THỰC HÀNH LAB 2

Named Entity Recognition (NER)

1. Quy định về việc nộp bài

- Thời gian và hình thức nộp: Trên Moodle.
- Bài nộp gồm: Các file cài đặt, file pdf tổng hợp kết quả, được bỏ vô chung một thư mục và nén lại thành một tập tin (.zip)
- Cách đặt tên file jupyter notebook: **NER_CâuX_Tên mô hình_MSSV.ipynb**

  (Với X là 1 hoặc 2, tương ứng với câu 1 và câu 2 trong bài thực hành)
- Một file pdf tổng hợp kết quả các mô hình trong câu 1, câu 2 và câu 3.
- Cách đặt tên bài nộp: **BTTH3_MSSV.zip**
- Lưu ý: Sai qui định thì sẽ nhận 0 điểm.

2. Nội dung thực hành

Câu 1: Sử dụng mô hình LSTM để giải quyết bài toán NER cho tiếng Anh.

- Sinh viên theo dõi bài hướng dẫn và cài đặt lại.
- Link tải bộ dataset: https://drive.google.com/drive/folders/1AGoS-p2Mf3v4NWWss_AwcU0T38cM2mjH?usp=sharing

  Thông tin ngắn gọn của bộ dữ liệu:
  - Số dòng dữ liệu: 1,048,575
  - Tổng số từ duy nhất: 35,179
  - Số thực thể: 9
  - Số lượng nhãn thực thể: 17

  ['B-per', 'I-per', 'B-tim', 'I-tim', 'B-art', 'I-art', 'B-geo', 'I-geo', 'B-eve', 'I-eve', 'B-nat', 'I-nat', 'B-org', 'I-org', 'B-gpe', 'I-gpe', 'O']

  Thông tin cơ bản của các thực thể:
  - geo = Thực thể địa lý (Geographical Entity)
  - org = Tổ chức (Organization)

- per = Người (Person)
- gpe = Thực thể địa chính trị (Geopolitical Entity)
- tim = Thời gian (Time indicator)
- art = Đồ tạo tác (Artifact)
- eve = Sự kiện (Event)
- nat = Hiện tượng tự nhiên (Natural Phenomenon)
- o = Khác (Other)

- Yêu cầu:
  - Thêm đánh giá mô hình theo độ đo F1-score (micro và macro)
  - Xuất ra kết quả của F1-score và Accuracy cho từng nhãn thực thể.
  - Lưu lại mô hình và bộ trọng số đã train.

Câu 2: Áp dụng mô hình trên để giải quyết bài toán NER cho tiếng Việt.

Dataset: COVID-19 Named Entity Recognition for Vietnamese

- Thông tin bộ dữ liệu và link tải: https://github.com/VinAIResearch/PhoNER_COVID19
- **Sử dụng bộ theo word.**
- Link paper: https://arxiv.org/abs/2104.03879
- Số thực thể: 10 thực thể.

Label	Definition
PATIENT_ID	Unique identifier of a COVID-19 patient in Vietnam. An PATIENT_ID annotation over "X" refers to as the X^{th} patient having COVID-19 in Vietnam.
PERSON_NAME	Name of a patient or person who comes into contact with a patient.
AGE	Age of a patient or person who comes into contact with a patient.
GENDER	Gender of a patient or person who comes into contact with a patient.
OCCUPATION	Job of a patient or person who comes into contact with a patient.
LOCATION	Locations/places that a patient was presented at.
ORGANIZATION	Organizations related to a patient, e.g. company, government organization, and the like, with structures and their own functions.
SYMPTOM&DISEASE	Symptoms that a patient experiences, and diseases that a patient had prior to COVID-19 or complications that usually appear in death reports.
TRANSPORTATION	Means of transportation that a patient used. Here, we only tag the specific identifier of vehicles, e.g. flight numbers and bus/car plates.
DATE	Any date that appears in the sentence.

Entity Type	Train	Valid.	Test	All
PATIENT_ID	3240	1276	2005	6521
PERSON_NAME	349	188	318	855
AGE	682	361	582	1625
GENDER	542	277	462	1281
OCCUPATION	205	132	173	510
LOCATION	5398	2737	4441	12576
ORGANIZATION	1137	551	771	2459
SYMPTOM&DISEASE	1439	766	1136	3341
TRANSPORTATION	226	87	193	506
DATE	2549	1103	1654	5306
# Entities in total	15767	7478	11735	34984
# Sentences in total	5027	2000	3000	10027

- Số lượng nhãn thực thể: 19

['B-PATIENT_ID', 'I-PATIENT_ID',

'B-NAME ', 'I-NAME ',

'B-AGE ',

'B-GENDER',

'B-JOB', 'I-JOB',

'B-LOCATION', 'I-LOCATION',

'B-ORGANIZATION', 'I-ORGANIZATION',

'B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE',

'B-TRANSPORTATION', 'I-TRANSPORTATION',

'B-DATE', 'I-DATE',

, 'O']

- Yêu cầu:
    - Đánh giá mô hình theo độ đo F1-score và Accuracy (micro và macro)
    - Xuất ra kết quả của F1-score và Accuracy cho từng nhãn thực thể.
    - Lưu lại mô hình và bộ trọng số đã train.

Câu 3: Cải thiện kết quả của câu 1, 2 bằng cách cài đặt thêm các mô hình khác.

- Cách đặt tên file jupyter notebook:

  NER_CâuX_Tên mô hình_MSSV.ipynb (Với X là 1 hoặc 2, tương ứng với câu 1 và câu 2 trong bài thực hành)

  VD: **NER_Câu1_BERT_MSSV.ipynb**

- Sinh viên có thể tham khảo và tự lựa chọn các mô hình muốn cái đặt có thể là các mô hình học chuyển tiếp (transfer learning), CNN, GRU, …
- Sinh viên chú ý nếu cần cài thêm thư viện nào thì phải có file **README.md** chú thích cho Giảng viên.