**Exercice 1 : Scraping de Données Textuelles**

**Objectif** : Apprendre à extraire des données textuelles du web pour construire un dataset.

**Tâches** :

1. Utiliser une bibliothèque de scraping comme BeautifulSoup ou Scrapy pour extraire les données des liens suivants :
   1. <https://burkinainfo.com> pour ce lien récupérer les titres des articles pour les 10 premières pages.
   2. <https://lefaso.net> : pour ce lien récupérer les titres des articles pour les 10 premières pages de la rubrique politique et société.
2. Chercher et récupérer les données sur les différentes attaques terroriste au Burkina de 2020 à 2025.
3. Nettoyer et stocker les données extraites dans un format structuré (CSV ou JSON).

**Exercice 2 : Introduction au Traitement du Langage Naturel (NLP)**

**Objectif** : Se familiariser avec les bibliothèques spaCy et NLTK.

**Tâches** :

1. Fusionnez les deux premiers dataset obtenu lors du scraping (a et b)
2. Charger le dataset de texte obtenu par la fusion.
3. Effectuer la tokenisation et le POS tagging (Part of Speech) sur le texte avec spaCy et NLTK.
4. Comparer les résultats entre les deux bibliothèques.

**Exercice 3 : Prétraitement des Données Textuelles**

**Objectif** : Apprendre les étapes de prétraitement des textes.

**Tâches** :

**Pour le dataset fusionnée et le dataset des attaques, effectuer les taches suivantes** :

* 1. Nettoyer le texte : suppression de la ponctuation, des stop words, et transformation en minuscules.
  2. Effectuer la tokenisation, le stemming et la lemmatisation.
  3. Analyser l'impact de chaque étape de prétraitement sur le texte.

**Exercice 4 : Exploration des Embeddings de Mots**

**Objectif** : Comprendre l’utilisation des word embeddings pour la représentation des mots.

**Tâches** :

* 1. Utiliser les embeddings de spaCy pour obtenir des vecteurs de mots.
  2. Visualiser les similarités entre différents mots à l'aide de PCA ou t-SNE.
  3. Expérimenter avec des mots similaires et observer les résultats des embeddings.

**Exercice 5 : Extraction d'Entités Nommées et Analyse Syntaxique**

**Objectif** : Apprendre à extraire les entités nommées et analyser la syntaxe d'un texte.

**Tâches** :

* 1. Utiliser spaCy pour détecter les entités nommées dans un texte.
  2. Analyser la structure syntaxique d'une phrase complexe.
  3. Extraire les relations entre entités et comprendre leur contexte.

**Exercice 6 : Classification de Texte avec Naive Bayes**

**Objectif** : Appliquer un modèle de machine learning Naive Bayes pour la classification de texte.

**Tâches** :

* 1. Préparer un dataset de classification binaire (ex. critiques positives vs négatives).
  2. Utiliser TF-IDF pour vectoriser le texte.
  3. Entraîner un modèle de classification Naive Bayes pour prédire la classe d'un texte.
  4. Évaluer la performance du modèle avec des métriques appropriées.

**Exercice 7 : Introduction à BERT pour le NLP**

**Objectif** : Apprendre à utiliser le modèle BERT pour des tâches de NLP.

**Tâches** :

* 1. Utiliser la bibliothèque transformers de Hugging Face pour charger un modèle pré-entraîné BERT.
  2. Appliquer BERT pour la tâche de classification de texte (ex. sentiment analysis) sur le dataset préparé.
  3. Comparer les performances de BERT avec celles des modèles traditionnels comme Naive Bayes.

**Exercice 8 : Génération de Texte avec GPT**

**Objectif** : Découvrir les capacités de génération de texte de GPT.

**Tâches** :

* 1. Utiliser la bibliothèque transformers pour charger un modèle GPT (comme GPT-3).
  2. Générer des textes basés sur des prompts spécifiques (ex. début d'un article, histoire courte).
  3. Analyser la qualité et la cohérence des textes générés.
  4. Discuter des applications potentielles et des défis éthiques liés à l'utilisation de modèles génératifs.

**Exercice 9 : Analyse de Sentiment**

**Télécharger le dataset ici** [**https://drive.google.com/file/d/1C-vY-J1iq4B4KGaRQwDzeIcf\_9PLEmED/view?usp=sharing**](https://drive.google.com/file/d/1C-vY-J1iq4B4KGaRQwDzeIcf_9PLEmED/view?usp=sharing)**.**

**Objectif** : Comparer les méthodes d'analyse de sentiment basées sur des lexiques à celles basées sur des modèles de machine learning.

**Tâches** :

* 1. Utiliser ce dataset pour Classifier les discours haineux en utilisant un modèle de machine learning entraîné pour l'analyse de sentiment.