Các đề tài khác

K1. Chuyển tự Hán-Nôm sang chữ Quốc ngữ cho các văn bản Y học cổ truyền (2 thành viên)

K2. Chuyển tự chữ Quốc ngữ sang chữ Hán-Nôm (2 thành viên)

K3. Nhận diện thực thể có trong các câu đối Hán-Nôm (2 thành viên)

K4. Nhận diện thực thể có trong các văn bản cổ Hán-Nôm (2 thành viên)

K5. Gán dấu câu (punctuation) cho các câu đối Hán-Nôm (2 thành viên)

K6. Gán dấu câu (punctuation) cho các văn bản cổ Hán-Nôm (2 thành viên)

K7. Giải nghĩa (interpretation) các câu đối chữ Hán ở VN bằng chữ Quốc ngữ (3 thành viên)

K8. Giải nghĩa (interpretation) văn bản cổ chữ Hán ở VN bằng chữ Quốc ngữ (3 thành viên)

K9. Tìm hiểu về mạng ngữ nghĩa từ vựng WordNet: cách tổ chức, cách truy cập, tra cứu nghĩa qua các hàm thư viện có sẵn của WordNet (1 thành viên)

K10. Tìm hiểu các công cụ duyệt (browse) cây ngữ nghĩa WordNet (1 thành viên)

K11. Tìm hiểu công cụ cho phép chỉnh sửa cây ngữ nghĩa WordNet (1 thành viên)

K12. Tìm hiểu về các mạng ngữ nghĩa từ vựng của tiếng Việt: đánh giá (1 thành viên)

K13. Tìm hiểu về hệ thống mạng ngữ nghĩa từ vựng toàn cầu GlobalNet (1 thành viên)

K14. Tìm hiểu công cụ cho phép xây dựng cây ngữ nghĩa WordNet cho một ngôn ngữ khác (vd: tiếng Việt) (1 thành viên)

K15. Tìm hiểu các hàm đo khoảng cách ngữ nghĩa giữa 2 synset ID của WordNet (2 thành viên)

K16. Xây dựng từ điển Anh-Anh-Việt có nhãn synset_ID dựa trên CSDL WordNet và từ điển Anh-Việt (2 thành viên)

K17. Tìm hiểu hệ thống phân tích ngữ nghĩa USAS (1 thành viên)

K18. Xây dựng công cụ gán nhãn ngữ nghĩa tiếng Việt trên UCREL (2 thành viên)

K19. Xây dựng từ điển Anh-Việt có nhãn UCREL dựa trên CSDL UCREL và từ điển Anh-Việt (2 thành viên)

K20. Khai thác LLMs để gán nhãn tự động cho văn bản Lịch sử Việt Nam

- Mục tiêu: Đề tài này nhằm mục đích giúp học viên hiểu được cách hoạt động của LLMs, qua đó sử dụng các kỹ thuật Prompting để khai thác khả năng của LLMs trên các tác vụ NLP.
- Yêu cầu:
  + Chuẩn bị gold dataset: Gold dataset cần được gán nhãn thủ công tối thiểu 500 samples.
  + Tìm hiểu và áp dụng các kỹ thuật Prompting dựa trên các đặc thù của văn bản đang được xử lý để LLMs (tối thiểu 2 models) gán nhãn tự động trên toàn bộ dataset.
  + Phân tích và so sánh kết quả (các độ đo) gán nhãn của LLMs với bộ gold dataset.
- Kết quả đầu ra:
  + Bộ dữ liệu đã được gán nhãn.
  + Một tập tin báo cáo bao gồm:
    ▪ Mô tả bài toán được chọn: định nghĩa, phạm vi, phương pháp, đánh giá.
    ▪ Quy trình gán nhãn thủ công: các công cụ được sử dụng và cách thức thực hiện.
    ▪ Quy trình gán nhãn bằng LLMs: các prompt được sử dụng
    ▪ Bảng kết quả
    ▪ Phân tích khó khăn và cách giải quyết trong quá trình thực hiện.
    ▪ Bài học kinh nghiệm
  + Source code
  + Tài liệu tham khảo
  + Số thành viên: 2

K21. Khảo sát các tác vụ NLP trong lĩnh vực Lịch sử Việt Nam

- Mục tiêu: Đề tài này nhằm mục đích giúp học viên có được một sự phân tích và đánh giá hiệu suất của mô hình ngôn ngữ tiên tiến hiện nay trên các tác vụ NLP trong lĩnh vực Lịch sử Việt Nam.
- Yêu cầu:
  + Chuẩn bị gold dataset: Gold dataset cần được gán nhãn thủ công tối thiểu 500 samples.
  + Tìm hiểu và chọn ra tối thiểu 3 models để chạy thực nghiệm trên tác vụ được chọn.

