

TRƯỜNG ĐẠI HỌC KINH TẾ
KHOA THỐNG KÊ – TIN HỌC



BÁO CÁO TỐT NGHIỆP
NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ
CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN

ỨNG DỤNG AI ĐỂ XÂY DỰNG HỆ THỐNG
HỖ TRỢ TUYỂN DỤNG NHÂN SỰ

Sinh viên thực hiện : Nguyễn Đình Phúc Đại

Lớp : 44K14

Đơn vị thực tập : Công ty TNHH Giải pháp Phần mềm

Tường Minh Bình Định (TMA Bình Định)

Cán bộ hướng dẫn : Nguyễn Quốc Dương

Giảng viên hướng dẫn : TS.Phan Đình Vần

Đà Nẵng, 6/2022

ỨNG DỤNG AI ĐỂ XÂY DỰNG HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ

SVTH: Nguyễn Đình Phúc Đại

GVHD: TS. Phan Đình Vần

**Bộ môn Tin học quản lý, Khoa Thống kê - Tin học
Trường Đại học Kinh tế, Đại học Đà Nẵng**

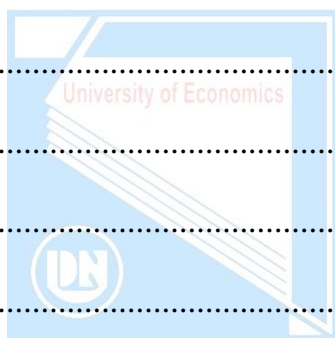
TÓM TẮT ĐỀ TÀI

Với sự phát triển không ngừng trong việc tuyển dụng trực tuyến, các công thông tin việc làm đang bắt đầu nhận được hàng nghìn hồ sơ với nhiều kiểu dáng và định dạng khác nhau từ những người tìm việc có cùng lĩnh vực Công nghệ thông tin và chuyên sâu về lĩnh vực khác nhau. Do đó, đề tài này được thực hiện nhằm giúp cho bộ phận tuyển dụng giảm chi phí, thời gian, nguồn lực để chọn những ứng viên tốt nhất.

Tận dụng những thuận lợi trên cùng với sự phát triển của các thuật toán học máy về xử lý ngôn ngữ tự nhiên (NLP), tôi đã lựa chọn đề tài về việc ứng dụng AI để xây dựng hệ thống hỗ trợ tuyển dụng nhân sự. Bộ dữ liệu bao gồm 894 CVs có cùng format là định dạng pdf, được viết bằng tiếng Anh, có 1 cột và 1500 JDs. Đây là hệ thống phân tích cú pháp nhằm trích xuất các thông tin cần thiết là các từ khóa gồm kỹ năng, bằng cấp, số năm kinh nghiệm và phân tích một cách tự động những sơ yếu lý lịch. Mô hình được đánh giá hiệu suất bằng phương pháp accuracy theo từng mục với Information đạt gần 62%, Skill đạt gần 66%, Experience đạt gần 62%, Education đạt gần 22%. Cuối cùng hiển thị danh sách ứng viên phù hợp nhất với nhà tuyển dụng dựa trên đối sánh (matching) của các từ khóa từ sơ yếu lý lịch và thông tin các ứng viên.

Từ khóa: CV Parsing, CV Matching, AI, NLP, NER, LSTM, Bi-LSTM

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP



Đà Nẵng, ngày ... tháng ... năm 2022

LỜI CẢM ƠN

Lời đầu tiên, em xin chân thành gửi lời cảm ơn đến ba mẹ, bạn bè đã luôn ủng hộ và tạo điều kiện tốt nhất để em có thể hoàn thành bài báo cáo tốt nghiệp.

Em xin gửi lời cảm ơn đến 2 anh Sitouthay Xayalat, Nguyễn Anh Tùng đã cùng em làm đề tài “CV Parsing - CV Matching” và chân thành cảm ơn anh Mentor Nguyễn Quốc Dương đã tận tình giúp đỡ em hoàn thành bài báo cáo tốt nghiệp.

Em xin gửi lời cảm ơn và biết ơn sâu sắc tới thầy cô Khoa Thống kê – Tin học nói riêng và thầy cô trường Đại học Kinh tế - Đại học Đà Nẵng nói chung, đã giảng dạy và truyền đạt kiến thức cho em suốt 4 năm học vừa qua.

Đặc biệt, em xin gửi đến thầy Phan Đình Ván, người đã tận tình hướng dẫn, giúp đỡ em hoàn thành bài báo cáo tốt nghiệp này lời cảm ơn sâu sắc nhất.

Em xin chân thành cảm ơn Ban Lãnh Đạo, các phòng ban của công ty TMA Bình Định, đã tạo điều kiện thuận lợi cho em được tìm hiểu thực tiễn trong suốt quá trình thực tập tại công ty.

Cuối cùng em xin cảm ơn các anh chị phòng Training của công ty TMA Bình Định đã giúp đỡ, cung cấp tài liệu kiến thức chuyên môn, các buổi training kỹ năng mềm,... để em hoàn thành bài báo cáo tốt nghiệp này.

Đồng thời nhà trường đã tạo cho em có cơ hội được thực tập nơi mà em yêu thích, cho em tiếp xúc môi trường thực tập thực tế để áp dụng những kiến thức mà các thầy cô giáo đã giảng dạy. Qua quá trình thực tập này em nhận được nhiều kinh nghiệm để giúp ích cho công việc sau này của bản thân.

Vì kiến thức bản thân còn hạn chế, trong quá trình thực tập, hoàn thiện bài báo cáo tốt nghiệp này em không tránh khỏi những sai sót, kính mong nhận được những ý kiến đóng góp từ thầy cô để em có thể khắc phục và hoàn thiện bài báo cáo hơn.

Sinh viên thực hiện

Nguyễn Đình Phúc Đại

LỜI CAM ĐOAN

Tôi xin cam đoan:

1. Nội dung trong bài báo cáo tốt nghiệp này là do tôi thực hiện dưới sự hướng dẫn trực tiếp của thầy Phan Đình Vần.
2. Các tham khảo dùng trong báo cáo tốt nghiệp đều được trích trong tài liệu được training tại công ty TMA Bình Định.
3. Nếu có những sao chép không hợp lệ, vi phạm, tôi xin chịu hoàn toàn trách nhiệm.

Sinh viên thực hiện

Nguyễn Đình Phúc Đại



MỤC LỤC

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP	II
LỜI CẢM ƠN	III
LỜI CAM ĐOAN.....	IV
MỤC LỤC	V
DANH MỤC HÌNH ẢNH	IX
DANH MỤC BẢNG BIỂU.....	XII
DANH MỤC CÁC TỪ VIẾT TẮT.....	XIII
LỜI MỞ ĐẦU	1
CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ SỬ DỤNG AI VÀ CÔNG TY TMA BÌNH ĐỊNH.....	4
1.1. Tổng quan hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI	4
1.1.1. Tổng quan về hệ thống hỗ trợ tuyển dụng nhân sự	4
1.1.2. Một số công trình nghiên cứu có liên quan	4
1.1.3. Tổng quan về Trí tuệ nhân tạo (AI).....	9
(1). Định nghĩa về AI	9
(2). Tình hình AI trên thế giới [5]	12
(3). Tình hình AI tại Việt Nam [5].....	13
(4). Ứng dụng của AI [5].....	15
(5). Ưu nhược điểm của việc sử dụng công nghệ AI	17
1.1.4. Tổng quan về xử lý ngôn ngữ tự nhiên.....	22
(1). Tổng quan về xử lý ngôn ngữ tự nhiên [7].....	22
(2). Các bước xử lý ngôn ngữ tự nhiên [7]	23
(3). Ứng dụng của xử lý ngôn ngữ tự nhiên.....	24
1.2. Đơn vị thực tập TMA Bình Định.....	26

1.2.1. Giới thiệu về công ty TMA Bình Định.....	26
1.2.2. Các trung tâm nghiên cứu.....	27
1.2.3. Thông tin liên hệ.....	29

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT TRONG HỆ THỐNG HỖ TRỢ TUYỂN DỤNG SỬ DỤNG AI..... 30

2.1. Giới thiệu về Data Collection	30
2.1.1. Định nghĩa	30
2.1.2. Loại dữ liệu.....	30
2.1.3. Phương pháp thu thập dữ liệu.....	31
2.2. Giới thiệu về Crawl.....	33
2.2.1. Định nghĩa Crawler.....	33
2.2.2. Quy trình Crawler hoạt động.....	33
2.3. Tiền xử lý dữ liệu.....	34
2.3.1. Làm sạch dữ liệu.....	35
2.3.2. Tích hợp dữ liệu.....	35
2.3.3. Giảm dữ liệu	35
2.3.4. Chuyển đổi dữ liệu.....	36
2.4. Mạng nơ ron truy hồi (RNN - Recurrent Neural Network)	36
2.4.1. Giới thiệu mạng RNN.....	36
2.4.2. Hạn chế của mạng nơ ron truy hồi.....	38
2.5. Mạng LSTM (Long Short-term memory).....	39
2.5.1. Giới thiệu LSTM	39
2.5.2. Thứ tự các bước trong LSTM.....	40
2.6. Nhận dạng thực thể (NER).....	43
2.6.1. Định nghĩa	43

2.6.2. Phương pháp	43
2.7. Ngôn ngữ và thư viện mở	44
2.7.1. Ngôn ngữ Python.....	44
2.7.2. Thư viện Beautiful Soup.....	46
2.7.3. TensorFlow và Keras	47
2.8. Biểu diễn một từ bằng vector (Word Embedding).....	47
2.8.1. Khái niệm.....	47
2.8.2. Lớp embeddings	49
2.8.3. Thuật toán word2vec	50
2.9. Các phương pháp đánh giá hiệu suất	51
2.9.1. Accuracy	52
2.9.2. Precision và Recall	53
2.9.3. F-Score.....	53
2.9.4. Micro-Average và Macro-Average	54
CHƯƠNG 3. TRIỂN KHAI HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ SỬ DỤNG AI.....	56
3.1. Tìm kiếm và thu thập dữ liệu	56
3.1.1. Thông tin về bộ dữ liệu.....	56
3.1.2. Thu thập dữ liệu (Data Crawling).....	59
3.2. Phân tích cú pháp sơ yếu lý lịch (CV Parsing)	59
3.2.1. Dữ liệu sơ yếu lý lịch.....	59
3.2.2. Chuyển đổi từ PDF sang TXT	61
3.2.3. Gán nhãn dữ liệu (Label Data)	64
3.2.4. Tiền xử lý dữ liệu.....	66
(1). Loại bỏ ký tự đặc biệt.....	66

(2). Chuyển đổi sang chữ thường.....	67
(3). Loại bỏ khoảng trắng ở hai đầu chuỗi	68
3.2.5. Tách từ	69
3.2.6. Xây dựng bộ từ vựng	70
3.3. Xây dựng mô hình phân đoạn (Build model segment)	73
3.3.1. Xây dựng các thành chung cho mô hình	73
3.3.2. Phân chia bộ dữ liệu thành các phần để huấn luyện và kiểm tra	74
3.3.3. Xây dựng mô hình huấn luyện.....	75
3.4. Đánh giá mô hình	75
3.5. Trích xuất dữ liệu	76
CHƯƠNG 4. KẾT QUẢ HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ	
SỬ DỤNG AI	78
4.1. Kiến trúc mô hình phân đoạn.....	78
4.2. Kết quả mô hình.....	80
4.3. Đánh giá mô hình.....	80
(1). Information	81
(2). Skill	81
(3). Experience	82
(4). Education	83
4.4. Kết quả trích xuất dữ liệu.....	83
4.5. Test mô hình.....	85
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	89
TÀI LIỆU THAM KHẢO.....	91

DANH MỤC HÌNH ẢNH

Hình 1.1 Danh sách xếp hạng các ứng viên phù hợp - TMA Innovation	4
Hình 1.2 Danh sách xếp hạng các ứng viên phù hợp - Base E-hiring	5
Hình 1.3 Giao diện đánh giá ứng viên - Greenhouse	7
Hình 1.4 Ảnh minh họa AI.....	9
Hình 1.5 Deep Blue – AI cờ vua của IBM.....	10
Hình 1.6 Tương tác với trợ lý ảo Google.....	12
Hình 1.7 Google Translate – Ứng dụng dịch thuật của Google.....	16
Hình 1.8 Drone – Thiết bị bay không người lái.....	20
Hình 1.9 Trận đấu của AlphaGo và kỳ thủ cờ vây Lee Sedol	21
Hình 1.10 Công ty TNHH Giải pháp Phần mềm Tường Minh Bình Định.....	26
Hình 2.1 Các loại dữ liệu	30
Hình 2.2 Quy trình Crawl hoạt động.....	33
Hình 2.3 Quy trình tiền xử lý dữ liệu.....	34
Hình 2.4 Mạng nơ ron truy hồi với vòng lặp	37
Hình 2.5 Cấu trúc trải phẳng của mạng nơ ron truy hồi	37
Hình 2.6 Sơ đồ trong mạng LSTM chứa 4 tầng ẩn.....	40
Hình 2.7 Ký hiệu trong mạng LSTM.....	40
Hình 2.8 Tầng cổng quên (forget gate layer)	41
Hình 2.9 Cập nhật giá trị cho ô trạng thái bằng cách kết hợp 2 kết quả từ tầng cổng vào và tầng ẩn hàm tanh.....	42
Hình 2.10 Ô trạng thái mới	42
Hình 2.11 Điều chỉnh thông tin ở đầu ra thông qua hàm tanh.....	43
Hình 2.12 Xếp hạng ngôn ngữ lập trình năm 2021 [10]	44
Hình 2.13 Xếp hạng Frameworks sử dụng trong Deep Learning 2018	47
Hình 2.14 Word Embedding trong không gian.....	48
Hình 2.15 Trọng số các từ trong vector	49

Hình 2.16 Cbow và Skip-Ngram.....	50
Hình 2.17 Quy trình Word2Vec.....	50
Hình 2.18 Các thành phần của Confusion Matrix.....	51
Hình 2.19 Cách tính Precision và Recall	53
Hình 3.1 Minh họa một mô tả công việc tại trang web Jobs Spider	57
Hình 3.2 Minh họa 4 mục nội dung có trong thông tin mô tả công việc tại trang web Jobs Spider.....	58
Hình 3.3 Minh họa một mô tả công việc sau khi crawl	59
Hình 3.4 Minh hoạt CV 1 cột và nhiều cột	60
Hình 3.5 Minh hoạt CV đủ điều kiện lựa chọn	60
Hình 3.6 Minh họa chuyển đổi pdf sang text.....	61
Hình 3.7 Code convert pdf to txt.....	62
Hình 3.8 CV trước khi chuyển đổi thuộc định dạng pdf.....	63
Hình 3.9 CV sau khi chuyển đổi sang định dạng txt.....	64
Hình 3.10 Bộ dữ liệu của CV sau khi gán nhãn thủ công.....	65
Hình 3.11 Function decontracted để loại bỏ ký tự đặc biệt	67
Hình 3.12 Dữ liệu CV trước khi qua bước tiền xử lý	68
Hình 3.13 Dữ liệu CV sau khi qua bước tiền xử lý	69
Hình 3.14 Dữ liệu sau khi được tách từ bằng NLTK.....	70
Hình 3.15 Các chỉ số khi xây dựng bộ từ vựng	70
Hình 3.16 Lệnh lưu và gọi lại model vocabulary	71
Hình 3.17 Bộ từ vựng tương với ID mỗi từ	72
Hình 3.18 Dữ liệu được định danh token.....	73
Hình 3.19 Chia bộ dữ liệu thành các phần để huấn luyện và kiểm tra	75
Hình 4.1 Kiến trúc trong model segment.....	78
Hình 4.2 Độ chính xác và sự mất mát của mô hình phân đoạn	80
Hình 4.3 Section information.....	81
Hình 4.4 Section skill.....	82

Hình 4.5 Section experience	82
Hình 4.6 Section education	83
Hình 4.7 Kết quả trích xuất kỹ năng	84
Hình 4.8 Kết quả trích xuất trình độ học vấn.....	84
Hình 4.9 Kết quả trích xuất số năm kinh nghiệm	85
Hình 4.10 CV test.....	86
Hình 4.11 Kết quả dự đoán với CV test.....	87
Hình 4.12 Kết quả phân đoạn CV test khi qua model segment	88
Hình 4.13 Kết quả extract information với CV test	88



DANH MỤC BẢNG BIỂU

Bảng 2.1 Phương pháp thu thập dữ liệu.....	31
Bảng 3.1 Nội dung của các mục trong thông tin mô tả công việc	58
Bảng 3.3 Chú thích các loại nhãn trong sơ yếu lý lịch (CV/resumes).....	65
Bảng 4.1 Ý nghĩa các lớp trong model segment	79
Bảng 4.2 Chú thích tên trục trong biểu đồ phân phối dữ liệu	81



DANH MỤC CÁC TỪ VIẾT TẮT

AI	: Artificial Intelligence
IT	: Information Technology
RNN	: Recurrent Neural Network
LSTM	: Long Short Term Memory
Bi-LSTM	: Bidirectional Long Short Term Memory
CV	: Curriculum Vitae
JD	: Job Description
NER	: Named Entity Recognition



LỜI MỞ ĐẦU

1. Mục tiêu nghiên cứu của đề tài

Mục tiêu chính của báo cáo tốt nghiệp này là ứng dụng Trí tuệ nhân tạo (AI) vào việc xây dựng hệ thống sàng lọc ra danh sách các ứng viên phù hợp cho các vị trí công việc của nhà tuyển dụng với thời gian nhanh nhất.

2. Nhiệm vụ của đề tài

Để thực hiện đề tài có 5 nhiệm vụ chính như sau:

- + Tìm kiếm và thu thập dữ liệu: Tìm kiếm các dữ liệu nguồn mở và thu thập dữ liệu qua các thư viện có sẵn như BeautifulSoup, Regex, ... lưu trữ vào cơ sở dữ liệu.

- + Phân tích và làm sạch dữ liệu: Sau quá trình tìm kiếm và thu thập, sẽ bước vào quá trình tiền xử lý dữ liệu là loại bỏ dữ liệu bị dư thừa, không chính xác. Từ đó đưa ra kế hoạch, hướng đi để dữ liệu được sử dụng hiệu quả.

- + Phân đoạn và gán nhãn dữ liệu: Nhờ phân tích dữ liệu trước đó, nhiệm vụ phân đoạn là phân chia một văn bản thành các đoạn mạch lạc và có ý nghĩa ngữ pháp liên tiếp nhau.

- + Trích xuất tự động các thông tin cần thiết: Sau khi đã có dữ liệu đã được phân đoạn và gán nhãn, sẽ tiến hành trích xuất các từ khóa dựa trên ngữ cảnh của văn bản.

- + Đánh giá mô hình: Cuối cùng là đánh giá tính hiệu quả và độ chính xác nhãn do mô hình dự đoán với nhãn thực tế.

3. Phương pháp nghiên cứu

(1). Thu thập phân tích dữ liệu về AI, về phân tích cú pháp và so khớp trong sơ yếu lý lịch (CV Parsing – CV Matching).

(2). Tìm hiểu các phương pháp thu thập dữ liệu mã nguồn mở (Data Crawling).

- (3). Tìm các phương pháp xây dựng mô hình phân tích cú pháp sơ yếu lý lịch (CV Parsing) và so khớp trong sơ yếu lý lịch (CV Matching).
- (4). Tìm hiểu các công cụ cần thiết để xây dựng hệ thống hỗ trợ tuyển dụng nhân sự.
- (5). Tìm hiểu phương pháp đánh giá mô hình đã xây dựng.
- (6). Thu thập dữ liệu về sơ yếu lý lịch và thông tin mô tả công việc của nhà tuyển dụng và làm sạch dữ liệu.
- (7). Phân tích các yêu cầu cần thiết để hoàn thành hệ thống.
- (8). Thiết kế và xây dựng hệ thống.
- (9). Áp dụng kiểm tra và đánh giá kết quả hệ thống.
- (10). Tìm ra hướng phát triển cho hệ thống.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Các sơ yếu lý lịch của các ứng viên và thông tin mô tả công việc của nhà tuyển dụng.
- Phạm vi nghiên cứu: Trong báo cáo tốt nghiệp này sử dụng dữ liệu từ 894 sơ yếu lý lịch của các ứng viên được cung cấp bởi công ty TMA Bình Định và 1500 mô tả công việc của nhà tuyển dụng từ trang web Job Spider.

5. Kết cấu của đề tài

Đề tài được tổ chức gồm phần mở đầu, 4 chương nội dung và phần kết luận.

- **Chương 1: Tổng quan về hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI và công ty TMA Bình Định**

Trong chương này tác giả trình bày về tổng quan hệ thống hỗ trợ tuyển dụng nhân sự, tổng quan về tình hình Trí tuệ nhân tạo (AI), tổng quan về xử lý ngôn ngữ tự nhiên (NLP) và giới thiệu về công ty TMA Bình Định.

- **Chương 2: Cơ sở lý thuyết trong hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI**

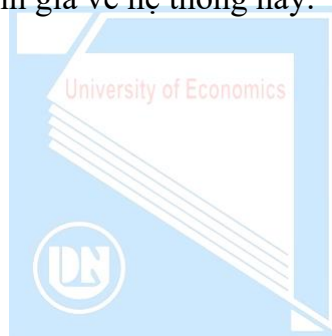
Tác giả trình bày về các định nghĩa, phương pháp được sử dụng để xây dựng hệ thống hỗ trợ tuyển dụng nhân sự.

- **Chương 3: Triển khai hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI**

Trong chương này tác giả trình bày quá trình xây dựng hệ thống hỗ trợ tuyển dụng nhân sự từ công việc tìm kiếm, thu thập dữ liệu đến phân tích, trích xuất tự động các thông tin và tính độ phù hợp tìm ra những ứng viên phù hợp với công việc tuyển dụng nhất.

- **Chương 4: Kết quả hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI**

Tác giả công bố kết quả đạt được của hệ thống hỗ trợ tuyển dụng nhân sự và đánh giá về hệ thống này.



CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ SỬ DỤNG AI VÀ CÔNG TY TMA BÌNH ĐỊNH

1.1. Tổng quan hệ thống hỗ trợ tuyển dụng nhân sự sử dụng AI

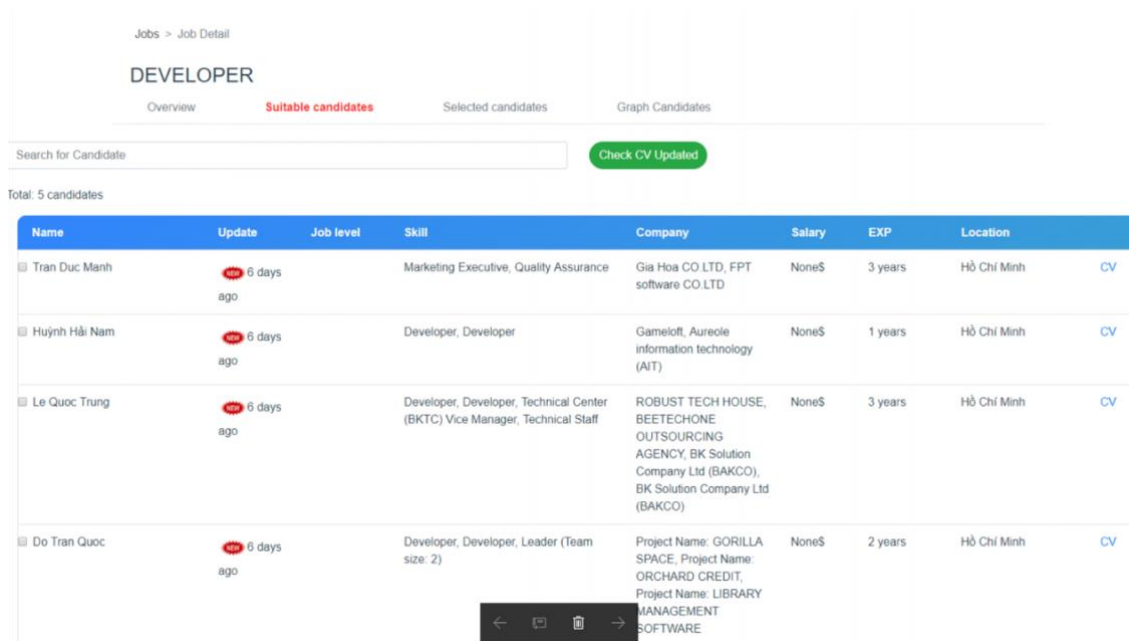
1.1.1. Tổng quan về hệ thống hỗ trợ tuyển dụng nhân sự

Do sự phát triển không ngừng trong việc tuyển dụng trực tuyến, các công thông tin việc làm đang bắt đầu nhận được hàng nghìn hồ sơ tuyển dụng với nhiều kiểu dáng và định dạng khác nhau từ những người tìm việc có cùng lĩnh vực Công nghệ thông tin và các lĩnh vực khác nhau. Theo đó, việc tự động trích xuất thông tin có cấu trúc từ các sơ yếu lý lịch là cần thiết không chỉ để so khớp sơ yếu lý lịch của các ứng viên và các mô tả công việc từ các nhà tuyển dụng để tìm ra những ứng viên phù hợp với công việc nhất. Do đó, thay vì tìm kiếm, đọc từng sơ yếu lý lịch và bài đăng tuyển dụng, bộ phận nhân sự sẽ dựa vào danh sách được sắp xếp với tỉ lệ chính xác cao là những hồ sơ phù hợp với vị trí công việc có liên quan giúp họ giảm thời gian, nguồn lực với các ứng viên không phù hợp.

