ToolBox

CentraleSupelec & ILLUIN Technology
12/1/2022 - 27/1/2022

Méthodes de NLP d'extraction
- Expression régulière
- Question Answering Extractif
- Question Answering Booléen
- Named Entity Recognition (NER)
- Extraction de relations entre entités
- NLI (Natural Language Inference)
- Topic Modelling
- Clustering
- Information Retrieval

Librairies python open-source

Méthodes de NLP d'extraction

Expression régulière

- Définition: Les expressions régulières sont des chaînes de caractères qui permettent, par l'utilisation de caractères spéciaux, de désigner un ensemble de chaînes de caractères.

Question Answering Extractif

- Définition: Le Question Answering extractif est une tâche du NLP qui consiste à trouver la réponse à une question posée dans un contexte (paragraphe) donné.

Question Answering Booléen

- Définition: Le Question Answering booléen est une tâche du NLP qui consiste à répondre à oui ou non à une question posée dans un contexte (paragraphe) donné.

Named Entity Recognition (NER)

- Définition: La NER est une tâche du NLP qui consiste à extraire des entités (pays, ville, adresses, nom, prénom, etc) dans du texte.

Extraction de relations entre entités

- Définition: L'extraction de relations est la tâche consistant à prédire les attributs et les relations des entités dans une phrase. Par exemple, dans la phrase "Barack Obama est né à Honolulu, Hawaï", un classificateur de relations vise à prédire la relation "bornInCity". L'extraction de relations est le composant clé pour la construction de graphes de connaissances de relations, et elle est d'une importance cruciale pour les applications de traitement du langage naturel telles que la recherche structurée, l'analyse des sentiments, la réponse aux questions et le résumé.

NLI (Natural Language Inference)

- Définition: La NLI est une tâche du NLP qui consiste à déterminer si une "hypothèse" est vraie (implication), fausse (contradiction), ou indéterminée (neutre) étant donné une "prémisse".

Topic Modelling

- Définition: Le Topic Modelling est une tâche du NLP non supervisée qui consiste à identifier les thématiques principales contenues dans des textes. Il est nécessaire d'établir en amont le nombre de thématiques à rechercher.

Clustering

- Définition: Le Clustering est une tâche de data science non supervisée qui consiste à regrouper des données homogènes dans un même cluster. Il est nécessaire d'établir en amont le nombre de clusters à rechercher.

Information Retrieval

- Définition: L'Information Retrieval est une tâche du NLP qui consiste à trouver les n documents ayant le plus de probabilités de répondre à la recherche d'un utilisateur.

Librairies python open-source

- transformers :
    - Lien : https://huggingface.co/transformers/
    - Description : Librairie permettant d'entraîner et d'utiliser des modèles de NLP dont l'architecture repose sur des transformers à l'état de l'art. De nombreux modèles pré-entraînés sont disponibles : https://huggingface.co/models.
    - Tâches : NER, QA, Sentence Classification, Token classification, etc.

- sentence-transformers :
    - https://github.com/UKPLab/sentence-transformers
- spacy :
    - Lien : https://spacy.io/
- scikit-learn :
    - Lien : https://scikit-learn.org/stable/
- gensim :
    - Lien : https://radimrehurek.com/gensim/index.html
- nltk :
    - Lien : https://www.nltk.org/
- rank_bm25 :
    - Lien : https://pypi.org/project/rank-bm25/
    - Tâches : Information Retrieval
- sentencepiece :
    - Lien : https://pypi.org/project/sentencepiece/
    - Description : sentence piece tokenizer