+ So sánh và đánh giá (thông qua độ đo) hiệu suất của các models so với baseline.

- Kết quả đầu ra:
  + Một tập tin báo cáo bao gồm:
    - Mô tả bài toán được chọn: định nghĩa, phạm vi, phương pháp, đánh giá.
    - Quy trình gán nhãn thủ công: các công cụ được sử dụng và cách thức thực hiện.
    - Mô tả models: phương pháp, cách thức hoạt động.
    - Bảng kết quả và phân tích ưu nhược điểm của mỗi model.
    - Phân tích khó khăn và cách giải quyết trong quá trình thực hiện.
    - Bài học kinh nghiệm
  + Source code
  + Tài liệu tham khảo
  + Số thành viên: 2

K22. Khảo sát các công cụ OCR cho dữ liệu chữ Hán-Nôm trong ảnh ngoại cảnh ở Việt Nam.

- Mục tiêu: Đề tài này nhằm mục đích giúp học viên tìm hiểu, phân tích và đánh giá hiệu suất của các mô hình OCR trên dữ liệu chữ Hán – Nôm ngoại cảnh ở Việt Nam.
- Yêu cầu:
  + Chuẩn bị gold dataset: Gold dataset cần được gán nhãn thủ công tối thiểu 300 samples.
  + Tìm hiểu và chọn ra tối thiểu 3 models để chạy thực nghiệm trên tác vụ được chọn.
  + So sánh và đánh giá (thông qua độ đo) hiệu suất của các models so với baseline.
- Kết quả đầu ra:
  + Một tập tin báo cáo bao gồm:
    - Mô tả bài toán được chọn: định nghĩa, phạm vi, phương pháp, đánh giá.
    - Quy trình gán nhãn thủ công: các công cụ được sử dụng và cách thức thực hiện.
    - Mô tả models: phương pháp, cách thức hoạt động.
    - Bảng kết quả và phân tích ưu nhược điểm của mỗi model.
    - Phân tích khó khăn và cách giải quyết trong quá trình thực hiện.
    - Bài học kinh nghiệm
  + Source code
  + Tài liệu tham khảo

+ Số thành viên: 2

K23. Xây dựng công cụ gán nhãn bán tự động cho các ảnh Hán Nôm ngoại cảnh:

- Mục tiêu: Đề tài này nhằm mục đích giúp học viên tìm hiểu quy trình xây dựng bộ dữ liệu OCR từ dữ liệu thô.
- Yêu cầu:
  + Tìm hiểu và lựa chọn 1 trong các công cụ/API có khả năng OCR cho chữ Hán-Nôm ngoại cảnh (gợi ý: PPOCR, Google Vision API, ChatGPT,..).
  + Chuẩn bị dữ liệu cần gán nhãn: tìm hiểu và thu thập dữ liệu từ các nguồn ineternet đáng tin cậy.
  + Xây dựng quy trình để gán nhãn tự động sử dụng công cụ/API đã lựa chọn.
  + Đề xuất phương pháp để kiểm tra tính đúng của bộ dữ liệu.
- Kết quả đầu ra:
  + Một tập tin báo cáo bao gồm:
    - Một tập dữ liệu (tối thiểu 5000 sample) rõ ràng và được gáng nhãn tương ứng bằng file .txt hoặc .json.
    - Tài liệu toàn diện trình bày chi tiết về quy trình tạo tập dữ liệu, nguyên tắc chú thích và mọi bước tiền xử lý đã thực hiện.
    - Ví dụ: Tệp README giải thích cách thu thập, chú thích và xử lý trước dữ liệu.
  + Source code
  + Tài liệu tham khảo.
  + Số thành viên: 3

K24. Xây dựng công cụ tạo sinh dữ liệu cho ảnh Hán – Nôm ngoại cảnh:

- Mục tiêu: Đề tài này nhằm mục đích giúp học viên tìm hiểu và triển khai các công cụ tạo sinh dữ liệu cho bài toán OCR ảnh ngoại cảnh Hán - Nôm
- Yêu cầu:
  + Tìm hiểu và mô tả về dữ liệu chữ Hán – Nôm ngoại cảnh.
  + Tìm hiểu và lựa chọn các phương pháp tạo sinh dữ liệu cho bài toán OCR trên chữ Hán – Nôm: tạo dữ liệu đa dạng hơn về màu sắc, font chữ, background,...
- Kết quả đầu ra:
  + Một tập tin báo cáo bao gồm:

        - Tài liệu toàn diện trình bày chi tiết về quy trình tạo tập dữ liệu, nguyên tắc chú thích và mọi bước tiền xử lý đã thực hiện.
        - Kết quả thực hiện
    + Source code
    + Tài liệu tham khảo.
    + Số thành viên: 2

K25. Xây dựng Mô hình NLP cho Văn bản Y học Tiếng Việt

- Mục tiêu: Xây dựng và huấn luyện mô hình cho văn bản y học Tiếng Việt đối với một trong các bài toán NLP sau đây:
    + Text Summarization
    + Question Answering
    + Information retrieval
    + Named entity recognition
    + Relation extraction
    + Text classification/Document Classification
    + Document Ranking
    + Topic modelling
    + Keyword Extraction
    + Machine translation
    + Hoặc các bài toán khác trong lĩnh vực NLP
- Yêu cầu cụ thể:
    + Thu thập Dữ liệu: Sử dụng dữ liệu y học Tiếng Việt từ các nguồn thu thập được.
    + Xây dựng Mô hình: Huấn luyện mô hình cho bài toán NLP đã chọn trên bộ dữ liệu đã thu thập .
    + Đánh giá Mô hình: Đánh giá hiệu suất mô hình trên tập dữ liệu kiểm tra với các độ đo cụ thể phù hợp cho bài toán đã chọn.
- Kết quả đầu ra:
    + Một mô hình NLP y học cho văn bản Tiếng Việt.
    + Báo cáo chi tiết về quá trình xây dựng, huấn luyện và đánh giá mô hình.
- Số thành viên: 2

K26. Xây dựng ứng dụng xác định hình ảnh chứa văn bản Hán - Nôm

Mô tả: Xây dựng ứng dụng xác định hình ảnh có chứa văn bản hoặc chữ Hán – Nôm hay không.

Input: Một hình ảnh

Output: Trả về giá trị 0 và 1 tương ứng với [Không chứa chữ Hán Nôm] và [Có chứa chữ Hán – Nôm]

Số thành viên: 1

K27. Xây dựng ứng dụng Phân loại hình ảnh Hán - Nôm

Mô tả: Trong văn bản Hán – Nôm hiện tại đang chia thành 3 loại cơ bản:

1. Văn bản thông thường
2. Văn bản hành chính
3. Văn bản ngoại cảnh

Xây dựng ứng dụng xác định loại của hình ảnh văn bản Hán - Nôm .

Input: Một hình ảnh

Output: Trả về giá trị 0, 1, 2 tương ứng với 3 loại văn bản đã mô tả.

Số thành viên: 1

K28. Xây dựng ứng dụng xác định chiều của văn văn bản Hán - Nôm

Mô tả: Trong văn bản Hán – Nôm hiện tại, cách viết đang chia thành 2 loại cơ bản:

1. Viết theo chiều ngang (Từ trái qua phải)
2. Viết theo chiều dọc (Từ trên xuống dưới)

Xây dựng ứng dụng xác định cách viết của văn bản Hán - Nôm .

Input: Một hình ảnh

Output: Trả về giá trị 0, 1 tương ứng với 2 cách viết đã mô tả

Số thành viên: 1

K29. Xây dựng ứng dụng phân loại văn băn Hán - Nôm

Mô tả: Trong văn bản Hán – Nôm hiện tại đang chia thành 3 loại cơ bản:

1. Văn bản thông thường
2. Văn bản hành chính
3. Văn bản ngoại cảnh

Với mỗi loại ở trên thì văn phong, cú pháp để viết sẽ khác nhau, ví dụ:

1. Văn bản thông thường: có Truyện Kiều, Truyện Lục Vân Tiên
2. Văn bản hành chính: có Sắc Phong, Chiếu Chỉ
3. Văn bản ngoại cảnh: có câu đối ở đình chùa

Xây dựng ứng dụng xác định loại của văn bản Hán - Nôm.

Input: Một đoạn văn bản Hán - Nôm

Output: Trả về giá trị 0, 1, 2 tương ứng với 3 loại văn bản đã mô tả.