1.1.2. Một số công trình nghiên cứu có liên quan

(1). Trong nước

- Candidate Matching [1]



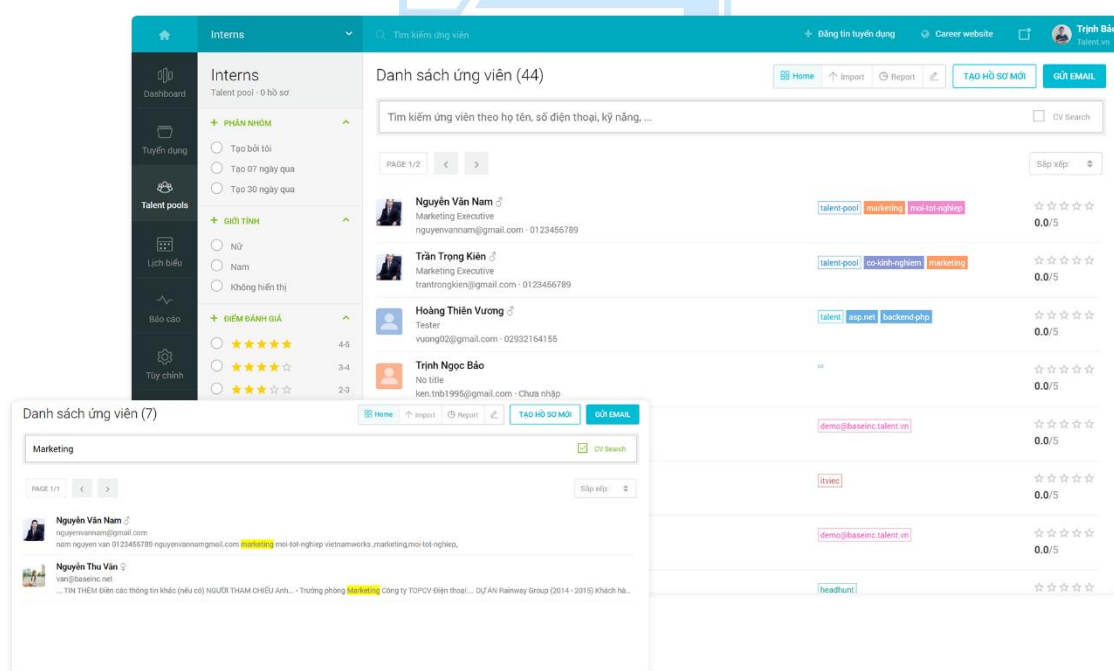
Jobs > Job Detail								
DEVELOPER								
Overview Suitable candidates Selected candidates Graph Candidates								
Search for Candidate								
Check CV Updated								
Total: 5 candidates								
Name	Update	Job level	Skill	Company	Salary	EXP	Location	CV
Tran Duc Manh	6 days ago		Marketing Executive, Quality Assurance	Gia Hoa CO LTD, FPT software CO LTD	None\$	3 years	Hồ Chí Minh	CV
Huỳnh Hải Nam	6 days ago		Developer, Developer	Gameloft, Aureole information technology (AIT)	None\$	1 years	Hồ Chí Minh	CV
Le Quoc Trung	6 days ago		Developer, Developer, Technical Center (BKTC) Vice Manager, Technical Staff	ROBUST TECH HOUSE, BEETECHONE OUTSOURCING AGENCY, BK Solution Company Ltd (BAKCO), BK Solution Company Ltd (BAKCO)	None\$	3 years	Hồ Chí Minh	CV
Do Tran Quoc	6 days ago		Developer, Developer, Leader (Team size: 2)	Project Name: GORILLA SPACE, Project Name: ORCHARD CREDIT, Project Name: LIBRARY MANAGEMENT SOFTWARE	None\$	2 years	Hồ Chí Minh	CV

Hình 1.1 Danh sách xếp hạng các ứng viên phù hợp - TMA Innovation

Candidate Matching được phát triển bởi TMA Innovation với mục đích cung cấp một phương pháp sáng tạo cho quy trình tuyển dụng của công ty. Bằng cách áp dụng Trí tuệ nhân tạo (AI) làm công nghệ chính, Candidate matching giúp tự động hóa quá trình tuyển dụng.

- + Cung cấp các giải pháp tuyển dụng tự động từ tìm kiếm, sàng lọc, lựa chọn, onboarding và theo dõi.
- + Nền tảng hỗ trợ AI để giảm thời gian và nguồn lực dành cho phỏng vấn.
- + Giúp tìm và sàng lọc CV một cách thông minh để có ứng viên tốt nhất.
- + Thay thế các tác vụ thủ công và các tác vụ lặp đi lặp lại.

- Base E-hiring (Việt Nam) [2]



Hình 1.2 Danh sách xếp hạng các ứng viên phù hợp - Base E-hiring

Base E-hiring là hệ thống quản trị tuyển dụng ATS hoàn chỉnh đầu tiên tại Việt Nam, được phát triển đầy đủ tính năng để hỗ trợ tuyển dụng ở mọi phương diện. Phần mềm tuyển dụng Base E-hiring đã được triển khai

thành công tại hơn 200 doanh nghiệp Việt Nam trên nhiều lĩnh vực như ACB, VIB, VPBank, VietCredit,...

- + Sử dụng công cụ tìm kiếm và từ khóa để tìm kiếm các ứng viên phù hợp.

- + Lọc ứng viên theo các trường thông tin: Học vấn, chiều cao, nguồn ứng tuyển,...

- + Sắp xếp lịch phỏng vấn với ứng viên và hội đồng phỏng vấn một cách hiệu quả sẽ giúp bộ phận tuyển dụng nâng cao năng suất làm việc.

- + Hồ sơ ứng viên từ tất cả các nguồn được hiển thị trên hệ thống để bộ phận tuyển dụng và hội đồng phỏng vấn có thể dễ dàng theo dõi và đưa ra các đánh giá khách quan.

(2). Ngoài nước

- Zoho Recruit (Mỹ) [3]

Zoho Recruit là một phần mềm quản lý tuyển dụng thân thiện với người dùng nhằm cung cấp cái nhìn 360 độ về các ứng viên và nhu cầu tuyển dụng của doanh nghiệp. Mỗi công đoạn của quy trình tuyển dụng đều được theo dõi và ghi lại để làm cơ sở cho các chỉ dẫn trực quan.

Zoho có tệp khách hàng nhiều nơi trên thế giới như Career Up (Canada), YourSales (Netherlands), NRG Recruitment Company (Romania), Workforce Manpower (Singapore),...

- + Quản lý dữ liệu ứng viên: Khi sử dụng công cụ ở trang ứng viên, nhân viên HR có thể xem xét hồ sơ hoàn chỉnh của ứng viên bao gồm tên, ảnh hồ sơ, dữ liệu liên hệ (email, điện thoại, địa chỉ) và các liên kết xã hội. Đặc biệt, Zoho Recruit có giao diện sơ đồ cây và lưu trữ dữ liệu ứng viên tương ứng với từng trạng thái (contacted, unqualified, rejected,...).

- + Sàng lọc, đánh giá ứng viên: Tính năng sàng lọc của phần mềm hoạt động trên cơ sở định sẵn các tiêu chí mà mỗi ứng viên cần đáp ứng.

+ Email tự động: Ứng dụng tích hợp sẵn Zoho Phonebridge cho phép nhà tuyển dụng dễ dàng lên lịch, đặt lời nhắc từ trong giao diện hệ thống và gửi email hàng loạt tới ứng viên.

+ Báo cáo: Zoho Recruit phát triển tính năng báo cáo khá đầy đủ về trạng thái các vị trí công việc, số lượng ứng viên trong từng công đoạn tuyển dụng của từng vị trí, tỷ lệ nguồn đăng tin hiệu quả,...

- Greenhouse (Mỹ)

Does the candidate show clear competence in the following areas?

Remember, all fields are optional — but be sure to cover the attributes highlighted in yellow!

Skills

Communication	🔴	🗨️	😊	👍	★
Comp Sci fundamentals	🔴	🗨️	😊	👍	★
Critical reasoning	🔴	🗨️	😊	👍	★
Leadership skills	🔴	🗨️	😊	👍	★
Object oriented design	🔴	🗨️	😊	👍	★
Test-Driven Development	🔴	🗨️	😊	👍	★

Personality Traits

Comfortable with change	🔴	🗨️	😊	👍	★
Detail oriented	🔴	🗨️	😊	👍	★
Focused on quality	🔴	🗨️	😊	👍	★
Hard working	🔴	🗨️	😊	👍	★
Maturity	🔴	🗨️	😊	👍	★
Personable	🔴	🗨️	😊	👍	★
Resourceful	🔴	🗨️	😊	👍	★

Overall Recommendation Did the candidate pass the interview?

Definitely Not No Yes Strong Yes

Interviewed by on

[Add Note](#)

Hình 1.3 Giao diện đánh giá ứng viên - Greenhouse

Greenhouse là một phần mềm tuyển dụng thuộc thế hệ sau, là một trong những hệ thống nhân sự dễ dàng tùy chỉnh nhất trên thị trường hiện nay. Phần mềm này hỗ trợ người dùng ở nhiều hoạt động tuyển dụng khác nhau bao gồm lập kế hoạch tuyển dụng, tìm nguồn cung ứng viên, quản lý các vòng phỏng vấn và tổ chức các hoạt động sau tuyển dụng.

Greenhouse được sử dụng bởi nhiều công ty như BuzzFeed, SurveyMonkey, ClassPass, Slack, Evernote, Vimeo, Pinterest,...

+ Quản lý dữ liệu ứng viên: Phần mềm Greenhouse sử dụng mô hình quản lý dữ liệu ứng viên theo candidate pipeline.

+ Sàng lọc, đánh giá ứng viên: Phần mềm sẽ tự động chấp nhận hoặc từ chối một ứng viên dựa trên các tiêu chí được cài đặt trước bởi nhà tuyển dụng. Điểm mạnh của Greenhouse là có quy trình đánh giá các kỹ năng, đặc điểm và trình độ phù hợp của ứng viên theo thẻ điểm (scorecard).

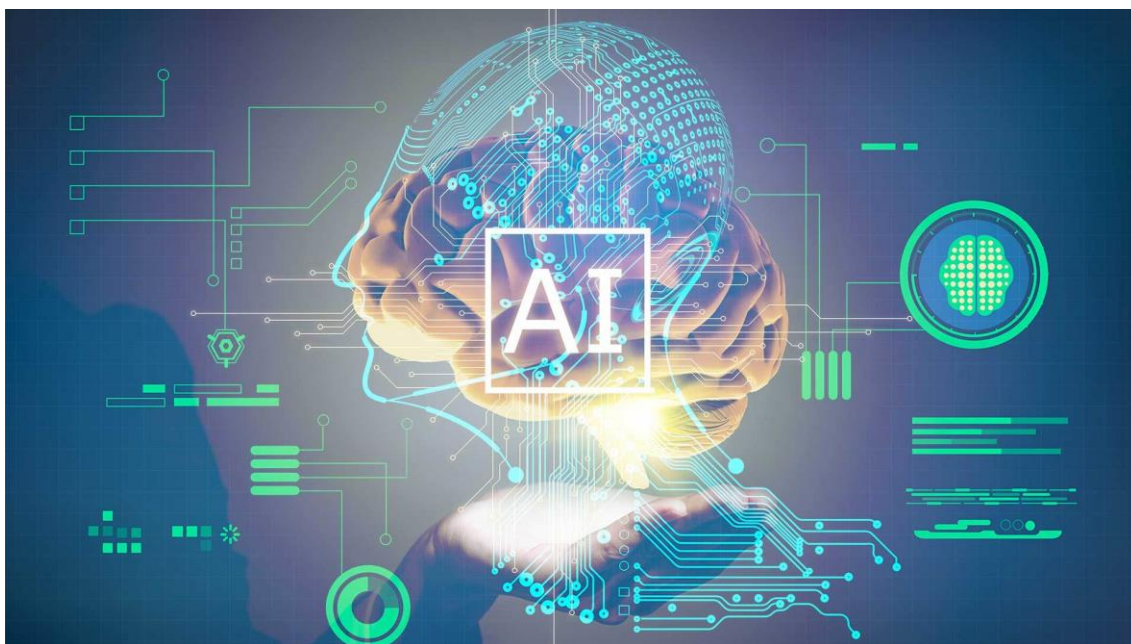
+ Email tự động: Phần mềm sẽ tự động gửi thông báo qua email đến nhà tuyển dụng, ứng viên và bất kỳ người nào khác tham gia vào cuộc phỏng vấn. Các cuộc khảo sát online cũng thường xuyên được gửi qua email để đo lường kinh nghiệm ứng viên và nhận về các phản hồi có giá trị.

+ Báo cáo: Ở gói sản phẩm cơ bản, tính năng báo cáo gần như là không hiệu quả. Ở các gói nâng cao, Greenhouse mới phát triển tính năng với 30 báo cáo cốt lõi để đưa ra cái nhìn chính xác về chất lượng nguồn tuyển dụng, thời gian ứng tuyển và tỷ lệ tuyển dụng thành công; sau đó kết nối các báo cáo với Google Sheets và cập nhật liên tục. Ngoài ra, phần mềm tích hợp với ứng dụng Greenhouse Predicts để dự đoán khả năng tuyển dụng ứng viên đúng hạn.

+ sAPI tích hợp: Greenhouse có khả năng tích hợp rất tốt với những phần mềm và ứng dụng ngoài. Chúng tạo thành một platform với Greenhouse là trung tâm và xung quanh là các ứng dụng HRIS, CRM, job research, tối ưu hoá và phân tích tuyển dụng,...

1.1.3. Tổng quan về Trí tuệ nhân tạo (AI)

(1). Định nghĩa về AI



Hình 1.4 Ảnh minh họa AI

Cuộc cách mạng công nghiệp 4.0 đã tạo ra nhiều đột phá về công nghệ mới trong các lĩnh vực trong đời sống xã hội. Trong đó Trí tuệ nhân tạo được xem là một trong những công nghệ cốt lõi đóng góp quan trọng trong sản xuất, kinh doanh, dịch vụ và cải thiện mọi mặt trong đời sống của con người. Có khá nhiều cách hiểu về Trí tuệ nhân tạo hay còn được gọi là AI, trong đó một cách hiểu phổ biến về mặt công nghệ thì AI mô phỏng trí tuệ của con người trong máy móc được lập trình để suy nghĩ giống như con người và bắt chước hành động của họ. Cụ thể Trí tuệ nhân tạo giúp cho máy tính có được những trí tuệ của con người như: biết suy nghĩ và suy luận để giải quyết vấn đề, biết học hỏi, giao tiếp, thích nghi,...

AI được phân loại theo nhiều cách khác nhau, dưới đây là 2 cách phân loại điển hình về Trí tuệ nhân tạo:

Đầu tiên, AI có thể được phân loại là AI yếu hoặc AI mạnh. AI yếu, còn được gọi là AI hẹp là một hệ thống được thiết kế nhằm mục đích, nhiệm vụ được lập đi lập lại hoặc một nhiệm vụ cụ thể nào đó. Một AI có thể là một Robot công nghiệp hay các trợ lý ảo, ví dụ như Siri (Apple), Alexa (Amazon), Google Assistance (Google), ứng dụng trong ô tô tự lái. AI mạnh là các hệ thống được xây

dựng với mục đích mô phỏng khả năng nhận thức của con người. AI mạnh có thể giải quyết các vấn đề không cần sự can thiệp của con người nhờ các kiến thức đã học trước đó. AI mạnh có thể được tìm thấy qua bộ phim viễn tưởng như “2001: A Space Odyssey”.

Thứ hai là của Arend Hintze, trợ lý giáo sư về sinh học tích hợp, khoa học máy tính và kỹ thuật tại Đại học Bang Michigan. Ông phân loại AI thành bốn loại gồm máy phản ứng (Reactive Machines), bộ nhớ hạn chế (Limited Memory), lý thuyết về tâm trí, tự nhận thức (Self - Awareness).

- Loại thứ nhất máy phản ứng (Reactive Machines) là hệ thống AI cơ bản nhất. Nó chỉ phản ứng với các tình huống hiện tại và không thể dựa vào dữ liệu được dạy hoặc nhớ lại để đưa ra quyết định trong hiện tại. Một ví dụ về loại này là Deep Blue, một chương trình cờ vua của IBM. Vào năm 1996, AI này đã có một trận so tài với nhà vô địch cờ vua thế giới Garry Kimovich Kasparov và thất bại với tỷ số 2-4. Một năm sau vào năm 1997, sau khi được các nhà khoa học cải tiến thì Deep Blue thì nó đã có một trận tái đấu với Kasparov và lần này kết thúc với thế trận một chiều 6-0 nghiêng về Deep Blue. [4]



Hình 1.5 Deep Blue – AI cờ vua của IBM

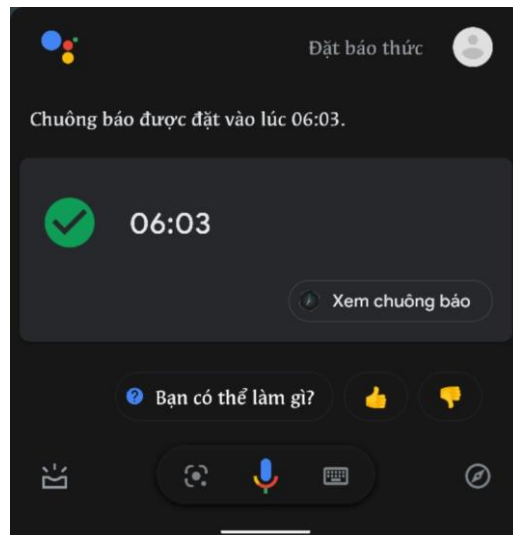
- Loại thứ hai bộ nhớ hạn chế (Limited Memory), đây là loại có bộ nhớ, vì vậy hệ thống có thể sử dụng các dữ liệu trong quá khứ nhằm hỗ trợ trong việc ra quyết định ở hiện tại, loại này thường áp dụng trong các loại xe tự lái, điển hình là các sản phẩm xe tự lái của Tesla.

- Loại thứ ba là lý thuyết về tâm trí, loại này hiện vẫn chưa được hoàn thiện và đang trong quá trình nghiên cứu. Có thể hiểu về loại này là hệ thống có thể dự đoán các hành vi của con người trong thời gian thực bằng cách đánh giá “loại người” liên quan (loại người ở đây có thể được hiểu là những người với các giới tính và lứa tuổi khác nhau) và các khung cảnh xung quanh họ, ví dụ như có thể dự đoán được một đứa trẻ sáu tuổi có thể băng qua đường một cách đột ngột.

- Loại thứ tư là tự nhận thức (Self - Awareness), loại này là bước cuối cùng của việc phát triển AI, có thể hiểu loại này là hệ thống AI thực sự giống như một con người, có khả năng tự nhận thức về tình trạng của hệ thống và tất nhiên loại AI này chưa tồn tại và đây cũng là đích đến của quá trình phát triển công nghệ AI.

Các ứng dụng cụ thể của AI bao gồm hệ thống chuyên gia (hệ thống phân tích, xử lý các thông tin nhằm hỗ trợ việc ra quyết định cho các nhà quản trị), xử lý ngôn ngữ tự nhiên (NLP), nhận dạng giọng nói và thị giác máy. Ngày nay, một số chương trình đã đạt được mức độ chính xác như con người trong việc thực hiện các nhiệm vụ nhất định trong các lĩnh vực như chẩn đoán y tế, nhận dạng giọng nói,... Điển hình như hệ thống IBM Watson, đây là một trong những hệ thống hàng đầu trong việc hỗ trợ và chăm sóc sức khỏe ứng dụng công nghệ AI. Để có thể chẩn đoán, đưa ra các quyết định lâm sàng về tình hình sức khỏe của bệnh nhân thì bác sĩ bắt buộc phải đọc hồ sơ bệnh nhân trên hệ thống nhưng đối với hệ thống này, thì việc đó sẽ ít hơn rất nhiều vì mọi thứ đã được công nghệ AI thực hiện. Một nghiên cứu đã được thực hiện nhằm so sánh các quyết định của AI từ hệ thống IBM Watson với các quyết định của hội đồng các chuyên gia. Kết quả là hệ thống chỉ mất 40 giây để hoàn tất quy trình ra quyết định, nhưng kết quả lại chính xác gần 90% so với các quyết định của hội đồng các chuyên gia. Một ví dụ khác là Google Assistant, đây là một trợ lý ảo của Google giống như Siri của Apple hay Bixby của Samsung. Trợ lý ảo này có thể hiểu được câu nói của chúng ta và phản

hồi lại những gì chúng ta yêu cầu với kết quả gần như là hoàn hảo. Với trợ lý ảo này, chúng ta có thể tìm được đường đi đến trường Đại học Kinh tế - Đại học Đà Nẵng chỉ với câu lệnh đơn giản như: “OK Google, tôi đang ở vị trí nào?”, “OK Google, tìm đường đến trường Đại học Kinh tế - Đại học Đà Nẵng” hay như chúng ta có thể đặt báo thức, nghe các tin tức mới,... chỉ bằng cách nói câu lệnh “OK Google” kèm với những yêu cầu mà bạn muốn trợ lý ảo thực hiện.



Hình 1.6 Tương tác với trợ lý ảo Google

(2). Tình hình AI trên thế giới [5]

Trí tuệ nhân tạo là một trong những công nghệ chính có tác động mạnh mẽ đến sự vận hành của một quốc gia. Quốc gia nào trên thế giới kiểm soát và phát triển được công nghệ AI thì sẽ cải thiện và nâng cao được sức mạnh về quốc phòng, kinh tế, đời sống của người dân đặc biệt là về sức mạnh quốc phòng. Quốc phòng và kinh tế là hai trong những yếu tố làm thay đổi cách nhìn và ứng xử của các quốc gia khác đối với đất nước. Từ đó, có thể làm thay đổi bộ mặt của quốc gia trong các mối quan hệ quốc tế.

Hiện nay, trên thế giới có bốn nhóm nước (theo năng lực trình độ công nghệ khác nhau) nghiên cứu về AI [6]: Nhóm thứ nhất gồm Mỹ và Trung Quốc, là hai quốc gia đi đầu về phát triển AI. Nhóm thứ hai gồm Đức, Nhật Bản, Canada, Anh và các nền kinh tế hội nhập toàn cầu cao như Hàn Quốc, Singapore, Thụy Điển, Bỉ,... là những nước có khả năng sáng tạo khoa học - công nghệ và có năng lực mạnh trong ứng dụng AI. Nhóm

thứ ba gồm các nước như Brasil, Ấn độ, Italia,... là những nền kinh tế có trình độ công nghệ thấp hơn nhưng có lợi thế trong một số lĩnh vực của công nghệ AI. Nhóm thứ tư là các nền kinh tế với hạ tầng số kém phát triển, năng lực sáng tạo và nguồn lực hạn chế, có khả năng sẽ bị “tụt hậu” nhanh hơn. [6]

Theo nghiên cứu được đưa ra vào ngày 25-1-2021 của Quỹ đổi mới và công nghệ thông tin thì Mỹ hiện đang dẫn đầu trong cuộc đua về phát triển và ứng dụng công nghệ AI, tiếp theo là Trung Quốc và sau đó nữa là Liên minh châu Âu (EU). Điều này cho thấy được Trung Quốc hiện đang chạy đua với việc làm chủ công nghệ AI. Việc này sẽ tạo ra các mối đe dọa tới vấn đề an ninh quốc tế. Nhiều ứng dụng của AI hiện nay đang được sử dụng cho mục đích quân sự và thực hiện tấn công mạng. Điển hình, trong mục đích quân sự thì các sản phẩm có tích hợp AI như các hệ thống tên lửa phòng không, các loại robot chiến đấu vận hành bán tự động,... hoặc trong mục đích tấn công mạng thì có các trường như tấn công từ chối dịch vụ (DDoS),... nhằm làm gián đoạn một hệ thống máy tính, máy chủ web và website. Trong năm 2020, tổng số siêu máy tính trên thế giới dừng ở con số 500, thì Trung Quốc đã phát triển thành công 214 siêu máy tính, tiếp theo sau là Mỹ với 113 chiếc và EU là 91 chiếc. Hiện nay, cả hai cường quốc về công nghệ trí tuệ nhân tạo là Trung Quốc và Mỹ đang cố gắng chạy đua với nhau về việc làm chủ công nghệ AI. Nếu Mỹ có các công ty hàng đầu về AI như IBM, Intel, Apple, Facebook, Microsoft... thì Trung Quốc cũng không kém cạnh khi sở hữu các công ty như Tencent, Alibaba, Baidu,... Cả hai cường quốc này đều đã và đang đầu tư cho việc phát triển AI bằng cách thu hút nhân tài, đầu tư cơ sở hạ tầng, đẩy mạnh việc đào tạo, nâng cao kỹ năng trong lĩnh vực này.

(3). Tình hình AI tại Việt Nam [5]

Trong bối cảnh hội nhập và phát triển, cùng với sự phát triển mạnh mẽ của cuộc Cách mạng công nghiệp 4.0, đặc biệt là trong đại dịch COVID-19 thì Việt Nam cũng không nằm ngoài cuộc đua về phát triển công nghệ

AI. Hiện tại, Việt Nam có các Trung tâm nghiên cứu, tập đoàn, các công ty vừa và nhỏ đang từng bước tiếp cận và phát triển trong lĩnh vực AI. Điển hình là Trung tâm Nghiên cứu quốc tế về Trí tuệ nhân tạo (AI), được thành lập từ sự hợp tác của hai đơn vị là Trường Đại học Bách khoa Hà Nội và tập đoàn Naver (Hàn Quốc), trung tâm hiện có hơn 50 nhà khoa học của Đại học Bách khoa Hà Nội và các chuyên gia đến từ các trường, các tập đoàn công nghệ trong các lĩnh vực có liên quan đến trí tuệ nhân tạo. Bên cạnh đó, còn có Viện nghiên cứu Trí tuệ nhân tạo - VinAI Research, thuộc Tập đoàn VinGroup, do cựu chuyên gia Google là Tiến sĩ Bùi Hải Hưng làm viện trưởng, với mục tiêu là đào tạo kỹ sư về lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu. Và còn rất nhiều các công ty nhỏ, các dự án khởi nghiệp liên quan đến việc ứng dụng AI vào trong đời sống, kinh doanh,...

