ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Báo cáo bài tập 4

Nhận diện thực thể tên (NER)

Conditional random field (CRF)

GV hướng dẫn: TS. Nguyễn Thị Quý

Họ và tên	*MSSV*	*Mã lớp*
Trần Hoàng Bảo Ly	21521109	CS231.N21.KHTN
Lê Thanh Minh	21520063	CS231.N21.KHTN
Nguyễn Quốc Trường	21521604	CS231.N21.KHTN
Lê Thu Hà	21520800	CS231.N21.KHTN
Trần Xuân Minh	21520352	CS231.N21.KHTN

Hồ Chí Minh, tháng 5 năm 2023

Mục lục

1 Giới thiệu bài toán

1.1 Giới thiệu chung

Nhận dạng thực thể có tên (Named Entity Recognition – NER) nhằm nhận biết các chuỗi từ trong văn bản là tên của một đối tượng nào đó, điển hình như tên người, tên tổ chức, tên địa danh, thời gian v.v. NER là nhiệm vụ đóng vai trò quan trọng trong các ứng dụng trích xuất thông tin, đã được quan tâm nghiên cứu trên thế giới từ đầu những năm 1990. Từ năm 1995, hội thảo quốc tế chuyên đề Hiểu thông điệp (Message Understanding Conference - MUC) lần thứ 6 đã bắt đầu tổ chức đánh giá các hệ thống NER cho tiếng Anh. Tại hội thảo CoNLL năm 2002 và 2003, các hệ thống NER cho tiếng Hà Lan, Tây Ban Nha, Đức và Anh cũng được đánh giá. Trong các tác vụ đánh giá này, người ta xét 4 loại thực thể có tên: tên người, tên tổ chức, tên địa danh và các tên khác. Gần đây, vẫn tiếp tục có các cuộc thi về NER được tổ chức, ví dụ GermEval 2014 cho tiếng Đức.

1.2 Nhiệm vụ

đánh giá khả năng nhận dạng các thực thể có tên thuộc một trong ba loại: tên người, tên tổ chức và tên địa danh

1.3 Dữ liệu

Dữ liệu là các bài báo, đăng trên các phương tiện truyền thông xã hội, không phải dữ liệu nhân tạo (do người làm dữ liệu sinh ra).Trong đó, ba loại thực thể có tên được xác định tương thích với các loại thực thể mô tả trong CoNLL2003.

1. Tên địa lí (Địa danh - Location) bao gồm các thực thể có toạ độ địa lí nhất định, ghi lại được trên bản đồ:
    - Tên gọi các hành tinh: Mặt Trăng, Mặt Trời, Trái Đất...
    - Tên gọi các thực thể mang yếu tố địa lí tự nhiên và địa lí lịch sử (quốc gia, vùng lãnh thổ, châu lục), các vùng quần cư (làng, thị trấn, thành phố, tỉnh, giáo khu, giáo xứ), các điểm kinh tế (vùng nông nghiệp, khu công nghiệp)
    - Tên gọi các thực thể tự nhiên (đèo, núi, dãy núi, rừng, sông, suối, hồ, biển, vịnh, vũng, eo biển, đại dương, thung lũng, cao nguyên, đồng bằng, khu bảo tồn thiên nhiên, bãi biển, khu sinh thái, v.v.)
    - Tên gọi các thực thể là công trình xây dựng, công trình kiến trúc công cộng (cầu, đường, cảng, đập, lâu đài, tháp, quảng trường,

bảo tàng, phòng trưng bày, hội trường, trường học, nhà trẻ, thư viện, bệnh viện, viện dưỡng lão, trung tâm y tế, nhà thờ, nhà xứ, tu viện, nhà ở, chung cư, kí túc xá, chợ, công viên, nhà hát, rạp chiếu phim, khu thể thao, bể bơi, trung tâm thanh thiếu niên, khu cắm trại, doanh trại quân đội, nhà máy, sân bay, nhà ga, nhà kho, bãi đỗ xe, sân chơi, nghĩa trang, ...)

- Tên gọi địa điểm, địa chỉ thương mại (hiệu thuốc, quán rượu, nhà hàng, khách sạn, câu lạc bộ đêm, các địa điểm tổ chức âm nhạc, ...)
- Một số địa danh trừu tượng khác (Vườn Địa Đàng, Sông Ngân, Cầu Ô Thước...).

