BÁO CÁO NGHIÊN CỨU HỌC TẬP TUẦN 7 VÀ TUẦN 8

1. Nghiên cứu trong việc đọc ảnh PDF xuát qua file Excel

* Sử dụng thư viện Apose-pdf để chuyển đổi dữ liệu trong file pdf qua file excel

A computer screen shot of a program

AI-generated content may be incorrect.  
- Kết quả: + Chỉ đọc được 4 trang free, tuy nhiên đọc không hề hiệu quả.

* Em đã cắt từng trang 1 của file pdf để có thể dùng apose-pdf chuyển từng trang pdf sang excel tuy nhiên đều dính footnote và không hề chuyển được dữ liệu qua.

+ Cần mua bản quyền để có thể sử dụng.

A screenshot of a computer

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.  
A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Sử dụng Tesseract-OCR để đọc ảnh   
   - Cần chuyển pdf -> ảnh -> Tesseract-ORC để đọc ảnh-> xuất dữ liệu qua file excel

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of text

AI-generated content may be incorrect.

* Kết quả:

+ Với đầu vào là báo cáo tài chính của VNM trong Q1 2025 thì đã đọc được thông tin, tuy nhiên tiếng Việt Tesseract-OCR xử lý còn khá nhiều lỗi.

A screenshot of a computer

AI-generated content may be incorrect.

1. Viết mã nguồn để tải file PDF từ trang web chứng khoán bằng python.

A computer screen shot of a program code

AI-generated content may be incorrect.

* Kết quả: Các trang web chứng khoán giờ đã tích hợp báo cáo tóm tắt trực tiếp trên trang web, tải file vẫn còn lỗi tuy nhiên có thể nghiên cứu thêm về cào file pdf bctc hoặc cào dữ liệu báo cáo tài chính tóm tắt từ trang web luôn.  
  -> Trong tuần 7 em đã tiếp tục nghiên cứu sâu hơn về cào file pdf bctc từ trang web chứng khoán, tuy đã chia nhỏ từng bước để nghiên cứu sâu hơn tuy nhiên kết quả vẫn chưa đạt được như mong đợi, mã nguồn liên tục thông báo về là không tìm thấy file pdf trong trang web.

A screenshot of a computer program

AI-generated content may be incorrect.