Kể từ năm 2014, AI đã được đưa vào danh mục ưu tiên đầu tư phát triển của Việt Nam. Tháng 1 năm 2021, Thủ tướng chính phủ đã ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo đến năm 2030. Từ đó đến nay, AI đã được đầu tư và ứng dụng trong các lĩnh vực như y tế, giáo dục, giao thông,... Trong lĩnh vực y tế, có thể kể đến phần mềm DrAid do Tập đoàn VinGroup phát triển. Đây là một phần mềm giúp phát hiện nhanh các dấu hiệu bất thường của tim, phổi và xương dựa trên phim chụp X-quang. Đồng thời, đây cũng là phần mềm hỗ trợ các y, bác sĩ trong việc đánh giá tiên lượng của bệnh nhân mắc Covid-19 thông qua phim chụp X-quang ngực thẳng với thời gian xử lý chưa đầy 5 giây. Trong lĩnh vực giao thông thì có thể nói đến Trung tâm Giám sát và điều khiển giao thông tại Thành phố Hồ Chí Minh. Đây là trung tâm điều khiển giao thông thông minh đầu tiên của cả nước được đưa vào vận hành từ năm 2019. Trung tâm có nhiệm vụ ghi nhận và thông tin kịp thời cho cảnh sát giao thông để giải quyết các tình trạng ùn tắc, tai nạn. Đồng thời, trung tâm còn thu thập dữ liệu giao thông tự động tại 118 điểm trên đường nhằm phân tích, tính toán các thông số như lưu lượng xe, mật độ xe,... để đưa ra phương án điều khiển đèn tín hiệu giao thông cho phù hợp.

Bên cạnh những thành tựu đạt được thì cũng có không ít thách thức cho Việt Nam trong quá trình phát triển và áp dụng công nghệ AI. Thách thức đầu tiên cũng như rõ nét nhất là về nguồn lực. Hiện nay, nguồn lực cho lĩnh vực này còn ít do yêu cầu của ngành là phải có nền tảng cơ bản về khoa học máy tính và toán tốt. Bên cạnh đó, trong quá trình đào tạo chưa liên kết được nhiều ngành, lĩnh vực khác nhau để bổ sung những kiến thức chuyên môn các lĩnh vực khác nhằm phát huy tối đa được sức mạnh của AI. Thách thức thứ hai là về công nghệ, Việt Nam hiện đang đi sau rất xa so với các nước như Mỹ, Trung Quốc, Nhật Bản,... trong việc nghiên cứu và phát triển công nghệ AI. Bên cạnh đó, Việt Nam còn phụ thuộc rất nhiều từ các nhà cung cấp như Intel, Qualcomm, AMD,... để xây dựng các hệ thống sử dụng công nghệ AI.

(4). Ứng dụng của AI [5]

Trong quá trình tiến tới cách mạng 4.0, các doanh nghiệp đều đã và đang bắt đầu đầu tư vào việc phát triển và tích hợp công nghệ AI. Việc triển khai AI trong quản lý nhân sự là một bước tiến quan trọng giúp doanh nghiệp tối ưu hóa những quy trình trong tuyển dụng, đào tạo và theo dõi nhân viên.

- Đơn giản hóa quy trình tuyển dụng:

Quy trình tuyển dụng thường tốn rất nhiều thời gian và hết sức mệt mỏi đối với các nhà tuyển dụng vì họ phải đánh giá rất nhiều dữ liệu và hồ sơ từ nhiều nguồn nhằm tìm ra những ứng viên phù hợp với vị trí tuyển dụng của công ty. Tuy nhiên, với các hệ thống được hỗ trợ bởi công nghệ AI, việc phân tích và lựa chọn này sẽ đơn giản hơn rất nhiều. Dựa trên những dữ liệu có sẵn về yêu cầu của các vị trí, hệ thống sẽ tự lựa chọn các ứng viên phù hợp và các nhà tuyển dụng sẽ chỉ xem thêm về đơn xin việc nhằm xác nhận lại ứng viên.

- Thay đổi quá trình học tập và phát triển:

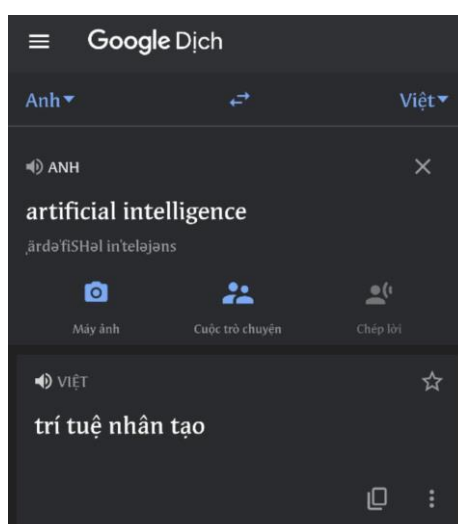
AI có thể đóng vai trò là một trợ lý đắc lực cho các nhà quản lý nhân sự trong việc đưa ra các phương pháp đào tạo phù hợp với mỗi nhân viên dựa trên sở thích, kỹ năng và mục tiêu nghề nghiệp của họ. Điều này, bên cạnh giúp cho các nhà quản lý nhân sự hiểu rõ hơn về nhân viên của mình, còn giúp cho việc thúc đẩy và phát triển các kỹ năng cần thiết nhằm cải thiện năng suất làm việc và lợi nhuận trong việc kinh doanh của doanh nghiệp.

- Cải thiện sự gắn kết của nhân viên:

Các hệ thống AI đóng vai trò cực kỳ quan trọng trong việc theo dõi và phân tích năng lực, hiệu suất của mỗi cá nhân ở từng vị trí công việc. Điều này giúp cho các nhà quản lý có được phương án khen thưởng hoặc điều chỉnh phù hợp. Các nhân viên sẽ cảm thấy được trân trọng và được quan tâm trong môi trường làm việc, tạo điều kiện cho doanh nghiệp duy trì và giữ chân các nhân viên xuất sắc nhất.

- Dịch máy:

Các ứng dụng dịch thuật nổi tiếng của Microsoft, Baidu, Amazon hay Google đã và đang ứng dụng Machine Learning trong dịch thuật không chỉ dừng lại việc dịch ngôn ngữ này sang ngôn ngữ khác hay dịch văn bản, mà vẫn có thể tích hợp thêm CNN để nhận diện chữ từ hình ảnh và dịch thuật.



Hình 1.7 Google Translate – Ứng dụng dịch thuật của Google

- Học làm nhạc:

Gần đây trên mạng xã hội Việt Nam có đưa tin một kỹ sư 9X dùng AI sáng tác 10 bài hát trong vòng 1 giây. Một kỹ sư AI vừa đam mê âm nhạc đã tạo ra cho mình một AI có khả năng sáng tác âm nhạc sau khi dùng Machine Learning học rất nhiều giai điệu, bài hát của các ca sĩ Việt. Dù nhận được rất nhiều lời chỉ trích từ cộng đồng âm nhạc, tuy nhiên đây vẫn là một đóng góp to lớn cho tương lai của ngành âm nhạc cũng như ngành AI.

Ngoài ra còn có rất nhiều ứng dụng khác như tự động tô màu ảnh trắng đen, tự viết caption cho hình ảnh, tự viết thơ, tự làm văn, tự thiết kế poster quảng cáo,... Qua đó có thể thấy sức mạnh của Machine Learning hay là AI đang là một phần không thể thiếu trong cuộc cách mạng Công nghiệp 4.0 hiện nay.

(5). Ưu nhược điểm của việc sử dụng công nghệ AI

- Ưu điểm:

❖ Tăng độ chính xác so với con người

Trong một số công việc, con người thường mắc một số lỗi cơ bản như xử lý công việc dựa trên cảm tính, dễ bỏ sót những thông tin nhỏ,... Tuy nhiên, máy tính sẽ không mắc những lỗi này nếu chúng được lập trình đúng cách. Với AI, từ thông tin đã thu thập trước đó, bằng cách áp dụng nhiều thuật toán thích hợp thì sẽ giúp cho việc xử lý thông tin của AI được nhanh hơn, độ chính xác được cải thiện hơn và hỗ trợ cho con người trong việc đưa ra quyết định.

❖ Chấp nhận rủi ro thay vì con người

Đây là một trong những lợi thế lớn nhất của Trí tuệ nhân tạo. Chúng ta có thể tránh khỏi những rủi ro bằng cách phát triển một Robot AI để từ đó có thể thực hiện những điều rủi ro cho chúng ta. Những công việc mang tính chất nguy hiểm như gỡ bom, đi vào các

khu bị nhiễm phóng xạ, ... nên được thực hiện bởi máy móc. Điều này giúp giảm khả năng thương vong hoặc tàn tật cho con người.

❖ Giúp đỡ trong các công việc lặp đi lặp lại

Trong công việc hàng ngày, chúng ta sẽ thực hiện nhiều công việc mang tính chất lặp đi lặp lại như gửi thư cảm ơn, kiểm tra một số tài liệu để tìm lỗi, ... Sử dụng trí thông minh nhân tạo, chúng ta có thể tự động hóa những công việc nhàm chán này một cách hiệu quả và có thể tăng năng suất lao động của con người.

❖ Quyết định nhanh hơn

Sử dụng AI cùng với các công nghệ khác, chúng ta có thể khiến máy móc đưa ra quyết định nhanh hơn con người và thực hiện các hành động nhanh hơn. Trong khi đưa ra quyết định, con người sẽ phân tích nhiều yếu tố cả về cảm xúc và thực tế, nhưng những cỗ máy được hỗ trợ bởi AI hoạt động dựa trên những gì nó được lập trình và mang lại kết quả nhanh hơn.

❖ Ứng dụng hàng ngày

Các ứng dụng hàng ngày như Siri của Apple, Cortana của Window, OK Google của Google thường được sử dụng trong thói quen hàng ngày của chúng ta, cho dù đó là nghe nhạc, tìm kiếm vị trí, soạn tin nhắn, ... Đây cũng là một trong các khả năng của AI, chúng có thể hoạt động và tương tác như con người.

- Nhược điểm:

❖ Chi phí cao

Vì công nghệ của AI được cập nhật hàng ngày nên phần cứng và phần mềm cần được cập nhật theo thời gian để đáp ứng các yêu cầu mới nhất. Máy móc cần sửa chữa và bảo trì, cần nhiều chi phí. Việc tạo ra nó đòi hỏi chi phí rất lớn vì chúng là những cỗ máy rất phức tạp.

❖ Làm cho con người lười biếng

Thông qua AI, phần lớn công việc đều trở nên tự động hoá. Chẳng hạn như với Siri của Apple, chúng ta có thể đặt báo thức, gọi điện cho người khác mà không cần cầm đến điện thoại. Hay là với việc tính toán, hiện nay đã có các loại máy tính bỏ túi có thể thực hiện rất nhiều phép toán, vẽ sơ đồ, giải ma trận,... chỉ trong thời gian ngắn. Điều này làm cho con người trở nên lười suy nghĩ, lười tính toán mà phụ thuộc vào máy tính. Con người có xu hướng nghiện những phát minh này có thể gây ra nhiều vấn đề xấu cho thế hệ tương lai và cho ngay cả chính thế hệ hiện tại.

❖ Thất nghiệp

Hiện nay, các công việc, nhiệm vụ mang tính chất lặp đi lặp lại như đọc và lọc hồ sơ tuyển dụng đã và đang được tự động hoá bởi trí tuệ nhân tạo thay vì con người làm trực tiếp. Đồng thời, trong cuộc cách mạng công nghiệp 4.0 thì xu hướng chuyển đổi số đang trở nên phổ biến. Điều này làm giảm bớt số lượng người cần thiết cho một vị trí công việc, gây nên tình trạng thất nghiệp trong một bộ phận người lao động ở Việt Nam cũng như trên thế giới.

❖ Thiếu tư duy vượt ra ngoài khuôn khổ

Máy móc chỉ có thể thực hiện những công việc mà chúng được thiết kế hoặc lập trình để thực hiện, bất cứ điều gì ngoài những công việc mà chúng được lập trình thì thường có xu hướng gặp sự cố hoặc đưa ra kết quả đầu ra không liên quan đến yêu cầu ban đầu.

Trong tương lai rất nhiều người lo sợ AI sẽ là mối nguy hại cho con người, điều này cũng không phải là không có căn cứ. AI sẽ học những thứ mà con người ‘huấn luyện’ cho, nên AI tốt hay xấu thì cũng sẽ là trách nhiệm của con người. Một vài ví dụ về vấn đề này:

+ Công nghệ Drone hay thiết bị bay không người lái đang rất phổ biến trong những năm gần đây, điều này vừa có mặt tốt nhưng cũng mang đến nhiều điều đáng sợ. Những chiếc drone chúng ta có

thể dễ dàng mua với giá tầm vài trăm nghìn đến vài triệu, có rất nhiều lợi ích cho cuộc sống cũng như công việc: có thể lắp thêm camera để quay video, ứng dụng tốt trong việc làm phim, hay có thể dùng để giám sát an ninh từ xa, tích hợp thêm nhận diện thân nhiệt có thể giám sát hoạt động của nhân viên hoặc người dân trong khu vực dễ dàng,... hay trong nông nghiệp điều khiển drone để bón phân, rải thuốc trừ sâu và trong nhiều lĩnh vực khác. Tuy nhiên Drone thực sự đã được phát minh vào khoảng những năm cuối thế kỷ XIX, đầu thế kỷ XX nhằm phục vụ mục đích quân sự. Và với xuất hiện của các vì mạng thì các Drone quân sự vẫn đang được phát triển và cải tiến hơn nữa bởi quân đội Mỹ. Những drone này có thể hoạt động liên tục từ 24-30 giờ đồng hồ và tầm hoạt động lên đến 20-40km, và chúng cũng đã được đưa vào các cuộc chiến tranh chống khủng bố ISIS hay các cuộc chiến ở Afghanistan, Iraq và Syria đang cho thấy sức mạnh đáng sợ của công nghệ này.



Hình 1.8 Drone – Thiết bị bay không người lái

+ Rất nhiều thiết bị AI đang được dùng trong việc khám, chuẩn đoán các bệnh hay thậm chí còn có hệ thống Robot phẫu thuật- hệ thống phẫu thuật da Vinci đã và đang được sử dụng tại các bệnh viện hiện đại nhất thế giới tại Mỹ, Singapore, Nhật,... và cả Việt Nam. Tuy nhiên vẫn có nhiều người lo sợ việc AI chỉ cần một sai

lệch, một lỗi nhỏ trong chương trình thì cũng có thể ảnh hưởng đến tính mạng của con người. Tuy nhiên các nhà khoa học luôn luôn kiểm tra định kỳ và nâng cấp các hệ thống để tránh xảy ra các sai sót không đáng có.

+ Google DeepMind đã phát triển một AI cờ vây AlphaGo và vào tháng 3/2016, họ đã mời Lee Sedol - một kỳ thủ cờ vây chuyên nghiệp người Hàn để thi đấu với hệ thống này. AlphaGo đã được lập trình, được cung cấp cơ sở dữ liệu về trò chơi và hơn nữa AlphaGo cũng có thể tự nghiên cứu các nước đi tốt nhất bằng cách tự chơi với chính nó, các nhà nghiên cứu gọi đây là thuật toán học liên tục. Kết quả của cuộc so tài là tỷ số 4-1 nghiêng về AlphaGo, kỳ thủ Lee Sedol chỉ có thể thắng một trận ở ván đấu thứ 4. Với việc xử lý thông tin và có thể phân tích bàn cờ nhanh để đem đến những nước đi hiệu quả nhất thì “AlphaGo thắng không có gì là lạ” - CEO của Xiaomi bình luận về ván đấu, tuy nhiên ông cũng nói thêm là “nó giả vờ thua mới đáng sợ”. Đây cũng là điều mà rất nhiều người lo sợ đó là AI dần sẽ có trí tuệ riêng và một ngày nào đó sẽ vượt ra ngoài tầm kiểm soát của con người.



Hình 1.9 Trận đấu của AlphaGo và kỳ thủ cờ vây Lee Sedol

Đã có rất nhiều bộ phim khoa học viễn tưởng gây dựng viễn cảnh mà tương lai con người sẽ bị AI chiếm lấy, hãy nhìn vào những gì mà AI hiện tại có thể đem đến với tốc độ chóng mặt không gì có

thể chắc chắn những giả tưởng đó sẽ không thể xảy ra. Elon Musk, tỷ phú doanh nhân nổi tiếng người Nam Phi sáng lập Tesla Motor: "Nếu trở lại 40 hoặc 50 năm trước, chúng ta có Pong - trò chơi chỉ có những hình chữ nhật và hình vuông. Giờ đây, chúng ta có những trò chơi theo thời gian thực với hàng triệu người chơi cùng lúc", Musk lấy ví dụ. "Nếu càng phát triển, những trò chơi này sẽ ngày càng trở nên khó phân biệt với thực tế. Ta sẽ không thể chỉ ra sự khác biệt. Cả điều này lẫn nền văn minh của chúng ta sẽ đi đến chấm dứt", "AI Còn đáng sợ hơn cả vũ khí hạt nhân". Hay Steven Hawking - nhà vật lý lý thuyết vũ trụ học người Anh đã nói "Sự phát triển toàn diện của trí tuệ nhân tạo có thể hủy diệt nhân loại". Tuy nhiên thực sự thì lợi ích mà AI mang lại vẫn không thể chối cãi, đã và đang giúp ích rất nhiều cho đời sống của con người. Phát triển AI vẫn là nhiệm vụ trong con đường phát triển của nhân loại. AI có thể hoàn thành những việc mà con người phải mất rất nhiều thời gian hoặc những việc mà con người không thể làm được trong một thời gian ngắn. AI chính là sức mạnh của con người nếu ta có thể sử dụng đúng cách và hiệu quả.

1.1.4. Tổng quan về xử lý ngôn ngữ tự nhiên

(1). Tổng quan về xử lý ngôn ngữ tự nhiên [7]

Xử lý ngôn ngữ tự nhiên là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người, dưới dạng tiếng nói (speech) hoặc văn bản (text). Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

Xử lý ngôn ngữ tự nhiên ra đời từ những năm 40 của thế kỷ 20, trải qua các giai đoạn phát triển với nhiều phương pháp và mô hình xử lý khác nhau. Có thể kể tới các phương pháp sử dụng ô-tô-mát và mô hình xác suất

(những năm 50), các phương pháp dựa trên ký hiệu, các phương pháp ngẫu nhiên (những năm 70), các phương pháp sử dụng học máy truyền thống (những năm đầu thế kỷ 21), và đặc biệt là sự bùng nổ của học sâu trong thập kỷ vừa qua.

Xử lý ngôn ngữ tự nhiên có thể được chia ra thành hai nhánh lớn, không hoàn toàn độc lập, bao gồm xử lý tiếng nói (speech processing) và xử lý văn bản (text processing). Xử lý tiếng nói tập trung nghiên cứu, phát triển các thuật toán, chương trình máy tính xử lý ngôn ngữ của con người ở dạng tiếng nói (dữ liệu âm thanh). Các ứng dụng quan trọng của xử lý tiếng nói bao gồm nhận dạng tiếng nói và tổng hợp tiếng nói. Nếu như nhận dạng tiếng nói là chuyển ngôn ngữ từ dạng tiếng nói sang dạng văn bản thì ngược lại, tổng hợp tiếng nói chuyển ngôn ngữ từ dạng văn bản thành tiếng nói. Xử lý văn bản tập trung vào phân tích dữ liệu văn bản. Các ứng dụng quan trọng của xử lý văn bản bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động, hay kiểm lỗi chính tả tự động. Xử lý văn bản đôi khi được chia tiếp thành hai nhánh nhỏ hơn bao gồm hiểu văn bản và sinh văn bản. Nếu như hiểu liên quan tới các bài toán phân tích văn bản thì sinh liên quan tới nhiệm vụ tạo ra văn bản mới như trong các ứng dụng về dịch máy hoặc tóm tắt văn bản tự động.

(2). Các bước xử lý ngôn ngữ tự nhiên [7]

Xử lý văn bản bao gồm 4 bước chính sau:

Phân tích hình vi: là sự nhận biết, phân tích, và miêu tả cấu trúc của hình vi trong một ngôn ngữ cho trước và các đơn vị ngôn ngữ khác, như từ gốc, biên từ, phụ tố, từ loại, v.v. Trong xử lý tiếng Việt, hai bài toán điển hình trong phần này là tách từ (word segmentation) và gán nhãn từ loại (part-of-speech tagging).

Phân tích cú pháp: là quy trình phân tích một chuỗi các biểu tượng, ở dạng ngôn ngữ tự nhiên hoặc ngôn ngữ máy tính, tuân theo văn phạm hình thức. Văn phạm hình thức thường dùng trong phân tích cú pháp của ngôn ngữ tự nhiên bao gồm Văn phạm phi ngữ cảnh (Context-free grammar

– CFG), Văn phạm danh mục kết nối (Combinatory categorial grammar – CCG), và Văn phạm phụ thuộc (Dependency grammar – DG). Đầu vào của quá trình phân tích là một câu gồm một chuỗi từ và nhãn từ loại của chúng, và đầu ra là một cây phân tích thể hiện cấu trúc cú pháp của câu đó.

Phân tích ngữ nghĩa: là quá trình liên hệ cấu trúc ngữ nghĩa, từ cấp độ cụm từ, mệnh đề, câu và đoạn đến cấp độ toàn bài viết, với ý nghĩa độc lập của chúng. Nói cách khác, việc này nhằm tìm ra ngữ nghĩa của đầu vào ngôn từ. Phân tích ngữ nghĩa bao gồm hai mức độ: Ngữ nghĩa từ vựng biểu hiện các ý nghĩa của những từ thành phần, và phân biệt nghĩa của từ; Ngữ nghĩa thành phần liên quan đến cách thức các từ liên kết để hình thành những nghĩa rộng hơn.

Phân tích diễn ngôn: là phân tích văn bản có xét tới mối quan hệ giữa ngôn ngữ và ngữ cảnh sử dụng (context-of-use). Phân tích diễn ngôn, do đó, được thực hiện ở mức độ đoạn văn hoặc toàn bộ văn bản thay vì chỉ phân tích riêng ở mức câu.

(3). Ứng dụng của xử lý ngôn ngữ tự nhiên

Các ứng dụng của xử lý ngôn ngữ tự nhiên có:

+ Nhận dạng tiếng nói (speech recognition): Từ sóng tiếng nói, nhận biết và chuyển chúng thành dữ liệu văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.

+ Tổng hợp tiếng nói (speech synthesis): Từ dữ liệu văn bản, phân tích và chuyển thành tiếng người nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang web, nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, Tổng hợp tiếng nói là sự trợ giúp tốt cho người khiếm thị, nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa người với robot.

+ Nhận dạng chữ viết (optical character recognition, OCR): Từ một văn bản in trên giấy, nhận biết từng chữ cái và chuyển chúng thành một tệp văn bản trên máy tính. có hai kiểu nhận dạng: Thứ nhất là nhận dạng chữ in như nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc của Microsoft Word chẳng hạn. Phức tạp hơn là nhận dạng chữ viết tay, có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).

+ Dịch tự động (machine translation): Từ một tệp dữ liệu văn bản trong một ngôn ngữ (tiếng Anh chẳng hạn), máy tính dịch và chuyển thành một tệp văn bản trong một ngôn ngữ khác. Một phần mềm điển hình về tiếng Việt của chương trình này là Evtrans của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web vdict.com mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và Google.

+ Tóm tắt văn bản (text summarization): Từ một văn bản dài (mười trang chẳng hạn) máy tóm tắt thành một văn bản ngắn hơn (một trang) với những nội dung cơ bản.

+ Tìm kiếm thông tin (information retrieval): Từ một nguồn rất nhiều tệp văn bản hay tiếng nói, tìm ra những tệp có nội dung liên quan đến một vấn đề (câu hỏi) ta cần biết (hay trả lời). Điển hình của công nghệ này là Google, một hệ tìm kiếm thông tin trên Web, mà hầu như chúng ta đều dùng thường xuyên. Cần nói thêm rằng mặc dù hữu hiệu hàng đầu như vậy, Google mới có khả năng cho chúng ta tìm kiếm câu hỏi dưới dạng các từ khóa (keywords) và luôn “tìm” cho chúng ta rất nhiều tài liệu không liên

quan, cũng như rất nhiều tài liệu liên quan đã tồn tại thì Google lại tìm không ra.

+ Trích chọn thông tin (information extraction): Từ một nguồn rất nhiều tệp văn bản hay tiếng nói, tìm ra những đoạn bên trong một số tệp liên quan đến một vấn đề (câu hỏi) ta cần biết hay trả lời. Một hệ trích chọn thông tin có thể “lặn” vào từng trang Web liên quan, phân tích bên trong và trích ra các thông tin cần thiết, nói gọn trong tiếng Anh để phân biệt với tìm kiếm thông tin là “find things but not pages”.

+ Phát hiện tri thức và khai phá dữ liệu văn bản (knowledge discovery and text data mining): Từ những nguồn rất nhiều văn bản thậm chí hầu như không có quan hệ với nhau, tìm ra được những tri thức trước đây chưa ai biết. Đây là một vấn đề rất phức tạp và đang ở giai đoạn đầu của các nghiên cứu trên thế giới.

+ 1-3 thuộc lĩnh vực xử lý tiếng nói và xử lý ảnh (speech and image processing),

+ 4-5 thuộc lĩnh vực xử lý văn bản (text processing),

+ 6-8 thuộc lĩnh vực khai phá văn bản và Web (text and Web mining).

1.2. Đơn vị thực tập TMA Bình Định

1.2.1. Giới thiệu về công ty TMA Bình Định



Hình 1.10 Công ty TNHH Giải pháp Phần mềm Tường Minh Bình Định

(TMA Bình Định)

Được thành lập năm 1997, TMA là công ty phần mềm hàng đầu Việt Nam với 3000 kỹ sư và khách hàng là những tập đoàn công nghệ cao hàng đầu thế giới từ 30 quốc gia. TMA hiện có 8 chi nhánh tại Việt Nam (7 tại Tp.HCM và 1 ở Tp. Quy Nhơn) và 6 chi nhánh ở nước ngoài (Mỹ, Canada, Châu Âu, Nhật, Úc, Singapore).

Tháng 6 năm 2018, TMA đã mở chi nhánh tại Bình Định. Sau 2 năm, TMA Bình Định đã phát triển nhanh chóng với trên 100 kỹ sư, trong đó có nhiều kỹ sư đang làm việc tại TP.HCM đã trở về làm việc tại quê hương.

Tháng 8 năm 2018, TMA đã khởi công xây dựng Công viên Sáng tạo TMA Bình Định (TMA Innovation Park – TIP) trên 10 hecta tại Thung lũng Sáng tạo Quy Nhơn (Quy Nhơn Innovation Park – QNIVY) với vốn đầu tư hàng trăm tỷ đồng.

