

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN HỆ THỐNG THÔNG TIN



NGUYỄN QUỐC ĐẠT 15110188

TRẦN THỊ TỎ UYÊN 15110361

CAO XUÂN NHÃN 15110266

ĐỀ TÀI:

**TÌM HIỂU WEB SCRAPING VÀ TOPIC ANALYSIS
ĐỂ PHÂN TÍCH CÁC THÔNG BÁO TUYỂN DỤNG**

KHÓA LUẬN TỐT NGHIỆP KỸ SƯ CNTT

GIÁO VIÊN HƯỚNG DẪN

ThS. QUÁCH ĐÌNH HOÀNG

KHÓA 2015- 2019

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN HỆ THỐNG THÔNG TIN



NGUYỄN QUỐC ĐẠT 15110188

TRẦN THỊ TỎ UYÊN 15110361

CAO XUÂN NHÃN 15110266

ĐỀ TÀI:

TÌM HIỂU WEB SCRAPING VÀ TOPIC ANALYSIS

ĐỂ PHÂN TÍCH CÁC THÔNG BÁO TUYỂN DỤNG

KHÓA LUẬN TỐT NGHIỆP KỸ SƯ CNTT

GIÁO VIÊN HƯỚNG DẪN

ThS. QUÁCH ĐÌNH HOÀNG

KHÓA 2015- 2019

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Họ và tên Sinh viên 1: Nguyễn Quốc Đạt MSSV 1: 15110188

Họ và tên Sinh viên 2: Trần Thị Tố Uyên MSSV 2: 15110361

Họ và tên Sinh viên 3: Cao Xuân Nhân MSSV 3: 15110266

Ngành: Công nghệ thông tin

Tên đề tài: Tìm hiểu web scraping và topic analysis để phân tích các thông báo tuyển dụng.

Họ và tên Giáo viên hướng dẫn: ThS. Quách Đình Hoàng

NHẬN XÉT

1. Về nội dung đề tài và khối lượng thực hiện:

.....

.....

.....

.....

.....

2. Ưu điểm:

.....

.....

.....

.....

.....

3. Khuyết điểm:

.....

.....

.....

.....

.....

4. Đề nghị cho bảo vệ hay không:
5. Đánh giá loại:
6. Điểm:

Tp. Hồ Chí Minh, Ngày ... tháng...năm 2019

Giáo viên hướng dẫn

(Ký) và ghi rõ họ tên)

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên Sinh viên 1: Nguyễn Quốc Đạt MSSV 1: 15110188

Họ và tên Sinh viên 2: Trần Thị Tố Uyên MSSV 2: 15110361

Họ và tên Sinh viên 3: Cao Xuân Nhân **MSSV 2:** 15110266

Ngành: Công nghệ thông tin

Tên đề tài: Tìm hiểu web scraping và topic analysis để phân tích các thông báo tuyển dụng.

Họ và tên Giáo viên phản biện:

NHẬN XÉT

1. Về nội dung đề tài và khối lượng thực hiện:

.....

.....

.....

.....

.....

2. Ưu điểm:

.....

.....

.....

.....

.....

3. Khuyết điểm:

.....

.....

.....

.....

.....

4. Đề nghị cho bảo vệ hay không:
5. Đánh giá loại:
6. Điểm:

Tp. Hồ Chí Minh, Ngày ... tháng ... năm 2019

Giáo viên phân biện

(Ký và ghi rõ họ tên)

LỜI CẢM ƠN

Một kỳ thực hiện khóa luận ngắn ngủi đã trôi qua nhưng để lại trong chúng tôi rất nhiều cảm xúc. Chúng tôi xin được gửi lời cảm ơn chân thành đến Thầy Quách Đình Hoàng. Thầy đã cung cấp cho chúng tôi tài liệu và hướng dẫn tận tình cho chúng tôi. Trong suốt quá trình thực hiện khóa luận, Thầy luôn theo dõi tiến độ và giải đáp, chia sẻ giúp chúng tôi vượt qua những khó khăn. Chúng tôi rất trân quý sự tâm huyết và trách nhiệm của Thầy trong công việc giảng dạy và truyền thụ kiến thức.

Chúng tôi cũng xin gửi lời cảm ơn sâu sắc đến Thầy Cô khoa Công nghệ thông tin-Đại học Sư phạm kỹ thuật TP.HCM đã truyền dạy kiến thức và hỗ trợ chúng tôi trong suốt quá trình học tập và thực hiện khóa luận. Chúng tôi xin cảm ơn mái trường Sư phạm kỹ thuật đã tạo nhiều điều kiện thuận lợi cho hoạt động học tập của sinh viên chúng tôi, đặc biệt là thư viện với nguồn tri thức vô tận. Chúng tôi cũng gửi lời cảm ơn chân thành đến các bạn của mình, các bạn khóa 15 ngành Công nghệ thông tin, các bạn khóa 15 chuyên ngành hệ thống thông tin và các anh chị em ngành Công nghệ thông tin. Chúng tôi cảm ơn những góp ý và chia sẻ quý giá từ tất cả các bạn. Cảm ơn sự động viên tinh thần từ các bạn để nhóm chúng tôi có thể giữ vững tinh thần và thực hiện khóa luận đúng tiến độ.

Những điều mà nhà trường, Thầy Cô và bạn bè mang đến cho chúng tôi, chúng tôi sẽ luôn ghi nhớ và thúc đẩy bản thân phát triển và hoàn thiện hơn nữa. Công việc nào chắc chắn cũng có khó khăn nhưng khổ luyện thành nhân, rõ ràng chúng tôi thấy bản thân đã phát triển thêm rất nhiều cả về kiến thức, tư duy, kỹ năng, cách làm việc và mối quan hệ ứng xử với mọi người. Chúng tôi cũng nhận thấy bản thân có những khuyết điểm và thiếu sót cần cố gắng cải thiện để tốt hơn, hướng tới mục tiêu lớn trong tương lai.

NGUYỄN QUỐC ĐẠT

TRẦN THỊ TỐ UYÊN

CAO XUÂN NHÃN

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP HCM
KHOA CÔNG NGHỆ THÔNG TIN

--0--

ĐỀ CƯƠNG LUẬN VĂN TỐT NGHIỆP

Họ và tên SV thực hiện 1: Nguyễn Quốc Đạt **MSSV:** 15110188
Họ và tên SV thực hiện 2: Trần Thị Tố Uyên **MSSV:** 15110361
Họ và tên SV thực hiện 3: Cao Xuân Nhãn **MSSV:** 15110266
Thời gian làm luận văn: Từ: 03/2019 Đến: 07/2019
Chuyên ngành: Hệ thống thông tin
Tên luận văn: Tìm hiểu web scraping và topic analysis để phân tích các thông báo tuyển dụng
Giáo viên hướng dẫn: Ths. Quách Đình Hoàng

NHIỆM VỤ CỦA LUẬN VĂN:

Nhiệm vụ của luận văn là thu thập dữ liệu cho một loại việc làm cụ thể trên trang web tuyển dụng nổi tiếng. Sau đó tiến hành phân tích theo chủ đề để phát hiện các yêu cầu chung, mức lương phổ biến, chính sách đãi ngộ,... tương ứng cho loại việc làm đó, đưa ra các từ khóa phổ biến liên quan đến vị trí công việc đang tìm kiếm, phán đoán được các kỹ năng mới, các xu hướng trong tương lai. Để đạt được điều đó, chúng tôi tập trung tìm hiểu một số vấn đề sau:

1. Tìm hiểu khái niệm web scraping và cấu trúc HTML
2. Tìm hiểu khái niệm topic analysis và các kỹ thuật phân tích theo chủ đề (PLSA, LDA)
3. Thực hiện công việc thu thập dữ liệu web tuyển dụng với các thư viện của R
4. Phân tích theo chủ đề đối với dữ liệu thu thập được để đưa ra các thông tin quan tâm
5. Đánh giá và giải thích kết quả

ĐỀ CƯƠNG VIẾT LUẬN VĂN:

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ, SƠ ĐỒ

DANH MỤC TỪ VIẾT TẮT

TÓM TẮT

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 GIỚI THIỆU BÀI TOÁN

1.2 MỤC TIÊU

1.3 BỐ CỤC CỦA BÁO CÁO

CHƯƠNG 2: TỔNG QUAN WEB SCRAPING

2.1 KHÁI NIỆM WEB SCRAPING

2.2 CẤU TRÚC TÀI LIỆU HTML

CHƯƠNG 3: TỔNG QUAN TOPIC ANALYSIS

3.1 CÁC KHÁI NIỆM

3.1.1 Khái niệm chủ đề

3.1.2 Khái niệm topic analysis và topic model

3.2 BÀI TOÁN TOPIC ANALYSIS

3.2.1 Đối tượng và dữ liệu quan tâm

3.2.2 Các bước thực hiện

3.3 CÁC THUẬT TOÁN PHÂN TÍCH CHỦ ĐỀ (TOPIC ANALYSIS)

3.3.1 Thuật toán PLSA

3.3.2 Thuật toán LDA

CHƯƠNG 4: THỰC HIỆN THU THẬP CÁC THÔNG BÁO TUYỂN DỤNG

4.1 ĐỐI TƯỢNG VÀ DỮ LIỆU QUAN TÂM

4.2 GIỚI THIỆU CÔNG CỤ R VÀ CÁC THƯ VIỆN

4.2.1 Giới thiệu ngôn ngữ R

4.2.2 Các gói thư viện hỗ trợ web scraping

4.3 THỰC HIỆN TẢI XUỐNG DỮ LIỆU VỚI CÔNG CỤ R

CHƯƠNG 5: ỨNG DỤNG TOPIC MODEL PHÂN TÍCH THÔNG BÁO TUYỂN DỤNG

5.1 CÁC CÔNG CỤ VÀ THƯ VIỆN CẦN THIẾT

5.1.1 Thư viện Tm

5.1.2 Thư viện Quanteda

5.1.3 Thư viện Topicmodels

5.2 DỮ LIỆU VÀ TIỀN XỬ LÝ

5.2.1 Mô tả dữ liệu

5.2.2 Tiền xử lý dữ liệu

5.3 THỰC HIỆN TOPIC ANALYSIS CÁC THÔNG BÁO TUYỂN DỤNG

CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ PHÂN TÍCH

6.1 BIỂU DIỄN KẾT QUẢ

6.2 GIẢI THÍCH KẾT QUẢ

KẾT LUẬN

TÀI LIỆU THAM KHẢO

Tiếng Việt

- [1] Nguyễn Trung Kiên, Lê Quang Minh, *Xây dựng ứng dụng thu thập dữ liệu tự động từ các Website*, Luận văn tốt nghiệp, Đại học Bách Khoa TP.HCM, 2009, trang 13-15.
- [2] Đào Minh Tùng, *Phân cụm đa mức web bằng k-means dựa trên chủ đề ẩn và thực nghiệm đánh giá*, Luận văn tốt nghiệp, Đại học Công nghệ- ĐHQGHN, 2011, trang 10-20-21.

Tiếng Anh

- [3] Simon Munzert, *Automated Data Collection with R- A Practical Guide to Web Scraping and Text Mining*, John Wiley & Sons Ltd, 2015, trang 292.
- [4] ChengXiang Zhai, Sean Massung, *Text Data Management and Analysis - A Practical Introduction to Information Retrieval and Text Mining*, ACM Books, 2016, trang 331.
- [5] Foster Provost, Tom Fawcett, *Data Science For Business- What You Need To Know About Data Mining And Data-Analytic Thinking*, O'Reilly Media, 2013, trang 252.

