TP1

Partie 1 | étude de cas *CoNLL 2003* :

1. Quelle type de tâche propose CoNLL 2003 ?
2. Quel type de données y a-t-il dans CoNLL 2003 ?
3. A quel besoin répond CoNLL 2003 ?
4. Quels types de modèles ont été entraînés sur CoNLL 2003 ?
5. Est un corpus monolingue ou multilingue ?

Partie 2 | En vous inspirant des informations que vous avez récupérées pour CoNLL 2003, définissez les besoins de votre projet:

1.dans quel besoin vous inscrivez vous ?

2.quel sujet allez vous traiter ?

3.quel type de tâche allez vous réaliser ?

4.quel type de données allez vous exploiter ?

5.où allez vous récupérer vos données ?

6.sont-elles libres d'accès ?

Réponses :

Partie 1 :

1. NER (Named Entity Recognition) , une tâche de Reconnaissance d'Entités Nommées, c'est une tâche de séquence étiquetée (sequence labeling).

2. Le format IOB, personnes (PER), organisations (ORG), lieux (LOC) et autres (MISC). Les phrases sont séparées par des lignes vides.

3. CoNLL-2003 répond au besoin de reconnaissance d'entités nommées dans des textes. Comme l'analyse de documents, l'extraction des mot-clés, etc.

4. Les modèles statistiques comme HMM, ou les modèles à base de réseaux neuronaux comme BERT et RNN. Ou les Conditional Random Fields (CRF). Ou les SVM ou encore des systèmes hybrides.

5. Il s'agit d'un corpus multilingue. Car il contient des données en anglais et en allemand.

Partie 2 :

1. L'objectif est de classifier automatiquement les critiques de films en deux catégories : positives et négatives.

2. Le sujet est la classification de sentiment (sentiment analysis), à partir de critiques cinématographiques en français, extraites du site Allociné.

3. C’est une tâche de classification binaire de texte. Chaque critique est représentée comme une séquence de texte, avec un label 1 (positive, note ≥ 3) ou 0 (négative, note < 3).

4. Les critiques textuelles de films rédigées par des utilisateurs, avec peu de mots sont au nombre de trois, contenant entre 70 et 100 caractères, tandis que les autres se situent entre 100 et 530 caractères.

5. Les données ont été récupérées à partir du site Allociné, avce le script *crawler.py* en utilisant *requests* et *BeautifulSoup*.

6. J’ai consulté le fichier *robots.txt* d’Allociné, et il n’intserdit pas le crawl de la section *critique*, donc c’est autorisé. Toutefois, Allociné ne propose pas d’API officielle pour accéder aux données, et ses conditions générales d’utilisation n’autorisent pas explicitement le scraping automatisé. Afin de rester dans un cadre éthique et légal, j’ai limité le nombre de critiques (environ 20), et aussi utiliser *time.sleep(1)* pour ne pas surcharger le serveur. Et je n’ai pas mis les données sur GitHub.