THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://www.youtube.com/watch?v=tY-ApIvycBs
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/tuanmc15/CS2205.CH183
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

- Họ và Tên: Mai Chấn Tuấn
- MSSV: 240101030

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 8/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 2
- Số câu hỏi QT của cả nhóm:
- Link Github:
  https://github.com/mynameuit/CS2205.xxx/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG THỊ GIÁC MÁY TÍNH VÀ MÔ HÌNH NGÔN NGỮ LỚN ĐỂ TỰ ĐỘNG HOÁ THÔNG MINH VIỆC PHÂN TÍCH VĂN BẢN HÀNH CHÍNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

LEVERAGE COMPUTER VISION AND LARGE LANGUAGE MODELS FOR SMART AUTOMATION OF ADMINISTRATIVE DOCUMENTS

TÓM TẮT *(Tối đa 400 từ)*

Tài liệu hành chính là một trong những loại tài liệu được quản lý chặt chẽ nhất trong nhiều lĩnh vực do vai trò quan trọng của chúng trong việc quản trị và vận hành. Tài liệu này thường do các giám đốc điều hành hoặc lãnh đạo doanh nghiệp ban hành và phục vụ các mục đích như công bố chính sách, phê duyệt ngân sách và chuẩn bị báo cáo. Độ thận trọng của việc sử dụng ngôn ngữ trong các tài liệu này vô cùng cao. Bất kỳ sự mơ hồ hoặc lỗi nhỏ nào cũng có thể dẫn đến hiểu sai, tạo ra kẽ hở trong quy trình hoặc rủi ro về việc khai thác lỗ hổng.

Những sai sót trong tài liệu hành chính có thể gây ra hậu quả nghiêm trọng, bao gồm tranh chấp pháp lý, quản lý tài chính sai lệch, bị phạt do vi phạm quy định và tổn hại danh tiếng. Một điều khoản không rõ ràng trong hợp đồng có thể tạo ra lỗ hổng dễ bị khai thác hoặc một báo cáo tài chính có lỗi diễn đạt có thể dẫn đến tính toán ngân sách sai và một chỉ thị chính sách không rõ ràng có thể khiến các bộ phận thực hiện một cách không đồng nhất. Trong các ngành có yêu cầu quản lý nghiêm ngặt như tài chính, y tế và chính phủ, những sai sót nhỏ có thể dẫn đến vi phạm, kéo theo các khoản phạt lớn hoặc gây gián đoạn hoạt động.

Chúng tôi đề xuất một giải pháp tích hợp công nghệ LLM và thị giác máy tính tiên tiến để phát hiện lỗi trong tài liệu hành chính nhằm ngăn chặn những sai sót đã đề cập. Kết quả mong đợi là một hệ thống phát hiện lỗi tự động với độ chính xác cao, đảm bảo tính chính xác, loại bỏ sự mơ hồ và đảm bảo các quy định được đề ra một cách chặt chẽ. Bằng cách triển khai giải pháp này, các tổ chức có thể nâng cao độ tin cậy của tài liệu hành chính, giảm

