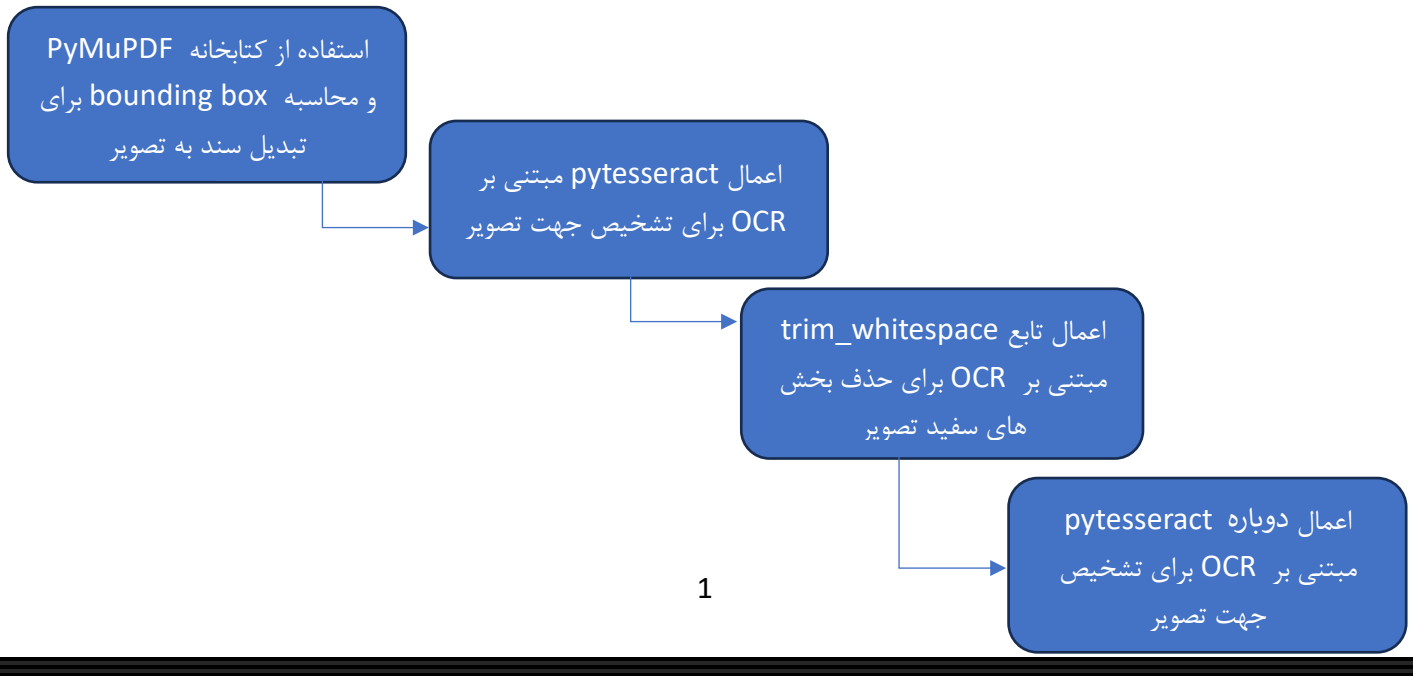
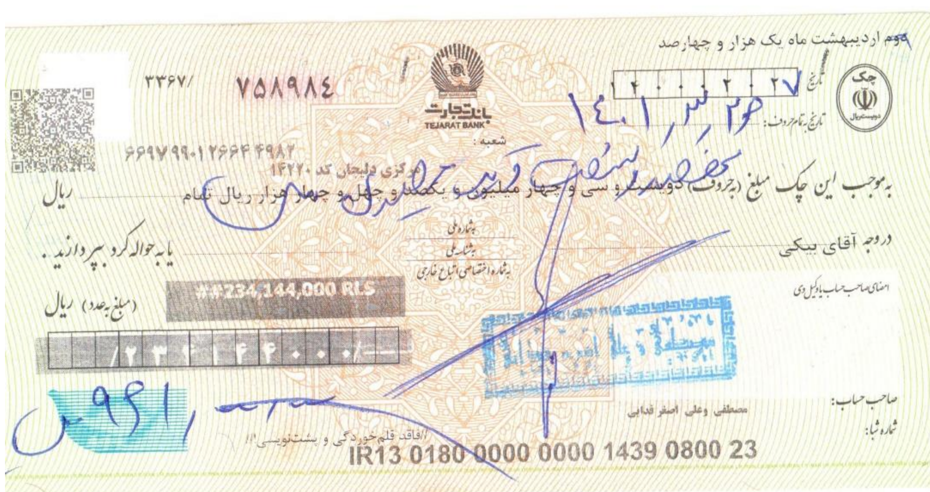
پایپلاین پیش پردازش دیتاست : (لینک گیت هاب)

