Named Entity Recognition

1. Quy trình xử lí

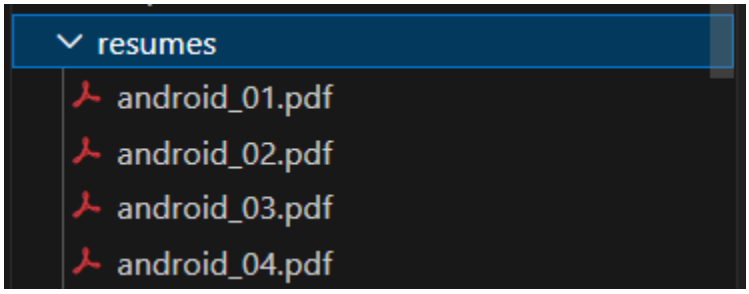
Quy trình giải quyết bài toán gồm các bước sau:

1. Chuyển văn bản dạng pdf sang văn bản thuần

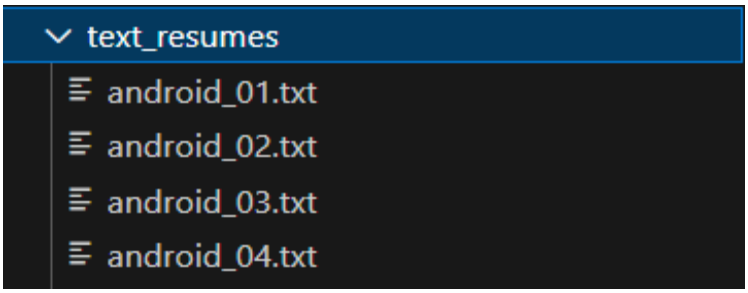
- Ở phần này, em sử dụng thư viện **pypdf** để đọc dữ liệu pdf.
- Bởi vì thông tin **NAME** và **EMAIL** đều nằm trong trang 1 trong tất cả các CV, do đó để giảm bớt quá trình gán nhãn và xử lí, em chỉ đọc dữ liệu pdf ở trang 1 và lưu thành các file txt tương ứng.

```python
def convert_pdf_to_text(file_path: str) -> None:
    """
    This function converts PDF file to string.
    Because email and name must appear on the first page so, return information only on the first page.

    Args:
        file_path: path of pdf file
    """

    reader = PdfReader(file_path)

    page = reader.pages[0] # only on first page
    text = page.extract_text()

    text = text.replace("\n"," ")
    text = text.replace("\r", " ")
    text = " ".join(text.split(" "))
    return text
```