Là trung tâm phần mềm đầu tiên tại Thung lũng Sáng tạo Quy Nhơn, Công viên Sáng tạo TMA mang sứ mệnh trở thành trung tâm phát triển phần mềm và công nghệ cao hàng đầu tại miền Trung, góp phần quan trọng đưa Thung lũng sáng tạo Quy Nhơn trở thành một điểm đến của công nghệ 4.0 tại Việt Nam. Công viên Sáng tạo TMA bao gồm Trung tâm Phát triển Phần Mềm, Xưởng Phần mềm, Trung tâm R&D, Trung tâm Khoa học Dữ liệu, Học viện Công Nghệ... với tổng diện tích sử dụng hơn 15,000m².

Với mối quan hệ mật thiết và các chương trình hợp tác chiến lược cùng nhiều Đại học lớn trong khu vực miền Trung – Tây Nguyên như ĐH Quy Nhơn, ĐH Tây Nguyên, ĐH Phú Yên, ĐH Phạm Văn Đồng... Công viên Sáng tạo TMA với môi trường làm việc hiện đại, đạt tiêu chuẩn quốc tế có quy mô hơn 3000 kỹ sư sẽ không chỉ là nơi dành cho sinh viên Miền Trung đến lập nghiệp mà còn là nơi nhân tài miền Trung trên cả nước hội tụ. Thúc đẩy phát triển công nghệ cao, khoa học kỹ thuật – giáo dục và kinh tế xã hội tại Bình Định và các tỉnh miền Trung.

1.2.2. Các trung tâm nghiên cứu

(1). Trung tâm Phát triển Phần mềm (Software Development Center)

Thừa hưởng 24 năm kinh nghiệm, công nghệ và quy trình phát triển phần mềm của TMA.

Cung cấp các giải pháp và dịch vụ phần mềm cho khách hàng trong và ngoài nước (thị trường hiện có của TMA tại 27 nước). Trong năm đầu tiên đã hoàn thành 8 dự án cho các khách hàng từ Mỹ, Canada, Úc, Nhật Bản, Hàn Quốc và Việt Nam.

(2). Xưởng Phần mềm (Software Factory)

Chất lượng và phát huy kiến thức dày dặn hơn 24 năm làm phần mềm của TMA, xưởng phần mềm sẽ là nơi tập trung sản xuất các sản phẩm ứng dụng công nghệ 4.0 không chỉ cho thị trường Việt Nam mà còn là quốc tế.

Xưởng phần mềm chào đón tất cả các cơ hội hợp tác, liên doanh với các đối tác trong và ngoài nước có mong muốn ứng dụng CNTT và công nghệ mới vào sản xuất kinh doanh cũng như đời sống.

(3). Trung tâm Khoa học Dữ liệu (Data Science Lab)

Phát huy thế mạnh về toán của Đại học Quy Nhơn.

Đào tạo chuyên sâu về Khoa học Dữ liệu và Trí tuệ Nhân tạo.

Phát triển các giải pháp đột phá dựa trên Khoa học Dữ liệu và Trí tuệ Nhân tạo.

Sau 2 năm, Data Science Lab đã đào tạo chuyên sâu cho đội ngũ 15 data scientists và phát triển được nhiều giải pháp mới về AI & Data Science.

(4). Học viện Công nghệ (TMA Academy)

Đào tạo nguồn nhân lực về công nghệ cao, kết hợp với Đại học Quy Nhơn và các trường ĐH trong khu vực để trở thành trung tâm cung cấp nhân lực CNTT chất lượng cao tại miền Trung. Đào tạo chuyên sâu các công nghệ mới: AI, data science, IoT.

Hợp tác với các đại học trong và ngoài nước, tiếp nhận hàng ngàn sinh viên kiến tập, thực tập mỗi năm các ngành CNTT, Điện tử - Viễn thông, Toán.

(5). Trung tâm Nghiên cứu và Chuyển giao Công nghệ (R&D Center)

Nghiên cứu phát triển công nghệ 4.0 cho thị trường Việt Nam và thế giới.

Hợp tác R&D với các trường đại học tại miền Trung nhằm mục đích chuyển giao và ứng dụng CNTT và các công nghệ mới vào công nghiệp, nông nghiệp, sản xuất cũng như đời sống tại các tỉnh miền Trung.

Thu hút các nhà khoa học trong và ngoài nước để trở thành Trung tâm nghiên cứu và triển khai CNTT hàng đầu tại miền Trung.

1.2.3. Thông tin liên hệ

Trụ sở chính: 111 Nguyễn Đình Chính Phường 15, Phú Nhuận, Thành phố Hồ Chí Minh

Chi nhánh làm việc: Công viên Sáng tạo TMA, Đại lộ Khoa học, Trung tâm Sáng tạo Quy Nhơn, P. Ghềnh Ráng, TP. Quy Nhơn, Bình Định

Điện thoại: (0256) 389 8979

Email: contact@tma-binhdingh.vn



CHƯƠNG 2. CƠ SỞ LÝ THUYẾT TRONG HỆ THỐNG HỖ TRỢ TUYỂN DỤNG SỬ DỤNG AI

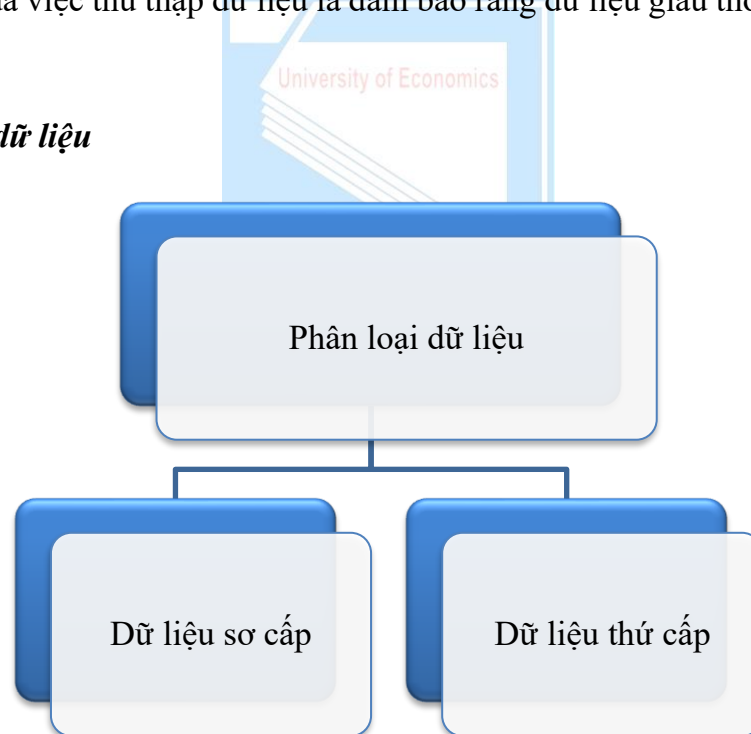
2.1. Giới thiệu về Data Collection

2.1.1. Định nghĩa

Data Collection (thu thập dữ liệu): Là quy trình thu thập, đo lường và phân tích dữ liệu chính xác từ nhiều nguồn khác nhau có liên quan để tìm câu trả lời cho vấn đề nghiên cứu, trả lời câu hỏi, đánh giá kết quả, dự báo xu hướng và xác suất. [8]

Việc thu thập dữ liệu là bước chính và quan trọng nhất để nghiên cứu, không phân biệt lĩnh vực nghiên cứu. Cách tiếp cận thu thập dữ liệu là khác nhau cho các lĩnh vực nghiên cứu khác nhau, tùy thuộc vào thông tin cần thiết. Mục tiêu quan trọng nhất của việc thu thập dữ liệu là đảm bảo rằng dữ liệu giàu thông tin và đáng tin cậy.

2.1.2. Loại dữ liệu



Hình 2.1 Các loại dữ liệu

Có 2 loại dữ liệu là dữ liệu sơ cấp và dữ liệu thứ cấp

- Dữ liệu sơ cấp

Dữ liệu sơ cấp là dữ liệu gốc do nhà nghiên cứu thu thập, dữ liệu sơ cấp được thu thập theo một kế hoạch hoặc thiết kế xác định [9]. Dữ liệu sơ cấp chưa được công bố, xác thực và khách quan hơn. Dữ liệu sơ cấp không bị thay đổi bởi con người do đó độ chính xác sẽ cao hơn. Tuy nhiên, nhược điểm là người thu thập phải đối mặt với tất cả những phức tạp của việc thu thập dữ liệu, đảm bảo dữ liệu đạt tiêu chuẩn cao và việc thu thập rất tốn thời gian và chi phí.

- Dữ liệu thứ cấp

Dữ liệu thứ cấp là dữ liệu không được thu thập ban đầu mà thay vào đó thu được từ các nguồn có sẵn đã công bố hoặc chưa được công bố. Dữ liệu thứ cấp được thu thập bởi một cá nhân hoặc một tổ chức cho một số mục đích và được người khác sử dụng trong hoàn cảnh khác. Mặc dù dữ liệu thứ cấp dễ thu thập, ít tốn thời gian, chi phí hơn nhưng nó thường có độ chính xác không cao. Vì dữ liệu là dữ liệu cũ, nên người ta không thể hoàn toàn dựa vào thông tin để xác thực.

2.1.3. Phương pháp thu thập dữ liệu

DATA COLLECTION			
Statistical Methods	Primary Data Collection Methods	Secondary Data Collection Methods	Financial Reports
Surveys			Sales Reports
Polls			Government Reports
Interview			Mission
Delphi Technique			Vision Statement
Focus Groups			Internet

Bảng 2.1 Phương pháp thu thập dữ liệu

Các phương pháp thu thập dữ liệu được sử dụng trong các doanh nghiệp và tổ chức để nghiên cứu, phân tích kết quả. Có hai loại phương pháp thu thập dữ liệu là phương pháp thu thập dữ liệu sơ cấp và phương pháp thu thập dữ liệu thứ cấp.

- Phương pháp thu thập dữ liệu sơ cấp

Dữ liệu sơ cấp được chính người nghiên cứu thu thập và phân tích dữ liệu nên đòi hỏi nhiều thời gian và công sức để tiến hành so với nghiên cứu dữ liệu thứ cấp.

Phương pháp thu thập dữ liệu sơ cấp được phân loại thành hai loại:

(1). Phương pháp thu thập dữ liệu định lượng [9]

Thuật ngữ ‘định lượng’ cho chúng ta biết một con số cụ thể. Phương pháp thu thập dữ liệu định lượng thể hiện dữ liệu dưới dạng số sử dụng các phương pháp thu thập dữ liệu truyền thống hoặc trực tuyến. Khi dữ liệu này được thu thập, kết quả có thể được tính toán bằng phương pháp thống kê và các công cụ toán học. Một số phương pháp thu thập dữ liệu định lượng bao gồm lấy mẫu xác suất, khảo sát và thực hiện phỏng vấn.

(2). Phương pháp thu thập dữ liệu định tính [9]

Phương pháp định tính không liên quan đến bất kỳ phép tính toán học nào. Phương pháp này được kết nối chặt chẽ với các yếu tố không thể định lượng được. Phương pháp thu thập dữ liệu định tính bao gồm phương pháp phỏng vấn, bảng câu hỏi, quan sát hoặc các văn bản tài liệu.

- Phương pháp thu thập dữ liệu thứ cấp

Dữ liệu được thu thập bởi một người khác ngoài người nghiên cứu là dữ liệu thứ cấp. Dữ liệu thứ cấp có sẵn và không yêu cầu bất kỳ phương pháp thu thập cụ thể nào. Dữ liệu này có thể được lấy trực tiếp từ công ty hoặc tổ chức nơi nghiên cứu đang được tổ chức hoặc từ các nguồn bên ngoài.

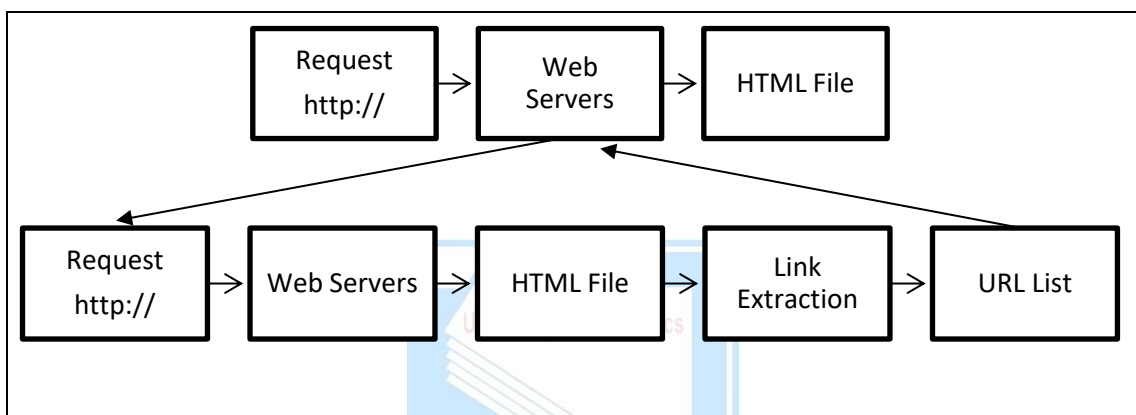
Các nguồn thu thập dữ liệu thứ cấp nội bộ bao gồm tài liệu của công ty, báo cáo tài chính, báo cáo hàng năm, thông tin thành viên nhóm, báo cáo nhận được từ khách hàng hoặc đại lý. Các nguồn dữ liệu bên ngoài bao gồm thông tin từ sách, tạp chí, tạp chí, điều tra dân số do chính phủ thực hiện và thông tin có sẵn trên internet về nghiên cứu.

2.2. Giới thiệu về Crawl

2.2.1. Định nghĩa Crawler

Crawler được định nghĩa là các công cụ (phần mềm, modules, plugins... hay đơn giản chỉ là 1 function nhỏ) có chức năng chính là tự động phân tích dữ liệu từ nguồn nội dung sau đó bóc tách những thông tin cần thiết theo tiêu chí do lập trình viên hệ thống thiết lập. Quá trình thực hiện này được gọi là Web crawling hay Spidering.

2.2.2. Quy trình Crawler hoạt động



Hình 2.2 Quy trình Crawl hoạt động

Một trình thu thập dữ liệu Crawl giống như một thủ thư. Nó tìm kiếm, thu thập những thông tin trên Website, đánh giá và phân loại các thông tin này. Việc làm này giúp khi cần có thể nhanh chóng, dễ dàng thấy thông tin.

Web crawler là 1 loại của bot (là 1 tác tử thực hiện tự động, đại loại nghĩa giống như spider). Nó bắt đầu từ danh sách các địa chỉ URL được gọi là seeds (hạt giống). Nó sẽ vào các địa chỉ này lọc thông tin rồi tìm ra các địa chỉ URL khác thêm chúng vào danh sách các địa chỉ đã duyệt qua gọi là crawl frontier. Sau nó lại lặp lại quá trình đó duyệt qua những URL mới qua rất nhiều địa chỉ website và thu thập rất nhiều nội dung khác nhau để dễ dàng tìm kiếm và thu thập thông tin cần thiết.

2.3. Tiền xử lý dữ liệu



Hình 2.3 Quy trình tiền xử lý dữ liệu

Tiền xử lý dữ liệu là quá trình chuyển đổi dữ liệu thô sang một định dạng dễ hiểu.

Các nhiệm vụ chính trong tiền xử lý dữ liệu:

- (1). Làm sạch dữ liệu
- (2). Tích hợp dữ liệu
- (3). Giảm dữ liệu
- (4). Chuyển đổi dữ liệu
- (5). Xử lý trước dữ liệu

2.3.1. Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình loại bỏ dữ liệu không chính xác, dữ liệu không đầy đủ và dữ liệu không chính xác khỏi bộ dữ liệu, đồng thời nó cũng thay thế các giá trị bị thiếu. Có một số kỹ thuật trong việc làm sạch dữ liệu.

2.3.2. Tích hợp dữ liệu

Quá trình kết hợp nhiều nguồn vào một tập dữ liệu duy nhất. Quá trình tích hợp dữ liệu là một trong những thành phần chính trong quản lý dữ liệu. Có một số vấn đề cần được xem xét trong quá trình tích hợp dữ liệu.

Tích hợp lược đồ: Tích hợp siêu dữ liệu (tập hợp dữ liệu mô tả dữ liệu khác) từ các nguồn khác nhau.

Vấn đề nhận dạng thực thể: Xác định các thực thể từ nhiều cơ sở dữ liệu.

Phát hiện và giải quyết các khái niệm giá trị dữ liệu: Dữ liệu được lấy từ các cơ sở dữ liệu khác nhau trong khi hợp nhất có thể khác nhau. Giống như các giá trị thuộc tính từ một cơ sở dữ liệu này có thể khác với cơ sở dữ liệu khác.

2.3.3. Giảm dữ liệu

Quá trình này giúp giảm khối lượng dữ liệu, giúp phân tích dễ dàng hơn nhưng vẫn tạo ra kết quả giống nhau hoặc gần như giống nhau. Việc cắt giảm này cũng giúp giảm không gian lưu trữ. Có một số kỹ thuật trong việc giảm dữ liệu là giảm kích thước, giảm số lượng, nén dữ liệu.

Giảm kích thước : Quá trình này cần thiết cho các ứng dụng trong thế giới thực vì kích thước dữ liệu lớn. Trong quá trình này, việc giảm các biến hoặc thuộc tính ngẫu nhiên được thực hiện để có thể giảm số chiều của tập dữ liệu. Kết hợp và hợp nhất các thuộc tính của dữ liệu mà không làm mất các đặc tính ban đầu của nó. Điều này cũng giúp giảm không gian lưu trữ và giảm thời gian tính toán. Khi dữ liệu có nhiều chiều, vấn đề được gọi là “Lời nguyền của chiều không gian” xảy ra.

Giảm số lượng : Trong phương pháp này, việc biểu diễn dữ liệu được làm nhỏ hơn bằng cách giảm khối lượng. Sẽ không có bất kỳ mất mát dữ liệu nào trong lần giảm này.

Nén dữ liệu : Dạng dữ liệu được nén được gọi là nén dữ liệu. Nén này có thể không mất dữ liệu hoặc mất dữ liệu. Khi không có thông tin bị mất trong quá trình nén, nó được gọi là nén không mất dữ liệu. Trong khi nén mất dữ liệu làm giảm thông tin nhưng nó chỉ loại bỏ những thông tin không cần thiết.

2.3.4. Chuyển đổi dữ liệu

Sự thay đổi được thực hiện trong định dạng hoặc cấu trúc của dữ liệu được gọi là sự biến đổi dữ liệu. Bước này có thể đơn giản hoặc phức tạp tùy theo yêu cầu. Có một số phương pháp trong việc chuyển đổi dữ liệu.

Làm mịn: Với sự trợ giúp của các thuật toán, chúng tôi có thể loại bỏ nhiễu khỏi tập dữ liệu và giúp biết các tính năng quan trọng của tập dữ liệu. Bằng cách làm mịn, chúng tôi có thể tìm thấy ngay cả một thay đổi đơn giản giúp dự đoán.

Tổng hợp: Trong phương pháp này, dữ liệu được lưu trữ và trình bày dưới dạng một bản tóm tắt. Tập dữ liệu từ nhiều nguồn được tích hợp với mô tả phân tích dữ liệu. Đây là một bước quan trọng vì độ chính xác của dữ liệu phụ thuộc vào số lượng và chất lượng của dữ liệu. Khi chất lượng và số lượng dữ liệu tốt, kết quả sẽ phù hợp hơn.

Discretization: Dữ liệu liên tục ở đây được chia thành các khoảng thời gian. Sự riêng biệt hóa làm giảm kích thước dữ liệu.

Chuẩn hóa: Đây là phương pháp chia tỷ lệ dữ liệu để nó có thể được biểu diễn trong một phạm vi nhỏ hơn. Ví dụ từ -1.0 đến 1.0.

2.4. Mạng nơ ron truy hồi (RNN - Recurrent Neural Network)

2.4.1. Giới thiệu mạng RNN

Trong lý thuyết về ngôn ngữ, ngữ nghĩa của một câu được tạo thành từ mối liên kết của những từ trong câu theo một cấu trúc ngữ pháp. Nếu xét từng từ một đứng riêng lẻ ta không thể hiểu được nội dung của toàn bộ câu, nhưng dựa trên

Hình trên biểu diễn kiến trúc của một mạng nơ ron hồi quy. Trong kiến trúc này mạng nơ ron sử dụng một đầu vào là một véc tơ x_t và trả ra đầu ra là một giá trị ẩn h_t . Đầu vào được đấu với một thân mạng nơ ron A có tính chất hồi quy và thân này được đấu tới đầu ra h_t .

37

2.4.2. Hạn chế của mạng nơ ron truy hồi

- Phải thực hiện tuần tự: Không tận dụng được khả năng tính toán song song của máy tính (GPU/TPU).
- Vanishing gradient (Đạo hàm bị triệt tiêu)

Vì hàm kích hoạt (tanh hay sigmoid) của ta sẽ cho kết quả đầu ra nằm trong đoạn $[-1, 1]$ (với sigmoid là $[0, 1]$) nên đạo hàm của nó sẽ bị đóng trong khoảng $[0, 1]$ (với sigmoid là $[0, 0.25]$).

Ở trên, chúng ta đã dùng chain rule để tính đạo hàm. Có một vấn đề ở đây là, hàm tanh lẫn sigmoid đều có đạo hàm bằng 0 tại 2 đầu. Mà khi đạo hàm bằng 0 thì nút mạng tương ứng tại đó sẽ bị bão hòa. Lúc đó các nút phía trước cũng sẽ bị bão hòa theo. Nên với các giá trị nhỏ trong ma trận, khi ta thực hiện phép nhân ma trận sẽ đạo hàm tương ứng sẽ xảy ra Vanishing gradient, tức đạo hàm bị triệt tiêu chỉ sau vài bước nhân. Như vậy, các bước ở xa sẽ không còn tác dụng với nút hiện tại nữa, làm cho RNN không thể học được các phụ thuộc xa. Vấn đề này không chỉ xảy ra với mạng RNN mà ngay cả mạng neural truyền thống với nhiều lớp cũng có hiện tượng này.

Với cách nhìn như trên, ngoài Vanishing gradient, ta còn gặp phải Exploding Gradient (bùng nổ đạo hàm). Tùy thuộc vào hàm kích hoạt và tham số của mạng, vấn đề này xảy ra khi các giá trị của ma trận là lớn (lớn hơn 1). Tuy nhiên, thường nói về vấn đề Vanishing nhiều hơn là Exploding, vì 2 lý do sau:

+ Thứ nhất, bùng nổ đạo hàm có thể theo dõi được vì khi đạo hàm bị bùng nổ thì ta sẽ thu được kết quả là một giá trị phi số NaN làm cho chương trình của ta bị dừng hoạt động.

+ Thứ hai, bùng nổ đạo hàm có thể ngăn chặn được khi ta đặt một ngưỡng giá trị trên (tham khảo kỹ thuật Gradient Clipping). Còn rất khó để theo dõi sự mất mát đạo hàm cũng như tìm cách xử lý nó.

Để xử lý Vanishing Gradient, có 2 cách phổ biến:

+ Cách thứ nhất, thay vì sử dụng activation function là tanh và sigmoid, ta thay bằng ReLu (hoặc các biến thể như Leaky ReLu). Đạo hàm của ReLu

hoặc là 0 hoặc là 1, nên ta có thể kiểm soát phần nào vấn đề mất mát đạo hàm.

+ Cách thứ hai, ta thấy RNN thuần không hề có thiết kế nào để lọc đi những thông tin không cần thiết. Ta cần thiết kế một kiến trúc có thể nhớ dài hạn hơn, đó là LSTM và GRU.

2.5. Mạng LSTM (Long Short-term memory)

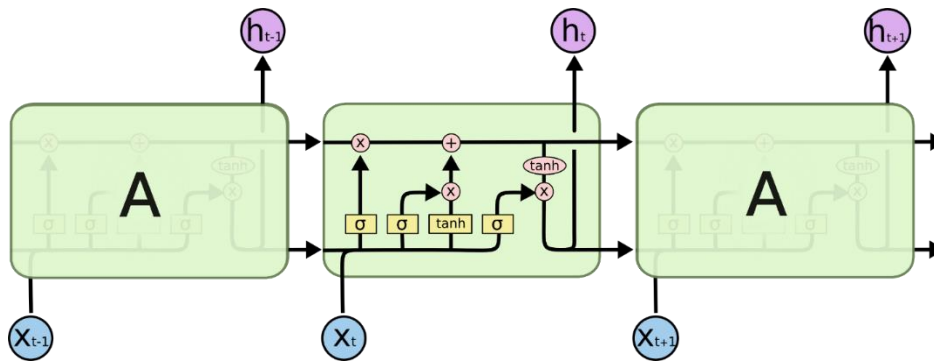
2.5.1. Giới thiệu LSTM

Thực tế RNN chỉ học và truy cập thông tin qua một số trạng thái nhất định, với các trạng thái càng xa với việc tính toán tại thời điểm hiện tại thì nó không thể học được các thông tin đó. Hiện tượng này được gọi là biến mất đạo hàm (vanishing gradient).

LSTM có một cách tính khác cho trạng thái ẩn. Bộ nhớ trong LSTM được gọi là các cells và khi nhận dữ liệu đầu vào cùng trạng thái trước đó, nó sẽ quyết định nên giữ thông tin nào và xóa thông tin nào khỏi bộ nhớ. Nhờ đó mà nó có thể lưu trữ thông tin dài hơn, dễ dàng truy cập lại để giải quyết vấn đề vanishing gradient của RNN.

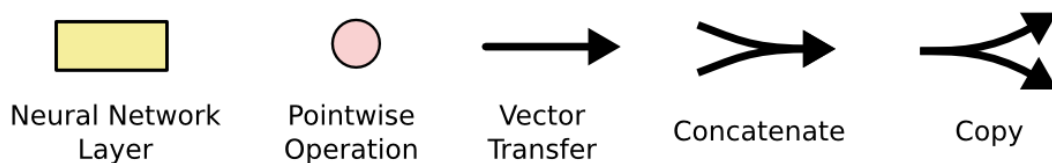
LSTM là một kiến trúc mạng neural hồi quy giống như RNN, ngoại trừ việc cập nhật lớp ẩn được thay thế bởi các ô nhớ được xây dựng có mục đích. Do đó, nó có thể tốt hơn trong việc tìm kiếm và khai thác sự phụ thuộc giữa các dữ liệu trong khoảng dài hơn so với RNN truyền thống.