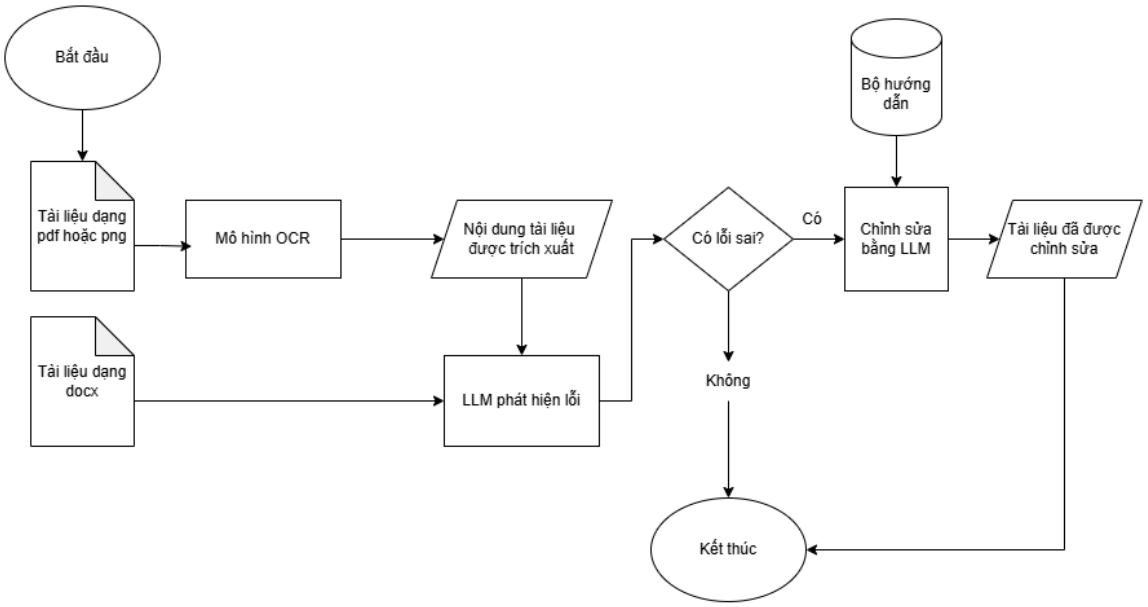
thiểu rủi ro pháp lý và vận hành, đồng thời tối ưu hóa quy trình rà soát tài liệu.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Văn bản hành chính là văn bản hình thành trong quá trình chỉ đạo, điều hành, giải quyết công việc của các cơ quan, tổ chức. Văn bản hành chính là dạng văn bản có cấu trúc, nội dung nhất quán, từ ngữ được sử dụng trong văn bản phải rõ ràng, chính xác tránh gây mập mờ, không chứa các nội dung đối lập. Văn bản hành chính cũng đòi hỏi hình thức trình bày khắt khe với các quy chuẩn.

Chính vì điều này, soạn thảo văn bản hành chính là một công việc khó khăn và đòi hỏi người thực hiện phải có kiến thức chuyên môn, hiểu biết về các quy định pháp luật, cũng như kỹ năng diễn đạt mạch lạc, súc tích. Ngoài ra, việc soạn thảo văn bản hành chính cần đảm bảo tính khoa học, logic, tuân thủ đúng thể thức và quy trình ban hành. Một văn bản hành chính được soạn thảo chính xác không chỉ giúp truyền đạt thông tin một cách hiệu quả mà còn thể hiện sự chuyên nghiệp, góp phần nâng cao hiệu quả quản lý, điều hành của cơ quan, tổ chức.

Trong bài báo này, chúng tôi đề xuất một phương pháp mới kết hợp công nghệ OCR và các mô hình ngôn ngữ lớn trong việc phân tích và sửa lỗi văn bản hành chính. Phương pháp này tận dụng khả năng nhận dạng ký tự quang học (OCR) để trích xuất nội dung từ các tài liệu hành chính dưới dạng số hóa, sau đó sử dụng các mô hình ngôn ngữ lớn (LLM) để phân tích, phát hiện và sửa lỗi và trả về văn bản đã được chỉnh sửa một cách tự động. Cách tiếp cận này không chỉ giúp cải thiện độ chính xác trong việc nhận diện văn bản mà còn nâng cao hiệu quả xử lý ngôn ngữ tự nhiên, giảm thiểu sai sót và tối ưu hóa quy trình soạn thảo, kiểm duyệt văn bản. Ngoài ra, chúng tôi cũng tiến hành đánh giá hiệu suất của phương pháp đề xuất trên tập dữ liệu thực tế, nhằm chứng minh tính hiệu quả và khả năng ứng dụng của hệ thống.



MỤC TIÊU

1. **Phân tích và xử lý nội dung tài liệu:** Ứng dụng Optical Character Recognition (OCR) và mô hình ngôn ngữ lớn (LLM) để đọc văn bản, phân tích cú pháp, ngữ pháp và ngữ nghĩa với độ chính xác cao.
2. **Đánh giá hiệu suất:** Sử dụng các chỉ số precision, recall, F1 score để đo lường hiệu quả của hệ thống.
3. **Triển khai hệ thống**: Sử dụng dịch vụ đám mây như AWS để triển khai Virtual Private Computer nhằm host 1 website cho phép người dùng tương tác với hệ thống một cách thuận tiện.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Nghiên cứu cách thức OCR hoạt động.

- **Mục tiêu:** Ứng dụng OCR trong việc đọc các văn bản đầu vào nhằm chuyển hóa các kiểu văn bản đặc biệt như PDF, … sang kiểu văn bản có cấu trúc. Việc này giúp hệ thống có thể xử lý nhiều loại văn bản hơn bên cạnh các kiểu văn bản cơ bản, có cấu

trúc khác như doc, docx, txt,...