2. Tên tổ chức (Organization) bao gồm các loại tên sau:

   - Các cơ quan chính phủ (các bộ ngành, uỷ ban nhân dân, hội đồng nhân dân, toà án, cơ quan báo chí, hội nghề nghiệp, đoàn thể chính trị, phòng ban, ...)
   - Công ti (ngân hàng, thị trường chứng khoán, hãng phim, nhà sản xuất, hợp tác xã, phòng ban, ...)
   - Các thương hiệu
   - Các tổ chức chính trị (các đảng phái chính trị, các tổ chức khủng bố, ...)
   - Các ấn phẩm (các tạp chí, báo)
   - Các công ti âm nhạc (ban nhạc, dàn nhạc, đội hợp xướng ...)
   - Các tổ chức công cộng (trường học, tổ chức từ thiện)
   - Các tổ chức khác của con người (câu lạc bộ thể thao, các hiệp hội, nhà hát, công ti, tôn giáo, tổ chức thanh niên...)

3. Tên người (Person) bao hàm các loại tên sau:

   - Tên, tên đệm và họ của một người
   - Tên động vật và các nhân vật hư cấu
   - Các bí danh

Một số ví dụ về dữ liệu:

- Tên địa lí: Thành phố Hồ Chí Minh, Núi Bà Đen, Sông Bạch Đằng...
- Tên tổ chức: Công ty Formosa, Nhà máy thủy điện Hòa Bình...
- Tên người: tên riêng trong "ông Lân", "bà Hà"...

1.4 Nhãn thực thể

Nhãn thực thể được gán theo cấu trúc BIO như định dạng dữ liệu phân cụm CoNLL. Có 11 nhãn: B-PER và I-PER cho tên người, B-ORG và I-ORG cho tên tổ chức, B-LOC và I-LOC cho tên địa danh và O cho các phần tử khác.

WORD	thủ tướng	Nguyễn	Xuân	Phúc
Nhãn thực thể	O	B-PEOPLE	I-PEOPLE	I-PEOPLE

Bảng 1: Bảng đánh nhãn thực thể

Lúc này câu dữ liệu đầu vào có dạng là: thủ tướng <ENAMEX TYPE="PEOPLE">Nguyễn Xuân Phúc</ENAMEX>.

2 Quy trình thực hiện

2.1 Chuẩn bị dữ liệu

Dữ liệu được lấy từ tập dữ liệu NERVLSP2018, đã được chia thành 3 phần train, test và validation.

Các bước chuẩn bị dữ liệu để tiến hành train CRF model.

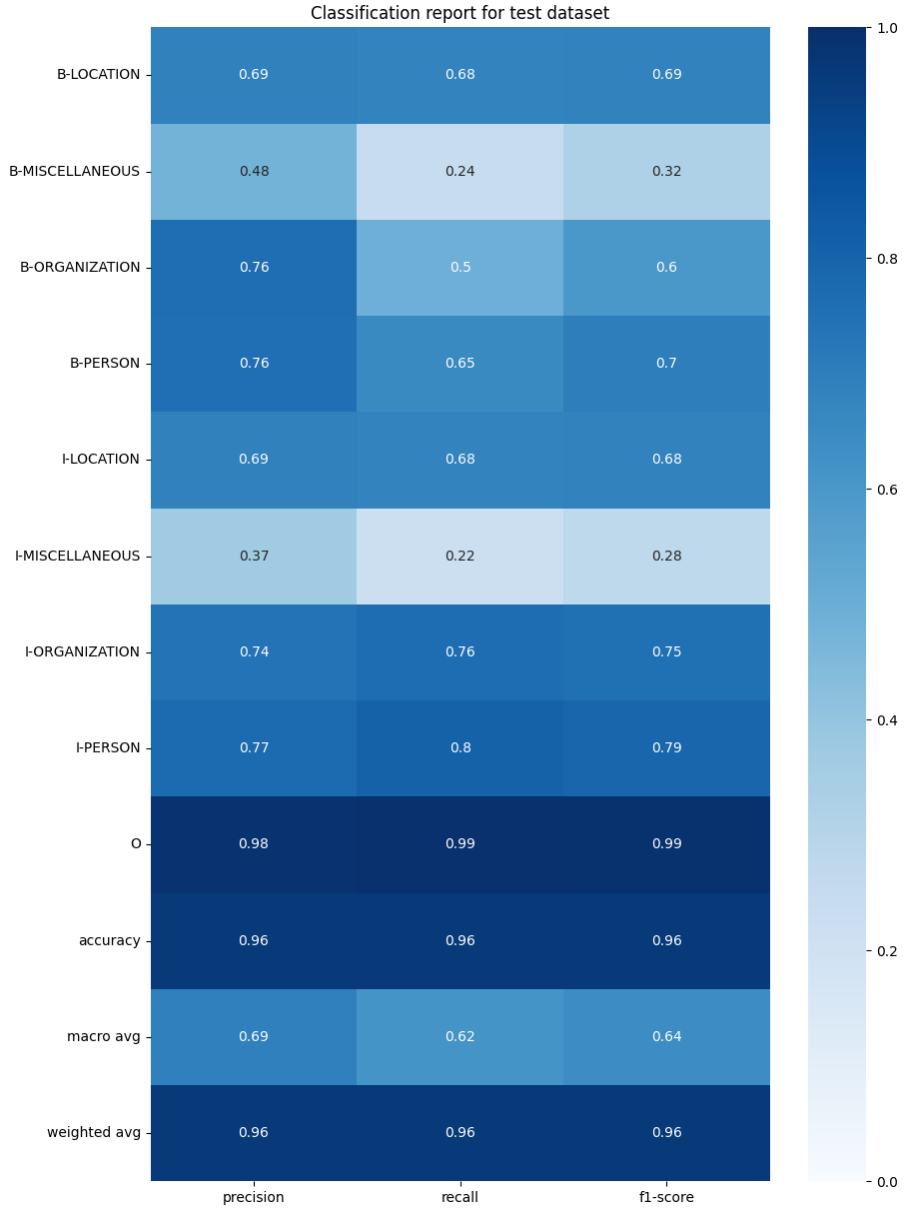
- Đọc dữ liệu từ các file riêng lẻ theo từng hàng.
- Tiến hành loại bỏ các cặp nhãn trong câu, đồng thời biểu diễn lại câu như Bảng 1 (bao gồm các bước tiền xử lý như: loại bỏ dấu câu, ghép các từ lại thành một từ cách nhau bằng dấu -).
- Sử dụng thư viên NLTK, tiến hành đánh post-tag chu từng từ.
- Tiến hành trích xuất các đặc trưng từ các từ bằng các đặc trưng như:
  - 'word.lower()': word.lower()
  - 'word[-3:]': word[-3:]
  - 'word[-2:]': word[-2:]
  - 'word.isupper()': word.isupper()
  - 'word.istitle()': word.istitle()
  - 'word.isdigit()': word.isdigit()
  - 'postag': postag