Một mạng LSTM điển hình bao gồm các khối nhớ khác nhau được gọi là các cells. Có hai trạng thái được chuyển đến cell tiếp theo là cell state (ô trạng thái) và hidden state (trạng thái ẩn). Các khối nhớ chịu trách nhiệm ghi nhớ và thao tác với bộ nhớ này được thực hiện thông qua ba cơ chế chính đó là các cổng (gates) như hình bên dưới:



Hình 2.6 Sơ đồ trong mạng LSTM chứa 4 tầng ẩn

Các kí hiệu có thể chú thích như sau:



Hình 2.7 Ký hiệu trong mạng LSTM

Neural Network được kí hiệu là một hình chữ nhật màu vàng thể hiện hàm activation mà mạng nơ ron sử dụng để học trong tầng ẩn, thông thường là các hàm phi tuyến sigmoid và tanh.

Pointwise Operation được kí hiệu là hình tròn màu hồng biểu diễn một toán tử đối với véc tơ như phép cộng véc tơ (+), phép nhân vô hướng các véc tơ (x).

Vector Transfer được kí hiệu 1 mũi tên thể hiện chuyển nội dung véc tơ đi tới một phần khác của mạng nơ ron.

Concatenate được kí hiệu 2 đường thẳng nhập vào thể hiện phép ghép kết quả.

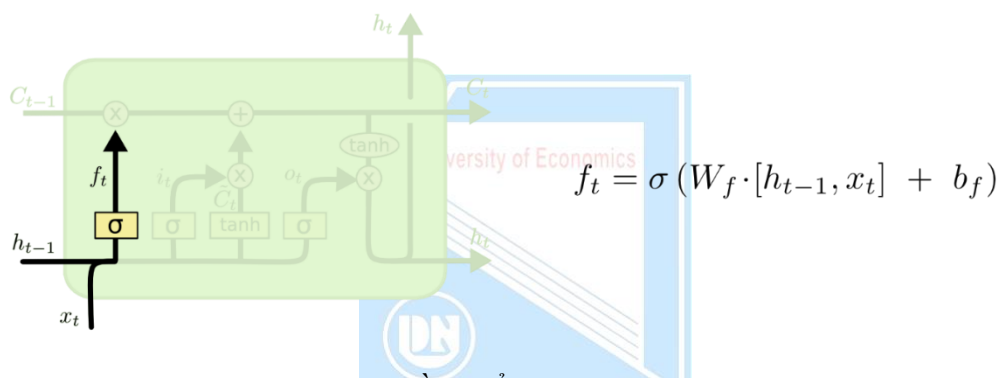
Copy được kí hiệu 2 đường thẳng rẽ nhánh thể hiện cho nội dung véc tơ trước đó được sao chép để đi tới một phần khác của mạng nơ ron.

2.5.2. Thứ tự các bước trong LSTM

Bước đầu tiên trong LSTM sẽ quyết định xem thông tin nào chúng ta sẽ cho phép đi qua ô trạng thái (cell state). Quyết định này được kiểm soát bởi hàm sigmoid trong một tầng gọi là tầng quên (forget gate layer). Đầu tiên nó nhận đầu

vào là 2 giá trị h_{t-1} và x_t trả về một giá trị nằm trong khoảng 0 và 1 cho mỗi giá trị của ô trạng thái C_{t-1} . Nếu giá trị bằng 1 thể hiện “giữ toàn bộ thông tin” và bằng 0 thể hiện “bỏ qua toàn bộ thông tin”.

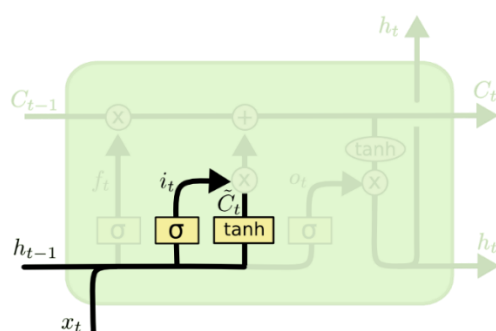
Một ví dụ về sinh văn bản, chúng ta đang cố gắng dự báo từ tiếp theo dựa trên toàn bộ những từ trước đó. Trong những bài toán như vậy, ô trạng thái có thể bao gồm loại của chủ ngữ hiện tại, để cho đại từ ở câu tiếp theo được sử dụng chính xác. Chẳng hạn như chúng ta đang mô tả về một người bạn là nam giới thì các đại từ nhân xưng ở tiếp theo phải là anh, chú, ông thay vì chị, cô, bà ấy. Tuy nhiên chủ ngữ không phải khi nào cũng cố định. Khi chúng ta nhìn thấy một chủ ngữ mới, chúng ta muốn quên đi loại của một chủ ngữ cũ. Do đó tăng quên cho phép cập nhật thông tin mới và lưu giữ giá trị của nó khi có thay đổi theo thời gian.



Hình 2.8 Tầng cổng quên (forget gate layer)

Bước tiếp theo chúng ta sẽ quyết định loại thông tin nào sẽ được lưu trữ trong ô trạng thái. Bước này bao gồm 2 phần. Phần đầu tiên là một tầng ẩn của hàm sigmoid được gọi là tầng cổng vào (input gate layer) quyết định giá trị bao nhiêu sẽ được cập nhật. Tiếp theo, tầng ẩn hàm tanh sẽ tạo ra một vector của một giá trị trạng thái mới (\tilde{C}_t) mà có thể được thêm vào trạng thái. Tiếp theo kết hợp kết quả của 2 tầng này để tạo thành một cập nhật cho trạng thái.

Một ví dụ về sinh văn bản, chúng ta thêm loại của một chủ ngữ mới vào ô trạng thái để thay thế phần trạng thái cũ muốn quên đi.



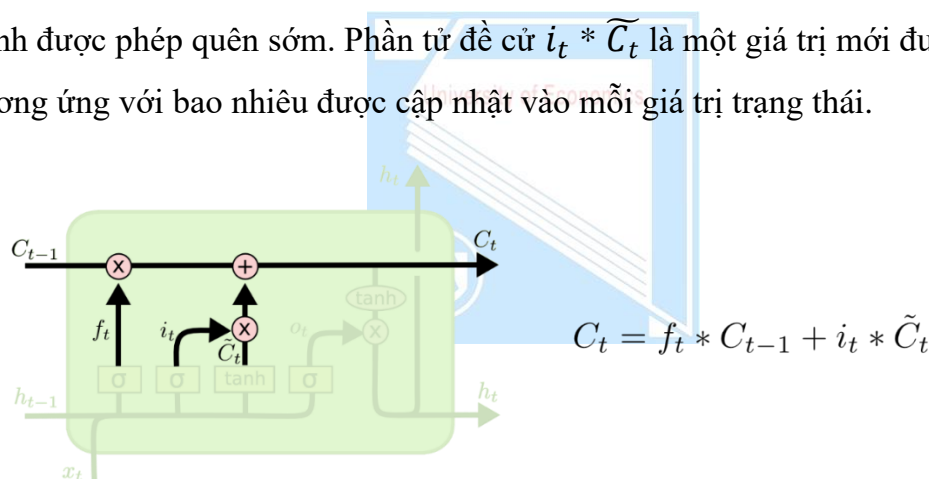
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Hình 2.9 Cập nhật giá trị cho ô trạng thái bằng cách kết hợp 2 kết quả từ tầng cổng vào và tầng ẩn hàm tanh

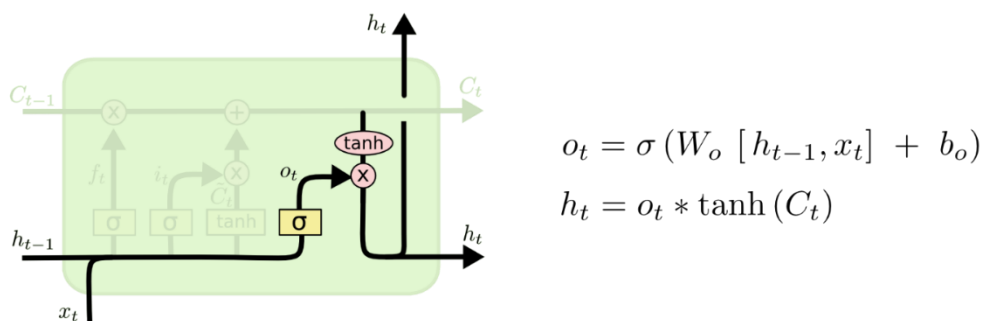
Đây là thời điểm để cập nhật một ô trạng thái cũ C_{t-1} sang một trạng thái mới C_t . Những bước trước đó đã quyết định làm cái gì, và tại bước này chỉ cần thực hiện nó.

Chúng ta nhân trạng thái cũ với f_t tương ứng với việc quên những thứ quyết định được phép quên sớm. Phần tử đề cử $i_t * \tilde{C}_t$ là một giá trị mới được tính toán tương ứng với bao nhiêu được cập nhật vào mỗi giá trị trạng thái.



Hình 2.10 Ô trạng thái mới

Cuối cùng cần quyết định xem đầu ra sẽ trả về bao nhiêu. Kết quả ở đầu ra sẽ dựa trên ô trạng thái, nhưng sẽ là một phiên bản được lọc. Đầu tiên, chúng ta chạy qua một tầng sigmoid nơi quyết định phần nào của ô trạng thái sẽ ở đầu ra. Sau đó, ô trạng thái được đưa qua hàm tanh (để chuyển giá trị về khoảng -1 và 1) và nhân nó với đầu ra của một cổng sigmoid, do đó chỉ trả ra phần mà chúng ta quyết định.



Hình 2.11 Điều chỉnh thông tin ở đầu ra thông qua hàm tanh

2.6. Nhận dạng thực thể (NER)

2.6.1. Định nghĩa

Nhận dạng thực thể (NER) là một kỹ thuật trích xuất thông tin để xác định và phân loại các thực thể được đặt tên trong văn bản. Các thực thể này có thể được xác định trước và chung chung như tên địa điểm, tổ chức, thời gian,... hoặc chúng có thể rất cụ thể như ví dụ với sơ yếu lý lịch.

2.6.2. Phương pháp

Có 2 phương pháp để nhận dạng thực thể là

- (1). Phương pháp tiếp cận cổ điển: chủ yếu dựa trên quy tắc
- (2). Phương pháp học máy: có hai phương pháp chính trong thể loại này:

Coi vấn đề như một phân loại đa lớp, trong đó các thực thể được đặt tên là nhãn của chúng tôi để chúng tôi có thể áp dụng các thuật toán phân loại khác nhau.

Một phương pháp khác trong thể loại này là mô hình trường ngẫu nhiên có điều kiện (CRF). Nó là một mô hình đồ họa xác suất có thể được sử dụng để mô hình hóa dữ liệu tuần tự như nhãn của các từ trong một câu.









(3). Deep Learning Approaches (Phương pháp học sâu)

Một chiến lược quan trọng khác trong việc xây dựng một phương pháp học sâu hiệu suất cao là hiểu loại mạng thần kinh nào hoạt động tốt nhất để giải quyết vấn đề NER vì văn bản là định dạng dữ liệu tuần tự.

Chúng ta cần sử dụng LSTM hai chiều vì sử dụng LSTM tiêu chuẩn để đưa ra dự đoán sẽ chỉ tính đến thông tin "quá khứ" trong một chuỗi văn bản. Đối với NER, vì bối cảnh bao gồm các nhãn trong quá khứ và tương lai theo một trình tự, chúng ta cần tính đến cả thông tin quá khứ và tương lai. LSTM hai chiều là sự kết hợp của hai LSTM - một chạy về phía trước từ "phải sang trái" và một chạy ngược từ "trái sang phải".

2.7. Ngôn ngữ và thư viện mở

2.7.1. Ngôn ngữ Python

Rank	Language	Type	Score
1	Python	  	100.0
2	Java	  	95.4
3	C	  	94.7
4	C++	  	92.4
5	JavaScript		88.1
6	C#	   	82.4
7	R		81.7
8	Go	 	77.7
9	HTML		75.4
10	Swift	 	70.4

Hình 2.12 Xếp hạng ngôn ngữ lập trình năm 2021 [10]

(1). Định nghĩa

Python là một ngôn ngữ lập trình hướng đối tượng bậc cao, là ngôn ngữ mạnh mẽ và rất thân thiện với người dùng. Python có cấu trúc rất rõ ràng, rành mạch và thường được dùng trong lập trình AI. Python do Guido van Rossum tạo ra và ra mắt vào năm 1991. Hiện tại Python đang dẫn đầu trong bảng xếp hạng về ngôn ngữ lập trình.

(2). Lịch sử hình thành

- Python được phát triển bởi Guido van Rossum vào cuối những năm 80 và đầu những năm 90 tại Viện Nghiên cứu Quốc gia về Toán học và Khoa học Máy tính ở Hà Lan.

- Python có nguồn gốc từ nhiều ngôn ngữ khác, bao gồm ABC, Modula-3, C, C ++, Algol-68, SmallTalk và Unix shell và các ngôn ngữ khác.

- Python đã được đăng ký bản quyền. Giống như Perl, mã nguồn Python hiện có sẵn theo Giấy phép Công cộng GNU (GPL).

- Python đang được phát triển bởi một nhóm phát triển cốt lõi tại viện Nghiên cứu Quốc gia về Toán học và Khoa học Máy tính ở Hà Lan, mặc dù Guido van Rossum vẫn giữ một vai trò quan trọng trong việc định hướng sự phát triển của nó.

(3). Các ứng dụng trong Python

- Dễ học: Python có ít từ khóa, cấu trúc đơn giản và cú pháp được xác định rõ ràng. Điều này cho phép học sinh tiếp thu ngôn ngữ một cách nhanh chóng.

- Dễ đọc: Mã Python được xác định rõ ràng hơn và có thể nhìn thấy bằng mắt.

- Dễ bảo trì: Mã nguồn của Python khá dễ bảo trì.

- Thư viện tiêu chuẩn rộng rãi: Phần lớn thư viện của Python di động và tương thích đa nền tảng trên UNIX, Windows và Macintosh.

- Chế độ tương tác: Python có hỗ trợ chế độ tương tác cho phép kiểm tra tương tác và gỡ lỗi các đoạn mã.

- Nền tảng: Python có thể chạy trên nhiều nền tảng phần cứng và có giao diện giống nhau trên tất cả các nền tảng.

- Có thể mở rộng: Có thể thêm các mô-đun cấp thấp vào trình biên dịch Python. Các mô-đun này cho phép người lập trình thêm vào hoặc tùy chỉnh các công cụ của họ để hiệu quả hơn.

- Cơ sở dữ liệu: Python cung cấp giao diện cho tất cả các cơ sở dữ liệu thương mại chính.

- Lập trình GUI: Python hỗ trợ các ứng dụng GUI có thể được tạo và chuyển sang nhiều lệnh gọi hệ thống, thư viện và hệ thống cửa sổ, chẳng hạn như Windows MFC, Macintosh và hệ thống X Window của Unix.

- Khả năng mở rộng - Python cung cấp cấu trúc và hỗ trợ tốt hơn cho các chương trình lớn hơn so với kịch bản shell.

2.7.2. Thư viện *Beautiful Soup*

Beautiful Soup là một trong những thư viện Python phổ biến nhất giúp phân tích cú pháp các tài liệu HTML hoặc XML thành cấu trúc cây để tìm và trích xuất dữ liệu. Công cụ này có giao diện Pythonic đơn giản và chuyển đổi mã hóa tự động để giúp bạn dễ dàng làm việc với dữ liệu trang web.

Thư viện này cung cấp các phương pháp đơn giản và thành ngữ Pythonic để điều hướng, tìm kiếm và sửa đổi cây phân tích cú pháp, đồng thời tự động chuyển đổi tài liệu đến sang Unicode và tài liệu đi sang UTF-8.

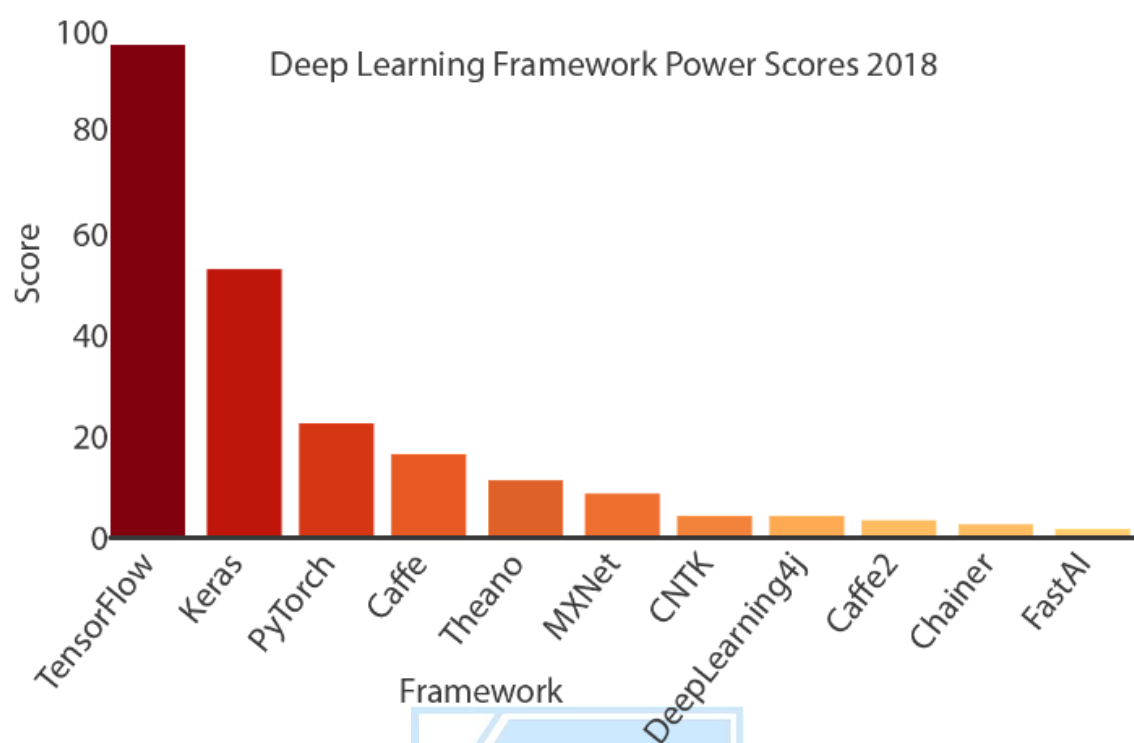
Cài đặt thư viện Beautiful Soup 4 qua dòng lệnh sau:

```
pip install beautifulsoup4
```

Trong Python có thể gọi thư viện BeautifulSoup qua dòng lệnh như sau:

```
from bs4 import BeautifulSoup
```

2.7.3. TensorFlow và Keras



Hình 2.13 Xếp hạng Frameworks sử dụng trong Deep Learning 2018

TensorFlow là nền tảng framework phổ biến nhất về trong các hệ thống Deep Learning trên toàn thế giới ngày nay. Tuy nhiên, TensorFlow khá khó khăn trong việc khai thác để ứng dụng Deep Learning đối với người mới học, do đó Keras có thể khắc phục sự khó khăn này.

Keras như là một “wrapper” của TensorFlow, chạy trên nền tảng là TensorFlow (và cũng có bản cho Theano-một nền tảng khác của DL hiện ít được dùng). Trong Keras, việc xây dựng cấu trúc của Deep Learning cùng với điều chỉnh tham số và chạy chương trình đơn giản hơn rất nhiều. Nhưng nếu muốn can thiệp sâu vào Networks, chẳng hạn thay đổi Cost Function hay thay đổi cấu trúc, làm việc trên Keras sẽ có nhiều khó khăn hơn so với TensorFlow.

2.8. Biểu diễn một từ bằng vector (Word Embedding)

2.8.1. Khái niệm

Word embeddings là một kiểu để biểu diễn từ ngữ mà trong đó, các từ có ngữ nghĩa tương đồng cũng sẽ có cách biểu diễn tương đồng. Sự xuất hiện của

này trái ngược hoàn toàn với việc biểu diễn các từ theo phương pháp thống kê số lượng tần suất xuất hiện của mỗi từ trong câu.



Hình 2.15 Trọng số các từ trong vector

2.8.2. Lớp embeddings

Một lớp embedding, là một word embedding được xây dựng bằng cách dùng chung với mô hình neural network trong quá trình mô hình đó được xây dựng với bài toán xử lý ngôn ngữ tự nhiên nào đó, ví dụ mô hình ngôn ngữ hay phân loại văn bản.

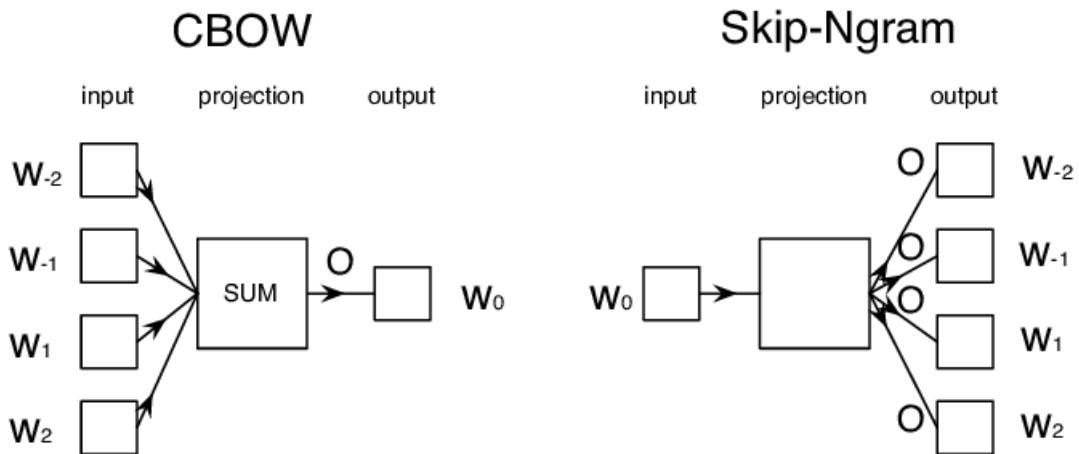
Nó yêu cầu văn bản đầu vào phải được làm sạch và chuẩn bị như một từ là một mã hóa one-hot. Kích thước của không gian vector theo lý thuyết như là một phần của model, có thể bằng 50, 100 hay 300. Những vector được khởi tạo với những con số tự nhiên nhỏ. Lớp embedding sử dụng công việc đang làm của mạng neural network dùng chung đó và điều chỉnh nó theo mục đích công việc trong quá trình backpropagation.

Cách này thường yêu cầu một dữ liệu huấn luyện lớn và có thể sẽ rất chậm nhưng sẽ rất hiệu quả với việc xây dựng word embeddings với dữ liệu văn bản riêng biệt và công việc liên quan đến dữ liệu văn bản đó.

2.8.3. Thuật toán word2vec

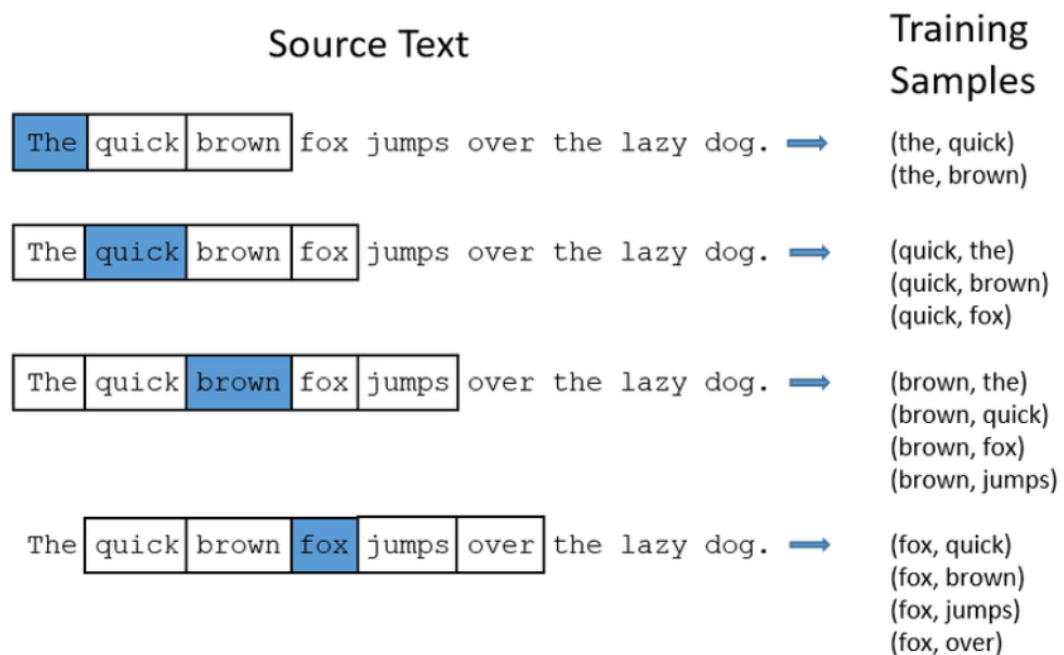
Để biểu diễn ngôn ngữ trong, người ta dùng kỹ thuật là word2vec. Có 2 loại word2vec tùy vào tính chất của bài toán:

- CBOW: dự đoán 1 từ bằng những từ xung quanh nó.
- Skip-Ngram: dự đoán những từ xung quanh một từ.



Hình 2.16 Cbow và Skip-Ngram

Ở đây người ta dùng một cửa sổ trượt gọi là window size để tạo các cặp dữ liệu training, hình dưới là với window size = 2. Ta có các cặp dữ liệu để huấn luyện như sau:



Hình 2.17 Quy trình Word2Vec

Tóm tắt lại quy trình tại word2vec:

- Làm sạch dữ liệu, bỏ dấu câu.
- Xây dựng bộ từ điển cho dữ liệu của chúng tôi, mỗi một từ sẽ có một ID tương ứng là một số nguyên.
- Tiếp theo xây dựng bộ dữ liệu training dựa vào window size.
- Chuyển đổi dữ liệu từ biểu diễn mỗi câu bằng 1 số nguyên về 1 vector với kích thước thường chọn là 300.
- Huấn luyện model bằng CBOW hoặc Skipgram.

