Stage Extraction d'information à partir de documents PDF complexes H/F

Test de compétences Python et NLP

Exercice 1 , source Codility. Difficulté facile

https://app.codility.com/programmers/trainings/5/parity_degree/

Exercice 2 , source Codility. Difficulté facile

https://app.codility.com/programmers/trainings/4/str_symmetry_point/

Exercice 3 , source Codility. Difficulté moyenne

https://app.codility.com/programmers/trainings/5/three_letters/

Exercice 4, source Codility. Difficulté moyenne

https://app.codility.com/programmers/lessons/15-caterpillar_method/min_abs_sum_of_two/

Exercice 5, source Codility. Difficulté élevée

https://app.codility.com/programmers/trainings/7/countries_count/

Exercice 6, source CEA-List LASTI. Difficulté moyenne

A partir du texte mentionné ci-dessous, utiliser une librairie NLP python pour extraire les entités nommées (incluant les noms de personnes, d'organisations ou d'entreprises, de lieux, les dates, etc.) présentes dans le texte.

Le texte à analyser est à récupérer d'une page Wikipedia Anglais :
https://en.wikipedia.org/wiki/Marie_Curie

Produire un résumé sémantique de l'information extraite :

- Restituer les 5 types d'entités les plus fréquents. (exemple d'entité : PERSON)
- Restituer les 5 mentions d'entités les plus fréquentes. (exemple de mention : Marie Curie)
- Restituer 5 cooccurrences de type d'entités les plus fréquentes avec leur mention.
  NB : On considère qu'il y a cooccurrence lorsque 2 entités apparaissent dans la même phrase .
  Quelques exemples parmi les plus intéressants pour l'extraction d'information :
  PERSON + DATE , PERSON + PERSON , PERSON + LOCATION , ORG + DATE , ORG + LOCATION , etc

Bonus 1 : implémentez 2 fois l'extraction d'information avec 2 librairies NLP différentes dont aymara/lima https://github.com/aymara/lima ,
Utilisez la version Python de Lima sur les textes en anglais https://pypi.org/project/aymara/ :

```
import aymara.lima
nlp = aymara.lima.Lima("ud-eng")
... etc
```

Bonus 2 : comparez les sorties des 2 librairies NLP sur l'extraction des entités nommées . On sélectionne une librairie pour devenir arbitrairement le producteur de la vérité terrain. La seconde librairie est évaluée sur cette base. Donner les mesures de Précision, Rappel et F1-mesure .
NB : Un mapping des types d'entité sera nécessaire a priori : vérifier que le type d'entité désignant PERSON est compatible entre les deux librairies testées. Si ce n'est pas le cas, faire un replacement automatique dans les résultats de la seconde librairie. Idem sur les autres types d'entités.

Bonus 3 : résolvez à l'échelle du document la normalisation des mentions d'entités (cf entity linking) , plusieurs méthodes sont possibles : via une simple comparaison des chaînes de caractères (cf distance d'édition) , ou éventuellement via un word-embedding sur les tokens de l'entité et un calcul de similarité.