue choisie par l'étudiant (Vérité de terrain) ;
- Les différentes traductions obtenues avec les outils de traductions recensés par l'étudiant.

4 Similarités et Distances avec scikit-learn

installer scikit-learn :

- `https://pypi.org/project/scikit-learn/`
- `https://scikit-learn.org/stable/install.html`

À partir de la documentation de scikit-learn : `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.DistanceMetric.html`

Vous vous entrainerez à faire tourner le programme suivant et vous analyserez autant que possible les résultats obtenus.

```
from sklearn.neighbors import DistanceMetric
from sklearn.feature_extraction.text import CountVectorizer


str3 = "Sur la planÃ¨te Mars il n'y a pas de plan B pour le climat"
str4 = "Sur la planete Mar il n'y a pas de plan b pour le clima"


dist = DistanceMetric.get_metric("jaccard")
V = CountVectorizer(analyzer='word')
X = V.fit_transform([str3, str4]).toarray()
distance_tab=dist.pairwise(X)

print(distance_tab)
```

- Vous utiliserez différentes métriques,
- Vous pouvez ajouter des phrases à comparer.
- faire des tests avec des phrases vraiment différentes pour tester que les métriques marquent bien que les phrases sont vraiment différentes
- faire des tests avec la même phrase exactement pour vérifier que les métriques montrent que les phrases sont bien similaires

Vous compléterez/modifierez ce programme pour stocker au format json :

- les phrases comparées,
- les noms des métriques utilisées,
- les résultats correspondants, et non pas les tableaux de résultats.