Nguồn khác

- [6] *Cross-validation of topic modelling, links <http://freerangestats.info/blog/2017/01/05/topic-model-cv>, 06/2019.*

KẾ HOẠCH THỰC HIỆN:

STT	Thời gian	Công việc	Ghi chú
1	01/03- 30/03	Tìm hiểu tài liệu và tài liệu liên quan	
		Tìm hiểu và viết báo cáo phần cấu trúc HTML	
		Tìm hiểu và viết báo cáo phần web scraping (khái niệm, các bước, vấn đề liên quan)	
2	20/03- 30/04	Xác định cấu trúc lưu trữ dữ liệu tải xuống	
		Code thu thập dữ liệu (web scraping)	
		Viết báo cáo cho phần thực hành tải Webscraping (giới thiệu trang web, các bước tiến hành, kết quả)	
		Tìm hiểu bài toán phân tích chủ đề và viết báo cáo về	

		bài toán topic analysis (input, output, bước thực hiện)	
		Tìm và viết báo giới thiệu một số trang web tuyển dụng phổ biến	
		Tìm hiểu thư viện Rvest, Xlm2, Tidyverse, Tm, Topicmodels	
3	15/04- 01/07	Tìm hiểu và viết báo cáo thuật toán PLSA, LDA (ý tưởng, giải thích, tính toán, nhận xét, so sánh)	
	20/04- 20/06	Code phân tích topic analysis và hàm thống kê keyword quan tâm	
4	10/05- 20/06	Thiết kế Slide phần giới thiệu đề tài và phần webscarping	
		Thiết kế Slide phần lí thuyết topic analysis và các thuật toán	
5	20/06-05/07	Viết báo cáo cho phần tiền xử lý và phân tích dữ liệu (mô tả dữ liệu, tiền xử lý, áp dụng thuật toán LDA, đưa ra nhận xét)	
		Thực hiện đánh giá, nhận xét và giải thích kết quả	
		Tổng kết đề tài và viết kết luận (điểm đạt được, chưa đạt và đề xuất cải thiện, hướng phát triển)	

TP.HCM, Ngày tháng năm 2019

Ý kiến của giáo viên hướng dẫn

(ký và ghi rõ họ tên)

Người viết đề cương

Nguyễn Quốc Đạt

Trần Thị Tố Uyên

Cao Xuân Nhãn

MỤC LỤC

LỜI CẢM ƠN.....	7
ĐỀ CƯƠNG LUẬN VĂN TỐT NGHIỆP	8
MỤC LỤC	12
DANH MỤC BẢNG BIỂU.....	14
DANH MỤC HÌNH VẼ, SƠ ĐỒ	15
DANH MỤC TỪ VIẾT TẮT	16
TÓM TẮT.....	17
CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI.....	19
1.1 GIỚI THIỆU BÀI TOÁN	19
1.2 MỤC TIÊU	19
1.3 BỐ CỤC CỦA BÁO CÁO	20
CHƯƠNG 2: TỔNG QUAN WEB SCRAPING	21
2.1 KHÁI NIỆM WEB SCRAPING.....	21
2.2 CẤU TRÚC TÀI LIỆU HTML	22
CHƯƠNG 3: TỔNG QUAN TOPIC ANALYSIS.....	26
3.1 CÁC KHÁI NIỆM	26
3.1.1 Khái niệm chủ đề	26
3.1.2 Khái niệm topic analysis và topic model	26
3.2 BÀI TOÁN TOPIC ANALYSIS	27
3.2.1 Đối tượng và dữ liệu quan tâm.....	27
3.2.2 Các bước thực hiện.....	28
3.3 CÁC THUẬT TOÁN PHÂN TÍCH CHỦ ĐỀ (TOPIC ANALYSIS)	29
3.3.1 Thuật toán PLSA.....	29
3.3.2 Thuật toán LDA	34

CHƯƠNG 4: THỰC HIỆN THU THẬP CÁC THÔNG BÁO TUYỂN DỤNG	41
4.1 ĐỐI TƯỢNG VÀ DỮ LIỆU QUAN TÂM.....	41
4.2 GIỚI THIỆU CÔNG CỤ R VÀ CÁC THƯ VIỆN	44
4.2.1 Giới thiệu ngôn ngữ R.....	44
4.2.2 Các gói thư viện hỗ trợ web scraping	46
4.3 THỰC HIỆN TẢI XUỐNG DỮ LIỆU VỚI CÔNG CỤ R.....	47
CHƯƠNG 5: ỨNG DỤNG TOPIC MODEL PHÂN TÍCH THÔNG BÁO TUYỂN DỤNG.....	49
5.1 CÁC CÔNG CỤ VÀ THƯ VIỆN CẦN THIẾT.....	49
5.1.1 Thư viện Tm.....	49
5.1.2 Thư viện Quanteda	50
5.1.3 Thư viện Topicmodels	50
5.2 DỮ LIỆU VÀ TIỀN XỬ LÝ	51
5.2.1 Mô tả dữ liệu	51
5.2.2 Tiền xử lý dữ liệu	51
5.3 THỰC HIỆN TOPIC ANALYSIS CÁC THÔNG BÁO TUYỂN DỤNG.....	52
CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ PHÂN TÍCH	54
6.1 BIỂU DIỄN KẾT QUẢ	54
6.2 GIẢI THÍCH KẾT QUẢ	57
KẾT LUẬN	60
TÀI LIỆU THAM KHẢO	62
PHỤ LỤC A	63
PHỤ LỤC B.....	65
PHỤ LỤC C.....	70
PHỤ LỤC D	71

DANH MỤC BẢNG BIỂU

Bảng 2.2.1: Các nhóm thẻ thông dụng trong cấu trúc HTML.....	23
Bảng 2.2.2: Mô tả chi tiết một số thẻ trong cấu trúc HTML.....	23
Bảng 2.2.3: Một số thuộc tính của thẻ body.....	24
Bảng 3.3.1.2.1: Ma trận tài liệu – từ vựng (document-term matrix).....	30
Bảng 3.3.1.2.2: Mỗi chủ đề là một phân bố xác suất trên tập từ vựng.....	30
Bảng 3.3.1.2.3: Mỗi tài liệu là một phân bố xác suất trên tập chủ đề	30
Bảng 3.3.2.1: Các điểm khác biệt quan trọng giữa PLSA và LDA.....	34
Bảng 3.3.2.2.1: Mỗi tài liệu là một phân phối Dirchlet trên tập chủ đề	37
Bảng 3.3.2.2.2: Mỗi chủ đề là một phân phối Dirchlet trên tập từ vựng	37
Bảng 3.3.2.2.3: Đại lượng trong mô hình LDA	38
Bảng 4.1: Một số trang web việc làm nổi tiếng tại Việt Nam và trên thế giới.....	41
Bảng 4.2.1.1: Một số thành phần trong R.....	44
Bảng 4.2.1.2: Một số hàm sẵn có thường dùng trong R.....	45
Bảng 4.2.2.1: Một số hàm phổ biến trong thư viện Rvest.....	46
Bảng 4.2.2.2: Một số hàm phổ biến trong thư viện Xlm2.....	46
Bảng 4.3.1: Các đối tượng lưu trữ thông tin trong quá trình web scraping	47
Bảng 4.3.2: Chi tiết cấu hình máy thực hiện web scraping và topic analysis	48
Bảng 5.1.1: Một số hàm phổ biến trong thư viện Tm	49
Bảng 5.1.2: Một số hàm phổ biến trong thư viện Quanteda	50
Bảng 5.1.2: Một số hàm phổ biến trong thư viện Quanteda	51
Bảng 5.3: Các tham số và giá trị tham số khi thực nghiệm với LDA	52
Bảng 6.1.1: Chi tiết tần suất của 50 từ có tần suất xuất hiện cao nhất	54
Bảng 6.1.2: Kết quả 40 chủ đề của của tập tài liệu khi phân tích với LDA	55
Bảng 6.1.3: Mức độ đóng góp của 40 chủ đề vào tập tài liệu Bigdata.csv	57
Bảng 6.2.1: Kết quả dự đoán tên chủ đề của một số topic	58
Bảng 6.2.2: Thống kê số lần xuất hiện của một số keyword trong các chủ đề	59

DANH MỤC HÌNH VẼ, SƠ ĐỒ

Hình 2.1: Xác định nguồn dữ liệu và dữ liệu quan tâm	21
Hình 2.2: Khung cấu trúc chung của một tài liệu HTML	23
Hình 3.1.2: Xác định chủ đề văn bản dựa trên ý tưởng phân bố xác suất của từ trong tập tài liệu	26
Hình 3.2.1: Input và output của mô hình chủ đề	27
Hình 3.2.2: Lược đồ thống kê tần số của từ theo định luật Zipf	29
Hình 3.3.1.2: Mối quan hệ sinh trong mô hình PLSA.....	31
Hình 3.3.1.3: Sơ đồ thuật toán EM.....	33
Hình 3.3.2.2.1: Đồ thị biểu diễn phân bố xác suất Dirichlet của biến ngẫu nhiên x trong 15 lần tạo ngẫu nhiên.....	36
Hình 3.3.2.2.2: Mối quan hệ sinh trong mô hình LDA	37
Hình 3.3.2.2.3: Minh họa phân phối Multinomial trong mô hình LDA.....	39
Hình 4.1.1: Trang kết quả tìm kiếm cho từ khóa Big data trên trang dice.com	43
Hình 4.1.2: Thông tin chi tiết vị trí Senior Big Data Architect trên trang dice.com.....	43
Hình 4.1.3: Cấu trúc lưu trữ cho các thông báo việc làm thu thập được.....	44
Hình 6.1.1: Biểu đồ thể hiện 50 từ có tần suất xuất hiện cao nhất khi phân tích	54
Hình 6.1.2: Biểu đồ thể hiện thông tin chi tiết chủ đề 35 (topic 35)	56
Hình 6.1.3: Mức độ đóng góp của các chủ đề vào các tài liệu.....	56

DANH MỤC TỪ VIẾT TẮT

EM	E xpectation M aximization
HTML	H yper T ext M arkup L anguage
LDA	L atent D irichlet A llocation
PLSA	P robabilistic L atent S emantic A nalysis
URL	U niform R esource L ocator
XML	e Xtensible M arkup L anguage

TÓM TẮT

Trong đề tài này, chúng tôi thực hiện việc thu thập dữ liệu trên web tuyển dụng dice.com liên quan đến từ khóa "big data" và tiến hành phân tích để khám phá chủ đề từ các thông báo việc làm để đưa ra các quan tâm của nhà tuyển dụng. Để làm được điều đó, chúng tôi tiến hành tìm hiểu nền tảng lý thuyết web scraping, topic analysis và cuối cùng là tiến hành thực nghiệm.

Đầu tiên chúng tôi tìm hiểu web scraping là quá trình thu thập và tải xuống dữ liệu quan tâm trên web một cách tự động, trong đó giới thiệu về cấu trúc HTML để xác định được vị trí dữ liệu cần lấy trong tài liệu HTML. Chúng tôi trình bày các bước để thu thập bao gồm: xác định nguồn dữ liệu quan tâm, công cụ tiến hành trích xuất dữ liệu, tải xuống và lưu trữ dữ liệu.

Tiếp theo chúng tôi giới thiệu cơ sở lý thuyết về bài toán topic analysis:

- + Input: tập tài liệu C và số k là số chủ đề cần xác định
- + Output: tập chủ đề θ và mức độ đóng góp các chủ đề vào tài liệu
- + Quá trình thực hiện topic analysis: lựa chọn dữ liệu, tiền xử lý, vector hóa, thu giảm số chiều và rút trích đặc trưng, phân tích và đánh giá kết quả.

Trong đó việc phân tích có thể áp dụng các thuật toán PLSA và LDA. Ý tưởng chung của mô hình PLSA và LDA đều xem mỗi tài liệu là một phân bố xác suất trên tập chủ đề, mỗi chủ đề là phân bố trên tập từ vựng. Các xác suất này là tham số chưa biết và cố gắng đi ước lượng các tham số sao cho khả năng sinh ra tài liệu là cao nhất. Việc ước lượng tham số trong mô hình PLSA sử dụng phương pháp maximum likelihood (thông qua thuật toán EM để tìm cực đại hàm likelihood). Việc ước lượng tham số trong mô hình LDA sử dụng phương pháp lấy mẫu Gibbs. Chúng tôi cũng nêu rõ điểm khác nhau giữa PLSA và LDA: số lượng từ trong một tài liệu, số lượng tham số, phân phối xác suất sử dụng, quá trình sinh văn bản.

Sau cùng chúng tôi tiến hành thực nghiệm bao gồm việc thu thập dữ liệu và phân tích chủ đề. Tại bước thu thập dữ liệu, chúng tôi giới thiệu các trang web việc làm nổi tiếng tại Việt Nam (Vietnamworks, Jobstreet, ITviec, TopCV,...) và trên thế giới (Linkedin, Indeed, Monster.com, Internships.com, Dice.com,...). Chúng tôi thu thập các thông báo việc làm liên quan đến từ khóa "big data" trên trang www.dice.com (do có cấu trúc HTML tương đối đồng nhất). Chúng tôi sử dụng các thư viện cung cấp sẵn trong

R (Rvest, Xml2) để tải xuống dữ liệu và lưu trữ dưới dạng file CSV. Kết quả là tệp CSV chứa 15276 thông báo việc làm (dung lượng 36MB) có 6 trường thông tin: job title, company name, job location, links, job description, job overview. Sau đó chúng tôi sử dụng các thư viện R sẵn có (Tm, Quanteda, Topicmodel) để tiền xử lý dữ liệu (ngắt câu, ngắt từ, xóa ký tự lạ hoặc các stopwords) và phân tích chủ đề cho trường dữ liệu job description. Thuật toán mà chúng tôi sử dụng là LDA với tham số đầu vào: $\alpha = 1.25$, $\beta = 0.01$, $k = 40$. Chúng tôi biểu diễn 50 từ có tần số xuất hiện cao nhất lên đồ thị như những từ khả năng cao đóng góp vào chủ đề. Kết quả phân tích chủ đề là 40 chủ đề trong đó liên quan nhiều đến: data, developer, software,... Dựa theo kết quả đó chúng tôi dự đoán tên cho từng chủ đề và biểu diễn mức độ đóng góp các chủ đề vào mỗi tài liệu. Chúng tôi cũng thống kê số lần xuất hiện các keyword mà chúng tôi quan tâm: sql, java, python, excel, hadoop, spark, azure, tableau,... và rút ra được một số nhận xét như: số lượng việc làm liên quan đến java và python nhiều hơn hẳn so với lượng việc làm liên quan đến r; việc thành thạo Excel cũng được quan tâm không kém so với việc nắm vững framework liên quan đến xử lý dữ liệu lớn như aws, hadoop, spark.

Kết thúc báo cáo, chúng tôi đưa ra các kết luận về đề những điểm đạt được, các hạn chế tồn đọng, các biện pháp cải thiện và hướng phát triển tiếp theo.

