COLLIN
Hugo

AYACHI
Yacine

BOURBON
Pierre

PERBET
Lucile

CHALLENGE IA

SISE CAMP

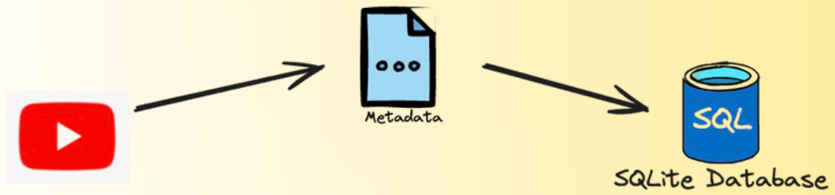
https://sise-camp.streamlit.app

OBJECTIFS DE L'APPLICATION

Effectuer des recherches dans la chaine youtube MASTER 2 SISE DATA SCIENCE

- Lancement de la vidéo à partir du prompt
- Obtenir un résumé de la vidéo
- Effectuer des quiz générés par LLM

SCRAPPING



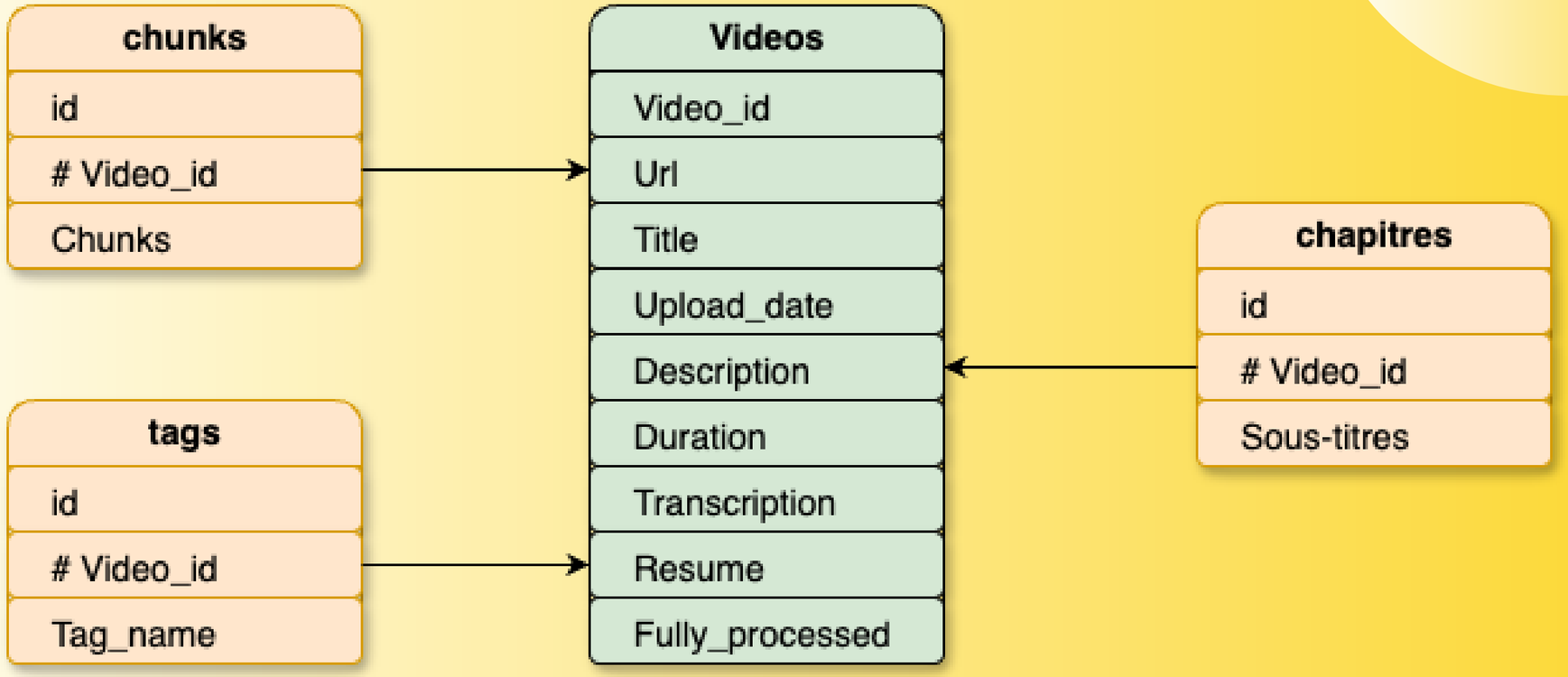
- Récupère tous les liens des vidéos de la chaîne SISE sur YouTube
- Enregistrement dans la base de données
- Ajout des métadonnées

294 vues 2 févr. 2025
Mise en œuvre d'une méthode de machine learning (une régression logistique) de la librairie MLlib via l'API Python PySpark (3.5.4) pour Spark. Les fichiers d'apprentissage et de test se présentent au format CSV usuel (avec en-tête, séparateur de colonne « ; », variable cible codée en chaînes de caractères). Nous mettons l'accent sur la préparation des données, assez spécifique, cruciale pour le bon fonctionnement du dispositif. La vidéo décrit dans ses aspects les plus schématiques les étapes d'entraînement sur l'échantillon d'apprentissage, de prédiction et d'évaluation sur l'échantillon test. L'idée est de disposer d'un exemple simple facile à reproduire.

MLlib : https://spark.apache.org/docs/latest/...
PySpark : https://spark.apache.org/docs/latest/...
Notebook et données : https://tutoriels-data-science.blogsp...

00:00 MLlib de Spark via PySpark
04:50 Organisation des données (fichiers CSV, train vs. test)
06:18 Environnement pour PySpark
09:15 Démarrage du notebook
09:58 Création d'une session
10:35 Chargement et inspection des données
12:11 Vectorisation des variables prédictives (VectorAssembler)
14:48 Recodage de la variable cible
15:42 DataFrame pour l'entraînement du modèle
16:00 Apprentissage - Régression logistique
17:04 Affichage des coefficients estimés dont l'intercept
17:51 Informations détaillées sur le modèle
18:50 Prédiction sur l'échantillon test
23:55 Confrontation : matrice de confusion
25:23 BinaryClassificationEvaluator de PySpark
27:00 Fin de session

BASE DE DONNÉES



TRANSCRIPTION

Extraction de l’audio

Transcription avec Whisper de Hugging Face (Whisper Large v3 Turbo)

Amélioration de texte avec Mistral AI (Mistral Large)

Bien, c'est parti. Dans cette vidéo, je
Bien, c'est parti. Alors, dans cette vid
Bien, c'est parti. Alors, dans cette vid
Bien, c'est parti. Dans cette vidéo, je
Bien, c'est parti. Dans cette vidéo, je
Bien, c'est parti. Alors, dans cette vid
Bien, c'est parti. Alors, dans cette vid
Bien, c'est parti. Dans cette vidéo, je
Voici le texte corrigé et amélioré : ---
Bien, c'est parti. Dans cette vidéo, je

MOTEUR DE RECHERCHE

Question

Mistral_Embed

Vector DB - Faiss

Text_Video

Chunk

Vector DB - Faiss

Chapitres

Chunk + Chapitre

SQL

SQLite DB

URL de la vidéo
Chapitre
Timestamp
Résumé LLM

GÉNÉRATION DU QUIZZ

Embeddings
du chunk des résultats de le recherche

LLM

Quizz

SISE CAMP

MERCI POUR VOTRE ÉCOUTE