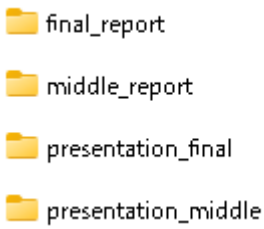
引継ぎ資料

各フォルダとファイルについて

- texのプロジェクト関連



- resources

データセット(テキスト，ベクトル）や，Doc2Vecの訓練済みモデル，提案モデルの訓練済みモデルが入っています．それとデータベースのバックアップもあります．

※データベースの構造については変更したほうが良いです（後述）

- その他ファイル

名前の通りです．

開発環境

OS: Windows11+WSL2(Ubuntu)

言語：Python3.10

仮想環境：pyenv+venv（dockerとかでいい気がする）

エディタ：VSCode+RemoteWSL

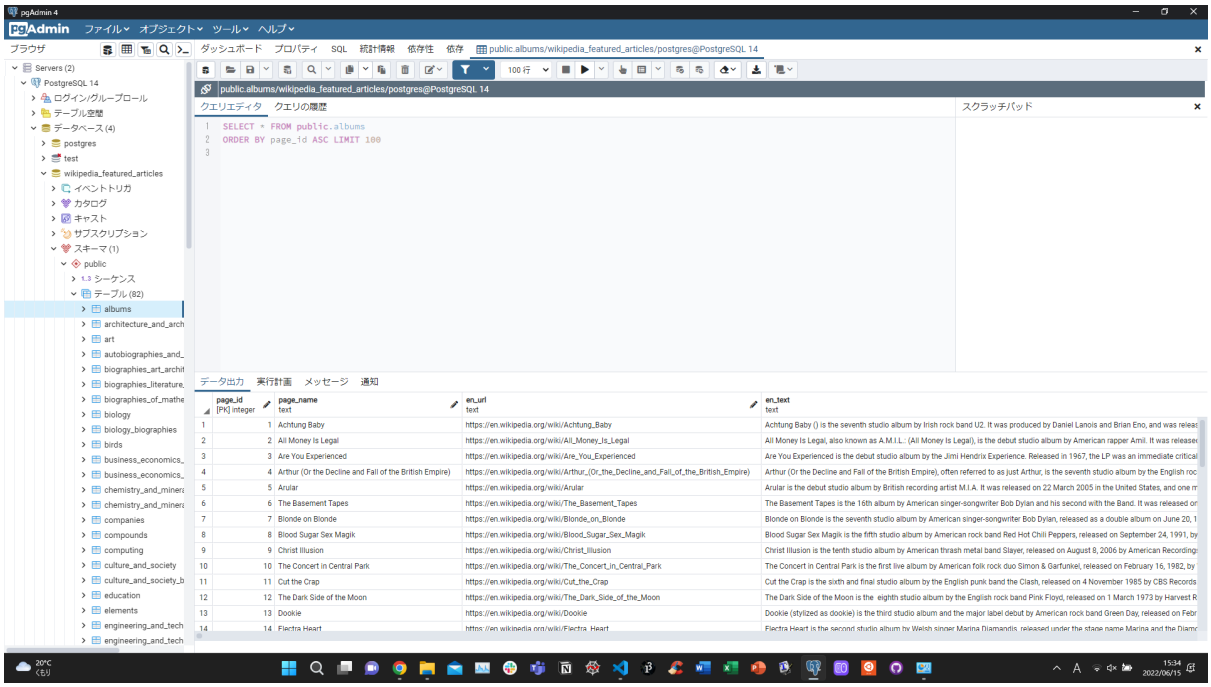
文書執筆環境：texlive on WSL+VSCode+latexworkshop(オーバーリーフで良いと思う）

各プログラムの説明

https://github.com/satabie/wikipedia-Data_to_DB

データベースにWikiFAの記事データを格納するプログラム．

使い方の詳細はREADME.mdへ



現在の状態だと，トピックごとにテーブルを分けているが，テーブルを分ける事のメリットよりもテーブルを分けたことで発生するデメリット(page_idが一意にならない)の方が大きいため，一つのテーブルにすべてのデータを格納するように変更されることを勧めます．

ただし変更すると他のプログラムに影響が出るため，そちらもまとめて変更が必要になります．(具体的にはデータベース操作を行っているプログラム)

https://github.com/satabie/train_wiki_doc2vec

doc2vecの訓練用プログラム

https://github.com/satabie/add-vector-to-db

↑で訓練したモデルを使って，データベースにデータを追加する

https://github.com/satabie/train-wikiFAmodel

提案モデルを訓練するプログラム