2.2 Train model

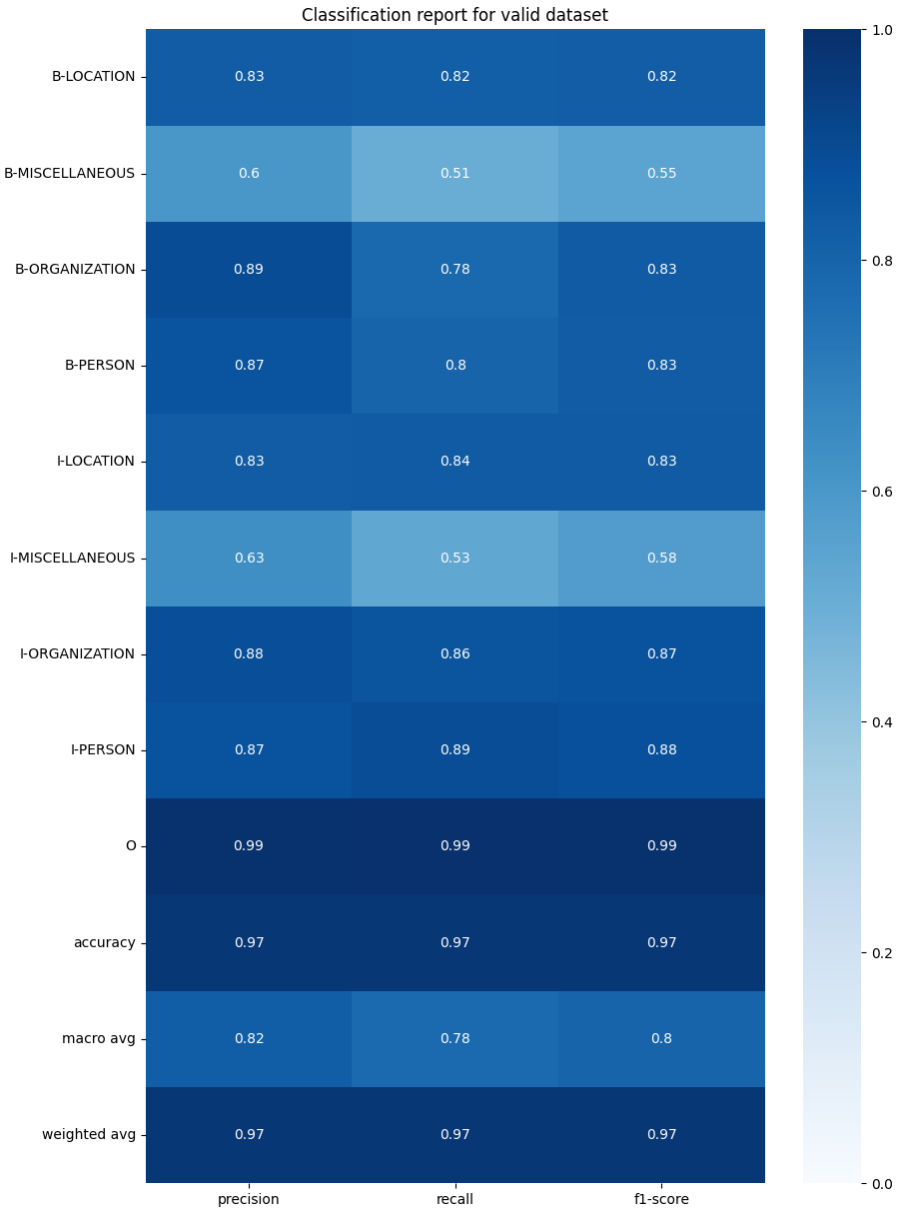
Sử dụng CRF(conditional random field) model từ thư viện pycrf-suite để tiến hành train model với các đặc trưng được trình bày ở trên.