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 GIỚI THIỆU BÀI TOÁN

Nghiên cứu tuyển dụng là một trong các hoạt động có ý nghĩa quan trọng đối với doanh nghiệp và xã hội. Việc khảo sát về yêu cầu cho một vị trí công việc, mức lương, chính sách chế độ đãi ngộ,... tốn nhiều nguồn lực, chi phí lớn. Quá trình thu thập và phân tích các khảo sát một cách thủ công cũng mất nhiều thời gian hơn và không bắt kịp xu hướng phát triển của thị trường việc làm trong tương lai. Vấn đề đặt ra ở đây là làm sao phát hiện được các yêu cầu chung cho một vị trí tuyển dụng, mức lương phổ biến cho vị trí công việc cũng như chế độ đãi ngộ đối với nhân viên như thế nào để vừa nhanh chóng, vừa chính xác và đạt được hiệu quả cao. Một phương án được đề xuất là sử dụng dữ liệu việc làm có sẵn từ Internet, kết hợp các công nghệ hỗ trợ giúp thu thập thông tin trên web tuyển dụng, thông qua thu thập và đánh giá tự động để tìm ra được các quan tâm của ứng viên và nhà tuyển dụng. Việc thu thập tự động giúp giảm thiểu sự sai sót và **nhọc nhằn** cho con người so với tiến hành khảo sát trực tiếp. Hơn nữa với tốc độ việc làm thay đổi chóng mặt, việc thu thập và phân tích tự động cho độ trễ thời gian thấp hơn, kết quả phù hợp hơn với tình hình hiện tại. Kết quả cũng mang ý nghĩa quan trọng tác động đến định hướng đào tạo sinh viên tại các trường đại học nhằm đáp ứng nhu cầu nhân lực tại các doanh nghiệp.

1.2 MỤC TIÊU

Nhiệm vụ của đề tài là thu thập dữ liệu cho một loại việc làm cụ thể trên trang web tuyển dụng nổi tiếng. Sau đó tiến hành phân tích theo chủ đề để phát hiện các yêu cầu chung, mức lương phổ biến, chính sách đãi ngộ,... tương ứng cho loại việc làm đó, đưa ra các từ khóa phổ biến liên quan đến vị trí công việc đang tìm kiếm, phán đoán được các kỹ năng mới, các xu hướng trong tương lai. Để đạt được điều đó, chúng tôi tập trung tìm hiểu một số vấn đề sau:

- + Tìm hiểu khái niệm web scarping và cấu trúc HTML
- + Tìm hiểu khái niệm topic analysis và các kỹ thuật phân tích theo chủ đề
- + Thực hiện công việc thu thập dữ liệu web tuyển dụng với các thư viện của R
- + Phân tích theo chủ đề đối với dữ liệu thu thập được để đưa ra các thông tin quan tâm
- + Đánh giá và giải thích kết quả

1.3 BỐ CỤC CỦA BÁO CÁO

Các phần còn lại của báo cáo khóa luận được tổ chức như sau:

Chương 2: Tổng quan về web scraping

Chương 3: Tổng quan topic analysis

Chương 4: Thực hiện thu thập các thông báo tuyển dụng

Chương 5: Ứng dụng topic model phân tích các thông báo tuyển dụng

Chương 6: Đánh giá và giải thích kết quả

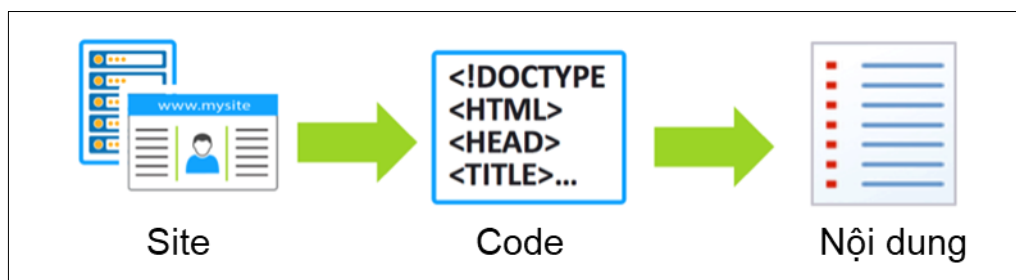
CHƯƠNG 2: TỔNG QUAN WEB SCRAPING

2.1 KHÁI NIỆM WEB SCRAPING

Sự ra đời mạng máy tính và môi trường world wide web (www) cung cấp nguồn thông tin và dữ liệu vô cùng phong phú. Con người có thể khai thác chúng để có được những thông tin giá trị phục vụ cho hoạt động kinh tế, chính sách, giáo dục, y tế,... Với nguồn dữ liệu lớn vượt quá khả năng xử lý của con người, các kỹ thuật thu thập và phân tích dữ liệu web dần được quan tâm, nghiên cứu và phát triển. Web scraping là một kỹ thuật phổ biến để thu thập dữ liệu trên web. Web scraping được hiểu là quá trình thu thập và tải xuống dữ liệu quan tâm có trên web một cách tự động [1]. Quá trình web scraping trải qua các giai đoạn chính:

- + Xác định nguồn dữ liệu và dữ liệu quan tâm
- + Sử dụng công cụ để tiến hành trích xuất dữ liệu quan tâm
- + Tải xuống và lưu trữ dữ liệu theo cấu trúc nhất định

Giai đoạn đầu tiên là xác định nguồn dữ liệu và dữ liệu quan tâm. Chúng ta cần xác định trang web mà ta đang hướng đến, nắm được cấu trúc của chúng và xác định dữ liệu ta quan tâm nằm ở đâu. Dữ liệu web có thể sẵn có để tải xuống như các tài liệu TXT, CSV, PDF, XLS, JPEG nhưng cũng có thể tồn tại ở dạng tài liệu khác như HTML.



Hình 2.1: Xác định nguồn dữ liệu và dữ liệu quan tâm

Tiếp sau đó chúng ta cần lựa chọn một công cụ phù hợp để tiến hành thu thập dữ liệu. Có rất nhiều công cụ hỗ trợ cung cấp giao diện thân thiện và thao tác đơn giản, phổ biến như: myTrama, import.io, kimonolabs, Tabula, Zamzar, cometDocs, Navigator extensions, Google spreadsheets IMPORTHTML,... Chúng ta cũng có thể sử dụng các ngôn ngữ phổ biến như Java, Python, R, Ruby, Node,... để lập trình công cụ thu thập

dữ liệu web. Cuối cùng là xác định cấu trúc lưu trữ và thực hiện tải xuống dữ liệu. Dữ liệu đầu ra có thể cấu trúc như TXT, CSV, JSON, XML, XLS,...

Các vấn đề lớn liên quan đến web scraping bao gồm [3]:

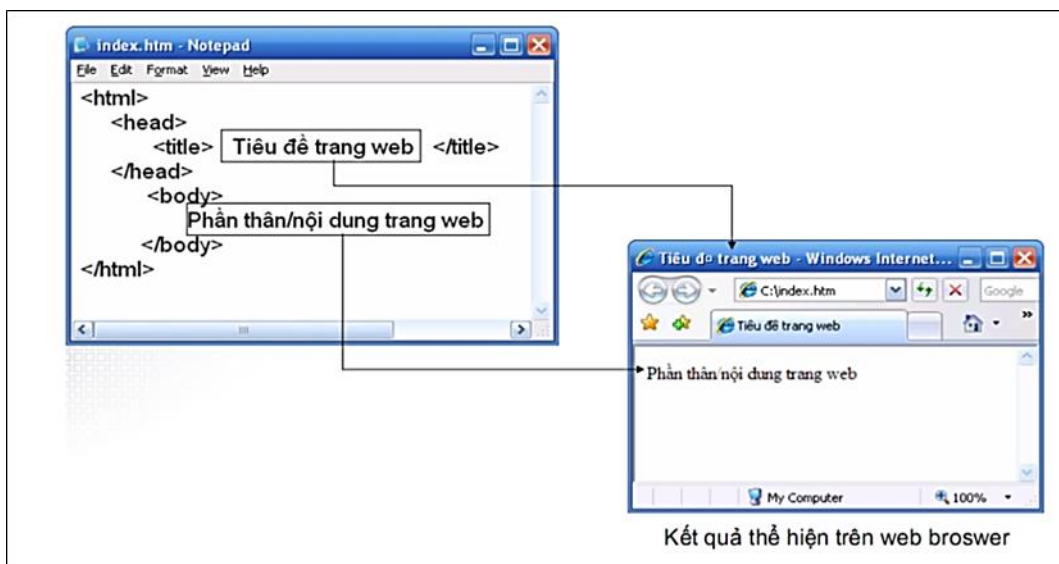
- + Cấu trúc của các web là không giống nhau và thường xuyên thay đổi, đặt ra yêu cầu cần có sự thay đổi về kỹ thuật để tương thích. Vấn đề làm sạch, tổ chức, xử lý dữ liệu cũng gặp nhiều khó khăn hơn do sự khác biệt về cấu trúc.
- + Thông tin thu thập có thể trùng lặp hoặc không còn phù hợp do có thể một vị trí được đăng tuyển trên nhiều trang khác nhau và nội dung thường xuyên được cập nhật, làm mới.
- + Vấn đề quyền sử dụng dữ liệu và tính an toàn cũng được đề cập đến trong các dự án về web scraping. Biện pháp an toàn đưa ra là tách biệt máy chủ lưu trữ dữ liệu tải xuống khỏi hệ thống lưu trữ của doanh nghiệp và thực hiện điều khiển từ xa.

Để có thể tiến hành bước thu thập dữ liệu web, trước hết chúng ta cần nắm về cấu trúc web để có các biện pháp kỹ thuật tương ứng. Trong đề tài này, chúng tôi chỉ tập trung vào thu thập dữ liệu web dạng HTML bởi sự phổ biến của loại web này.

2.2 CẤU TRÚC TÀI LIỆU HTML

HTML (HyperText Markup Language) là ngôn ngữ đánh dấu siêu văn bản dùng các thẻ (tags) để định dạng dữ liệu và tạo khung hoặc bảng cho trang web. HTML cho phép chúng ta tạo ra các trang phối hợp hài hòa giữa văn bản thông thường với hình ảnh, âm thanh, video, các mối liên kết đến các trang siêu văn bản khác... Tài liệu HTML được thiết kế để phân phối nội dung trên môi trường web. Tập hợp các tài liệu HTML cùng các tài liệu khác tạo thành website cung cấp thông tin về một tổ chức hoặc cá nhân. Theo quy ước, tất cả các tệp mã nguồn của trang siêu văn bản phải có đuôi là .html hoặc .htm. Chúng ta có thể truy cập đến một tài liệu HTML thông qua URL - một địa chỉ web và là một chuẩn xác định tài liệu web trên Internet.

Một tài liệu HTML được tạo nên từ nhiều thành phần HTML. Một thành phần HTML được đánh dấu bằng một cặp thẻ mở (<keyword>) và thẻ đóng (</keyword>). Các thành phần HTML có thể cấu trúc phân cấp hình cây, trong đó thành phần mẹ chứa nhiều thành phần con lồng bên trong nó. Khung cấu trúc của một tài liệu HTML được thể hiện như hình 2.2.



Hình 2.2: Khung cấu trúc chung của một tài liệu HTML

Các thẻ dùng để báo cho trình duyệt cách thức trình bày văn bản trên màn hình hoặc dùng để chọn một môi liên kết đến các trang hoặc một đoạn chương trình khác. Lệnh sẽ tác động vào đoạn văn bản nằm giữa hai thẻ. Các nhóm thẻ phổ biến được liệt kê trong bảng 2.2.1.

Bảng 2.2.1: Các nhóm thẻ thông dụng trong cấu trúc HTML

Thẻ	Ví dụ điển hình
Thẻ định dạng trang	<body>
Thẻ định dạng văn bản	, <p>, , <i>, <u>,...
Thẻ tạo siêu liên kết (hyperlink)	<a>
Thẻ định dạng danh sách	, ,
Thẻ chèn hình ảnh	

Chi tiết một số thẻ trong cấu trúc HTML và ví dụ được thể hiện trong bảng 2.2.2.

Bảng 2.2.2: Mô tả chi tiết một số thẻ trong cấu trúc HTML

Thẻ và thuộc tính	Ý nghĩa	Ví dụ
<sup>	Định dạng chỉ số trên	ax^{2}
<sub>	Định dạng chỉ số dưới	$H_{2}O$
<h1>,...,<h6>	Định dạng tiêu đề từ kích thước 1 đến 6 (tiêu đề 1 có kích thước lớn nhất)	<h1>Tiêu đề 1</h1> →Tiêu đề 1
<pre>	Bỏ qua định dạng của các thẻ	Xin chào

	HTML bên trong (không bỏ qua ký tự khoảng trắng, tab và xuống dòng)	<code>các bạn</code> → <code>Xin chào</code> các bạn
<code>
</code>	Kết thúc dòng hiện tại và chuyển sang dòng mới	<code><p>Dòng 1
Dòng 2</p></code> → Dòng 1 Dòng 2
<code></code>	Thiết lập font chữ cho văn bản	<code></code>
<code></code>	Định dạng chữ in đậm	<code>Đoạn văn bản in đậm</code>
<code><i></code> hay <code></code>	Định dạng chữ in nghiêng	<code><i>Đoạn văn bản in nghiêng</i></code>
<code><u></code>	Định dạng chữ in gạch dưới	<code><u>Đoạn văn bản in gạch dưới</u></code>
<code><strike></code>	Định dạng chữ in gạch ngang chữ	<code><strike>Đoạn văn bản in gạch ngang chữ</strike></code>

Mỗi thẻ có thể có một hoặc nhiều thuộc tính đi kèm. Ví dụ một số thuộc tính của thẻ body như trong bảng 2.2.3.