Hình 1: Hàm chuyển file pdf sang text



Hình 2: Thư mục chứa các CV dạng pdf

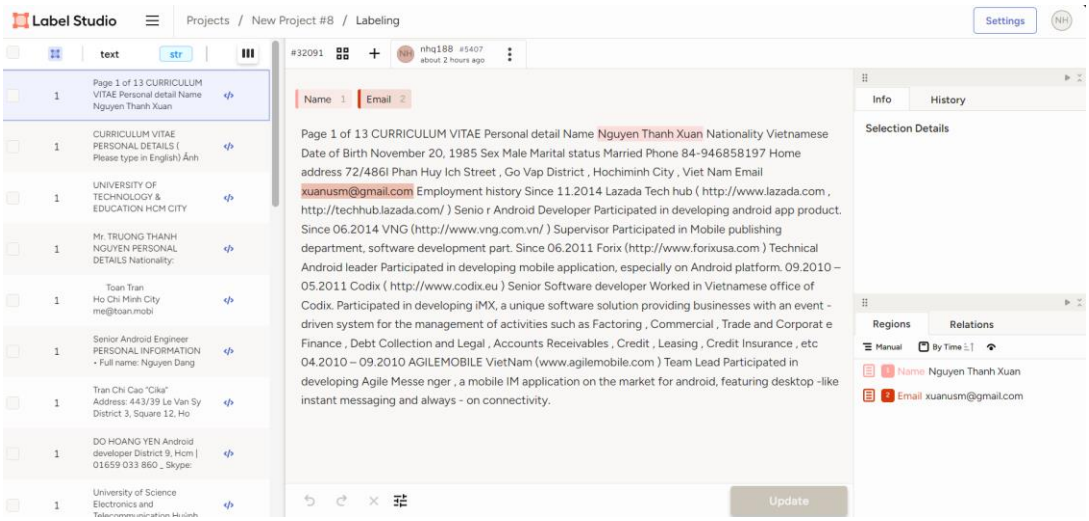


Hình 3: Thư mục chứa các CV dạng txt.

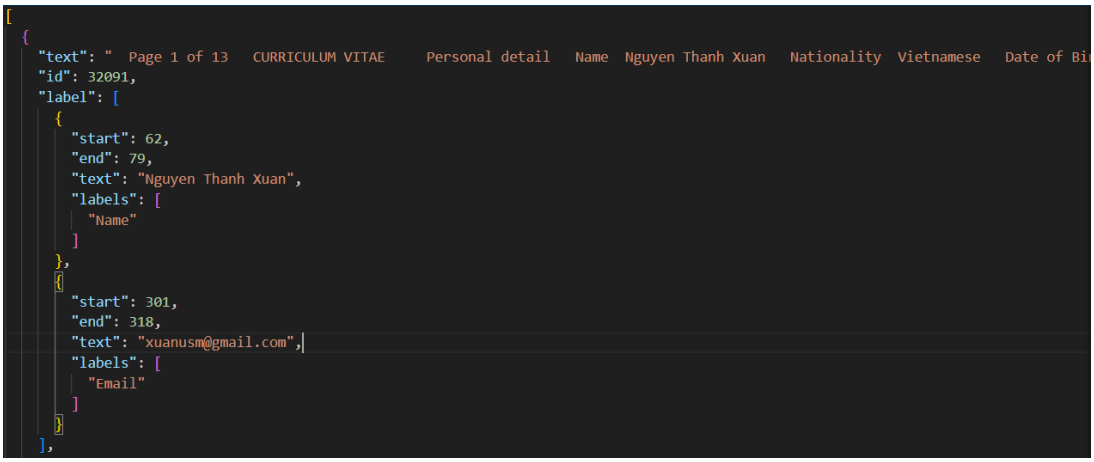
Hình 4: Trang 1 của CV android_01 dưới dạng txt.

2. Gán nhãn dữ liệu

Dữ liệu ban đầu chưa được gán nhãn, do đó em sử dụng tool gán nhãn là label-studio để gán nhãn Name và Email giúp thực hiện bài toán.



Hình 5: Giao diện gán nhãn dữ liệu của label studio.



Hình 6: Một mẫu dữ liệu được gán nhãn.