دیتاست موجود از اسکن چک ها شامل فایل های pdf ای بود که هر صفحه شامل یک چک بود هر چند بعضی صفحات هم خالی بودند ...
این کد با استفاده از کتابخانه‌های PyMuPDF و PIL پیاده‌سازی شده است و قابلیت حذف فضاهای سفید اطراف تصاویر را دارد.
بخش اصلی این کد تابع trim_whitespace است که مراحل زیر را برای حذف بخش سفید هر صفحه انجام می دهد:

ابتدا تصویر ورودی را به gray scale تبدیل میکنیم . سپس تصویر را معکوس یا invert می کنیم و این کار به تشخیص بهتر نواحی غیر سفید کمک می‌کند. بعد یک (Bounding Box)را برای نواحی غیر سیاه در تصویر معکوس محاسبه می کنیم . این جعبه کوچک‌ترین مستطیلی است که تمامی نواحی غیر سفید را در بر می‌گیرد. در اخر تصویر ورودی بر اساس جعبه محصور کننده برش داده می‌شود تا فضای سفید اطراف آن حذف شود. بعد از اعمال تبدیل های مد نظر بر روی pdf های ورودی ، تصاویر را در پوشه جدیدی ذخیره می کنیم .

سپس از کتابخانه pytesseract برای تشخیص جهت درست تصاویر تبدیل شده استفاده میکنیم این کتابخانه بر اساس OCR متن فارسی و انگلیسی جهت درست چک های مربوطه را تشخیص و اعمال می کند (البته طبیعتا برای همه چک ها درست عمل نمی کند به ویژگی مواردی که skewness دارند)
در مرحله بعدی بدلیل اینکه تابع trim_whitespace تمام بخش های سفید اطراف چک را حذف نکرده است در اینجا دوباره بر اساس OCR متن چک ها ، bounding box هایی را تخمین می زنیم تا حذف دقیق تری انجام شود (همچنین بدلیل اینکه با این روش بخشی از محتوای شامل QR code که دارای متن نبودند پاک می شد ۱۵۰ پیکسل از هر طرف کمتر حذف کردیم تا هیچ بخشی از اطلاعات چک های حذف نشوند)
بعد از طی کردن تمام مراحل قبلی در نهایت دوباره کد مربوط به تعیین orientation درست تصویر با کتابخانه pytesseract اعمال می شود که در خروجی نهایی فقط ۵ از ۱۱۰ تصویر جهتشان درست انتخاب نشده است!





تصویر بالا سمت راست خروجی تبدیل PDF به تصویر است و هیچ برشی روی آن ایجاد نشده است
تصویر بالا سمت راست خروجی تبدیل PDF به تصویر است و با تابع trim-whitespace و ایجاد bounding box بخش های سفید حذف شده اند.

تصویر پایین خورجی نهایی پایپلاین است که خروجی های سفید کناره تصویر را حذف کرده است همچنین با کتابخانه pytesseract با استفاده از متن چک تصویر را چرخانده است.

[لینک از نمونه تصاویر پیش پردازش شده چک]