Bảng 2.2.3: Một số thuộc tính của thẻ body

Thuộc tính	Ý nghĩa	Ví dụ
<code>bcolor="color"</code>	Thiết lập màu nền cho trang web	<code><body bcolor="#00FF00"></code>
<code>background="url"</code>	Thiết lập ảnh nền cho trang web	<code><body background="image.jpg"></code>
<code>leftmargin="num"</code> <code>topmargin="num"</code>	Thiết lập lề trái và lề trên cho trang web	<code><body leftmargin="0" topmargin="0"></code>
<code>text="color"</code>	Thiết lập màu chữ của văn bản, kể cả các đề mục	<code><body text="#FF0000"></code>
<code>alink="color"</code> <code>vlink="color"</code> <code>link="color"</code>	Xác định màu sắc cho các siêu liên kết trong văn bản (active link, visited link, link)	<code><body alink="#FF0000" vlink="#00FF00">link=</code> <code>"#0000FF"></code>

Các thẻ siêu liên kết (hyperlink) cho phép người dùng duyệt từ trang web này đến trang web khác. Cú pháp tạo thẻ hyperlink:

` Label `

Trong đó:

- + URL là địa chỉ của trang liên kết.
- + Label có thể là text, button, hình ảnh,...

Ngoài ra còn một số thẻ khác liên quan đến các loại tài liệu video, âm thanh, hình ảnh,... mà chúng tôi xin phép không trình bày trong báo cáo này.

3.2 BÀI TOÁN TOPIC ANALYSIS

3.2.1 Đối tượng và dữ liệu quan tâm

Topic model khám phá k chủ đề từ một tập các tài liệu cho trước [4], cụ thể:

Input:

- + Tập tài liệu C gồm n tài liệu: $C = \{d_1, d_2, d_3, \dots, d_n\}$.
- + Một số k được định sẵn là số chủ đề muốn khám phá.

Output:

- + Tập chủ đề θ gồm k chủ đề: $\theta = \{\theta_1, \theta_2, \theta_3, \dots, \theta_k\}$. Mỗi chủ đề θ_i là một phân bố xác suất **xuất** trên tập từ vựng V :

$$\sum_{w \in V} p(w|\theta_i) = 1$$

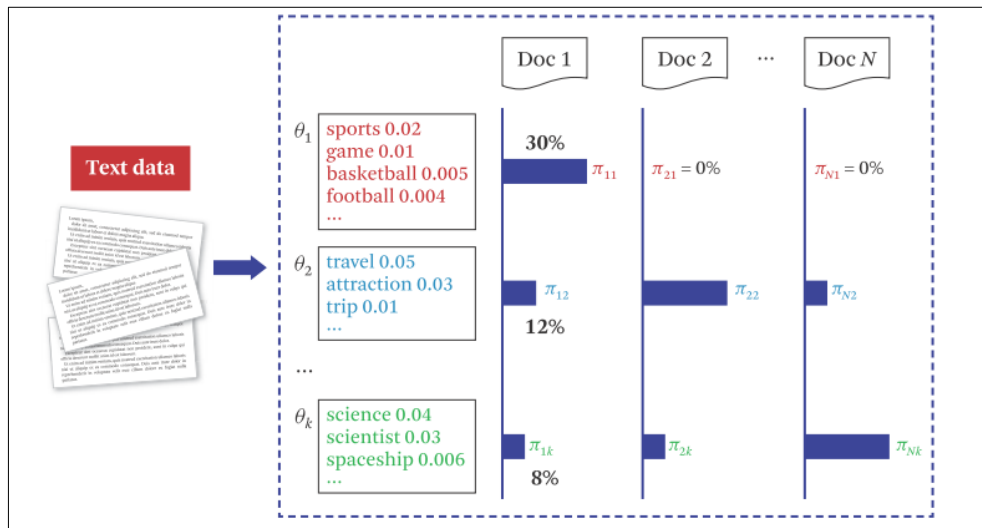
Tập từ vựng $V = \{w_1, w_2, w_3, \dots, w_m\}$ chứa tất cả các từ riêng biệt có trong tập C . Mỗi chủ đề θ_i đều bao gồm tất cả các từ trong tập từ vựng V với các xác suất khác nhau.

- + Với mỗi tài liệu d_i có mức độ đóng góp của các chủ đề $\{\pi_{i1}, \pi_{i2}, \pi_{i3}, \dots, \pi_{ik}\}$:

$$\sum_{j=0}^k \pi_{ij} = 1$$

Trong đó $\pi_{ij} = p(\theta_j|d_i)$ chính là mức độ đóng góp của chủ đề θ_j vào tài liệu d_i .

Hình 3.2.1 là một minh họa trực quan input/output của mô hình chủ đề.



Hình 3.2.1: Input và output của mô hình chủ đề

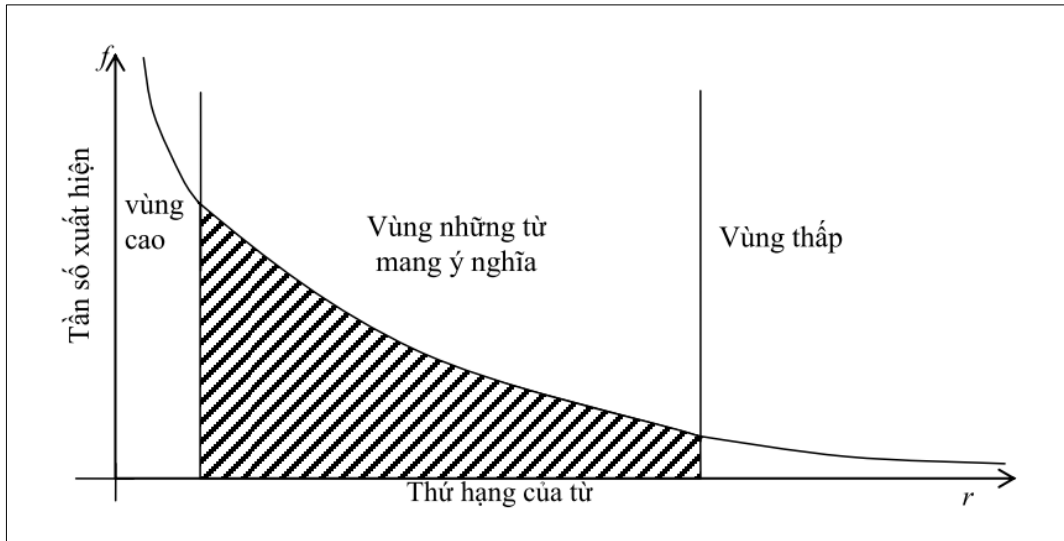
3.2.2 Các bước thực hiện

Bài toán phân tích theo chủ đề có thể được giải quyết theo các bước:

- Bước 1: **Lựa chọn tập dữ liệu** mà chúng ta cần tiến hành phân tích để xác định chủ đề.
- Bước 2: **Tiền xử lý dữ liệu**: bao gồm các công việc ngắt câu, ngắt từ, loại bỏ ký tự đặc biệt, các từ không đúng đắn,... Chúng ta cũng loại bỏ các stopwords như giới từ, từ nối vì chúng không thể hiện được chủ đề văn bản. Đưa các từ cùng gốc hoặc biến thể của chúng về từ gốc [2].
- Bước 3: **Số hóa văn bản thành vector**: ta xem mỗi văn bản được biểu diễn bằng một vector trong không gian nhiều chiều với mỗi chiều tương ứng với một mục từ riêng biệt trong tập tài liệu sau tiền xử lý. Mỗi thành phần vector mang một giá trị cho biết tần số của từ đó trong tài liệu.
- Bước 4: **Thu giảm số chiều và rút trích đặc trưng**: vì số lượng mục từ nhiều (số chiều không gian lớn) có thể ảnh hưởng kết quả phân tích nên chúng ta cần thu giảm số chiều trước khi phân tích. Tiến hành loại bỏ các từ có tần suất thấp hoặc từ có tần suất cao nhưng vô nghĩa vì chúng không thể hiện được chủ đề của tập dữ liệu.
- Bước 5: **Áp dụng kỹ thuật phân tích theo chủ đề** để rút trích các từ thể hiện được chủ đề của tập dữ liệu. Dựa trên các từ rút trích được, xác định ra k chủ đề được đề cập trong tập tài liệu và mức độ đóng góp của chủ đề vào tài liệu.

Một kỹ thuật đơn giản để xác định chủ đề của tài liệu là dựa vào tần số xuất hiện của các từ trong tài liệu đó. Các từ có tần **xuất** xuất hiện cao nhất sẽ được lựa chọn là chủ đề của tập tài liệu. **Tuy nhiên phương** pháp này tồn tại nhiều khuyết điểm:

- + Chủ đề không rõ ràng do chỉ được thể hiện thông qua một từ duy nhất.
- + Dễ xảy ra trường hợp các từ có mức độ bao phủ rộng lớn được xem là chủ đề nhưng thực tế chúng không mang ý nghĩa đại diện như các từ “*a*”, “*the*”, “*of*”,... trong Tiếng Anh (tham khảo hình 3.2.2).
- + Các từ khác nhau về hình thức nhưng giống hoặc có sự tương đồng về ý nghĩa dẫn đến sự giảm số lượng các chủ đề riêng biệt.



Hình 3.2.2: Lược đồ thống kê tần số của từ theo định luật Zipf

Chúng ta có thể sử dụng các kỹ thuật phổ biến như PLSA, LDA để khắc phục một số khuyết điểm của phương pháp trên. Các kỹ thuật sẽ được trình bày chi tiết hơn ở mục 3.3.

3.3 CÁC THUẬT TOÁN PHÂN TÍCH CHỦ ĐỀ (TOPIC ANALYSIS)

3.3.1 Thuật toán PLSA

3.3.1.1 Ý tưởng mô hình PLSA

Ý tưởng của PLSA (Probabilistic Latent Semantic Analysis) xem tập tài liệu bao gồm nhiều văn bản, mỗi văn bản là một phân phối các chủ đề, mỗi chủ đề là một phân phối từ. Mô hình PLSA tìm cách ước lượng các tham số (các xác suất) sao cho khả năng sinh ra cả tập tài liệu là cao nhất. Để thực hiện điều đó, tác giả đề xuất PLSA đã áp dụng thuật toán EM để xác định các tham số sao cho hàm log-likelihood đạt cực đại.

3.3.1.2 Giải thích ý tưởng

Giả sử tập tài liệu C gồm n tài liệu, $C = \{d_1, d_2, d_3, \dots, d_n\}$. Mỗi tài liệu d_i gồm nhiều từ. Nếu bỏ qua các quy định về ngữ pháp, trật tự từ, cấu trúc câu, ta định nghĩa tập từ vựng V là một tập hợp các từ có trong tập tài liệu C . Mỗi từ trong tập từ vựng là một khóa riêng biệt không trùng lặp (còn gọi là thuật ngữ). Từ tập từ vựng này, nếu lựa chọn các từ với xác suất và tần suất khác nhau ta thu được các tài liệu. Các tài liệu được xem như một túi các từ mà không quan trọng đến thứ tự xuất hiện các từ [5].

Gọi $c(w_j, d_i)$ là tần số xuất hiện của từ w_j trong văn bản d_i . Giá trị này hoàn toàn có thể xác định được khi biết tập tài liệu C và tập từ vựng V . Minh họa được thể hiện trong bảng 3.3.1.2.1.

Bảng 3.3.1.2.1: Ma trận tài liệu – từ vựng (document-term matrix)

	$w_{j=1}$	$w_{j=2}$	$w_{j=3}$...	$w_{j=m}$
$d_{i=1}$	$c(w_1, d_1)$	$c(w_2, d_1)$	$c(w_3, d_1)$		$c(w_m, d_1)$
$d_{i=2}$	$c(w_1, d_2)$	$c(w_2, d_2)$	$c(w_3, d_2)$		$c(w_m, d_2)$
$d_{i=3}$	$c(w_1, d_3)$	$c(w_2, d_3)$	$c(w_3, d_3)$		$c(w_m, d_3)$
...					
$d_{i=n}$	$c(w_1, d_n)$	$c(w_2, d_n)$	$c(w_3, d_n)$		$c(w_m, d_n)$

Giải sử ta cần tìm k chủ đề "ẩn", tức là chỉ biết $\theta = \{\theta_1, \theta_2, \theta_3, \dots, \theta_k\}$ mà không biết chính xác θ_i nói về điều gì. Ta xem mỗi θ_i là một phân bố xác suất trên tập từ vựng V , tức là mỗi chủ đề θ_i đều bao gồm tất cả các từ trong tập từ vựng với xác suất mỗi từ là khác nhau và được thể hiện như trong bảng 3.3.1.2.2. Xác suất $p(w_j|\theta_i)$ càng cao thì khả năng cao chủ đề θ_i rất liên quan đến từ w_j .

