Knowledge Extraction from Unstructured Text Using Machine Learning Techniques

Objectif:.

L'objectif de cet exercice est d'explorer et de mettre en œuvre des approches d'apprentissage automatique pour extraire des connaissances structurées (entités, relations, faits) à partir de données textuelles non structurées. Les étudiants utiliseront des techniques de traitement du langage naturel (NLP), combinées à des méthodes d'apprentissage supervisé et non supervisé, afin d'identifier et d'extraire des informations pertinentes à partir de contenu textuel brut

Dataset Options:

Choisir les datasets suivants et autres types de données à chercher (à partir de KAGGLE, GIthub)

1. **Articles Wikipedia** (e.g., figures historiques, topics scientifiques)
2. **News Articles Dataset** (e.g., COVID-19, geopolitical news, etc.)
3. **Commentaires des clients Dataset** (Amazon, Yelp, etc.)
4. **Abstracts Scientifiques** (from arXiv or PubMed).
5. **Autres types de data sets**

Tâches:

Partie A: Preprocessing et Representation Text

- Nettoyage et normalisation du texte (suppression bruit, ponctuation, etc.)
- Appliquer tokenization, POS tagging, lemmatization
- Convertir texte à une representation numerique en utilisant:
    - TF-IDF ou
    - Word embeddings (e.g., Word2Vec, GloVe) ou
    - Transformer-based embeddings (BERT, etc.)

Partie B: Entity Recognition et Classification

Appliquer Named Entity Recognition (NER) en utilisant :

    - Rule-based methods (e.g., spaCy)
    - ML models (e.g., CRF, BiLSTM-CRF)
- Classifier les entités en catégories (e.g., Person, Organization, Date, Product)
- Evaluation des performances en utilisant des métriques : precision, recall, F1-score

---

Partie C: Relation Extraction et la construction du Knowledge Graph (KG)

Extraction sémantique des relations entre entités (e.g., "X works at Y", "X born in Y")

- Utiliser dependency parsing ou bien ML-based relation classification
- (Optional) Fine-tune un modèle de transformer (e.g., BERT for RE)
- Construire un **mini knowledge graph** en utilisant les outils comme :
    - NetworkX, Neo4j, ou RDF triples
- Visualiser le réseau entity-relation

Partie D: Analyse et Discussion

- Discussion :
    - Challenges dans cette extraction
    - Comparison entre rule-based and ML-based techniques
    - Ambiguïté, polysémie, et autres problèmes linguistiques
- Réflexion du comment la connaissance extraite peut être utilisée dans des applications réelles (e.g., question answering-QA, recommandation, search)
- Utiliser an LLM pour faire les tâches précédentes pour construire un KG et faire la comparaison des résultats avec les approches de la partie C avec différents critères.

Part E: Rapport, Code, présentation orale

- Remettre :
    - un documented Jupyter notebook/ Python script
    - A Rapport (max 5 pages) :
        - Methodologie
        - Resulats et visualizations
        - Analyse critique

Deadline: 19 décembre 2025

Tools & Libraries:

- Python, Jupyter Notebook
- NLTK, spaCy, Scikit-learn
- Hugging Face Transformers
- NetworkX or Neo4j
- Pandas, Matplotlib/Plotly for visualization