Số thành viên: 1

K30. Xử lý vấn đề Out-of-vocabulary cho bài toán dịch nghĩa Hán văn sang tiếng Việt hiện đại.

- Mục tiêu: Nhằm để tăng chất lượng đầu ra của mô hình dịch nghĩa.
- Tài liệu tham khảo:
  + http://tcci.ccf.org.cn/conference/2019/papers/35.pdf
  + https://www2.statmt.org/moses/?n=Advanced.OOVs
- Kết quả đầu ra: Một chương trình có thể ứng dụng vào mô hình dịch tự động Ancient Chinese – Vietnamese với đầu ra có số lượng oov giảm so với mô hình cơ bản chưa xử lý OOV
- Số thành viên: 2

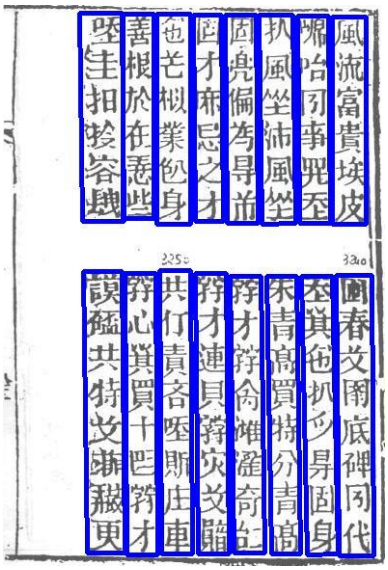
K31. Xây dựng ứng dụng OCR Detection và Recognition văn bản Hán - Nôm dựa trên Document AI của Google.

- **Mô tả:** Khai thác API Document AI của Google. Đây là một API cung cấp nhiều chức năng đọc văn bản đa ngôn ngữ. Có thể khai thác để OCR cho văn bản Hán - Nôm.
- **Input:** Một hình ảnh chứa văn bản Hán - Nôm
- **Output:** File results.txt chứa tọa độ bounding box và annotation của bounding box mà ứng dụng trả về. Ví dụ nội dung file results.txt: **[[588, 75], [627, 75], [627, 392], [588, 392]] 思及於其先惟乃家葉茂聲宏錫命盍推**
- Link demo từ Document AI: https://cloud.google.com/document-ai/docs/try-docai
- Số thành viên: 3

K32. Xây dựng ứng dụng tính toán độ đo cho giai đoạn Detection văn bản Hán - Nôm

- **Mô tả:** Trong giai đoạn Detection để đánh giá khả năng phát hiện văn bản của model so với labels ban đầu đã được gán nhãn, chúng ta sẽ dựa vào ba độ đo: Recall, Precision, H-mean (F1 - score). Hãy tìm hiểu thuật toán các độ đo, viết ứng dụng tính toán ba độ đo trên.

- **Input:** Label.txt, Test.txt. Trong đó Label.txt chứa tọa độ các bounding box đúng của một hình ảnh văn bản Hán - Nôm đã được gán nhãn (ground truth). Tests.txt chứa tọa độ các bounding box mà model trả ra (tọa độ bounding box Test.txt model trả ra Học viên có thể tự mô phỏng)
- Số thành viên: 1



Hình mô tả các bounding box

- Ví dụ nội dung:

  Label.txt

  [

  [[337, 9], [380, 9], [380, 233], [337, 233]],

  [[342, 276], [386, 276], [386, 576], [342, 576]],

  [[301, 7], [342, 7], [342, 230], [301, 230]],

  [[305, 276], [350, 276], [350, 574], [305, 574]],

  [[263, 6], [308, 6], [308, 230], [263, 230]],

  [[268, 273], [313, 273], [313, 575], [268, 575]],

  [[226, 6], [270, 6], [270, 230], [226, 230]],

  [[230, 274], [274, 274], [274, 575], [230, 575]],

  [[192, 5], [233, 5], [233, 231], [192, 231]],

  [[194, 273], [238, 273], [238, 573], [194, 573]],

  [[154, 4], [196, 4], [196, 231], [154, 231]],

  [[157, 274], [200, 274], [200, 575], [157, 575]],

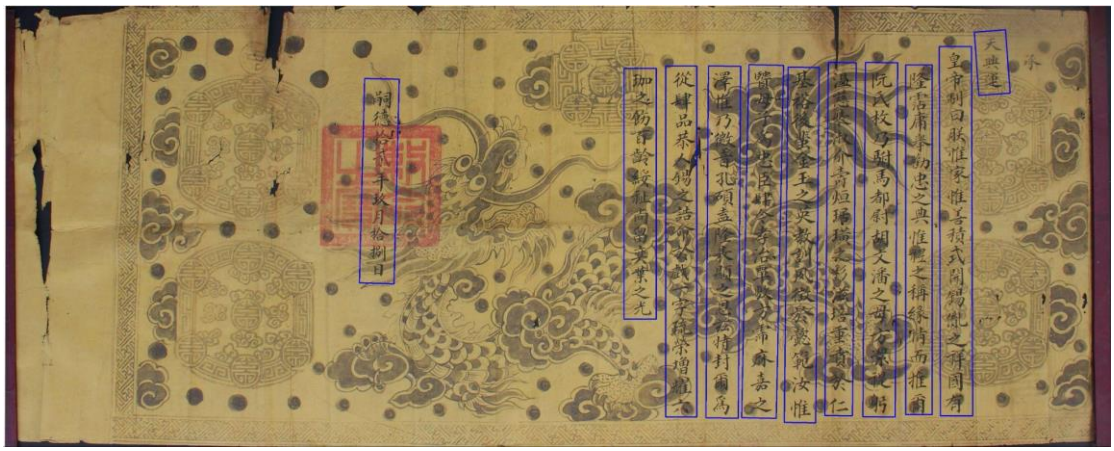
[[120, 4], [161, 4], [161, 227], [120, 227]],
[[122, 272], [165, 272], [165, 574], [122, 574]],
[[78, 6], [127, 6], [127, 226], [78, 226]],
[[79, 269], [130, 269], [130, 574], [79, 574]]
]
Test.txt
[
[[345, 280], [382, 280], [384, 571], [348, 571]],
[[341, 12], [376, 12], [376, 227], [341, 227]],
[[309, 281], [341, 281], [347, 569], [315, 570]],
[[305, 11], [337, 11], [339, 226], [307, 226]],
[[273, 281], [304, 281], [308, 569], [278, 569]],
[[265, 10], [300, 10], [303, 226], [268, 227]],
[[235, 281], [267, 281], [271, 568], [239, 568]],
[[230, 10], [262, 10], [264, 226], [232, 226]],
[[199, 278], [231, 278], [234, 569], [202, 569]],
[[197, 10], [226, 10], [228, 224], [199, 224]],
[[162, 275], [194, 275], [196, 569], [164, 569]],
[[159, 9], [194, 9], [194, 227], [159, 227]],
[[126, 278], [158, 278], [160, 570], [128, 570]],
[[83, 276], [123, 276], [126, 568], [87, 568]],
[[79, 8], [123, 8], [125, 224], [82, 224]]
]

- **Output:** Kết quả ba độ đo Recall, Precision, H-mean (F1 - score)
- Tham khảo: https://www.linkedin.com/pulse/precision-recall-f1-score-object-detection-back-ml-basics-felix/

K33. Xây dựng ứng dụng tính toán độ đo cho giai đoạn Recognition văn bản Hán - Nôm

- **Mô tả:** Trong giai đoạn Recognition để đánh giá độ chính xác khả năng recognize của model so với ground truth, chúng ta sẽ dựa vào ba độ đo: Accuracy. Hãy tìm hiểu thuật toán các độ đo, viết ứng dụng tính toán độ đo trên.

- **Input:** Label.txt, Test.txt. Trong đó Label.txt chứa nội dung của văn bản đúng của một hình ảnh văn bản Hán - Nôm đã được gán nhãn. Test.txt chứa nội dung văn bản mà model trả ra (nội dung Test.txt model trả ra Học viên có thể tự mô phỏng)
- Ví dụ nội dung:



