Unstructured 0.13.7

UNSTRUCTURED	Outil d'ingestion de données pour LLM	
		Multiplateforme
	Développé par :	Unstructured
Open source, Apache-2.0	Personne de contact :	katy.fokou@smals.be

Fonctionnalités

Unstructured -
structurées dans le but de construire des applications utilisant les grands modèles de language typiquement de type RAG (questions-réponses).
sous différents formats provenant de sources diverses. Il ou plateforme *low-code* en en version open-source.

Les fonctionnalités clés sont les suivantes :

-
  les fichiers Powerpoint, Word, Excel, XML ou les pages Web (html).
- Division du texte en blocs (*chunking*) avec différentes méthodes, taille fixe, par titre, par fenêtre
- Génération *embeddings* ou représentations vectorielles de blocs de texte par intégration de modèles de tierce-partis.
- -
  source).

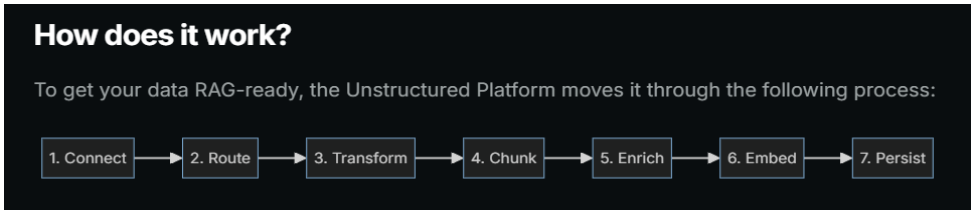

Fig1: Pipeline de traitement de données

Conclusions & Recommandations

Unstructured est un outil open-source incontournable pour le traitement de données pour leur utilisation dans des applications basées sur les LLM (*large language model*). La plateforme est très versatile et de nombreuses fonctionnalités intéressantes sont disponibles dans la version gratuite.

Tests & Résultats

Pour évaluer -source de
é
et de tables.

```
from IPython.core.display import display, HTML
display(HTML(tables[1].metadata.text_as_html))
```

```
/tmp/ipykernel_455318/1760248645.py:1: DeprecationWarning: Importing dis
ython display
  from IPython.core.display import display, HTML
```

Metrics	AlfWorld		ScienceWorld		HotPotQA		FEVER	
	SR(%)	EX($)	AR	EX($)	SR(%)	EX($)	SR(%)	EX($)
Z-CoT					0.01	0.95	0.39	1.07
F-CoT	0.43	98.60	16.58	272.22	0.32	5.73	0.61	2.25
CoT-SC	0.57	105.37	15.24	274.33	0.33	7.86	0.62	321
SayCan	0.60	113.61	12.36	125.71				
ReAct	0.57	152.18	15.05	356.03	0.34	66.00	0.63	22.20
Reflexion	0.71	220.17	1939	72448	0.39	112.49	0.68	26

Fig1. Table extraite d'un document PDF

Le premier test consiste en la validation de la fonction « partition » pour extraction de textes de documents et la fonction de *chunking*. La fonction « partition » peut être utilisée en spécifiant le format du document au préalable. Les documents sous format PDF, html, msg ont été testés et correctement traités .
Pour les documents PDF en particulier, plusieurs réglages sont possibles dont
,
etc extraction sans-faute (fig1) de documents PDF ainsi que la détection de titres et liens hypertexte de page html.
pas très fiable, beaucoup de morceaux de textes non pertinents sont identifiés comme titres. De même, la
En plus des fonctions spécifiques de partition, Unstructured propose une fonction générique capable de détecter automatiquement le format du docume
appliqués à des documents de différents formats et de différentes langues sont tous positifs.

La deuxième série de tests porte sur
(fig2). Le pipeline mis en place contient les éléments suivants : connexion à la source, extraction, chunking, *embedding* et stockage dans une base de données vectorielle (Chroma). Le pipeline

python ou en ligne de commande. Unstructured propose par défaut de multiples connecteurs pour les source de données brutes et les destinations de données traitées. Le pipeline est facile à mettre en

données se fait rapidement.

```
if __name__ == "__main__":
    Pipeline.from_configs(
        context=ProcessorConfig(work_dir=False, num_processes=4),
        indexer_config=LocalIndexerConfig(input_path=LOCAL_FILE_INPUT_DIR),
        downloader_config=LocalDownloaderConfig(),
        source_connection_config=LocalConnectionConfig(),
        partitioner_config=PartitionerConfig(
            partition_by_api=False,
            strategy="hi_res",
            hi_res_model_name='detectron2_mask_rcnn',
            additional_partition_args={
                "split_pdf_page": True,
                "split_pdf_allow_failed": True,
                "split_pdf_concurrency_level": 15
            }
        ),
        chunker_config=ChunkerConfig(chunking_strategy="by_title", chunk_max_characters=1000),
        embedder_config=EmbedderConfig(embedding_provider="langchain-huggingface",
                                       embedding_model_name="intfloat/multilingual-e5-large"),
        destination_connection_config=ChromaConnectionConfig(
```

Fig2. Ex. de configuration d'un pipeline

Conditions d'utilisation & Budget

Unstructured est disponible comme bibliothèque Python open-source et sous licence Apache ou, sous
à 10$ par 1000 pages.