Bảng 3.3.1.2.2: Mỗi chủ đề là một phân bố xác suất trên tập từ vựng

	$w_{j=1}$	$w_{j=2}$	$w_{j=3}$...	$w_{j=m}$
$\theta_{i=1}$	$p(w_1 \theta_1)$	$p(w_2 \theta_1)$	$p(w_3 \theta_1)$		$p(w_m \theta_1)$
$\theta_{i=2}$	$p(w_1 \theta_2)$	$p(w_2 \theta_2)$	$p(w_3 \theta_2)$		$p(w_m \theta_2)$
$\theta_{i=3}$	$p(w_1 \theta_3)$	$p(w_2 \theta_3)$	$p(w_3 \theta_3)$		$p(w_m \theta_3)$
...					
$\theta_{i=k}$	$p(w_1 \theta_k)$	$p(w_2 \theta_k)$	$p(w_3 \theta_k)$		$p(w_m \theta_k)$

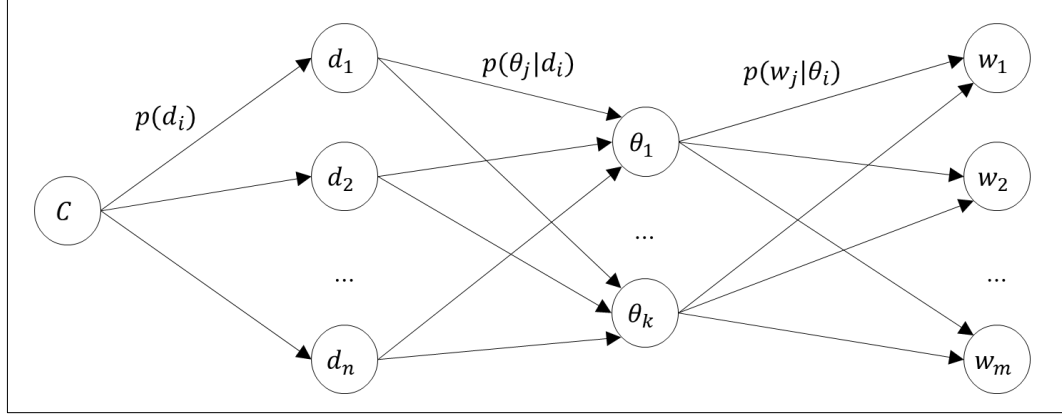
Ta cũng xem mỗi tài liệu d_i là một phân bố xác suất trên tập chủ đề θ , tức là mỗi tài liệu d_i đều bao gồm tất cả các chủ đề với xác suất mỗi chủ đề là khác nhau và được thể hiện trong bảng 3.3.1.2.3. Xác suất $p(\theta_j|d_i)$ càng cao thì khả năng cao θ_j là chủ đề được đề cập nhiều trong tài liệu d_i .

Bảng 3.3.1.2.3: Mỗi tài liệu là một phân bố xác suất trên tập chủ đề

	$\theta_{j=1}$	$\theta_{j=2}$	$\theta_{j=3}$...	$\theta_{j=k}$
$d_{i=1}$	$p(\theta_1 d_1)$	$p(\theta_2 d_1)$	$p(\theta_3 d_1)$		$p(\theta_k d_1)$
$d_{i=2}$	$p(\theta_1 d_2)$	$p(\theta_2 d_2)$	$p(\theta_3 d_2)$		$p(\theta_k d_2)$

$d_{i=3}$	$p(\theta_1 d_3)$	$p(\theta_2 d_3)$	$p(\theta_3 d_3)$		$p(\theta_k d_3)$
...					
$d_{i=n}$	$p(\theta_1 d_n)$	$p(\theta_2 d_n)$	$p(\theta_3 d_n)$		$p(\theta_k d_n)$

Nội dung trên được minh họa rõ hơn trong hình 3.3.1.2.



Hình 3.3.1.2: Mối quan hệ sinh trong mô hình PLSA

Theo đó, các ràng buộc cần **thỏa** mãn bao gồm:

$$\sum_{j=1}^{|V|} p(w_j|\theta_i) = 1; \quad \sum_{j=1}^k p(\theta_j|d_i) = 1$$

Khi đó, ta dễ dàng suy luận:

i) Xác suất sinh một từ trong tài liệu:

$$p(w_j|d_i) = \sum_{\theta_l \in \theta} p(w_j|\theta_l) p(\theta_l|d_i)$$

ii) Xác suất sinh một tài liệu trong tập:

$$p(d_i) = \prod_{w_j \in V} [p(w_j|d_i)]^{c(w_j, d_i)} = \prod_{w_j \in V} \left[\sum_{\theta_l \in \theta} p(w_j|\theta_l) p(\theta_l|d_i) \right]^{c(w_j, d_i)}$$

iii) Xác suất sinh tập tài liệu:

$$p(C) = \prod_{d_i \in C} p(d_i) = \prod_{d_i \in C} \prod_{w_j \in V} \left[\sum_{\theta_l \in \theta} p(w_j|\theta_l) p(\theta_l|d_i) \right]^{c(w_j, d_i)}$$

Vậy PLSA biểu diễn tập tài liệu thông qua mô hình xác suất được đặc trưng bởi cặp $(X, p(X))$. Trong đó:

- + Tập các quan sát X chính là tập dữ liệu C , bao gồm nhiều tham số: $p(w_j|\theta_i)$ giúp xác định chủ đề, $p(\theta_j|d_i)$ giúp xác định mức độ bao phủ các chủ đề đối với các tài liệu, các tham số này là chưa biết và cần tìm.
- + Tập các phân bố xác suất **xuất** trên X chính là $p(C)$, cho biết xác suất **xuất** sinh tập tài liệu C (khả năng sinh ra tập tài liệu).

Theo suy diễn Bayes, kết quả phân bố xác suất trên tập X có thể cho ta các kết luận về phần tử trong X . Áp dụng vào bài toán, việc xác định các tham số chưa biết của C có thể đạt được khi xác suất của mô hình (giá trị $p(C)$) đạt giá trị tối đa.

3.3.1.3 Biến đổi hàm likelihood và log-likelihood

Ta định nghĩa:

- + Hàm likelihood: hàm khả năng ước tính giá trị đầu ra (tức xác suất $p(C)$).
- + Log-likelihood: lấy logarit hàm likelihood nhằm giảm độ phức tạp và dễ tính toán.
- + Các giá trị tham số làm cho hàm likelihood hoặc log-likelihood đạt cực đại chính là các giá trị tham số cần tìm.

Theo đó ta có hàm likelihood ước tính giá trị cho xác suất $p(C)$:

$$L = \prod_{d_i \in C} \prod_{w_j \in V} \left[\sum_{\theta_l \in \theta} p(w_j|\theta_l) p(\theta_l|d_i) \right]^{c(w_j, d_i)}$$

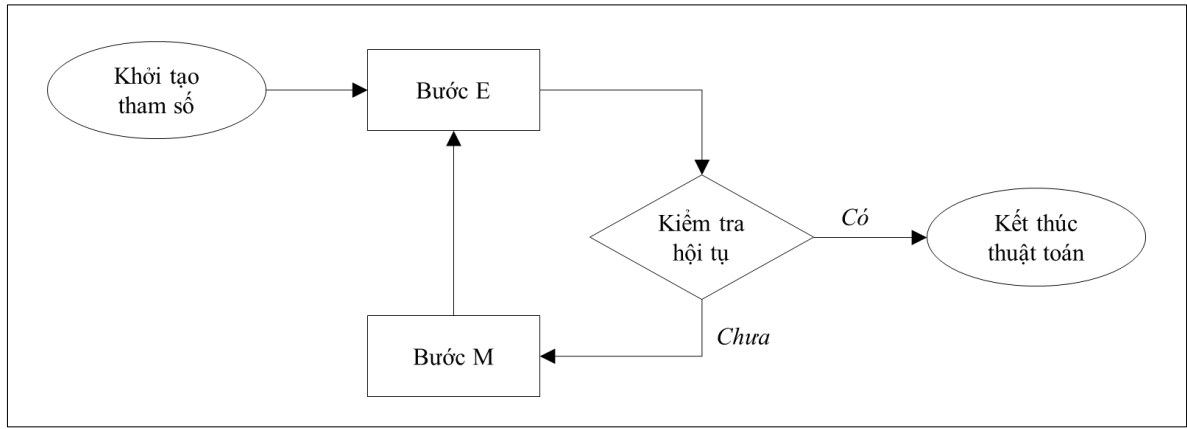
Lấy log-likelihood, biến đổi xấp xỉ và thu gọn ta được:

$$\log L = \sum_{d_i \in C} \sum_{w_j \in V} c(w_j, d_i) \log \left[\sum_{\theta_l \in \theta} p(w_j|\theta_l) p(\theta_l|d_i) \right]$$

Chi tiết biến đổi xin tham khảo phụ lục C. Để tìm cực đại hàm log-likelihood tác giả áp dụng thuật toán EM.

3.3.1.4 Thuật toán EM

Thuật toán expectation maximization (EM) là thuật toán cực đại hóa kỳ vọng nhằm ước lượng tham số cho các biến không xác định được cho mô hình thống kê. Sơ đồ minh họa thuật toán EM được thể hiện trong hình 3.3.1.3.



Hình 3.3.1.3: Sơ đồ thuật toán EM

Giải thích sơ đồ thuật toán EM:

- + Khởi tạo các tham số của mô hình: chọn ngẫu nhiên các giá trị tham số cho mô hình nằm trong đoạn $[0,1]$.
- + Bước E: tính kỳ vọng cho hàm likelihood dựa trên các tham số.
- + Bước M: cập nhật giá trị các tham số của mô hình.
- + Kiểm tra hội tụ: kết quả tính được sau nhiều lần lặp dần hội tụ hoặc không còn thay đổi nữa thì dừng thuật toán.

Áp dụng EM cho bài toán của chúng tôi:

- + Trước hết, khởi tạo ma trận tham số $(p(w_j|\theta_i), p(\theta_j|d_i))$ với giá trị phần tử trong đoạn $[0,1]$ và **thỏa**:

$$\sum_{j=1}^{|V|} p(w_j|\theta_i) = 1; \quad \sum_{j=1}^k p(\theta_j|d_i) = 1$$

Gọi biến ẩn $z_{d,w} \in \{1, 2, \dots, k\}$ mô tả phép gán chủ đề, $p(z_{d,w} = j)$ mô tả xác suất từ w của tài liệu d được sinh từ chủ đề j .

- + Bước E: Tính $p(z_{d,w} = j)$

$$p(z_{d,w} = j) = \frac{p(w|\theta_j) p(\theta_j|d)}{\sum_{l=1}^K p(w|\theta_l) p(\theta_l|d)}$$

- + Bước M: Ước lượng lại $p(w|\theta_j)$ và $p(\theta_j|d)$

$$p(w|\theta_j) = \frac{\sum_{d \in C} c(w, d) p(z_{d,w} = j)}{\sum_{i=1}^K \sum_{d \in C} c(w, d) p(z_{d,w} = i)}$$

$$p(\theta_j|d) = \frac{\sum_{w \in V} c(w, d) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w', d) p(z_{d,w'} = j)}$$

- + Quá trình này luân phiên cho đến khi giá trị hội tụ hoặc không còn thay đổi nữa. Kết quả là xác định được các $p(w|\theta_j)$ và $p(\theta_j|d)$ tức xác định được các chủ đề và mức độ đóng góp của chủ đề vào tài liệu.

3.3.1.5 Nhận xét về mô hình PLSA

Mặc dù mô hình PLSA sử dụng mô hình xác suất để biểu diễn tập tài liệu và ước lượng các giá trị tham số một cách tối ưu. **Tuy nhiên mô** hình PLSA còn tồn tại các thiếu sót:

- + Khi số lượng tài liệu trong tập tài liệu tăng dần đến các tham số xác suất tăng tuyến tính gây overfitting.
- + Không thể xác định được các tham số xác suất cho những tài liệu mới, nằm ngoài tập huấn luyện.

Một thuật toán cải tiến của PLSA là LDA giúp khắc phục các hạn chế còn tồn đọng của PLSA.

3.3.2 Thuật toán LDA

3.3.2.1 Ý tưởng mô hình LDA

Dựa trên ý tưởng của PLSA, mô hình LDA (Latent Dirichlet Allocation) xem tập tài liệu bao gồm nhiều văn bản, mỗi văn bản là một phân phối các chủ đề, mỗi chủ đề là một phân phối từ. Mô hình LDA cũng tìm cách ước lượng xấp xỉ các tham số (các xác suất) sao cho khả năng sinh ra cả tập tài liệu là cao nhất. Để thực hiện điều đó, tác giả đề xuất LDA đã tiến hành lấy mẫu Gibbs để xác định các tham số. So sánh hai mô hình LDA và mô hình PLSA ta rút ra được một số điểm khác biệt quan trọng và được làm rõ trong bảng 3.3.2.1.

Bảng 3.3.2.1: Các điểm khác biệt quan trọng giữa PLSA và LDA

Tiêu chí	PLSA	LDA
Số lượng từ trong một tài liệu	Mỗi tài liệu chứa tất cả các từ trong tập từ vựng với xác suất khác nhau	Số lượng từ trong mỗi tài liệu là khác nhau với xác suất khác nhau được rút trích từ tập từ vựng
Số lượng tham số	Số lượng tham số lớn: $n.(k-1) + k.(V-1)$ và tăng tuyến tính theo số lượng tài liệu trong tập tài liệu	Số lượng tham số ít hơn so với mô hình PLSA: $k + V$ và không phụ thuộc vào số lượng tài liệu trong tập tài liệu

Phân phối xác suất	Sử dụng phân phối đơn giản cho phân phối từ và phân phối chủ đề	Sử dụng phân phối Dirichlet cho phân phối từ và phân phối chủ đề
Quá trình sinh văn bản (tham khảo mục 3.3.1.2 và 3.3.2.2)	Sinh phân phối từ phụ thuộc quá trình sinh phân phối chủ đề	Sinh phân phối từ độc lập với quá trình sinh phân phối chủ đề

Trong mục 3.3.2.2 chúng tôi giải thích rõ hơn ý tưởng của mô hình LDA.