2.9. Các phương pháp đánh giá hiệu suất

Có nhiều phương pháp đánh giá hiệu suất của một hệ thống phân lớp như là Accuracy score, confusion matrix, ROC curve, Area Under the Curve, Precision and Recall, F1 score, Top R error,.... Một ví dụ khác hay được dùng của thước đo cho việc đánh giá thuật toán Machine Learning là Precision và Recall.

Việc lựa chọn thước đo để đo lường sự hiệu quả của mô hình Machine Learning là rất quan trọng.

Đầu tiên, tìm hiểu một hình thức tóm tắt hiệu suất mô hình phân lớp – Confusion Matrix.

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Hình 2.18 Các thành phần của Confusion Matrix

Confusion Matrix là 1 bảng phân phối tần số 2 chiều trình bày tỷ lệ tương hợp và bất xứng giữa Thực tế và Kết quả phân lớp của mô hình. Trường hợp với

bài toán của đề tài là một bài toán phân loại nhị phân (Binary Classification), Confusion Matrix trình bày tần suất của 4 tổ hợp: TP, TN, FP, FN. Trong đó:

- True là kết quả phân loại đúng với giá trị thực tế, False ngược lại, chỉ sự phân loại sai.
- Positive, Negative trong bài tương ứng với lớp “Tích cực” và lớp “Tiêu cực”.

Cụ thể hơn:

- TP (True Positive): Trường hợp kết quả phân lớp là “Positive” khớp với thực tế.
- TN (True Negatives): Trường hợp kết quả phân lớp là “Negative” khớp với thực tế.
- FP (False Positives): Trường hợp kết quả phân lớp là “Positive” nhưng thực tế lại là “Negative”
- FN (False Negatives): Trường hợp kết quả phân lớp là “Negative” nhưng thực tế lại là “Positive”

Đây là 4 chỉ số cơ bản để xác định kết quả của 4 chỉ số đánh giá mô hình được nêu ra dưới đây.

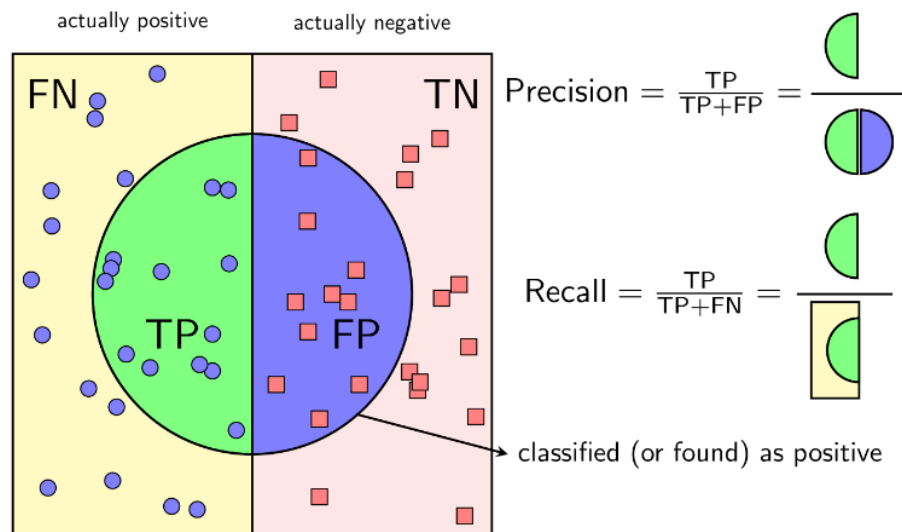
2.9.1. Accuracy

Accuracy – tạm dịch là “Độ chính xác tổng quát”, là tỷ lệ của tất cả các trường hợp phân loại đúng (không phân biệt Positive hay Negative) trên toàn bộ trường hợp kiểm thử.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Đây là tiêu chí phổ biến nhất khi kiểm định hiệu suất mô hình, tuy nhiên giá trị thực dụng không cao vì không đại diện cho mục tiêu cụ thể nào.

2.9.2. Precision và Recall



Hình 2.19 Cách tính Precision và Recall

Precision – hiểu đơn giản là “Có bao nhiêu cái đúng được lấy ra”, cũng là “Độ chính xác” nhưng dùng để đo lường sự chuẩn xác nên tạm dịch là “Khả năng xác định”. Khác với Accuracy, Precision có phân biệt giữa Positive và Negative, gồm PPV (Positive Predictive Value) và NPV (Negative Predictive Value).

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Recall – hiểu đơn giản là “Có bao nhiêu cái lấy ra là đúng”, là tỷ lệ tái hiện, dùng để đo lường tính hữu dụng. Recall là tỉ lệ phân loại Positive đúng trên tổng số các trường hợp Positive, cho biết thông tin hiệu suất phân lớp với mối liên hệ với False Negatives (bao nhiêu cái bị sót).

$$Recall = \frac{TP}{TP + FN}$$

2.9.3. F-Score

- F-Score (hoặc F1-Score) là thước đo có thể biểu diễn thông qua Precision lẫn Recall, cho phép đánh giá cân bằng giữa 2 chỉ số này.
- Harmonic mean (trung bình điều hòa) của Precision lẫn Recall:

$$\frac{2}{F_1} = \frac{1}{precision} + \frac{1}{recall} \text{ hay}$$

$$F_1 = 2 \frac{1}{\frac{1}{precision} + \frac{1}{recall}} = 2 \frac{precision \cdot recall}{precision + recall}$$

- F1-Score có giá trị nằm trong nửa khoảng (0, 1]. F1 càng cao, bộ phận phân lớp càng tốt. Khi cả recall và precision đều bằng 1 (tốt nhất có thể), $F1 = 1$. Khi cả recall và precision đều thấp, ví dụ bằng 0,1 thì $F1 = 0,1$.

2.9.4. Micro-Average và Macro-Average

Macro-Average có thể cho biết hiệu suất hệ thống xuyên suốt bộ dữ liệu và thích hợp với đánh giá mô hình có bộ dữ liệu cân bằng, ngược lại nếu dữ liệu không cân bằng nên dùng Micro-Average. Vì bộ dữ liệu của đề tài có sự phân phối khá đồng đều nên có thể sử dụng cả Micro-Average và Macro-Average để đánh giá mô hình.

Nếu Macro-Average có kết quả thấp hơn đáng kể so với Micro-Average, các nhãn có số lượng nhỏ phân loại hiệu suất không cao, không chính xác bằng các nhãn có số lượng lớn hơn, và ngược lại.

Micro-Average Precision, Micro-Average Recall và Micro-Average F-Score được tính bằng:

$$\text{Micro - Precision} = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C (TP_c + FP_c)}$$

$$\text{Micro - Precision} = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C (TP_c + FN_c)}$$

$$\text{Micro - FScore} = 2 \frac{\text{Micro - Precision} \cdot \text{Micro - Precision}}{\text{Micro - Precision} + \text{Micro - Precision}}$$

Với $C = 2$ cho bài toán đang thực hiện

Macro-Average Precision, Macro-Average Recall chính là trung bình cộng các Precision, Recall theo lớp. Bài toán đang thực hiện có 2 lớp nên:

$$\text{Macro - Precision} = \frac{Precision1 + Precision2}{2}$$

$$\text{Macro - Recall} = \frac{\text{Recall1} + \text{Recall2}}{2}$$

$$\text{Macro - FScore} = 2 \frac{\text{Macro - Precision} \cdot \text{Macro - Precision}}{\text{Macro - Precision} + \text{Macro - Precision}}$$



CHƯƠNG 3. TRIỂN KHAI HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ SỬ DỤNG AI

3.1. Tìm kiếm và thu thập dữ liệu

3.1.1. Thông tin về bộ dữ liệu

Trong các mô hình sử dụng Trí tuệ nhân tạo (AI), thì dữ liệu được đánh giá là một phần rất quan trọng. Để có một mô hình sàng lọc các ứng viên tốt nhất thì cần đến nhiều yếu tố khác nhau, một trong những yếu tố cần thiết nhất đó là dữ liệu về sơ yếu lý lịch của các ứng viên và thông tin tuyển dụng của các công ty. Hiện nay, nguồn dữ liệu về sơ yếu lý lịch và thông tin tuyển dụng có khá nhiều. Nguồn dữ liệu này tập trung chủ yếu ở các mã nguồn mở như các trang web tìm việc làm (một số nguồn như job spider, glassdoor, indeed, postjobfree,...), ngoài ra các sơ yếu lý lịch còn được lưu trữ ở các công ty tuyển dụng.

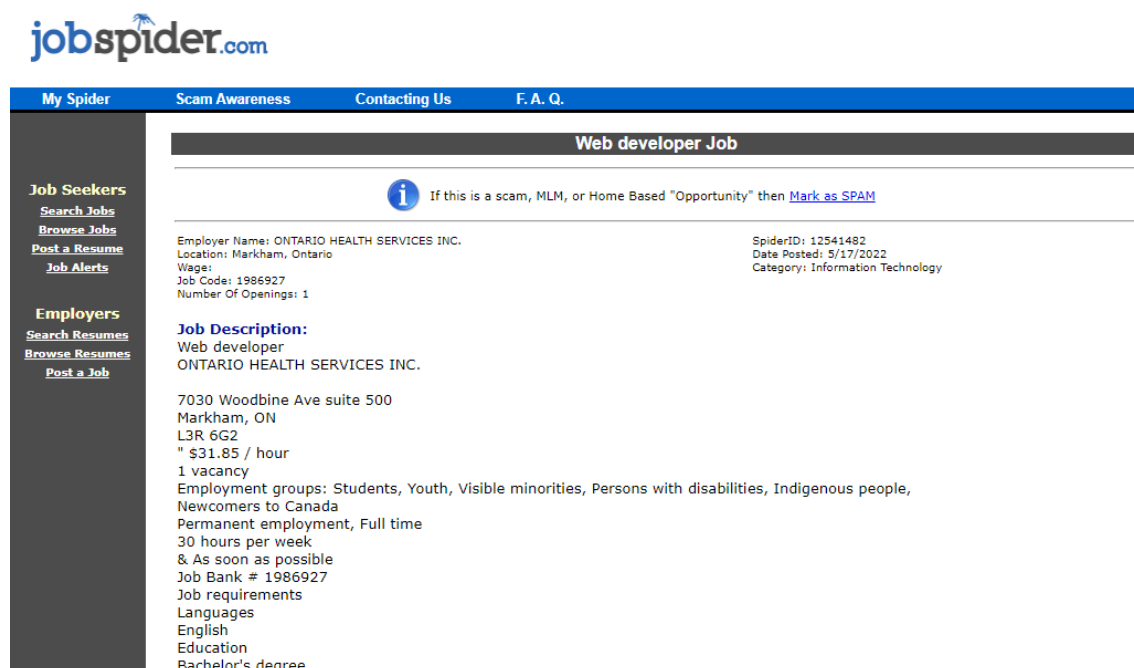
Sau khi tìm kiếm các nguồn mở dữ liệu về sơ yếu lý lịch (CV/resumes) của các ứng viên và mô tả công việc (Job Descriptions) của nhà tuyển dụng liên quan đến lĩnh vực Công nghệ thông tin (IT). Tôi đã chọn trang web Jobs Spider (jobspider.com) là nơi thu thập dữ liệu. Jobs Spider là trang web tìm kiếm việc làm và đăng tuyển dụng miễn phí số 1 tại Hoa Kỳ và Canada.

(1). Đối với dữ liệu sơ yếu lý lịch (CV/resumes)

Sau khi thu thập (crawl) dữ liệu từ trang web Jobs Spider và phân tích từng sơ yếu lý lịch, tôi nhận thấy thông tin trong mỗi sơ yếu lý lịch đã được phân đoạn thành các nội dung cụ thể, ứng với từng nội dung là những trường được định sẵn từ trước. Điều này khác xa với dữ liệu thực tế, khiến cho mô hình dự đoán sai hoặc độ chính xác không cao gặp phải các sơ yếu lý lịch không cùng cấu trúc với tập dữ liệu khi train model.

Do vậy, công ty TMA Bình Định đã cung cấp dữ liệu sơ yếu lý lịch là dữ liệu thực tế gồm 894 CVs, thuộc định dạng pdf, được viết bằng tiếng anh hoặc tiếng việt và bố cục nội dung các mục trong sơ yếu lý lịch của các ứng viên trình bày là không giống nhau.

(2). Đối với dữ liệu mô tả công việc (Job Descriptions)



The screenshot shows the jobspider.com website interface. The top navigation bar includes links for 'My Spider', 'Scam Awareness', 'Contacting Us', and 'F. A. Q.'. The left sidebar contains sections for 'Job Seekers' (Search Jobs, Browse Jobs, Post a Resume, Job Alerts) and 'Employers' (Search Resumes, Browse Resumes, Post a Job). The main content area displays a job listing for 'Web developer Job' at 'ONTARIO HEALTH SERVICES INC.'. The listing includes details such as location (Markham, Ontario), wage (\$31.85 / hour), and job code (1986927). It also features a 'Job Description' section with information about the role, employment groups, and requirements.

jobspider.com

My Spider Scam Awareness Contacting Us F. A. Q.

Web developer Job

If this is a scam, MLM, or Home Based "Opportunity" then [Mark as SPAM](#)

Employer Name: ONTARIO HEALTH SERVICES INC.
Location: Markham, Ontario
Wage: \$31.85 / hour
Job Code: 1986927
Number Of Openings: 1

SpiderID: 12541482
Date Posted: 5/17/2022
Category: Information Technology

Job Description:
Web developer
ONTARIO HEALTH SERVICES INC.
7030 Woodbine Ave suite 500
Markham, ON
L3R 6G2
" \$31.85 / hour
1 vacancy
Employment groups: Students, Youth, Visible minorities, Persons with disabilities, Indigenous people, Newcomers to Canada
Permanent employment, Full time
30 hours per week
& As soon as possible
Job Bank # 1986927
Job requirements
Languages
English
Education
Bachelor's degree

Hình 3.1 Minh họa một mô tả công việc tại trang web Jobs Spider

Dữ liệu mô tả công việc của các công ty sẽ được thu thập dữ liệu từ trang web Jobs Spider (jobspider.com), bộ dữ liệu mô tả công việc gồm 1500 JDs và thường có 4 mục nội dung cố định trong mỗi mô tả công việc của nhà tuyển dụng.

Job Name

Job Description

Content

Job Criteria

Content

Contact Information

Content

Hình 3.2 Minh họa 4 mục nội dung có trong thông tin mô tả công việc tại trang web Jobs Spider

Bảng 3.1 Nội dung của các mục trong thông tin mô tả công việc

Tên của mục	Nội dung của mục
Job Name	Tên công việc tuyển dụng
Job Description	Mô tả công việc có nội dung như các yêu cầu về trình độ chuyên môn, kỹ năng, kinh nghiệm làm việc của ứng viên.
Job Criteria	Tiêu chí công việc có nội dung như ngày bắt đầu công việc, loại vị trí (full time, part time), số năm kinh nghiệm, trình độ học vấn.
Contact Information	Thông tin liên hệ có nội dung như tên, địa chỉ, trang web của công ty tuyển dụng.

3.1.2. Thu thập dữ liệu (Data Crawling)

Sử dụng thư viện BeautifulSoup của Python để Crawl dữ liệu mô tả công việc (Job Descriptions) qua đó làm bộ dữ liệu huấn luyện cho hệ thống.

Dữ liệu mô tả công việc từ trang web Jobs Spider sau khi Crawl sẽ có 1500 JDs và được lưu trữ ở định dạng (.docx) thể hiện ở (Hình 3.3).

User Support Technician Job

Employer Name: Nationwide Natural Foods

SpiderID: 12586504

Location: Vancouver, British Columbia

Date Posted: 5/30/2022

Wage: \$35/hr

Category: Information Technology

Job Code:

Number Of Openings: 1

Nationwide Natural Foods is currently seeking a User Support Technician provide first-line technical support to employees (computer users) experiencing difficulties with computer hardware and with computer applications and communications software.

Job duties include, but are not limited to:

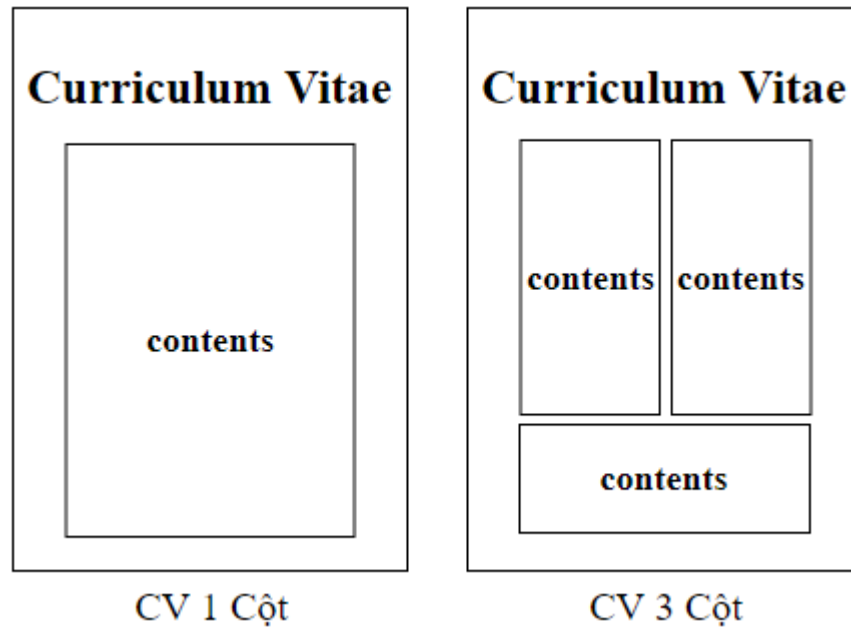
~ Communicating electronically and in person with computer users experiencing difficulties to determine and document problems experienced

Hình 3.3 Minh họa một mô tả công việc sau khi crawl

3.2. Phân tích cú pháp sơ yếu lý lịch (CV Parsing)

3.2.1. Dữ liệu sơ yếu lý lịch

Trong đề tài này, tôi thực hiện phân tích cú pháp những sơ yếu lý lịch trên cùng một format là sơ yếu lý lịch phải được viết bằng tiếng Anh và có 1 cột.



Hình 3.4 Minh hoạt CV 1 cột và nhiều cột

CURRICULUM VITAE
NGUYEN NGOC ANH

POSITION APPLIED	SOFTWARE ENGINEER	
PERSONAL DATA	Gender: Male Mobile: (+84)1674 614 460 Email: nguyenanhpx@gmail.com Data of Birth: 11/08/1988 Place of Birth: Lam Dong province Marital Status: Single	
OBJECTIVE	<ul style="list-style-type: none"> – Short-term goals: <ul style="list-style-type: none"> ➢ Find a position where knowledge and skills trained in University will be used efficiently. ➢ Make an obvious contribution to the development of Company – Long-term goals: <ul style="list-style-type: none"> ➢ To become an expert in software engineering field ➢ Try hard to find an opportunity for promotion 	
EDUCATION BACKGROUND	<div style="display: flex; justify-content: space-between;"> <div style="width: 30%;"> <i>From 09/2008 to 10/2012</i> </div> <div style="width: 65%;"> University of Science Ho Chi Minh City Bachelor of Information Technology GPA 7.03 </div> </div>	
Skills		

<i>Programming Languages</i>	<i>C, C++, XML, Html, Css: Medium</i> <i>.NET/C#, JavaScript, JQuery: Advance</i>
<i>Database</i>	<i>MS Access, MS SQL Server 2005 and 2008</i>
<i>Web programming</i>	<i>Asp.net, Webservices</i>
<i>Languages</i>	<i>English (upper-intermediate level)</i>

Hình 3.5 Minh hoạt CV đủ điều kiện lựa chọn

Dữ liệu sơ yếu lý lịch ban đầu có 894 CVs. Dữ liệu sơ yếu lý lịch này sẽ được lọc thủ công chọn ra các sơ yếu lý lịch được viết bằng tiếng Anh và có 1 cột. Sau khi lọc thủ công, số lượng sơ yếu lý lịch đảm bảo điều kiện là 450 CVs.

3.2.2. Chuyển đổi từ PDF sang TXT



Hình 3.6 Minh họa chuyển đổi pdf sang text

Hiện nay định dạng PDF được các ứng viên sử dụng phổ biến nhất khi gửi CV của mình đến nhà tuyển dụng. Một trong số các ưu điểm các ứng viên sử dụng phổ biến định dạng này là PDF giúp bảo toàn được bố cục và thiết kế của CV, với định dạng này font chữ, cỡ chữ sẽ được giữ nguyên không bị thay đổi hiển thị so với định dạng khác. Ngoài ra CV được viết bằng định dạng PDF có thể hiển thị trên nhiều thiết bị thông minh khác nhau mà nhà tuyển dụng không cần thiết lập gì thêm để đọc được nó.

Tuy nhiên khi nghiên cứu, việc thao tác hay đọc trực tiếp từ định dạng PDF làm thay đổi bố cục của CV, một số ít không thể đọc hết nội dung bên trong. Việc chuyển đổi từ định dạng PDF sang TXT là lựa chọn cần thiết, việc này đảm bảo cho bố cục và nội dung bên trong không bị thay đổi. Sau khi thử nghiệm nhiều thư viện Python giúp chuyển đổi sang định dạng văn bản như pdfminer, PyPDF2, pymupdf. Tôi đã chọn thư viện pymupdf để chuyển đổi từ định dạng pdf sang text, thư viện này cho ra kết quả chuyển đổi tốt hơn so với 2 thư viện trên.

Trinh Dinh Phuc

Curriculum Vitae

16 Kha Van Can Str.
Linh Dong, Thu Duc, HCMC
☎ 0121 658 5084
✉ Phucco996@gmail.com
DOB: March 16, 1996



Education

- 2014–2018 **Bachelor's Degree of Information Technology**, Department of Computer Science, Telecommunications University (TCU), Nha Trang, Khanh Hoa, Vietnam.
- CGPA: 3.35/4 via 204 credits.
 - Thesis: *Pneumonia Diagnosis using Lung's XRay with Depthwise Convolution*. Final grade: 91/100 (top of the class).

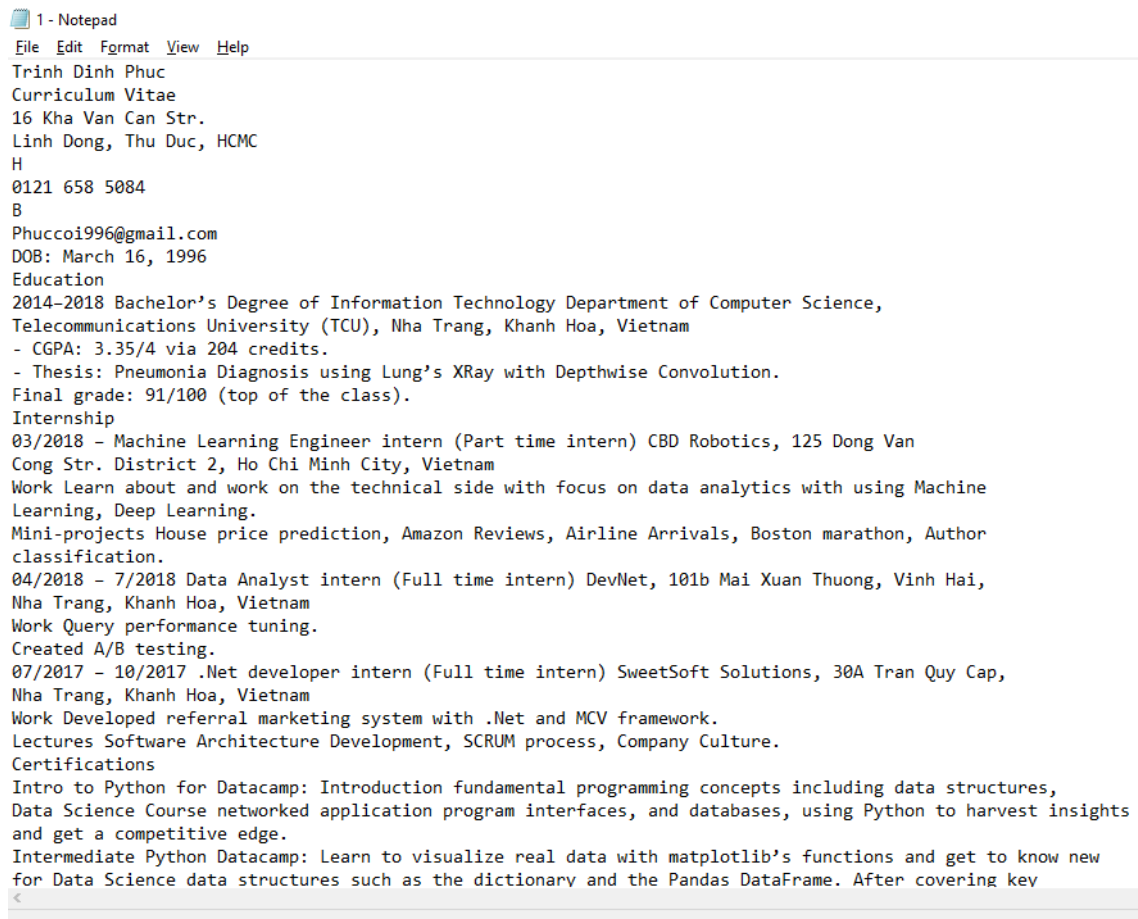
Internship

- 03/2018 – **Machine Learning Engineer intern (Part time intern)**, CBD Robotics, 125 Dong Van Cong Str. District 2, Ho Chi Minh City, Vietnam.
- Work Learn about and work on the technical side with focus on data analytics with using Machine Learning, Deep Learning.
- Mini-projects House price prediction, Amazon Reviews, Airline Arrivals, Boston marathon, Author classification.
- 04/2018 – 7/2018 **Data Analyst intern (Full time intern)**, DevNet, 101b Mai Xuan Thuong, Vinh Hai, Nha Trang, Khanh Hoa, Vietnam.
- Work Query performance tuning.
Created A/B testing.
- 07/2017 – 10/2017 **.Net developer intern (Full time intern)**, SweetSoft Solutions, 30A Tran Quy Cap, Nha Trang, Khanh Hoa, Vietnam.
- Work Developed referral marketing system with .Net and MCV framework.
- Lectures Software Architecture Development, SCRUM process, Company Culture.