- **Phương pháp:** Nghiên cứu các tài liệu hiện có về công nghệ OCR, đồng thời kết hợp thực hành để đẩy nhanh quá trình nắm vững kỹ thuật này, qua đó rút ngắn thời gian tìm hiểu và ứng dụng vào hệ thống.

Nội dung 2: Sử dụng LLM để phân tích và sửa lỗi văn bản

- **Mục tiêu:** Từ output của mô hình OCR, sử dụng mô hình LLM để phân tích, phát hiện, và sửa lỗi.
- **Phương pháp:** Áp dụng mô hình ngôn ngữ lớn (LLM) để phân tích văn bản hành chính dựa trên một bộ hướng dẫn (guideline) được xây dựng nhằm chuẩn hóa các tiêu chí về hình thức và nội dung. Sau đó huấn luyện hoặc điều chỉnh mô hình LLM để nhận diện, đánh giá và phân tích văn bản theo các tiêu chí này. Phương pháp này cho phép mô hình tự động phát hiện sai sót, đề xuất chỉnh sửa và đảm bảo tính nhất quán trong quá trình soạn thảo văn bản hành chính.

Nội dung 3: Triển khai hệ thống trên dịch vụ đám mây

- **Mục tiêu:** Triển khai hệ thống trên dịch vụ đám mây nhằm giữ hệ thống luôn sẵn sàng. Hệ thống phải chịu tải cao, chống sập, có khả năng hồi phục sau sự cố,… Đồng thời cung cấp cho người dùng một giao diện để có tương tác dễ dàng hơn.
- **Phương pháp:** Sử dụng dịch vụ điện toán đám mây của AWS để triển khai hệ thống, trong đó máy chủ được triển khai trên nền tảng Amazon EC2 (Elastic Compute Cloud) trong môi trường Virtual Private Cloud (VPC). Giải pháp này giúp đảm bảo tính linh hoạt, khả năng mở rộng và bảo mật cho hệ thống.

Nội dung 4: Đánh giá hệ thống

- **Mục tiêu:** Đánh giá nhằm chứng minh mức độ hiệu quả mà hệ thống mang lại. Sử dụng các chỉ số rõ ràng nhằm tăng mức độ tin tưởng của người dùng.
- **Phương pháp:** Sử dụng các chỉ số như precision, recall, accuracy, F1 score, false positive rate, false negative rate,...

KẾT QUẢ MONG ĐỢI

- Nắm được phần cốt lõi và có thể áp dụng OCR vào hệ thống một cách thuần thục.
- Kết hợp được OCR và LLM nhằm tạo ra một ứng dụng có thể phân tích văn bản hành

chính

- Ứng dụng có thể sử dụng trong thực tế, giúp rút ngắn thời gian soạn thảo, giảm thiểu lỗi xuống mức tối đa.
- Ứng dụng có thể được ứng dụng linh hoạt tuỳ vào bộ hướng dẫn cụ thể của từng tổ chức.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] R. Liu and N. B. Shah, "ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing," arXiv.org, 2023. https://arxiv.org/abs/2306.00622

[2] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv:2004.05150 [cs]*, Dec. 2020, Available: https://arxiv.org/abs/2004.05150

[3] G. Kim *et al.*, "OCR-Free Document Understanding Transformer," *Lecture Notes in Computer Science*, pp. 498–517, Jan. 2022, doi: https://doi.org/10.1007/978-3-031-19815-1_29.

[4] J. Wu, Z. Wu, R. Li, A. Hasan, Y. Kim, J. P. Y. Cheung, T. Zhang, and H. Wu, "Integrating Knowledge Retrieval and Large Language Models for Clinical Report Correction," *Arxiv.org*, 2023. https://arxiv.org/html/2406.15045v2 (accessed Feb. 27, 2025).

[5] C. Amrhein and S. Clematide, "Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods," *Journal for Language Technology and Computational Linguistics (JLCL)*, vol. 33, no. 1, pp. 49–76, 2018, doi: https://www.zora.uzh.ch/id/eprint/162394/1/AmrheinClematide2018.pdf.

[6] J. Evershed and K. Fitch, "Correcting noisy OCR," May 2014, doi: https://doi.org/10.1145/2595188.2595200.