3.3.2.2 Giải thích ý tưởng

Giả sử tập tài liệu C gồm n tài liệu, $C = \{d_1, d_2, d_3, \dots, d_n\}$. Tập từ vựng $V = \{w_1, w_2, w_3, \dots, w_M\}$ với M là số lượng từ trong tập từ vựng. Mỗi tài liệu d_i có số lượng từ khác nhau, ký hiệu là m ($m \leq M$). Trong mô hình LDA, số lượng từ trong một tài liệu phù hợp hơn với thực tế. Số lượng tham số trong mô hình LDA ít hơn so với mô hình PLSA. Trong mô hình PLSA do sử dụng phân bố xác suất đơn giản nên kết quả phân tích chưa tốt:

- + Đối với mỗi tài liệu, do được sinh từ tất cả các chủ đề với xác suất được chọn ngẫu nhiên dễ dẫn đến sự sai lệch rất ít giữa các xác suất. Điều này dẫn đến tài liệu sinh ra nói về các chủ đề với mức độ ngang nhau trong khi thực tế mỗi tài liệu chỉ tập trung thể hiện mạnh một vài chủ đề.
- + Đối với một chủ đề, do được sinh từ tất cả các từ trong tập từ vựng với xác suất được chọn ngẫu nhiên dễ dẫn đến sự sai lệch rất ít giữa các xác suất. Điều này dẫn đến chủ đề sinh ra có một ý nghĩa tương đối rộng (do bao gồm nhiều từ có mức độ ngang nhau), trong khi thực tế mỗi chủ đề chỉ thể hiện một ý nghĩa sâu hơn liên quan đến những từ có mức độ liên quan cao.

Giải pháp cho vấn đề này là cần đưa ra một phân bố xác suất có phương sai lớn (tức các giá trị xác suất có độ lệch lớn). Phân phối mà tác giả đề xuất mô hình LDA đã sử dụng là phân phối Dirichlet.

Phân phối Dirichlet được hiểu như sau:

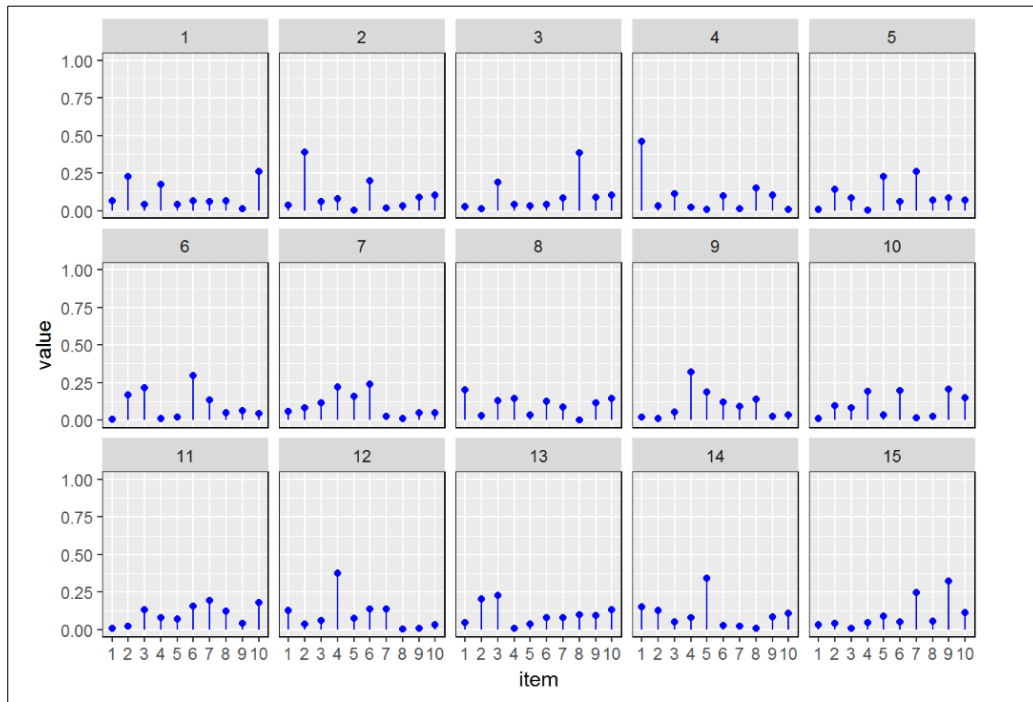
Cho vào tham số $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_k\}$ là tập số dương hoặc có thể là một phân bố xác suất đơn giản. Phân phối Dirichlet của biến ngẫu nhiên x theo tham số α , ký hiệu: $x \sim \text{Dir}(\vec{\alpha})$. Trong đó:

- + $x = \{x_1, x_2, x_3, \dots, x_k\}; \sum_{i=1}^k x_i = 1$

$$+ f(x, \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i-1} = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1}; \text{ với } \Gamma(n) = (n-1)! \text{ và } n \text{ là số nguyên;}$$

Kết quả phân phối Dirichlet được hiểu như một phân phối đã cường độ hóa mức sai lệch các giá trị xác suất trong tham số α . Chúng tôi lấy một ví dụ để dễ hình dung, chúng tôi chọn tham số $\alpha = 1, \alpha = \{\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{10}\}$. Tạo 15 phân phối Dirichlet cho cùng biến ngẫu nhiên $x: x \sim Dir(\vec{\alpha})$. Trong hình 3.3.2.2.1 biểu diễn phân bố xác suất Dirichlet của biến ngẫu nhiên x trong 15 lần tạo ngẫu nhiên.

Hình 3.3.2.2.1: Đồ thị biểu diễn phân bố xác suất Dirichlet của biến ngẫu nhiên x trong 15 lần tạo ngẫu nhiên



Chúng ta có một số nhận xét quan trọng về phân phối Dirichlet:

- + α_i càng cao thì xác suất x_i càng cao.
- + Nếu các α_i bằng nhau thì các xác suất đối xứng; đặc biệt $\alpha_1 = \alpha_2 = \dots = \alpha_k = 1$ thì các xác suất x_i là bằng nhau.
- + Nếu $\alpha_i < 1$ các xác suất ở hai đầu đạt cực đại; nếu $\alpha_i > 1$ các xác suất vùng giữa đạt cực đại.

Áp dụng phân phối Dirichlet vào bài toán:

Cho vào tham số $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_k\}$ là một phân bố xác suất đơn giản. Mỗi tài liệu d_i có phân phối chủ đề $\pi_i \sim Dir(\vec{\alpha})$ hay $d_i: \{\pi_{i1}, \pi_{i2}, \pi_{i3}, \dots, \pi_{ik}\}$ và được mô tả như trong bảng 3.3.2.2.1.

Bảng 3.3.2.2.1: Mỗi tài liệu là một phân phối Dirchlet trên tập chủ đề

	$\theta_{j=1}$	$\theta_{j=2}$	$\theta_{j=3}$	\dots	$\theta_{j=k}$
$d_{i=1}$	π_{11}	π_{12}	π_{13}		π_{1k}
$d_{i=2}$	π_{21}	π_{22}	π_{23}		π_{2k}
$d_{i=3}$	π_{31}	π_{32}	π_{33}		π_{3k}
\dots					
$d_{i=n}$	π_{n1}	π_{n2}	π_{n3}		π_{nk}

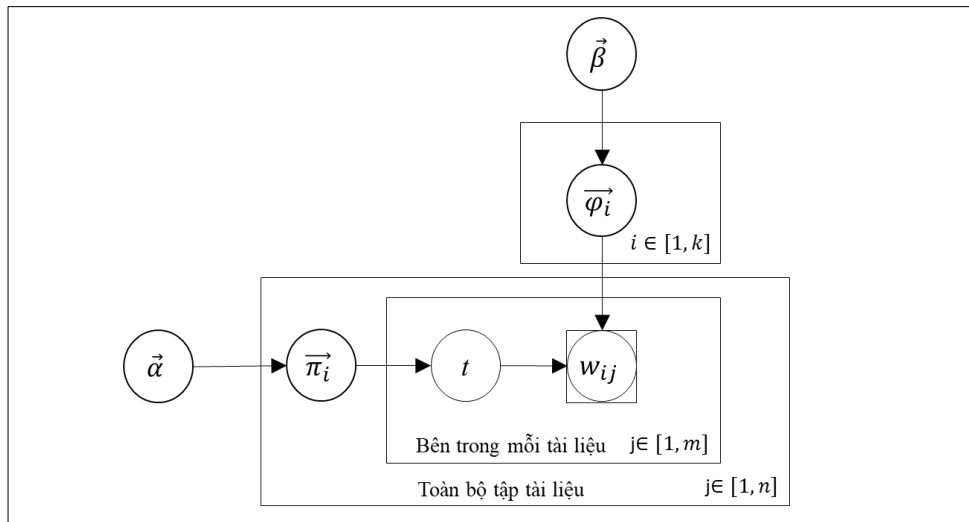
Cho vào tham số $\beta = \{\beta_1, \beta_2, \beta_3, \dots, \beta_k\}$ là một phân bố xác suất đơn giản. Mỗi chủ đề θ_i có phân phối từ $\varphi_i \sim \text{Dir}(\vec{\beta})$ hay $\theta_i: \{\varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \dots, \varphi_{ik}\}$ và được mô tả như trong bảng 3.3.2.2.2.

Bảng 3.3.2.2.2: Mỗi chủ đề là một phân phối Dirchlet trên tập từ vựng

	$w_{j=1}$	$w_{j=2}$	$w_{j=3}$	\dots	$w_{j=M}$
$\theta_{i=1}$	φ_{11}	φ_{12}	φ_{13}		φ_{1M}
$\theta_{i=2}$	φ_{21}	φ_{22}	φ_{23}		φ_{2M}
$\theta_{i=3}$	φ_{31}	φ_{32}	φ_{33}		φ_{3M}
\dots					
$\theta_{i=k}$	φ_{k1}	φ_{k2}	φ_{k3}		φ_{kM}

Quá trình sinh một văn bản trong mô hình LDA được minh họa trong hình 3.3.2.2.2. Mỗi khối vuông thể hiện quá trình lặp.

Hình 3.3.2.2.2: Mối quan hệ sinh trong mô hình LDA



Trong hình 3.3.2.2.2, quá trình sinh phân phối từ độc lập với quá trình sinh phân phối chủ đề. Các đại lượng trong mô hình LDA được mô tả chi tiết trong bảng 3.3.2.2.3.

Bảng 3.3.2.2.3: Đại lượng trong mô hình LDA

Đại lượng	Mô tả đại lượng
$\vec{\alpha}$	Tham số đầu vào cho phân phối Dirichlet của tài liệu
$\vec{\beta}$	Tham số đầu vào cho phân phối Dirichlet của chủ đề
$\vec{\pi}_i$	Phân phối chủ đề ứng với mỗi tài liệu d_i
$\vec{\varphi}_i$	Phân phối từ ứng với chủ đề θ_i
w_{ij}	Từ w_j của tài liệu d_i
k	Số lượng chủ đề khám phá
m	Số lượng từ trong tài liệu d_i
M	Số lượng từ trong tập từ vựng V
t	Chỉ số chủ đề (index of topic) được lựa chọn
n	Số lượng tài liệu trong tập tài liệu

Mã giả thuật toán LDA [2]:

```

for all  $\theta_i \in \{\theta_1, \theta_2, \theta_3, \dots, \theta_k\}$  do
{
    Sinh chủ đề  $\theta_i$  có phân phối từ:  $\varphi_i \sim Dir(\vec{\beta})$ 
    hay  $\theta_i: \{\varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \dots, \varphi_{ik}\}$ 
}
for all  $d_i \in \{d_1, d_2, d_3, \dots, d_n\}$  do
{
    1. Sinh tài liệu  $d_i$  có phân phối chủ đề:  $\pi_i \sim Dir(\vec{\alpha})$ 
    hay  $d_i: \{\pi_{i1}, \pi_{i2}, \pi_{i3}, \dots, \pi_{ik}\}$ 
    2. Tài liệu  $d_i$  có độ dài:  $m \sim Poiss(\gamma)$ 
    hay  $d_i = \{w_{i1}, w_{i2}, w_{i3}, \dots, w_{im}\}$ 
    3. for all  $w_{ij} \in \{w_{i1}, w_{i2}, w_{i3}, \dots, w_{im}\}$ 
    {
        Sinh chỉ số chủ đề (index of topic) của từ  $w_{ij}$ :  $t \sim Mult(\vec{\pi})$ 
        Sinh từ  $w_{ij}$ :  $w_{ij} \sim Mult(\vec{\varphi}_t)$ 
    }
}

```

Phân phối Multinomial được hiểu như sau:

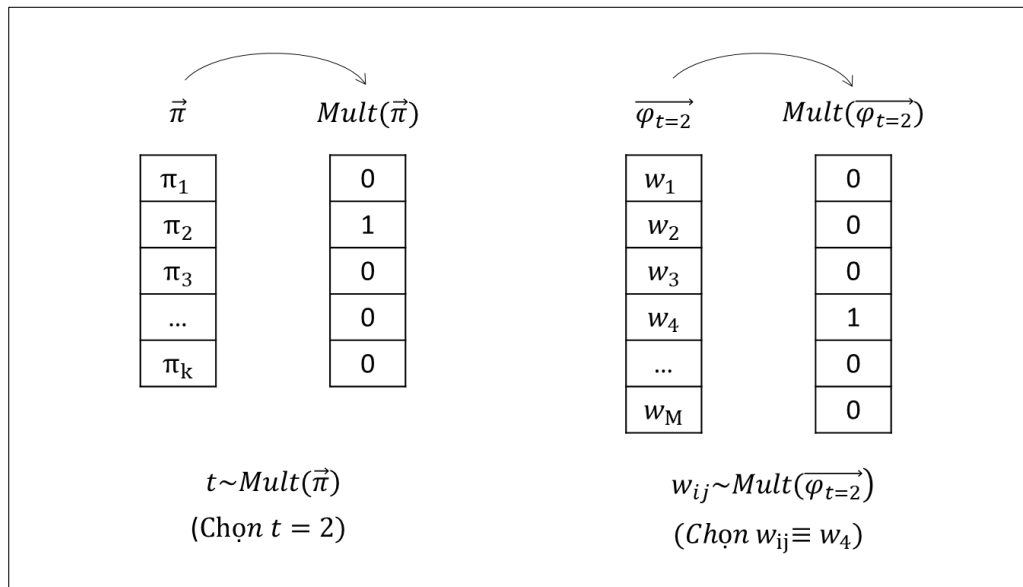
Cho vào tham số $p = \{p_1, p_2, p_3, \dots, p_k\}$ là một phân bố xác suất. Phân phối Multinomial của biến ngẫu nhiên x theo tham số p , kí hiệu: $x \sim Mult(\vec{p})$. Trong đó:

$$+ \quad x = \{x_1, x_2, x_3, \dots, x_k\}; \sum_{i=1}^k x_i = N \quad (N \text{ là số phép thử})$$

$$+ f(x, N, p) = \begin{cases} \frac{n!}{x_1! x_2! \dots x_n!} p_1^{x_1} p_2^{x_2} \dots p_n^{x_n}, & \sum_{i=1}^n x_i = N \\ 0, & \sum_{i=1}^n x_i \neq N \end{cases}$$

Kết quả phân phối Multinomial cho biết khả năng xảy ra đối với các giá trị xác suất của tham số p trong N lần thử. Trong mô hình LDA có số lần thử $N=1$, tức chỉ lựa chọn duy nhất một giá trị trong các giá trị của tham số và được minh họa trong hình 3.3.2.2.3.

Hình 3.3.2.2.3: Minh họa phân phối Multinomial trong mô hình LDA



3.3.2.3 Ước lượng tham số thông qua áp dụng lấy mẫu Gibbs

Ý tưởng Gibbs: Ý tưởng của phương pháp lấy mẫu Gibbs cũng là phương pháp ước lượng tham số như EM.

Điểm khác biệt giữa thuật toán VEM và Gibbs:

+ VEM:

+ Gibbs:

Việc sử dụng thuật toán nào là do chúng ta lựa chọn dựa trên sự phù hợp và sự khả thi.

Chúng tôi lựa chọn lấy mẫu Gibbs cho mô hình LDA với một số lý do:

+

+

Phương pháp lấy mẫu Gibbs dựa trên nền tảng toán học cao cấp (cụ thể) và cần nhiều nguồn lực và thời gian hơn mới có thể hiểu rõ. Ở đây chúng tôi chỉ dừng lại ở việc trình bày ý tưởng của phương pháp lấy mẫu Gibbs cũng như lý do lựa chọn nó cho mô hình LDA còn chi tiết việc lấy mẫu Gibbs cho bài toán xin tham khảo tài liệu [2].

3.3.2.4 Nhận xét về mô hình LDA

Mô hình LDA có một số điểm nổi bật so với mô hình PLSA:

- + Số lượng tham số ít hơn PLSA và không tăng tuyến tính khi số lượng tài liệu trong tập tài liệu tăng lên.
- + LDA thể hiện được tầm quan trọng của một số chủ đề trong tập tài liệu cũng như tầm quan trọng của một số từ trong chủ đề thông qua độ sai lệch lớn giữa các giá trị xác suất trong phân phối.
- + Quá trình sinh văn bản phù hợp hơn và kết quả sinh ra văn bản tốt hơn. Xác định được các tham số xác suất cho những tài liệu mới nằm ngoài tập huấn luyện (do phân phối chủ đề độc lập với quá trình sinh từ trong tài liệu).

Tuy nhiên, mô hình LDA còn tồn tại một số vấn đề:

- + Kết quả các chủ đề khác nhau được tạo ra nếu thứ tự dữ liệu đào tạo bị xáo trộn.
- + Kết quả phân tích khó áp dụng cho việc phân loại, tìm kiếm tài liệu.
- + Chưa xét tới mối quan hệ tương đồng và quan hệ cú pháp giữa các từ trong câu để dự đoán chủ đề phù hợp hơn.


CHƯƠNG 4: THỰC HIỆN THU THẬP CÁC THÔNG BÁO TUYỂN DỤNG

4.1 ĐỐI TƯỢNG VÀ DỮ LIỆU QUAN TÂM

Đối tượng mà chúng tôi quan tâm là các thông báo việc làm trên các website tuyển dụng. Một số trang web việc làm nổi tiếng ở Việt Nam và trên thế giới được liệt kê như bảng 4.1.

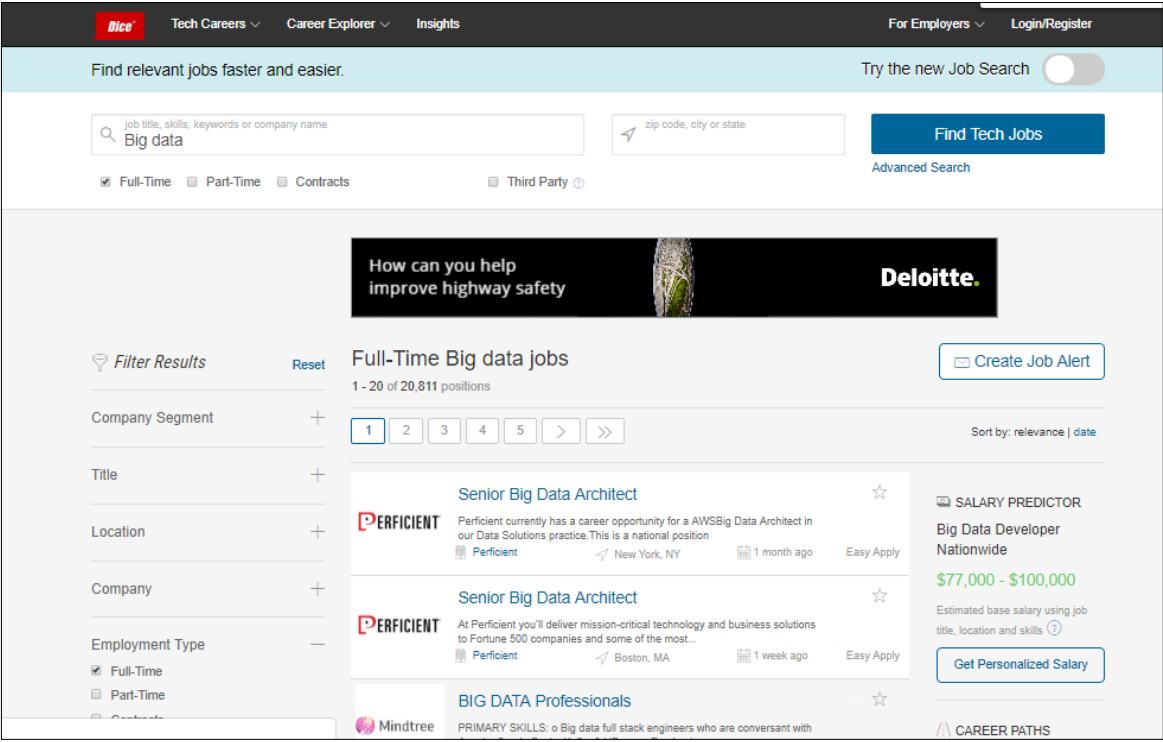
Bảng 4.1: Một số trang web việc làm nổi tiếng tại Việt Nam và trên thế giới

Vùng	Website	Mô tả
Việt Nam	<ul style="list-style-type: none"> - Vietnamworks - Careerbuilder 	Trang tìm việc dành cho người đi làm nhiều kinh nghiệm, nhân viên cấp cao, người nước ngoài,... http://www.vietnamworks.com/ http://careerbuilder.vn/
Việt Nam	<ul style="list-style-type: none"> - Jobstreet - Việc làm 24h - Tìm việc nhanh - Mywork - Careerlink 	Trang tìm việc phù hợp nhiều đối tượng. http://Jobstreet.vn https://vieclam24h.vn/ http://www.timviecnhanh.com/ https://mywork.com.vn/ https://www.careerlink.vn/
Việt Nam	<ul style="list-style-type: none"> - ITViec - TopDev 	Trang tìm việc dành cho người tìm việc về công nghệ thông tin. http://itviec.com/ https://topdev.vn
Việt Nam	<ul style="list-style-type: none"> - TopCV - Internship.edu - Ybox.vn 	Trang tìm việc dành cho sinh viên mới ra trường, tìm cơ hội thực tập, việc làm bán thời gian. http://itviec.com/ https://www.internship.edu.vn/ http://ybox.vn/
Thế giới	Linked-in	Mạng xã hội tìm việc lớn nhất thế giới, người dùng đăng tải kinh nghiệm và kỹ năng làm việc trên trang cá nhân. Nhà tuyển dụng cũng thường sử dụng website này để kết nối với những ứng viên tiềm năng. https://www.linkedin.com/

Thế giới	<ul style="list-style-type: none"> - Indeed - SimplyHired 	<p>Trang việc làm đầy đủ những thông tin như danh sách công ty, danh sách công việc và các trang tin tức. Công cụ tìm kiếm nâng cao của trang web giúp người dùng có thể tìm việc theo địa điểm, mức lương, hay ngành nghề.</p> <p>https://indeed.com</p> <p>https://www.simplyhired.com/</p>
Thế giới	<ul style="list-style-type: none"> - Monster.com - Glassdoor.com - Idealist.org 	<p>Ngoài đăng tin tuyển dụng, trang web còn có các bài hướng dẫn về viết CV, phỏng vấn, giới thiệu bản thân,..</p> <p>Trang web còn cung cấp các nhận xét của mọi người về một công ty/ tổ chức họ đã từng làm.</p> <p>https://www.monster.com.vn</p> <p>https://www.glassdoor.com</p> <p>https://www.idealists.org</p>
Thế giới	Internships.com	<p>Đây là website tuyển dụng thực tập sinh lớn nhất tại Mỹ bao gồm những công việc từ parttime đến các công việc fulltime.</p> <p>https://www.internships.com/</p>
Thế giới	USAJobs.com	<p>Đây là website cung cấp những việc làm thuộc chính phủ Mỹ cung cấp hàng ngàn công việc từ Bộ quốc phòng cho đến Sở giao thông vận tải.</p> <p>https://www.usajobs.gov/</p>
Thế giới	Dice.com	<p>Website này phù hợp cho những người tìm việc như những  sư, lập trình viên.</p> <p>https://www.dice.com/</p>

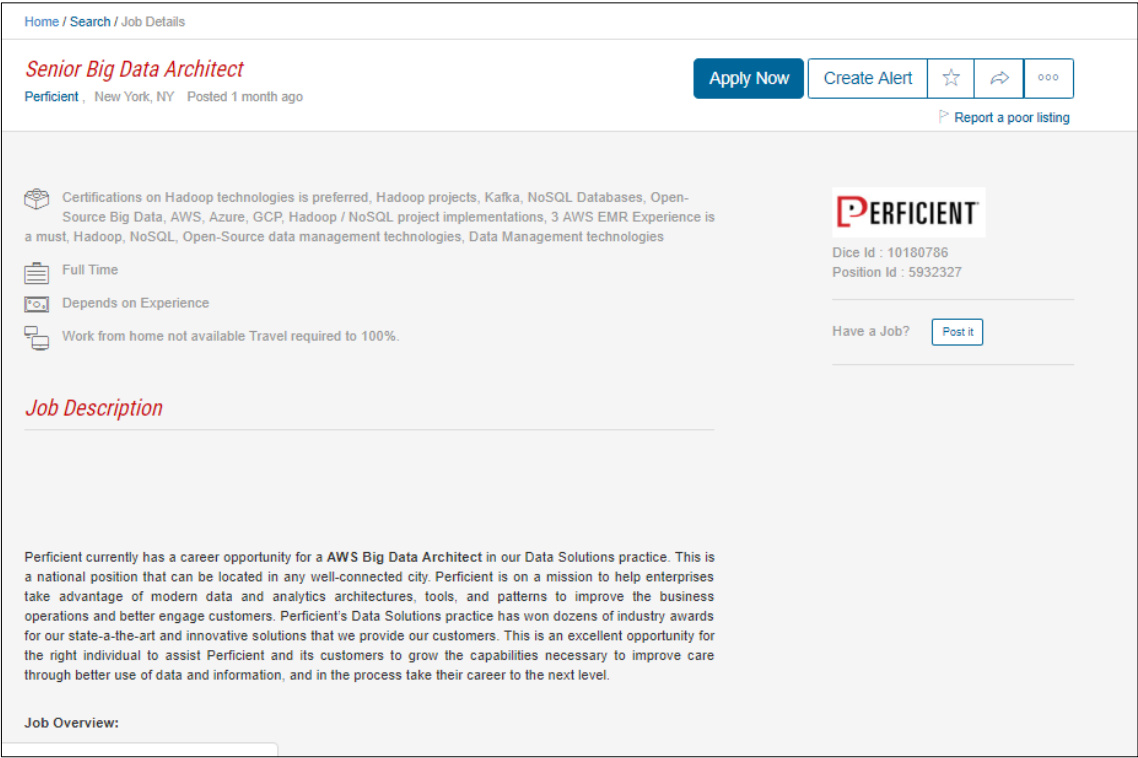
Vì sự phức tạp về cấu trúc cũng như sự không đồng nhất của các web tuyển dụng nên biện pháp kỹ thuật áp dụng cũng rất phức tạp. Do một số giới hạn nên chúng tôi chỉ tập trung thu thập dữ liệu từ một số web tuyển dụng có cấu trúc tương đối đồng nhất để thuận tiện cho quá trình web scraping. Trang web chúng tôi tiến hành thu thập là <https://www.dice.com>. Dữ liệu mà chúng tôi quan tâm trong một thông báo việc làm bao gồm thông tin vị trí công việc, mô tả công việc và một số thông tin khác như tên tổ chức, nơi làm việc,... Chúng tôi giới hạn phạm vi thu thập dữ liệu trong khoảng 15000 thông báo việc làm có liên quan đến một từ khóa cho trước. Chúng tôi lấy một ví dụ

cho từ khóa “Big data”. Trang kết quả tìm kiếm các thông báo việc làm liên quan đến từ khóa được hiển thị như hình 4.1.1.