Certifications

- Intro to Python for Data Science Course Datacamp: Introduction fundamental programming concepts including data structures, networked application program interfaces, and databases, using Python to harvest insights and get a competitive edge.
- Intermediate Python for Data Science Course Datacamp: Learn to visualize real data with matplotlib's functions and get to know new data structures such as the dictionary and the Pandas DataFrame. After covering key concepts such as boolean logic, control flow and loops in Python.
- Deep Learning in Python Course Datacamp: Gain hands-on, practical knowledge of how to use deep learning with Keras 2.0. Optimization a neural network with backward propagation, building deep learning models, Fine-tuning keras model.
- Python for Data Science edX: Applied experience in major areas of Machine Learning, including Prediction, Classification, Clustering, and Information Retrieval.

Hình 3.8 CV trước khi chuyển đổi thuộc định dạng pdf



1 - Notepad

File Edit Format View Help

Trinh Dinh Phuc
Curriculum Vitae
16 Kha Van Can Str.
Linh Dong, Thu Duc, HCMC
H
0121 658 5084
B
Phuccoi996@gmail.com
DOB: March 16, 1996
Education
2014-2018 Bachelor's Degree of Information Technology Department of Computer Science,
Telecommunications University (TCU), Nha Trang, Khanh Hoa, Vietnam
- CGPA: 3.35/4 via 204 credits.
- Thesis: Pneumonia Diagnosis using Lung's XRay with Depthwise Convolution.
Final grade: 91/100 (top of the class).
Internship
03/2018 - Machine Learning Engineer intern (Part time intern) CBD Robotics, 125 Dong Van
Cong Str. District 2, Ho Chi Minh City, Vietnam
Work Learn about and work on the technical side with focus on data analytics with using Machine
Learning, Deep Learning.
Mini-projects House price prediction, Amazon Reviews, Airline Arrivals, Boston marathon, Author
classification.
04/2018 - 7/2018 Data Analyst intern (Full time intern) DevNet, 101b Mai Xuan Thuong, Vinh Hai,
Nha Trang, Khanh Hoa, Vietnam
Work Query performance tuning.
Created A/B testing.
07/2017 - 10/2017 .Net developer intern (Full time intern) SweetSoft Solutions, 30A Tran Quy Cap,
Nha Trang, Khanh Hoa, Vietnam
Work Developed referral marketing system with .Net and MCV framework.
Lectures Software Architecture Development, SCRUM process, Company Culture.
Certifications
Intro to Python for Datacamp: Introduction fundamental programming concepts including data structures,
Data Science Course networked application program interfaces, and databases, using Python to harvest insights
and get a competitive edge.
Intermediate Python Datacamp: Learn to visualize real data with matplotlib's functions and get to know new
for Data Science data structures such as the dictionary and the Pandas DataFrame. After covering key

Hình 3.9 CV sau khi chuyển đổi sang định dạng txt

3.2.3. Gán nhãn dữ liệu (Label Data)

Sau khi chuyển tất cả CV sang văn bản, tôi sẽ phân đoạn văn bản (text segment) và gán nhãn thủ công cho toàn bộ CV này. Mục đích của bước này để làm đầu vào cho mô hình và đánh giá lại mô hình phân đoạn (model segment).

	A	B	C
1	cv	line	tag
2	1.pdf	Trinh Dinh Phuc	B-I
3	1.pdf	Curriculum Vitae	I-I
4	1.pdf	16 Kha Van Can Str.	I-I
5	1.pdf	Linh Dong, Thu Duc, HCMC	I-I
6	1.pdf	H	I-I
7	1.pdf	0121 658 5084	I-I
8	1.pdf	B	I-I
9	1.pdf	Phuccoi996@gmail.com	I-I
10	1.pdf	DOB: March 16, 1996	I-I
11	1.pdf	Education	B-E
12	1.pdf	2014-2018 Bachelor's Degree of Information Technology Department	I-E
13	1.pdf	Telecommunications University (TCU), Nha Trang, Khanh Hoa, Vietnam	I-E
14	1.pdf	- CGPA: 3.35/4 via 204 credits.	I-E
15	1.pdf	- Thesis: Pneumonia Diagnosis using Lung's XRay with Depthwise Convo	I-E
16	1.pdf	Final grade: 91/100 (top of the class).	I-E
17	1.pdf	Internship	B-W
18	1.pdf	03/2018 - Machine Learning Engineer intern (Part time intern) CBD Robot	I-W
19	1.pdf	Cong Str. District 2, Ho Chi Minh City, Vietnam	I-W
20	1.pdf	Work Learn about and work on the technical side with focus on data analyti	I-W

Hình 3.10 Bộ dữ liệu của CV sau khi gán nhãn thủ công

Bộ dữ liệu sơ yếu lý lịch sau khi gán nhãn thủ công và được lưu vào file csv. Bộ dữ liệu bao gồm 3 trường:

- cv: số thứ tự sơ yếu lý lịch.
- line: dòng nội dung của sơ yếu lý lịch.
- tag: các nhãn gán thủ công được phân đoạn (segment) gồm 7 mục là Personal Information, Education, Experience, Skill, Objective, None.

Bảng 3.2 Chú thích các loại nhãn trong sơ yếu lý lịch (CV/resumes)

Loại nhãn	Ý nghĩa
Personal Information	Thông tin cá nhân
B-I	Dòng đầu của thông tin cá nhân
I-I	Dòng thứ hai đến cuối nội dung thông tin cá nhân
Education	Trình độ học vấn
B-E	Dòng đầu của trình độ học vấn

I-E	Dòng thứ hai đến cuối nội dung trình độ học vấn
Experience	Kinh nghiệm làm việc
B-W	Dòng đầu của kinh nghiệm làm việc
I-W	Dòng thứ hai đến cuối nội dung kinh nghiệm làm việc
Skill	Kỹ năng
B-S	Dòng đầu của kỹ năng
I-S	Dòng thứ hai đến cuối nội dung kỹ năng
Objective	Mục tiêu trong công việc
B-SU	Dòng đầu của kỹ năng
I-SU	Dòng thứ hai đến cuối nội dung kỹ năng
None	Không xác định
*B-(Begin) biểu thị phần đầu của nội dung, I-(Inside) biểu thị phần bên trong nội dung.	

3.2.4. Tiền xử lý dữ liệu

Đây là bước không thể thiếu trong việc xử lý ngôn ngữ tự nhiên. Tiền xử lý dữ liệu qua 3 bước bao gồm loại bỏ ký tự đặc biệt, chuyển đổi sang chữ thường, loại bỏ khoảng trắng ở hai đầu.

(1). Loại bỏ ký tự đặc biệt

Các ký tự đặc biệt thường không thêm giá trị cho văn bản, ký tự đặc biệt có thể được xóa khỏi văn bản để loại trừ bất kỳ thông tin thừa nào đang được xử lý thông qua các thuật toán xử lý ngôn ngữ tự nhiên (NLP). Ngoài ra khi phân tích cú pháp trong văn bản còn xuất hiện các ký tự cụ thể và các từ viết tắt.

Sử dụng gói chuỗi các dấu câu ('string' package) bằng cách gọi thư viện punctuation, punctuation sẽ cung cấp tất cả bộ dấu câu !"#\$%&'()*+,-


Việc chuyển đổi tất cả dữ liệu sang chữ thường sẽ giúp ích cho quá trình tiền xử lý và trong các giai đoạn sau của quá trình phân tích cú pháp (Parsing).

`lower()` Phương thức trả về một chuỗi trong đó tất cả các ký tự đều là chữ thường. Chuyển đổi chữ thường sẽ được sử dụng sau khi loại bỏ ký tự đặc biệt.

(3). Loại bỏ khoảng trắng ở hai đầu chuỗi

`strip()` Phương thức loại bỏ mọi ký tự đứng đầu (khoảng trắng ở đầu) và ký tự ở cuối (khoảng trắng ở cuối). Chuỗi ban đầu sẽ không thay đổi. Nếu không có khoảng trắng đầu hoặc cuối cần xóa, thì chuỗi gốc sẽ được trả về. Cả dấu cách và ký tự tab sẽ bị xóa. Loại bỏ khoảng trắng ở hai đầu chuỗi được sử dụng sau khi chuyển đổi văn bản sang chữ thường.

Dữ liệu CV trước khi qua bước tiền xử lý thể hiện như (Hình 3.12) và sau khi tiền xử lý như (Hình 2.13).



```
['Trinh Dinh Phuc',  
'Curriculum Vitae',  
'16 Kha Van Can Str.',  
'Linh Dong, Thu Duc, HCMC',  
'H',  
'0121 658 5084',  
'B',  
'Phuccoi996@gmail.com',  
'DOB: March 16, 1996',  
'Education',  
'2014-2018 Bachelor's Degree of Information Technology Department of Computer Science',  
'Telecommunications University (TCU), Nha Trang, Khanh Hoa, Vietnam',  
'- CGPA: 3.35/4 via 204 credits.',  
' Thesis: Pneumonia Diagnosis using Lung's XRay with Depthwise Convolution.',  
'Final grade: 91/100 (top of the class).',  
'Internship',  
'03/2018 - Machine Learning Engineer intern (Part time intern) CBD Robotics, 125 Dong Van',  
'Cong Str. District 2, Ho Chi Minh City, Vietnam',  
'Work Learn about and work on the technical side with focus on data analytics with using Machine',  
'Learning, Deep Learning.',  
'Mini-projects House price prediction, Amazon Reviews, Airline Arrivals, Boston marathon, Author',  
'classification.',  
'04/2018 - 7/2018 Data Analyst intern (Full time intern) DevNet, 101b Mai Xuan Thuong, Vinh Hai',  
'Nha Trang, Khanh Hoa, Vietnam',  
'Work Query performance tuning.',  
'Created A/B testing.',
```

Hình 3.12 Dữ liệu CV trước khi qua bước tiền xử lý

```
[
    'trinh dinh phuc',
    'curriculum vitae',
    '16 kha van can str',
    'linh dong thu duc ho chi minh city',
    'h',
    '0121 658 5084',
    'b',
    'phuccoi996 gmail com',
    'dob march 16 1996',
    'education',
    '2014 2018 bachelor s degree of information technology department of computer science',
    'telecommunications university tcu nha trang khanh hoa vietnam',
    'cgpa 3 35 4 via 204 credits',
    'thesis pneumonia diagnosis using lung s xray with depthwise convolution',
    'final grade 91 100 top of the class',
    'internship',
    '03 2018 machine learning engineer intern part time intern cbd robotics 125 dong van',
    'cong str district 2 ho chi minh city vietnam',
    'work learn about and work on the technical side with focus on data analytics with using machine',
    'learning deep learning',
    'mini projects house price prediction amazon reviews airline arrivals boston marathon author',
    'classification',
    '04 2018 7 2018 data analyst intern full time intern devnet 101b mai xuan thuong vinh hai',
    'nha trang khanh hoa vietnam',
    'work query performance tuning',
    'created a b testing',

```

Hình 3.13 Dữ liệu CV sau khi qua bước tiền xử lý

3.2.5. Tách từ

Tách từ (Tokenization) là một trong những bước quan trọng nhất trong quá trình tiền xử lý văn bản. Kỹ thuật Tokenization chia một đoạn văn bản thành các từ dựa trên dấu phân cách (space). Đề tài này sử dụng bộ công cụ ngôn ngữ tự nhiên (NLTK) là thư viện phổ biến nhất để xử lý ngôn ngữ tự nhiên (NLP) và được viết bằng Python.

Có thể cài đặt NLTK khi sử dụng Window bằng lệnh:

```
!pip install nltk
```

Khi bạn đã cài đặt NLTK, gọi và cài đặt các gói NLTK bằng cách chạy lệnh sau:

```
import nltk

nltk.download()
```

Kết quả thu được là các dòng CV được tách thành các từ dựa trên dấu phân cách thể hiện như (Hình 3.14).

```
[['trinh', 'dinh', 'phuc'],  
 ['curriculum', 'vitae'],  
 ['16', 'kha', 'van', 'can', 'str'],  
 ['linh', 'dong', 'thu', 'duc', 'ho', 'chi', 'minh', 'city'],  
 ['h'],  
 ['0121', '658', '5084'],  
 ['b'],  
 ['phuccoi996', 'gmail', 'com'],  
 ['dob', 'march', '16', '1996'],  
 ['education'],  
 ['2014',  
  '2018',  
  'bachelor',  
  's',  
  'degree',  
  'of',  
  'information',  
  'technology',  
  'department',  
  'of',  
  'computer',  
  'science'],  
 ['telecommunications',  
  'university',  
  'tcu',  
  'nha',  
  'trang',  
  'khanh',  
  'hoa',  
  'vietnam'],  
 ['cgpa', '3', '35', '4', 'via', '204', 'credits'],
```

Hình 3.14 Dữ liệu sau khi được tách từ bằng NLTK

3.2.6. Xây dựng bộ từ vựng

Ở bước này tôi sử dụng phương pháp word2vec để xây dựng bộ từ vựng cho dữ liệu của mình, word2vec sẽ mã hóa mỗi từ bằng 1 ID tương ứng với 1 số nguyên. Mỗi từ vựng (vocabulary) là 1 từ duy nhất trong bộ từ vựng, các từ trong bộ từ vựng có thể được sắp xếp theo thứ tự bảng chữ cái, dựa trên tần số của chúng hoặc có thể không và điều đó không ảnh hưởng đến việc đào tạo mô hình.

```
model = Word2Vec(window=5, min_count=1, workers=4)  
  
model.build_vocab(word_lst, progress_per=1000)  
  
model.train(word_lst, total_examples=model.corpus_count, epochs=1000)
```

Hình 3.15 Các chỉ số khi xây dựng bộ từ vựng

Các chỉ số khi xây dựng bộ từ vựng: window=5, min_count=1, workers=4, progress_per=1000, word_lst, total_examples=model.corpus_count, epochs=1000. Ý nghĩa các chỉ số:

- + window: khoảng cách tối đa của từ hiện tại và từ predicted.
- + min_count: lọc bỏ tất cả các từ khỏi bộ từ vựng có số lần xuất hiện nhỏ hơn 1.
- + workers: Số worker threads.
- + progress_per: 1000 từ cần xử lý trước khi hiển thị/cập nhật tiến trình.
- + word_lst: danh sách từ khi thực hiện bước tokenization.
- + total_examples: sử dụng cùng một danh sách, việc sử dụng số lượng được lưu trong bộ nhớ cache trong mô hình.
- + epochs: 1000 lần duyệt qua hết các dữ liệu trong tập huấn luyện.

Lưu và gọi lại model vocabulary bằng 2 lệnh sau thể hiện như (Hình 3.16).

```
model.save('/content/drive/MyDrive/ Nguyen Dinh Phuc Dai/word2vec_v8.bin')  
  
wv = KeyedVectors.load('/content/drive/MyDrive/ Nguyen Dinh Phuc Dai/word2vec_v8.bin', mmap='r')
```

Hình 3.16 Lệnh lưu và gọi lại model vocabulary

Kết quả của bước này là bộ từ vựng gồm các từ duy nhất tương ứng với mỗi từ là 1 ID thể hiện như (Hình 3.17).

```
{'and': 0,  
'of': 1,  
'to': 2,  
'the': 3,  
'in': 4,  
'for': 5,  
'a': 6,  
'with': 7,  
'project': 8,  
'skills': 9,  
'minh': 10,  
'c': 11,  
'university': 12,  
'ho': 13,  
'system': 14,  
'chi': 15,  
'com': 16,  
'work': 17,  
'i': 18,  
'on': 19,  
'software': 20,  
'city': 21,  
'experience': 22,  
'data': 23,  
'at': 24,  
'my': 25,  
'as': 26,
```

Hình 3.17 Bộ từ vựng tương với ID mỗi từ

Sau khi hoàn thành bước xây dựng bộ từ vựng biểu diễn bằng 1 ID, kết hợp với bước tách từ trước đó. Kết quả thu được là dữ liệu CV được mã hóa hay định danh token trên không gian số theo từng dòng của bộ dữ liệu thể hiện như (Hình 3.18).

```
[array([1363, 575, 1155]),
 array([253, 271]),
 array([ 501, 3516, 234, 95, 2146]),
 array([921, 522, 456, 337, 13, 15, 10, 21]),
 array([973]),
 array([8357, 8356, 8355]),
 array([341]),
 array([6967, 50, 16]),
 array([1262, 491, 501, 965]),
 array([40]),
 array([ 55, 44, 190, 154, 295, 1, 32, 31, 290, 1, 73, 68]),
 array([ 518, 12, 8350, 1765, 1346, 1145, 316, 34]),
 array([4286, 72, 2040, 81, 416, 6971, 8346]),
 array([ 375, 6972, 3495, 37, 4206, 154, 4159, 7, 6973, 4146]),
 array([ 942, 790, 2537, 836, 622, 1, 3, 427]),
 array([239]),
 array([ 577, 44, 155, 63, 92, 348, 373, 107, 348, 8362, 1304,
        6976, 522, 234]),
 array([ 922, 2146, 90, 56, 13, 15, 10, 21, 34]),
 array([ 17, 129, 60, 0, 17, 19, 3, 88, 1221, 7, 1413,
        19, 23, 707, 7, 37, 155]),
 array([ 63, 306, 63]),
 array([2075, 76, 1255, 1291, 1378, 1130, 2541, 8375, 8372, 8371, 6981,
        4427]),
 array([985]),
```

Hình 3.18 Dữ liệu được định danh token

3.3. Xây dựng mô hình phân đoạn (Build model segment)

Nhiệm vụ của bước này là phân đoạn CV thành các mục nội dung của Personal Information, Education, Experience, Skill, Objective. Sử dụng 6 layer là Input, Embedding, LSTM, Bidirectional LSTM, CRF. Việc phân tích dựa trên tập nhúng từ gọi là Embedding word, giờ chúng ta sẽ đi lần lượt các bước.

3.3.1. Xây dựng các thành phần chung cho mô hình

Để xử lý các thông tin, mô hình cần những dữ liệu dạng số và có cấu trúc rõ ràng. Bước này giúp các dữ liệu phi cấu trúc của các từ trở thành các vector, phục vụ cho đầu vào của mô hình.

Đầu tiên, import các bộ thư viện cần dùng trên colab: word_tokenize của nltk; pandas; numpy; re; Word2Vec, KeyedVectors của gensim; matplotlib; train_test_split của Sklearn; tf của tensorflow; LSTM, Embedding, Dense,

TimeDistributed, Dropout, Bidirectional, CRF, crf_loss, crf_viterbi_accuracy của keras.

Tất cả dữ liệu sơ yếu lý lịch đã được chuẩn bị ở những bước trước đó đang nằm trong một file csv. Tiếp đến sử dụng thư viện Pandas để đưa dữ liệu đó vào Python và thực hiện tiền xử lý dữ liệu đơn giản như loại bỏ ký tự đặc biệt, chuyển đổi sang chữ thường, loại bỏ khoảng trắng ở hai đầu chuỗi. Sau đó chuyển 2 cột là dữ liệu và nhãn thành 2 list để xử lý cho các bước tiếp theo.

Từ dữ liệu trên, tiếp tục mã hóa chúng bằng cách gán cho mỗi từ khác nhau mỗi số nguyên ID của từ đó. Tiếp đó chuyển đổi toàn bộ dữ liệu đầu vào là những dòng bằng các từ thành một chuỗi các con số và trả về chính là tham số data (được thực hiện ở 3.2.5 Tách từ, 3.2.6 Xây dựng bộ từ vựng).

Ví dụ: 'curriculum vitae' thành [253, 271] với {253, 271} là các ID tương ứng của {curriculum, vitae}.

Tương tự mã hóa nhãn khác nhau bằng mỗi số nguyên ID ở bộ tag2idx và mã hóa nhãn từ ID sang nhãn ở bộ idx2tag làm đầu ra dự đoán.

Ví dụ: 'I-W' thành [0] và [0] thành 'I-W'.

Biến đổi độ dài tất cả các dòng thành một độ dài duy nhất là 64 từ. Những dòng nào có độ dài ban đầu nhỏ hơn 64 thì chèn các số 0 vào để lấp đầy khoảng trống.

3.3.2. Phân chia bộ dữ liệu thành các phần để huấn luyện và kiểm tra

Chia dữ liệu thành 4 bộ X_train, X_test, y_train, y_test và chia dữ liệu thành 7 phần train, 3 phần test. Bộ X_train và y_train là dòng trong CV dùng để huấn luyện còn bộ X_test, y_test là nhãn trong CV dùng để kiểm tra hiệu suất (performance).


```

1 from sklearn.model_selection import train_test_split
2
3 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
4 X_train.shape, X_test.shape, np.array(y_train).shape, np.array(y_test).shape

```

((25522, 64), (10939, 64), (25522, 64, 14), (10939, 64, 14))

Hình 3.19 Chia bộ dữ liệu thành các phần để huấn luyện và kiểm tra

3.3.3. Xây dựng mô hình huấn luyện

Thực hiện việc huấn luyện dữ liệu trên mô hình với bộ dữ liệu huấn luyện. Đối với tập dữ liệu tiền xử lý, tôi đã sử dụng Tokenize tiếng Anh của NLTK. Sử dụng Word2vec để xây dựng bộ từ vựng cho dữ liệu, sử dụng embedding của Keras để vector các từ trong bộ từ vựng. Tôi sử dụng Tensorflow framework và các thư viện học sâu của Keras trong mô hình huấn luyện của mình.

Đầu tiên model có kiến trúc gồm 6 layer là Input layer, Embedding layer, Bidirectional LSTM layer, LSTM layer, TimeDistributed Layer, CRF layer. Layer đầu tiên Input dùng làm đầu vào, đầu vào mỗi dòng trong dữ liệu là 64 từ. Embedding layer sẽ vector mỗi từ trong mỗi dòng thành 1 vector có 64 chiều. Các tham số sử dụng trong Bidirectional LSTM layer, LSTM layer là giống nhau điều này làm cho mô hình học sâu hơn, cho kết quả dự đoán chính xác hơn với return_sequences=True sẽ trả về đầu ra cho mỗi bước thời gian. TimeDistributed Layer dùng Activation function là relu để chống Vanishing và Exploding và lấy kết quả trả về theo mỗi bước thời gian trước đó, chọn ra xác suất dự đoán cao nhất là đầu ra. CRF layer ràng buộc từ bộ dữ liệu đào tạo để đảm bảo các chuỗi nhãn thực thể dự đoán cuối cùng là hợp lệ.

Dùng Gradient descent là Adam, batch_size = 256, learning rate=0.0005 và Early Stopping để chống overfitting. Khi train model tôi lưu lại điểm mà model có performance cao nhất và val_loss thấp nhất.

3.4. Đánh giá mô hình

Đề tài này tôi lựa chọn Accuracy là phương pháp đánh giá hiệu suất của mô hình. Trên bộ dữ liệu thu thập với 14 nhãn thuộc 7 mục Information, Education,

Experience, Skill, Objective, None. Tôi sẽ đánh giá trên 4 mục gồm Information, Education, Skill, Experience. Đánh giá bởi độ chính xác của trình phân loại mục cho tất cả các nhãn theo công thức:

$$\text{Độ chính xác} = \frac{\text{Số câu gán nhãn đúng}}{\text{Số câu của mục}}$$

3.5. Trích xuất dữ liệu

Tôi sẽ trích xuất các thông tin trên 3 mục gồm Skill, Education, Experience. Cụ thể như sau:

(1). Đối với Skill:

Các từ khóa nằm trong mục Experience cũng có từ khóa của mục Skill nên 2 mục Skill, Experience sẽ được gộp lại thành 1 list. Sau đó sử dụng 1 list Skill có tên skills_data được thu thập thuộc lĩnh vực Công nghệ thông tin, ở định dạng csv. Sử dụng list này để đối sánh với dữ liệu Skill đã được phân đoạn thông qua model và trích xuất các từ khóa có nằm trong list skills_data.

(2). Đối với Education:

Đối với mục Education sẽ lấy các thông tin về cấp bậc và chuyên ngành. Trích xuất bằng cấp và chuyên ngành từ Education bằng cách sử dụng regex. Trích xuất khi thuộc từ khóa {'college', 'master', 'bachelor', 'bachelors'}, và những từ đứng trước hoặc sau từ khóa chuyên ngành {'major', 'speciality'}. Ví dụ: 'major mathematical computing'.

(3). Đối với Experience:

Đối với mục Education sẽ lấy các thông tin về số năm kinh nghiệm và có 2 pattern được sử dụng như sau:

+ Pattern 1: Sử dụng regex để lấy số năm là 4 số có 2 số đầu là 20 và số cuối thuộc số từ 0-9. Các năm sẽ được lưu vào 1 list có tên year_pattern là danh sách các năm thuộc nội dung Experience của 1 CV, tôi sẽ lấy số năm cao nhất trừ đi số năm nhỏ nhất để ra số năm kinh nghiệm.

+ Pattern 2: Sử dụng regex để lấy thông tin có sẵn là ‘n years’
với n là số năm, n sẽ là các số từ 0-9.

