Code Usage Instructions

Question 1: KYC and AML

Hangyul Son
hson@connect.ust.hk

Set-Up

1. Set up Google Cloud Console according to the link.
2. Set up Google Vision API according to the link.
    a. Create 'credentials.json' file
3. Set up Google Drive API (pydrive) according to the link
    a. Create: 'client_secrets.json' file
4. Install python packages from 'requirement.txt' file.

Files

- **main.py**: Backbone of the entire KYC process.
- **google_drive.py**: Download HKID and address proof images from google drive
- **extract_id.py**: Extract text from the downloaded HKID image using Google Vision API.
- **extract_address.py**: Extract text from the downloaded address proof image using Google Vision API.
- **verify_regex.py:** Verify the extracted texts using regex.
- **database.py:** Save the verified texts to MySQL database

Code Explanation

1. The code starts and ends at main.py.
2. google_drive.py downloads two images, a sample HKID card image, and a sample bank statement image. A packaged version of the Google Drive API, pydrive, has been used for downloading the images.
3. After the images have been fully downloaded, main.py calls extract_id.py or extract_address.py to perform text extraction from the images.
    a. For HKID, the code automatically retrieves the '**Name**,' '**ID**,' '**Date of Birth**,' and '**Date of Issue**' from the HKID image.
    b. For address proofs, the code detects one of the three districts in Hong Kong: Kowloon, New Territories, and Hong Kong Island (I have assumed a KYC system based in Hong Kong). 4 lines of address are extracted from the address proof.
4. Saved the extracted text into a JSON format.
5. The extracted text is verified to ensure it is in the right format using regex expressions (Verification is done only for text extracted from the HKID for now) using the file verify_regex.py.
6. Once the text has been verified, save the text to MySQL database, to HKID and ADDRESS tables using database.py.

Notes

- Using the OpenCV library to apply denoising/thresholding turned out to worsen the performance of text extraction. Therefore, the code has been commented out.