3. Mô hình sử dụng

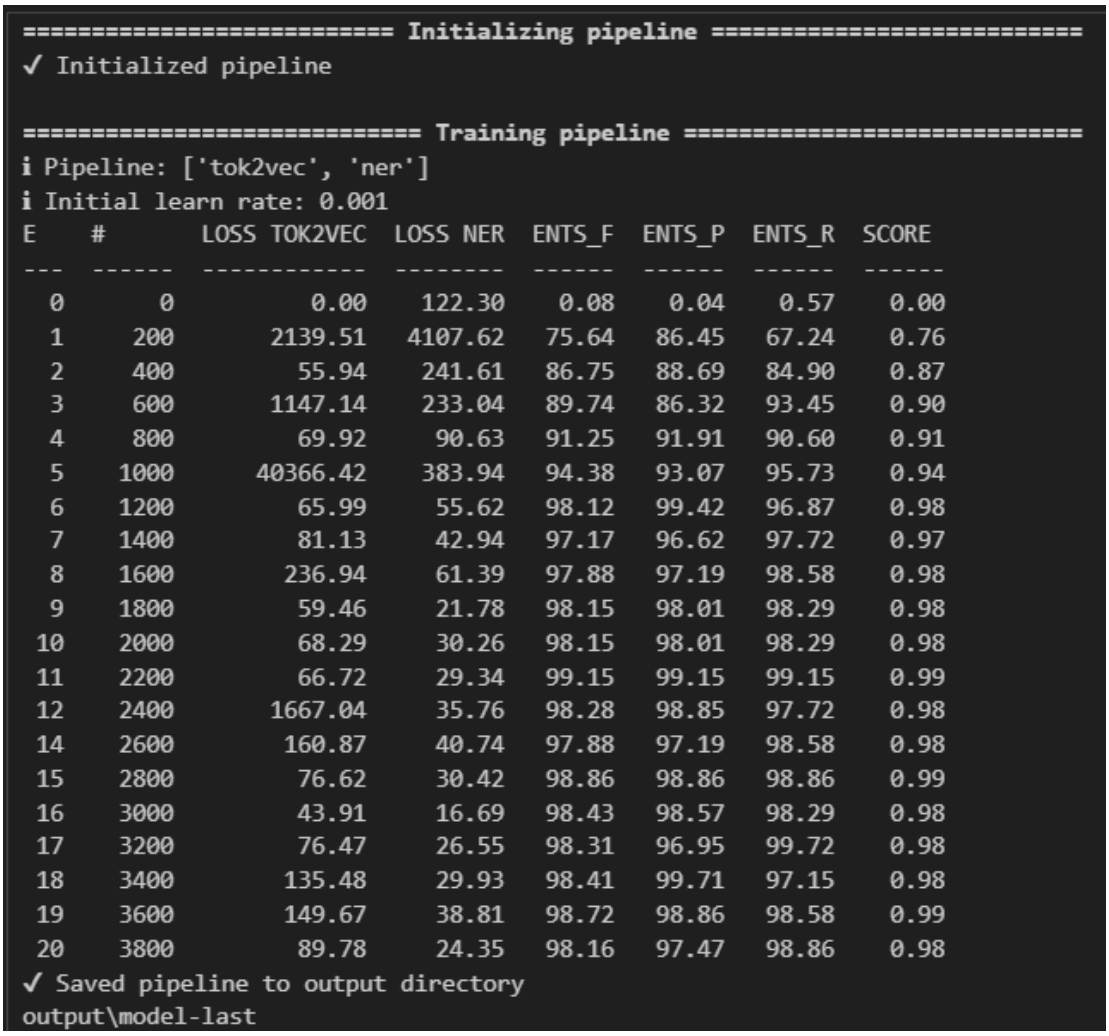
Sử dụng mô hình đa ngôn ngữ cho tác vụ NER của Spacy là **xx_ent_wiki_sm.**

```
import spacy
from spacy.tokens import DocBin
from tqdm import tqdm

nlp = spacy.load("xx_ent_wiki_sm")
doc_bin = DocBin() # create a DocBin object
```

Hình 7: Load mô hình xx_ent_wiki_sm.

4. Huấn luyện mô hình

Mô hình được huấn luyện trên batch_size=128 và epoch=20.