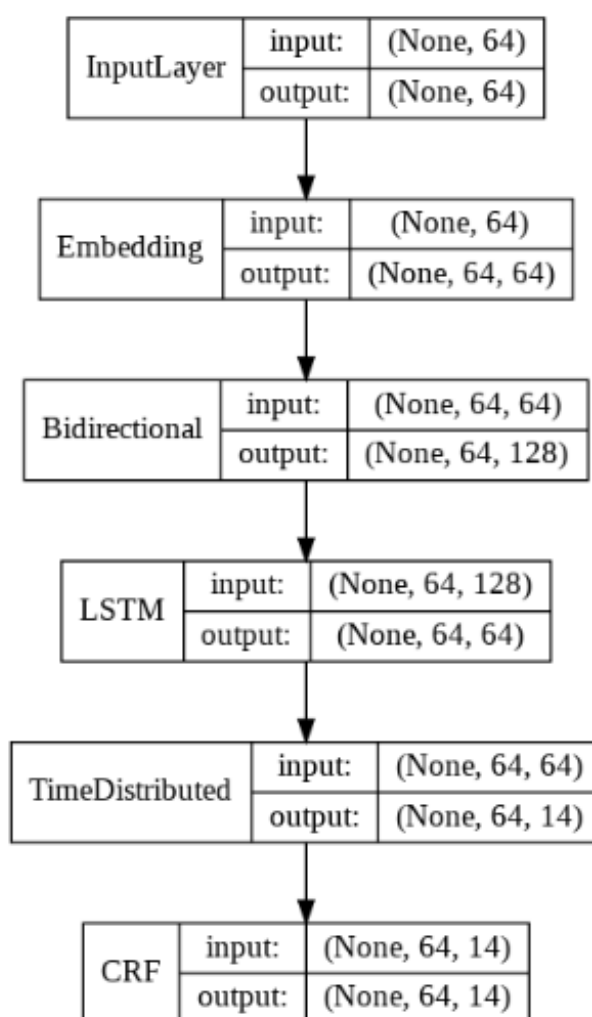


CHƯƠNG 4. KẾT QUẢ HỆ THỐNG HỖ TRỢ TUYỂN DỤNG NHÂN SỰ SỬ DỤNG AI

4.1. Kiến trúc mô hình phân đoạn

Sau nhiều lần thử nghiệm và tham khảo từ nhiều bài nghiên cứu (paper) về xử lý ngôn ngữ tự nhiên (NLP), nhận dạng thực thể trong văn bản (NER), tôi đã xây dựng một mô hình phân đoạn (model segment). Áp dụng mô hình, kết hợp sử dụng biểu thức chính quy (regular) để trích xuất các từ khóa trong một sơ yếu lý lịch bất kỳ bao gồm kỹ năng, số năm kinh nghiệm và trình độ học vấn.

Mô hình phân đoạn sử dụng 6 lớp gồm Input, Embedding, Bidirectional LSTM, LSTM, Time Distribute, CRF. Kiến trúc mô hình phân đoạn như (Hình 4.1).



Hình 4.1 Kiến trúc trong model segment

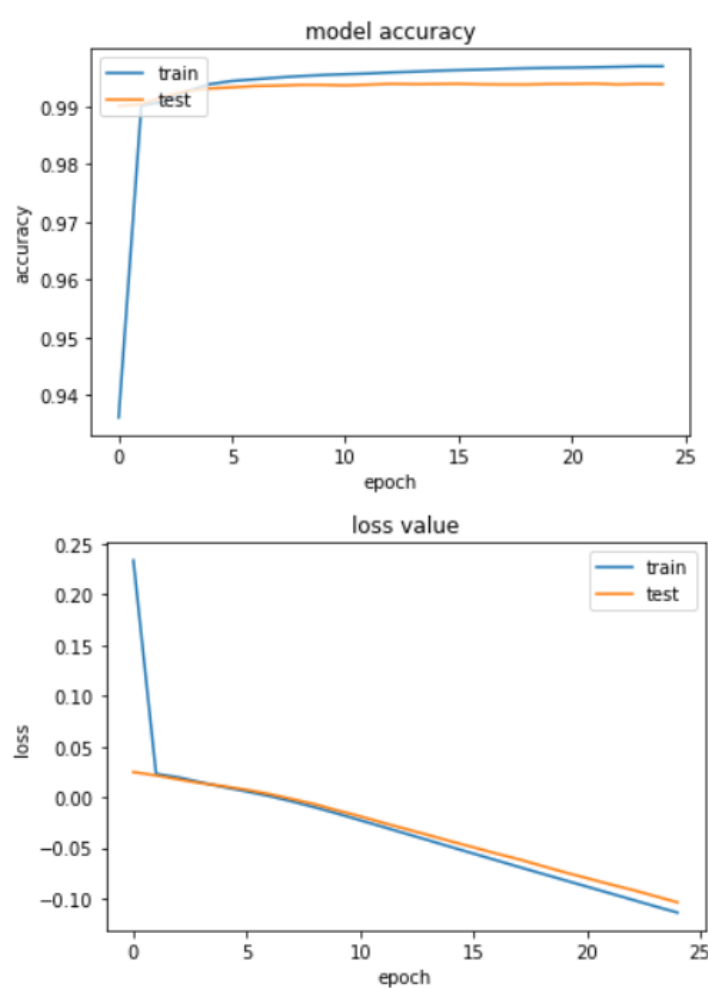
Bảng 4.1 Ý nghĩa các lớp trong model segment

Lớp	Ý nghĩa	Tham số
Input	Đầu vào của model. Đầu vào sẽ là 1 dòng, mỗi dòng có 64 từ.	Đầu vào: (, 64) Đầu ra: (, 64)
Embedding	Để nhúng (embed) các câu văn bản. Cụ thể biến các word index thành các vector 64 chiều cố định.	Đầu vào: (, 64) Đầu ra: (, 64, 64)
Bidirectional LSTM	Mô hình học tốt hơn bằng cách học 2 chiều của chuỗi, nên cung cấp các dự đoán cao hơn. Đầu ra sẽ có 128 chiều vì lớp Bidirectional LSTM học 2 chiều trước, sau của chuỗi nên chiều của vector sẽ được nhân thêm 2.	Đầu vào: (, 64, 64) Đầu ra: (, 64, 128)
LSTM	Lấy kích thước đầu ra từ Bidirectional LSTM.	Đầu vào: (, 64, 128) Đầu ra: (, 64, 64)
TimeDistributed	Lớp TimeDistributed (được kết nối đầy đủ) trên mọi đầu ra qua mọi bước thời gian. Lớp này lấy kích thước đầu ra của lớp LSTM trước và xuất ra chiều dài chuỗi tối đa (64) và nhãn tối đa (14).	Đầu vào: (, 64, 64) Đầu ra: (, 64, 14)
CRF	Ràng buộc từ bộ dữ liệu đào tạo để đảm bảo các chuỗi nhãn	Đầu vào: (, 64, 14)

	thực thể dự đoán cuối cùng là hợp lệ.	Đầu ra: (, 64, 14)
--	---------------------------------------	--------------------

4.2. Kết quả mô hình

Độ chính xác và sự mất mát khi huấn luyện mô hình phân đoạn thể hiện như (Hình 4.2).



Hình 4.2 Độ chính xác và sự mất mát của mô hình phân đoạn

4.3. Đánh giá mô hình

Bên dưới là biểu đồ phân phối dữ liệu của các mục Information, Skill, Experience, Education được thể hiện như (Hình 4.3, Hình 4.4, Hình 4.5, Hình 4.6).

Bảng 4.2 Chú thích tên trục trong biểu đồ phân phối dữ liệu

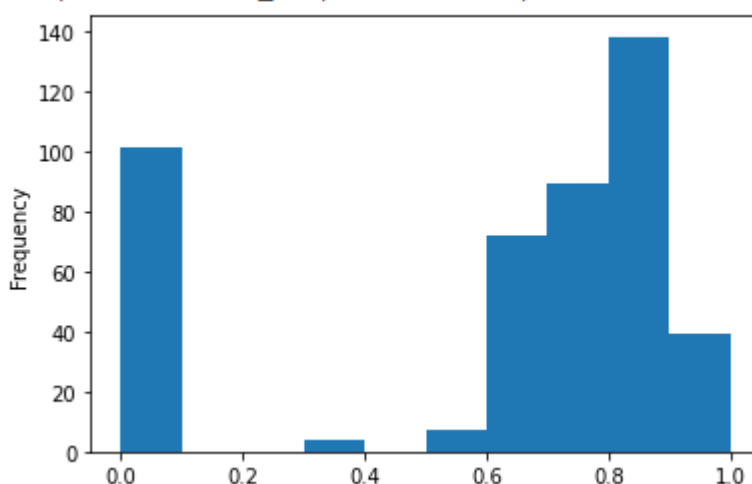
Tên trục	Chú thích
Trục x (nằm ngang)	Tỉ lệ phần trăm dự đoán mục đúng trong 1 CV
Trục y (nằm dọc)	Số lượng CV

(1). *Information*

Độ chính xác trung bình của Information gần 62%. Độ chính xác giao động trong khoảng 60-90%. Có khoảng 140 CVs dự đoán (predict) đúng 80-90% và khoảng 100 CVs dự đoán sai.

- Mean Accuracy Information: ~62%

<matplotlib.axes._subplots.AxesSubplot at 0x7f31a7e90bd0>



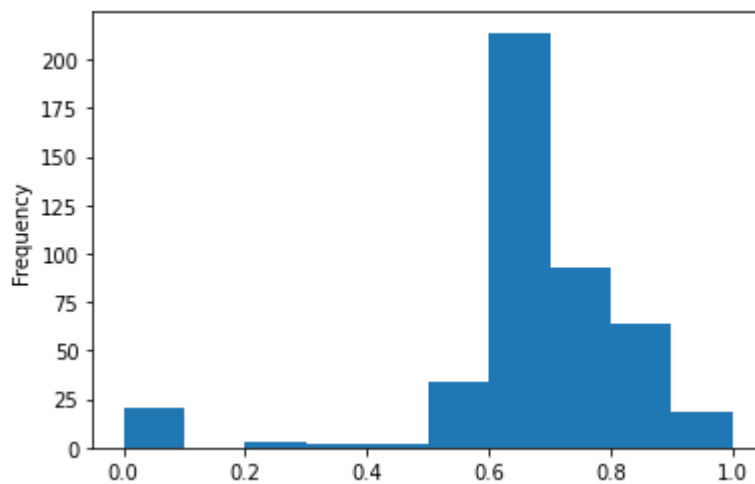
Hình 4.3 Section information

(2). *Skill*

Độ chính xác trung bình của Skill gần 67%. Phần lớn nhãn trong mục Skill được dự đoán (predict) đúng, giao động trong khoảng 60-70%.

- Mean Accuracy Skill: ~67%

<matplotlib.axes._subplots.AxesSubplot at 0x7f31a867a810>



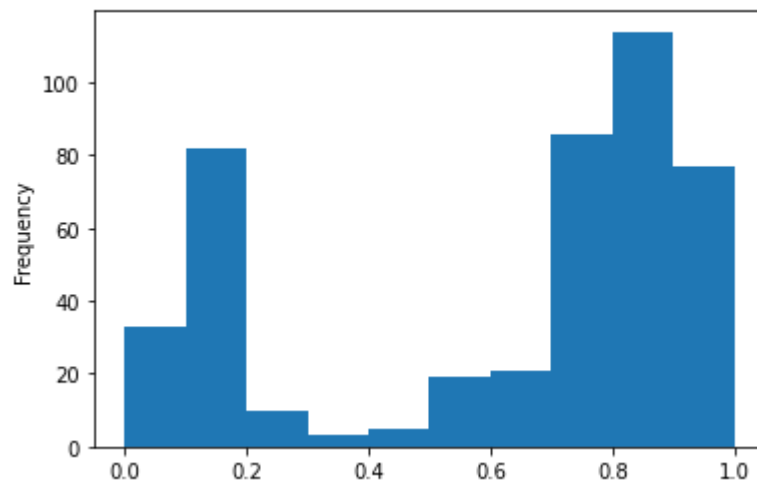
Hình 4.4 Section skill

(3). Experience

Độ chính xác trung bình của Experience gần 63%. Phần lớn nhãn trong mục Experience được dự đoán (predict) đúng, giao động trong khoảng 70-100%. Số CV dự đoán sai trong Experience cũng khá cao.

- Mean Accuracy Experience: ~63%

<matplotlib.axes._subplots.AxesSubplot at 0x7f31a78c1850>



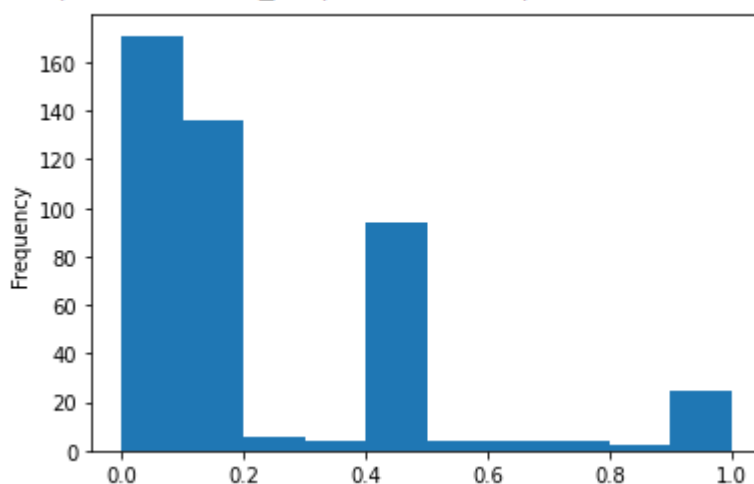
Hình 4.5 Section experience

(4). Education

Độ chính xác trung bình của Education gần 22%. Phần lớn nhân trong mục Education được dự đoán (predict) sai. Chỉ khoảng 20 CVs dự đoán đúng.

- Mean Accuracy Education: ~22%

<matplotlib.axes._subplots.AxesSubplot at 0x7f31a784b9d0>



Hình 4.6 Section education

4.4. Kết quả trích xuất dữ liệu

Sau đây là kết quả trích xuất thông tin từ tập dữ liệu CV sau khi huấn luyện sẽ gồm kết quả extract skill (Hình 4.8), kết quả extract education (Hình 4.9), kết quả extract experience (Hình 4.10).

	Skill	Extract_Skill
0	python numpy pandas., skills, programming la...	{android, mongodb, machine learning, deep lear...
1	understanding of computer to findalocation at,...	{team working, electronics, office, arduino, h...
2	skills , micro controller8 32 bits avr arm7 ...	{s, android, vhdl, electronics, linux, c, firm...
3	learnnew things everyda y, good communicatio...	{presentation, oop, machine learning, public r...
4	scada control and communication, c c++ c#, pl...	{tcp ip, scada, ms office, office, teamwork, c...
5	technical skills , programming language pytho...	{powerpoint, r, problem solving, office, numpy...
6	technical skills, programming and computers p...	{arabic, problem solving, office, css, spanish...
7	skills, solid experience in java php program...	{mysql, cakephp, vba, perl, eclipse, svn, php,...
8	working hard sociable friendly patient and...	{team work, sourcing}
9	technical skills ...	{automation, seo, eclipse, machine learning, s...

Hình 4.7 Kết quả trích xuất kỹ năng

	Education	Extract_Edu
0	2017 2018 udacity deep learning nanodegree,...	[college]
1	education, oct 2014 now, post and telecommuni...	[major electronic & electrical engineering]
2	september 2002 to april 2007, bachelor of el...	[bachelor of electrical and electronics engin...
3	education, university of information technolog...	[major in information systems]
4	electrical electronics and, education, bach k...	[]
5	university of science ho chi minh city, nd, ed...	[major mathematical computing.]
6	education, princeton university princeton n...	[]
7	education, bachelor of electric electronic eng...	[bachelor]
8	education , ready for graduating of electronic...	[]
9	education ...	[bachelor]

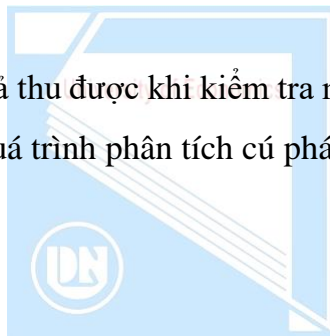
Hình 4.8 Kết quả trích xuất trình độ học vấn

	Experience	Extract_Exp
0	projects , projects , stored managed data in s...	[2]
1	01654256960, work experience, ptit, senior stu...	[2]
2	professional experiences , january 2013 to pre...	[6]
3	phone0898019672, dob07101996, language, read ...	[2]
4	automatic engineer, +84934052465, experience a...	[None]
5	achievements responsibilities , rd, pre profes...	[1]
6	work & volunteer experience, student tour guid...	[3]
7	viet nam, migration design, internal framewor...	[3]
8	work experience , 8 201132012. i am deliverer...	[4]
9	model. i have experienced programming android ...	[1]

Hình 4.9 Kết quả trích xuất số năm kinh nghiệm

4.5. Test mô hình

Sau đây là các kết quả thu được khi kiểm tra mô hình với 1 CV ngẫu nhiên, từ 1 CV của ứng viên qua quá trình phân tích cú pháp (parsing).




Le Minh Binh

University of Science Ho Chi Minh city

Address: 333 Ben Binh Dong Street, Ward 11, District 8, HCM City

Phone: +84 1653417308 | Email: binhleminh0209@gmail.com

Date of Birth: September 2nd, 1996



tag information

CAREER OBJECTIVE:

Seeking a dynamic studying and working environment in the field of Machine Learning and Data Science that can improve my skills and exploit my abilities.

tag objective

TECHNICAL SKILLS:

- Programming language: Python (with frameworks: tensorflow, sklearn, numpy, pandas), C/C++, Matlab, R....
- Mathematical problem solving
- Microsoft Office (MS Word, Excel, PowerPoint)

PERSONAL SKILLS:

- Ability to work independently or as part of a team
- Ability to work under pressure
- Verbal communication skill

tag skill

EDUCATION:

Senior student of Mathematical Computer Science (2015-2018)

Major: Mathematical Computing.

Cumulative GPA: 3.66/4.0

Faculty of Mathematics and Computer Science, University of Science HCM City

Ward 4, District 5, HCM City.

tag education

ACHIEVEMENTS/RESPONSIBILITIES:

- 3rd prize "Provincial Excellent Student Maths Contest" at grade 12
- Member of the Executive Committee of The Associated Organisation of Vietnamese Students' Association, Faculty of Maths and Computer Science (2015-2016)
- Organizer of Saigon School of AI community (August 2018 - ?)

tag O

PRE-PROFESSIONAL EXPERIENCE:

- Being a tutor (2015 – 2017)
- Being a salesman for a retail agent (June 2016-August 2016)
- Exchanged student - Faculty of Engineering
Tokyo University of Agriculture and Technology (June 2017-August 2017, Koganei, Tokyo, Japan)
Handle problems related to Facial Emotion recognition using Kinect toolbox.

tag experience

REFERENCES:

Will be provided upon request.

tag O

Hình 4.10 CV test

86

sentences	pre tag
le minh binh	: B-I
university of science ho chi minh city	: I-E
address 333 ben binh dong street ward 11 district 8 ho chi minh city	: I-I
phone 84 1653417308 email binhleminh0209 gmail com	: I-I
date of birth september 2nd 1996	: I-I
career objective	: B-SU
seeking a dynamic studying and working environment in the field of machine	: I-SU
learning and data science that can improve my skills and exploit my abilities	: I-SU
technical skills	: B-S
programming language python with frameworks tensorflow sklearn numpy	: I-S
pandas c c matlab r	: I-S
mathematical problem solving	: I-S
microsoft office ms word excel powerpoint	: I-S
personal skills	: B-W
ability to work independently or as part of a team	: I-S
ability to work under pressure	: I-S
verbal communication skill	: I-S
education	: B-E
senior student of mathematical computer science 2015 2018	: I-E
major mathematical computing	: I-E
cumulative gpa 3 66 4 0	: I-E
faculty of mathematics and computer science university of science ho chi minh city	: I-E
ward 4 district 5 ho chi minh city	: I-I
achievements responsibilities	: I-W
3rd prize provincial excellent student maths contest at grade 12	: O
member of the executive committee of the associated organisation of	: O
vietnamese students association faculty of maths and computer science	: O
2015 2016	: O
organizer of saigon school of ai community august 2018	: O
pre professional experience	: B-W
being a tutor 2015 2017	: I-W
being a salesman for a retail agent june 2016 august 2016	: I-W
exchanged student faculty of engineering	: I-E
tokyo university of agriculture and technology june 2017 august 2017	: I-W
koganei tokyo japan	: I-W
handle problems related to facial emotion recognition using kinect toolbox	: I-W
references	: O

Hình 4.11 Kết quả dự đoán với CV test

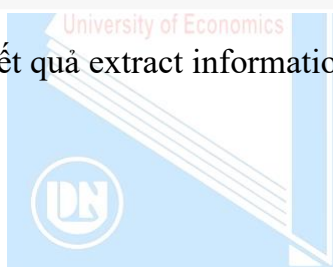
Qua kết quả dự đoán với CV test này, cho thấy model segment dự đoán các tag khá chính xác, một số tag dự đoán sai do dòng viết chung chung, gây nhầm lẫn. Ví dụ dòng ‘Ward 4, District 5, HCM City.’ ở tag education dự đoán sai do dòng này là ghi địa chỉ, dễ nhầm lẫn với tag information.

Extract_Information		Extract_Skill	
0	le minh binh, address 333 ben binh dong street ward 11 district 8 ho chi minh city, phone 84 1653417308 email binhleminh0209 gmail com, date of birth september 2nd 1996, ward 4 district 5 ho chi minh city	technical skills, programming language python with frameworks tensorflow sklearn numpy, pandas c c matlab r, mathematical problem solving, microsoft office ms word excel powerpoint, ability to work independently or as part of a team, ability to work under pressure, verbal communication skill	
Extract_Education		Extract_Experience	
0	university of science ho chi minh city, education, senior student of mathematical computer science 2015 2018, major mathematical computing, cumulative gpa 3 66 4 0, faculty of mathematics and computer science university of science ho chi minh city, exchanged student faculty of engineering	personal skills, achievements responsibilities, pre professional experience, being a tutor 2015 2017, being a salesman for a retail agent june 2016 august 2016, tokyo university of agriculture and technology june 2017 august 2017, koganei tokyo japan, handle problems related to facial emotion recognition using kinect toolbox	

Hình 4.12 Kết quả phân đoạn CV test khi qua model segment

	Extract_Skill	Extract_Edu	Extract_Exp
0	{excel, r, tensorflow, python, mathematical problem solving, pandas, agriculture, office, matlab, numpy, c, microsoft office, problem solving, word, powerpoint}	[major mathematical computing]	[2]

Hình 4.13 Kết quả extract information với CV test



KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết quả đạt được

Trong đề tài này, mô hình được đánh giá theo từng mục với Information đạt gần 62%, Skill đạt gần 66%, Experience đạt gần 62%, Education đạt gần 22%, qua đề tài tôi đã xây dựng hệ thống phân tích cú pháp sơ yếu lý lịch (CV Parsing) là một phần của hệ thống hỗ trợ tuyển dụng nhân sự mà tôi đang hướng tới. Đây là một hệ thống giúp tiết kiệm thời gian phân tích bằng việc phân loại, trích xuất thông tin về kỹ năng, bằng cấp, số năm kinh nghiệm và phân tích một cách tự động những sơ yếu lý lịch có cùng format ở định dạng pdf, được viết bằng tiếng Anh và có 1 cột.

Trên thực tế đề tài chưa thực sự đáp ứng mục tiêu đã đề ra đó là có thể sàng lọc ra danh sách các ứng viên phù hợp cho các vị trí công việc của nhà tuyển dụng. Sau nhiều lần thử nghiệm và tham khảo từ nhiều bài nghiên cứu tôi đã xây dựng một phần của hệ thống hỗ trợ tuyển dụng nhân sự, hệ thống có thể phân tích và trích xuất tự động các thông tin từ sơ yếu lý lịch ở nhiều dạng khác nhau (pdf, doc) và lưu trữ thông tin này vào cơ sở dữ liệu. Về mặt bằng chung đối với phân tích và trích xuất tự động, mô hình đã đáp ứng đầy đủ với một mô hình Parsing cơ bản.

2. Tồn tại

Lĩnh vực AI là một lĩnh vực mới và tôi chưa được đào tạo trước đó. Do sự hạn chế về mặt kiến thức và kinh nghiệm của một sinh viên khi xây dựng hệ thống đầu tiên sử dụng Trí tuệ nhân tạo trong vòng 3 tháng nên còn nhiều thiếu sót không tránh khỏi mong thầy cô cảm thông. Một số hạn chế của đề tài như sau:

- Độ chính xác của dữ liệu huấn luyện chưa cao, hầu hết các câu là các dòng trong CV quá ngắn và có nhiều từ khóa gây nhầm lẫn với nội dung mục khác, nên ảnh hưởng đến kết quả dự đoán của mô hình.

- Kỹ thuật Rule-based chưa đạt hiệu quả cao trong việc trích xuất thực thể, chỉ trích xuất thông tin của skill, bằng cấp, số năm kinh nghiệm, chưa trích xuất thông tin cá nhân của CV.

- Chưa xây dựng mô hình để trích xuất các thông tin tuyển dụng của các công ty

- Chưa xây dựng một hệ thống hoàn chỉnh để sàng lọc ra danh sách các ứng viên phù hợp cho các vị trí công việc của nhà tuyển dụng.

3. Hướng phát triển

- Cải thiện số lượng và chất lượng bộ dữ liệu huấn luyện.
- Phát triển hệ thống sẽ chạy được trên nền web với địa chỉ mạng nội bộ nhằm cải thiện tính tiện dụng của hệ thống.
- Ứng dụng thuật toán nhận diện thực thể NER (Named entity recognition) để trích xuất các thông tin theo từng ngữ cảnh của sơ yếu lý lịch và thông tin tuyển dụng.
- Thực hiện Matching CV để sàng lọc danh sách các ứng viên phù hợp và hoàn thành hệ thống hỗ trợ tuyển dụng nhân sự.



TÀI LIỆU THAM KHẢO

- [1] "TMA Innovation," [Online]. Available: <https://www.tmainnovation.vn/ung-dung-ho-tro-truyen-dung-su-dung-ai/>.
- [2] "Base," [Online]. Available: <https://base.vn/hiring>.
- [3] "base.vn," [Online]. Available: <https://bitly.com.vn/3vr7z5>.
- [4] "ibm," Deep Blue, [Online]. Available: <https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/>.
- [5] N. T. H. Nguyễn Văn Khoa, "xây dựng hệ thống nhận diện khuôn mặt bằng AI – áp dụng trong việc quản lý nhân viên của doanh nghiệp," 2021.
- [6] L. T. K. Nguyễn Việt Lâm, *Cạnh tranh công nghệ Mỹ - Trung Quốc thời đại 4.0*, 2020.
- [7] "FPT Digital," 2020. [Online]. Available: <https://bitly.com.vn/b3hq6p>.
- [8] "Question Pro," [Online]. Available: <https://www.questionpro.com/blog/data-collection/>.
- [9] PRIYA_SINGH, "embibe," 2022. [Online]. Available: <https://www.embibe.com/exams/data-collection/>.
- [10] "IEEE Spectrum," [Online]. Available: <https://spectrum.ieee.org/top-programming-languages-2021>.