🧩 Mini-projet 1 — Reuters-21578 (Actualités économiques)

📚 Description

- Ensemble de **21 578 dépêches économiques** issues de l'agence Reuters, catégorisées par thèmes (*acq*, *earn*, *grain*, *money-fx*, etc.).
- Corpus classique pour l'apprentissage et la recherche d'information.

📂 Accès

- Déjà inclus dans `nltk`:

```
import nltk
from nltk.corpus import reuters
nltk.download('reuters')
```

- Fichiers stockés localement après téléchargement (`nltk_data/corpora/reuters/`).

📄 Format

- Chaque fichier = une dépêche (`.sgm`)
- Métadonnées : `reuters.fileids()`, `reuters.categories(fileid)`

🔗 Source

- NLTK Reuters-21578 documentation
- Licence : usage éducatif libre (provenance Reuters, 1987)

🧩 Mini-projet 2 — arXiv Physics (Articles scientifiques)

📚 Description

- Métadonnées et résumés d'articles scientifiques arXiv (titre, résumé, catégorie, date).
- Sous-ensemble : *physics*, *cs*, *math*, *astro-ph*.

📂 Accès

Télécharger l'échantillon libre :

- 🔗 Kaggle – arXiv Metadata Dataset (Cornell)
  Fichier : `arxiv-metadata-oai-snapshot.json` (~7 GB, format JSON)

Créer un échantillon :

```
import pandas as pd, json
rows = []
with open('arxiv-metadata-oai-snapshot.json') as f:
    for i,line in enumerate(f):
        if i>10000: break
        d=json.loads(line)
        rows.append((d['title'],d['abstract'],d['categories']))
df = pd.DataFrame(rows, columns=['title','abstract','categories'])
```

Licence

- CC BY 4.0 (Cornell University, arXiv.org)

Mini-projet 3 — Wikinews EN (Actualités internationales)

Description

- Articles anglophones issus de **Wikinews**, média libre collaboratif de la Wikimedia Foundation.
- Données ouvertes et multithématiques (politique, environnement, sport, etc.).

Accès

- Wikimedia Dumps
  → télécharger le fichier `enwikinews-latest-pages-articles.xml.bz2`
- Extraction simplifiée possible avec `wikiextractor` :

```
pip install wikiextractor
python -m wikiextractor.WikiExtractor enwikinews-latest-pages-
articles.xml.bz2 -o output/
```

ou un jeu nettoyé :
Kaggle – English Wikinews Dataset (CSV)

Licence

- CC-BY-SA 3.0 (Wikimedia Foundation)

🧩 Mini-projet 4 — PubMed Open Access Subset (Corpus biomédical)

📚 Description

- Résumés et métadonnées d'articles biomédicaux accessibles en open access (NLM / NIH).
- Champs : `title`, `abstract`, `mesh_terms`, `journal`, etc.

🗂️ Accès

- 🔗 NCBI PubMed Central Open Access Subset
  (archives XML compressées)

Échantillon CSV libre :
🔗 Kaggle – PubMed Open Access Subset

💡 Conseil

Extraire 10 000 abstracts max.

```
import pandas as pd
df = pd.read_csv("pubmed_200k_train.csv").sample(10000)
```

🏷️ Licence

- CC-BY (National Library of Medicine, NIH)

🧩 Mini-projet 5 — Wikipedia FR (Articles culturels)

📚 Description

- Articles francophones issus de la catégorie "Arts et Culture" (musique, peinture, littérature…).
- Contenu sous licence libre Wikimedia.

🗂️ Accès

- 🔗 Wikipedia FR Dumps
  → fichier `frwiki-latest-pages-articles.xml.bz2`
  Extraction :

```
python -m wikiextractor.WikiExtractor frwiki-latest-pages-
articles.xml.bz2 -o output/
```

- **Alternative :**
  🔗 Kaggle – French Wikipedia Articles Dataset (CSV)

🏷️ Licence

- CC-BY-SA 3.0 (Wikimedia Foundation)

🧩 Mini-projet 6 — Stack Exchange Data Science (Forum technique)

📚 Description

- Données textuelles issues du forum **Data Science Stack Exchange** (questions, réponses, tags).
- Thèmes : apprentissage automatique, statistique, IA, Python, etc.

🗂️ Accès

- 🔗 Archive Stack Exchange officielle
  → fichier : `datascience.stackexchange.com.7z`
- Extraction des champs `Title`, `Body`, `Tags` avec `BeautifulSoup`.

Ou jeu simplifié :

- 🔗 Kaggle – Data Science Stack Exchange Posts

🏷️ Licence

- CC-BY-SA 4.0 (Stack Exchange Inc.)