3 Đánh giá hiệu suất thực hiện

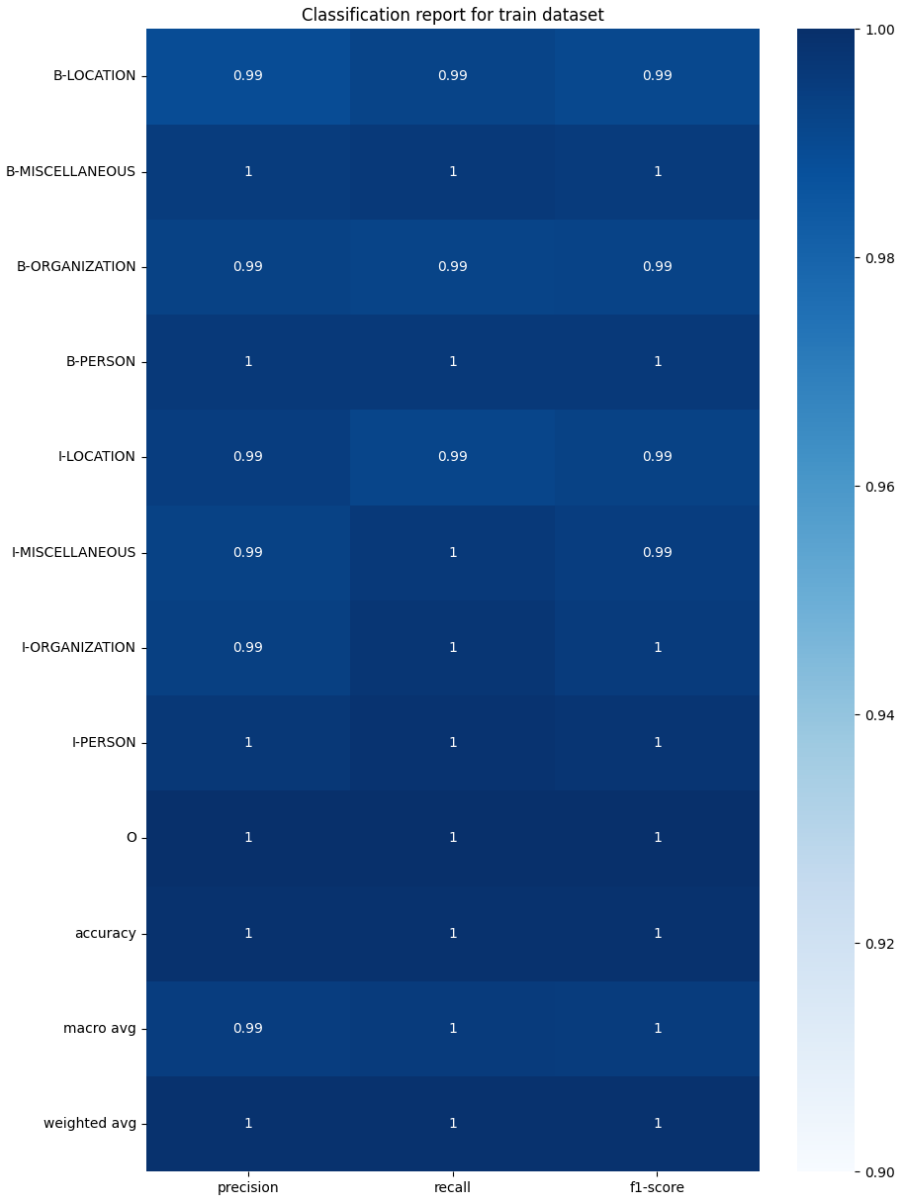
Kết quả đạt được được trình bày dưới dạng heatmap lần lượt trên các tập, test, validation và train:



Hình 1: Kết quả trên tập validation



Hình 2: Kết quả trên tập test



Hình 3: Kết quả trên tập train

Phần đánh giá hiệu suất thực hiện đã được bao gồm trong notebook để tiện theo dõi các bước và cách thức thực hiện cũng như kết quả thu được tương ứng.

Link đến github của assignment tại đây