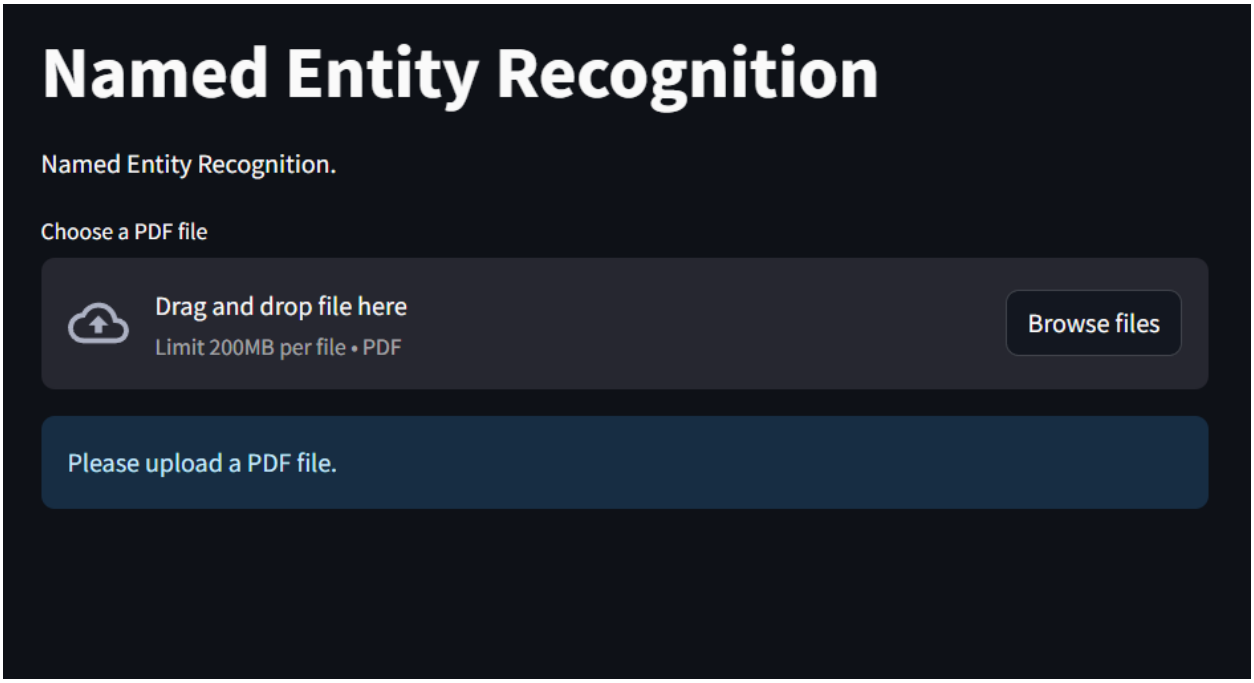


```
============================ Initializing pipeline ============================
✓ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00    122.30    0.08    0.04    0.57    0.00
  1     200       2139.51   4107.62   75.64   86.45   67.24    0.76
  2     400         55.94    241.61   86.75   88.69   84.90    0.87
  3     600       1147.14    233.04   89.74   86.32   93.45    0.90
  4     800         69.92     90.63   91.25   91.91   90.60    0.91
  5    1000      40366.42    383.94   94.38   93.07   95.73    0.94
  6    1200         65.99     55.62   98.12   99.42   96.87    0.98
  7    1400         81.13     42.94   97.17   96.62   97.72    0.97
  8    1600        236.94     61.39   97.88   97.19   98.58    0.98
  9    1800         59.46     21.78   98.15   98.01   98.29    0.98
 10    2000         68.29     30.26   98.15   98.01   98.29    0.98
 11    2200         66.72     29.34   99.15   99.15   99.15    0.99
 12    2400       1667.04     35.76   98.28   98.85   97.72    0.98
 14    2600        160.87     40.74   97.88   97.19   98.58    0.98
 15    2800         76.62     30.42   98.86   98.86   98.86    0.99
 16    3000         43.91     16.69   98.43   98.57   98.29    0.98
 17    3200         76.47     26.55   98.31   96.95   99.72    0.98
 18    3400        135.48     29.93   98.41   99.71   97.15    0.98
 19    3600        149.67     38.81   98.72   98.86   98.58    0.99
 20    3800         89.78     24.35   98.16   97.47   98.86    0.98
✓ Saved pipeline to output directory
output\model-last
```

