Hướng dẫn sử dụng tool đánh giá API văn bản hành chính

1. Chuẩn bị môi trường

Công cụ được viết bằng ngôn ngữ lập trình Python phiên bản >= 3.6

Cài đặt các thư viện:

```
$ cd voffice-eval
$ pip install -r requirements.txt
```

2. Phương pháp đánh giá

Đề xuất sử dụng 2 phương pháp đánh giá: (1) Độ chính xác theo ký tự AOC đối với các trường thông tin là chuỗi ký tự dài, (2) Độ chính xác phận loại ACC đối với các trường thông tin có tập giá trị là giới hạn.

5 trường thông tin được đánh giá theo độ chính xác theo ký tự gồm: Trích yếu nội dung văn bản, số ký hiệu văn bản, cơ quan gửi, người ký, ngày văn bản.

Giả sử chuỗi ký tự đúng là X, chuỗi ký tự do AI dự đoán là Y. Công thức tính độ chính xác theo ký tự được định nghĩa như sau:

`AOC(X, Y) = 2 * (Số ký tự giống nhau của X và Y) / (Số ký tự của X + Số ký tự của Y)`

2 trường thông tin được đánh giá theo độ chính xác phân lớp: Hình thức văn bản, Độ khẩn.

Công thức tính độ chính xác phân lớp được định nghĩa như sau:

`ACC = Số mẫu dự đoán đúng / Tổng số mẫu`

3. Chương trình đánh giá

Chương trình đánh giá nhận đầu vào gồm các dữ liệu sau:

- Thư mục chứa các ảnh test
- File nhãn chuẩn (gọi là GroundTruth) là các giá trị đúng của các trường thông tin cần trích xuất được gán nhãn bởi con người (tham khảo file "gt.csv" trong thư mục ***voffice-eval***)

Chương trình đánh giá sẽ gửi các ảnh trong thư mục test đến API để lấy các kết quả dự đoán của AI và so sánh với giá trị trong GroundTruth.

Để sử dụng chương trình đánh giá, chạy câu lệnh sau:

```
$ python voffice-eval-tool.py --gt_file path-to-groundtruth-file --img_path
path-to-testset
```

Trong đó:

- path-to-groundtruth-file: File nhãn chuẩn

- path-to-testset: Thư mục chứa các ảnh test

Ví dụ:

```
$ python voffice-eval-tool.py --gt_file gt.csv --img_path testset
```

Lưu ý: tệp gt.csv và thư mục testset được gửi đính kèm trong mail.

4. Kết quả

Sau khi chạy chương trình đánh giá, một tệp kết quả có tên như sau được tạo ra "**voffice-report-YYYYmmdd-HHMM.csv**"

Tệp kết quả gồm các cột sau:

- 'VBHC_PATH1', 'VBHC_PATH2', 'VBHC_PATH3', 'VBHC_PATH4': đường dẫn tới các file ảnh của một văn bản
- 'TYPE_PRED', 'TYPE_GT', 'TYPE_ACC': kết quả dự đoán, nhãn đúng và ACC của trường Hình thức văn bản
- 'TITLE_PRED', 'TITLE_GT', 'TITLE_AOC': kết quả dự đoán, nhãn đúng và AOC của trường Trích yếu nội dung văn bản
- 'CODE_PRED', 'CODE_GT', 'CODE_AOC': kết quả dự đoán, nhãn đúng và AOC của trường Số ký hiệu văn bản
- 'OFFICESENDER_PRED', 'OFFICESENDER_GT', 'OFFICESENDER_AOC': kết quả dự đoán, nhãn đúng và AOC của trường Cơ quan gửi
- 'PRIORITY_PRED','PRIORITY_GT', 'PRIORITY_ACC': kết quả dự đoán, nhãn đúng và ACC của trường Độ khẩn
- 'SIGNER_PRED', 'SIGNER_GT', 'SIGNER_AOC': kết quả dự đoán, nhãn đúng và AOC của trường Người ký
- 'DOCUMENTDATE_PRED', 'DOCUMENTDATE_GT', 'DOCUMENTDATE_AOC': kết quả dự đoán, nhãn đúng và AOC của trường Ngày văn bản
- JSON_RESULT: kết quả trả về của API
- TIME: thời gian API xử lý request

Thực hiện tính trung bình AOC/ACC của từng trường ta được kết quả cuối cùng.

5. Hướng dẫn mở tệp csv

File GroundTruth và File kết quả được lưu dưới định dạnh csv vì vậy khi mở trực tiếp bằng excel sẽ bị lỗi.

Để không bị lỗi xin làm theo hướng dẫn sau:

https://itz.vn/blog/cach-mo-file-csv-trong-excel-khong-bi-loi-font-tieng-viet-a71.html

1. Mở Excel, chuyển sang tab Data
2. Chọn "Import data from text file"
3. Trong màn hình xuất hiện, chọn cách ngăn chia file là "Delimited", còn bảng mã chọn "UTF-8". Nhớ là phải chọn UTF-8 chứ không lỗi sẽ không hết đâu
4. Trong màn hình kết tiếp, bỏ chọn ô Tab và chọn vào ô Comma. Ở đây chúng ta đang nói cho Excel biết rằng các trường trong file được ngăn cách bằng dấu phẩy, Excel cứ theo đó mà tách cột
5. Nhấn Next, chọn ô để hiển thị dữ liệu, có thể để mặc định cũng được, nhấn OK thêm lần nữa là xong.