Hình ví dụ

Label.txt

[[[[4144, 140], [4233, 135], [4245, 359], [4157, 363]], '天興運'],

[[[3994, 201], [4077, 200], [4082, 1702], [3999, 1702]], '皇帝制日朕惟家惟善積式開錫之國有'],

[[[2834, 264], [2927, 265], [2917, 1732], [2824, 1731]], '從肆品今錫之誥命欽家茲榮增雄上'],

[[[3005, 274], [3094, 274], [3089, 1731], [3000, 1731]], '澤惟乃徽孔頒盍隆之公特封爾'],

[[[3167, 274], [3265, 274], [3265, 1722], [3167, 1722]], '今肆今治平欽之庥嘉之'],

[[[3676, 274], [3774, 275], [3759, 1722], [3661, 1721]], '阮民枚乃馬都尉胡文潘之母欽'],

[[[3338, 288], [3436, 289], [3426, 1741], [3328, 1741]], '後金王之教風欽懿範汝'],

[[[3838, 279], [3921, 279], [3921, 1726], [3838, 1726]], '隆庸舉勸忠之典惟之稱緣情推爾'],

[[[2658, 293], [2741, 293], [2741, 1306], [2658, 1306]], '加之飭百論祉尚之'],

[[[3505, 308], [3588, 308], [3588, 1717], [3505, 1717]], '展命頒之彩重賁於一'],

[[[1552, 337], [1670, 339], [1654, 1165], [1536, 1162]], '嗣德拾貳年玖月拾捌日']]

Test.txt

[[[[4119, 115], [4251, 105], [4271, 374], [4138, 384]], '天興運'],

[[[3970, 176], [4107, 176], [4112, 1746], [3975, 1746]], '皇帝制曰朕惟家惟善積式開錫胤之祥國有'],

[[[3828, 254], [3940, 254], [3940, 1741], [3828, 1741]], '隆庸舉勸忠之典惟禮之稱緣情而推爾'],

[[[3657, 244], [3794, 245], [3783, 1751], [3646, 1750]], '阮民枚乃駙馬都尉胡文潘之母方潔禔躬'],

[[[3500, 244], [3618, 245], [3602, 1746], [3485, 1745]], '渥慈朱命烜之彩欽培重於化'],

[[[3329, 264], [3461, 265], [3446, 1776], [3313, 1774]], '基拾後金玉之典赦封民欽登範汝惟'],

[[[3163, 253], [3310, 256], [3279, 1757], [3132, 1754]], '贊肆今尊忠臣肆今孝拾覃敕方節麻嘉之'],

[[[2991, 259], [3128, 260], [3118, 1756], [2981, 1755]], '澤惟乃徽寵碩盍隆表顯恩茲特封爾為'],

[[[2810, 259], [2942, 259], [2942, 1751], [2810, 1751]], '從肆品恭錫之誥命於戲今宋疏榮增耀之'],

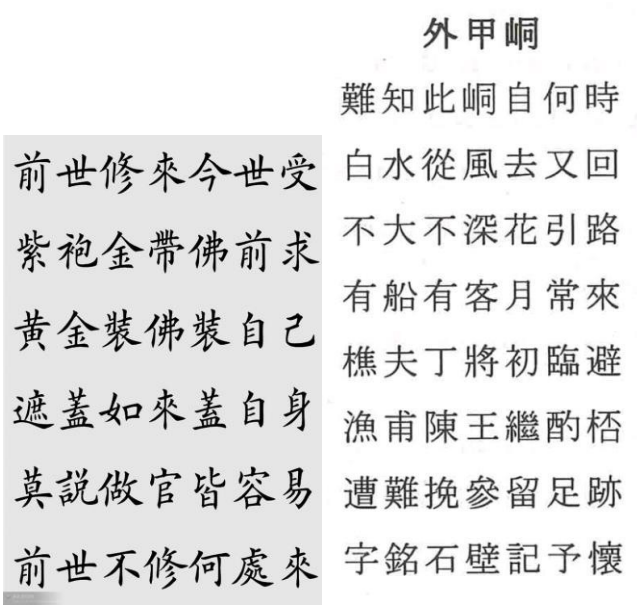
[[[2633, 269], [2770, 269], [2770, 1335], [2633, 1335]], '珈之飾百齡綏祉尚留奕葉之光'],

[[[1533, 312], [1680, 315], [1658, 1185], [1511, 1181]], '嗣德拾貳年玖月拾捌日']]

- **Output:** % Accuracy được tính toán giữ labels và test
- **Tham khảo:** https://www.docuclipper.com/blog/ocr-accuracy/
- Số thành viên: 1

K34. Phát triển tool tự động thu thập văn bản Hán - Nôm hoặc văn bản Trung Quốc dưới dạng hình ảnh từ internet. Nội dung văn bản thu thập được viết theo phương ngang như cách viết hiện đại.

Ví dụ một số hình ảnh:



Output: bộ dữ liệu hình ảnh văn bản Hán - Nôm viết theo phương ngang.

Số thành viên: 1