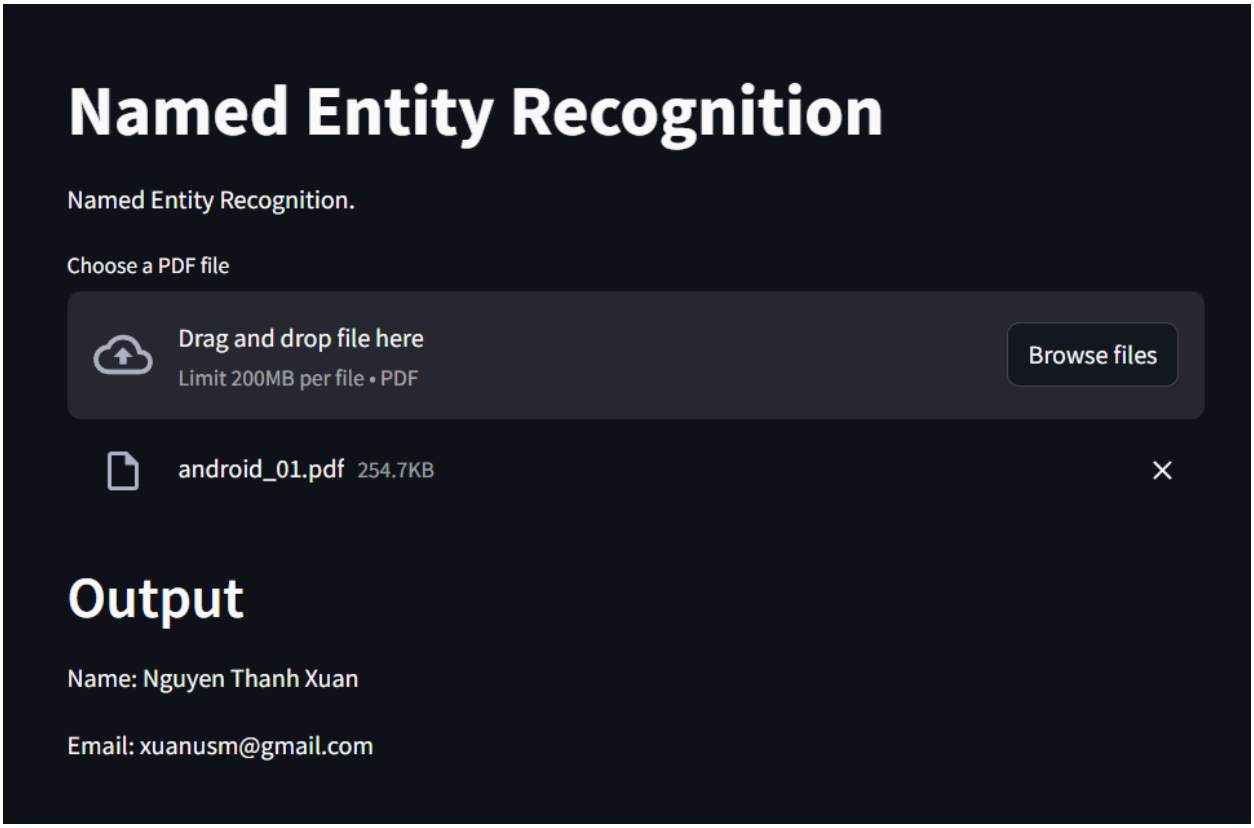
Hình 8: Kết quả trên từng epoch khi huấn luyện mô hình.

2. Giao diện demo

Nhấn chọn file pdf cần trích xuất NAME và EMAIL



Kết quả:



3. Hạn chế và hướng phát triển

Trong project này, em đã xây dựng mô hình NER cho bài toán trích xuất NAME và EMAIL trong CV. Tuy nhiên do lượng dữ liệu còn hạn chế và mô hình còn đơn giản nên kết quả khi trích xuất còn chưa đạt độ chính xác cao.

Do đó, em sẽ thử nghiệm các mô hình ngôn ngữ được huấn luyện trước như RoBERTa, BERT. Đồng thời tiếp tục bổ sung dữ liệu để mô hình khái quát tốt hơn các trường hợp.

Ngoài ra, khi gán nhãn dữ liệu, em nhận thấy có file pdf được scan, pipeline hiện tại không thể giải quyết các trường hợp này. Vì vậy, tiếp theo em sẽ sử dụng các mô hình OCR để trích xuất các văn bản đó.