Hình 4.1.1: Trang kết quả tìm kiếm cho từ khóa Big data trên trang dice.com

Giả sử chúng ta chọn xem một trang thông báo việc làm “Senior Big Data Architect”, nội dung chi tiết công việc được hiển thị như hình 4.1.2.



Hình 4.1.2: Thông tin chi tiết vị trí Senior Big Data Architect trên trang dice.com

Một điểm cần lưu ý là với cùng nội dung thông báo việc làm có thể xuất hiện ở nhiều trang web tuyển dụng nên việc thu thập có thể dẫn đến những tập dữ liệu có nội dung trùng lặp. Bên cạnh đó nội dung thông báo tuyển dụng thường xuyên được cập nhật nên một số dữ liệu thu thập không còn phù hợp. Có nhiều cách để xử lý vấn đề trên nhưng yêu cầu cao về kỹ thuật, để đơn giản trong đề tài này chúng tôi sẽ xử lý chúng một cách thủ công bằng cách loại bỏ đi những tập dữ liệu không phù hợp.

Chúng tôi thực hiện lưu trữ dữ liệu thu thập được dưới dạng text trong file CSV với các trường thông tin: job_title, company_name, job_location, job_description,, job_overview, links và được minh họa như hình 4.1.3.

	job_title	company_name	job_location	job_description	Job_Overview	links
1	Big Data Developer 20	Louisiana Economic	Baton Rouge, LA	Experience with technologies suc	As a member of the wc	/jobs/detail/Big-Data-Developer-204723
2	Senior Big Data Develc	Louisiana Economic	Baton Rouge, LA	Experience with technologies suc	IBM is market leader in	/jobs/detail/Senior-Big-Data-Developer-
3	Senior Big Data Archite	Perficient	New York, NY	Certifications on Hadoop technol	Qualified and intereste	/jobs/detail/Senior-Big-Data-Architect-P
4	AWS Big Data Architect	Perficient	New York, NY	Demonstrate broad solutions tec	This is a very urgent rei	/jobs/detail/AWS-Big-Data-Architect-Per
5	Big Data Architect	Mindtree Limited	Dallas, TX	Review design to make sure desig	Driven by values	/jobs/detail/Big-Data-Architect-Mindtree
6	Big Data Engineer for	Citigroup	Tampa, FL	Experience in Financial Industry is	NA NA	/jobs/detail/Big-Data-Engineer-for-Strate
7	Life Sciences & Health	Deloitte	New York, NY	An advanced degree around spec	We want job seekers e	/jobs/detail/Life-Sciences-%26-Healthcar
8	Big Data Lead/Solutor	Citigroup	New Castle, DE	Hands on Experience with open s	Job Skills/Qualification	/jobs/detail/Big-Data-Lead%26%2347Sol
9	Big Data Production A	UnitedHealth Group	Basking Ridge, NJ	NA NA	NA NA	/jobs/detail/Big-Data-Production-Applic
10	Life Sciences & Health	Deloitte	New York, NY	An advanced degree around spec	We want job seekers e	/jobs/detail/Life-Sciences-%26-Healthcar
11	Big Data Engineer	JPMorgan Chase & C	Lewisville, TX	Experiencewith Oracle databases	At JPMorgan Chase & C	/jobs/detail/Big-Data-Engineer-JPMorgan
12	SRE Infrastructure Eng	JPMorgan Chase & C	Jersey City, NJ	NA NA	NA NA	/jobs/detail/SRE-Infrastructure-Engineer
13	Big Data Analyst - Tele	UnitedHealth Group	Schaumburg, IL	NA NA	NA NA	/jobs/detail/Big-Data-Analyst-%26%2345
14	Big Data - Java Softwar	JPMorgan Chase & C	Wilmington, DE	Experience in developing Machin	At JPMorgan Chase &	/jobs/detail/Big-Data-%26%2345-Java-Sc
15	Big Data Subject Matter	Deloitte	McLean, VA	NA NA	We want job seekers	/jobs/detail/Big-Data-Subject-Matter-Ex
16	Software Developer - E	State Farm	Bloomington, IL	Job may require irregular work hc	#LI-LF1 SFARM Be Part	/jobs/detail/Software-Developer-%26%2
17	Deloitte US Innovator	Deloitte	New York, NY	Excellent knowledge of UML and	Deloitte is led by a pu	/jobs/detail/Deloitte-US-Innovation-%26
18	Big Data Subject Matter	Deloitte	McLean, VA	NA NA	We want job seekers	/jobs/detail/Big-Data-Subject-Matter-Ex

Hình 4.1.3: Cấu trúc lưu trữ cho các thông báo việc làm thu thập được

Sau khi đã xác định các đối tượng và dữ liệu quan tâm, chúng tôi giới thiệu công cụ R và các thư viện có sẵn để tiến hành web scraping.

4.2 GIỚI THIỆU CÔNG CỤ R VÀ CÁC THƯ VIỆN

4.2.1 Giới thiệu ngôn ngữ R

R là một ngôn ngữ lập trình cấp cao được sử dụng nhiều cho các bài toán về dữ liệu như tính toán thống kê, phân tích dữ liệu, khám phá tri thức và đồ thị... Giống như các ngôn ngữ lập trình khác, R cũng bao gồm các kiểu dữ liệu, các phép toán, các cấu trúc điều khiển, các hàm tự định nghĩa được trình bày trong bảng 4.2.1.1.

Bảng 4.2.1.1: Một số thành phần trong R

Thành phần	Chi tiết
Các kiểu dữ liệu	Luận lý (true/false), số nguyên, số thực, số phức, ký tự, chuỗi ký tự, vectơ, danh sách, ma trận, khung dữ liệu (data frame),...

Các phép toán	Cộng (+), trừ (-), nhân (*), chia (/), chia lấy phần dư (%%), chia lấy phần nguyên (%/%), nghịch đảo (!), lũy thừa (^), nhân ma trận (%*%), phép gán (<-), so sánh (<, <=, ==, >, >=, &, &&, ,)
Các cấu trúc điều khiển	<i>if</i> (điều kiện) {lệnh} <i>else</i> {lệnh};
Vòng lặp	<i>for</i> (biến <i>in</i> khoảng giá trị) {lệnh};
Các hàm tự định nghĩa	tên hàm <- <i>function</i> (danh sách đối số) {thân hàm};

Tất cả các công việc thực hiện trên R thông qua các hàm sẵn có bằng cách truyền vào các tham số và nhận về kết quả. Trong R cung cấp rất nhiều thư viện và các hàm thực hiện hầu hết các yêu cầu đặt ra và chúng ta cũng có thể tự xây dựng các hàm để thực hiện công việc (tham khảo bảng 4.1.2).

Bảng 4.2.1.2: Một số hàm sẵn có thường dùng trong R

Tên hàm	Ý nghĩa	Ví dụ
c()	Tạo ra vector	x <- c(12, 54, 13, 10, 5)
seq(a,b,d)	Tạo ra dãy số từ a đến b với bước nhảy d	x <- seq(6, 20, by = 0.5)
sum()	Tính tổng giá trị của vector	T <- sum(x)
read.table()	Đọc dữ liệu từ một tập tin có chứa khung dữ liệu	
write.table()	Viết dữ liệu sang tập tin để lưu trữ	
data.frame()	Tạo khung lưu trữ dữ liệu lưu trữ các trường	df <- data.frame(A,B,...)
rbind()	Cộng gộp dữ liệu từ các khung	df0<-rbind(df0, df)
paste0()	Kết hợp các vector	
source()	Thực hiện một function từ một nguồn khác	source('func1.R')

Bên cạnh đó còn rất nhiều hàm hỗ trợ cho các tính toán thống kê có sẵn trong R như: summary(), sample(), dnorm(), pnorm(), qnorm(), rnorm(), dunif(), punif(), qunif(), runif(), mean(), var(), sd(), cov(), cor(), lm(),... R cũng có các hàm hỗ trợ hiển thị dữ liệu với bốn phương pháp hiển thị dữ liệu, trong đó thường dùng là box plot (đồ thị hộp của 1 thuộc tính hiển thị giá trị nhỏ nhất, trung vị, lớn nhất, bách phân 25% và 75%) và histogram (tổ chức đồ để hiển thị thông tin về phân bố dữ liệu của một thuộc tính).

Các hàm hỗ trợ hiển thị dữ liệu với hai mức:

+ Mức cao (vẽ toàn bộ đồ thị bằng 1 lời gọi hàm duy nhất) bao gồm: `barplot()`, `boxplot()`, `contour()`, `coplot()`, `hist()`, `pairs()`, `persp()`, `plot()`, `pie()`.

+ Mức thấp (được sử dụng để thêm các thông tin vào đồ thị đã vẽ trước) bao gồm: `abline()`, `axis()`, `legend()`, `lines()`, `points()`, `polygon()`, `symbols()` và `text()`.

4.2.2 Các gói thư viện hỗ trợ web scraping

4.2.2.1 Thư viện Rvest

Thư viện Rvest chuyên hỗ trợ và xử lý việc lấy dữ liệu từ các trang web HTML bằng ngôn ngữ R. Một số hàm phổ biến trong thư viện Rvest như được mô tả trong bảng 4.2.2.1.

Bảng 4.2.2.1: Một số hàm phổ biến trong thư viện Rvest

STT	Các hàm phổ biến	Chức năng của hàm	Danh sách tham số	Kết quả trả về
1	<code>rvest::read_nodes()</code>	Đọc chi tiết các thành phần HTML	x, css, xpath	
2	<code>rvest::html_nodes(Xpath=...)</code>	Trích xuất các phần tử trong thẻ HTML bằng cách sử dụng bộ chọn XPath và CSS hoặc trích xuất từ 1 link url.	x, css, xpath	Trả về phần tử thẻ đầu tiên nó tìm thấy, nếu không thì là “missing”
3	<code>rvest::html_attr()</code>	Trích xuất thuộc tính từ thẻ html	X, name (tên thuộc tính, class, href,...)	Trả về nội dung thuộc tính đã khai ở name
4	<code>rvest::html_text()</code>	Trích xuất văn bản từ thẻ html	x, trim=FALSE	Nội dung text của thẻ html

4.2.2.2 Thư viện Xlm2

Thư viện Xlm2 chuyên hỗ trợ và xử lý công việc liên quan đến HTML, XLM trong R. Một số hàm phổ biến trong thư viện Xlm2 như được mô tả trong bảng 4.2.2.2.

Bảng 4.2.2.2: Một số hàm phổ biến trong thư viện Xlm2

STT	Các hàm phổ biến	Chức năng của hàm	Danh sách tham số	Kết quả trả về
1	<code>xlm2::read_xml()</code>	Đọc xml	x, encoding, as_html, options	Nội dung xml

2	xml2::read_html()	Đọc html	x, encoding, options	Nội dung html
3	xml2::download_xml()	Download xml	url, file, quite, mode, handle	File xml
4	xml2::download_html()	Download html	url, file, quite, mode, handle	File html
5	xml2::xml_structure()	Hiện thị cấu trúc của xml	x, indent	Cấu trúc xml
6	xml2::html_structure()	Hiện thị cấu trúc html	x,indent	Cấu trúc html
7	xml2::xml_text()	Trích xuất hoặc sửa đổi text	x, trim, value	Charactor vector
8	as_list()	Ép xml node thành kiểu list	x, ns	Xml kiểu list

4.3 THỰC HIỆN TẢI XUỐNG DỮ LIỆU VỚI CÔNG CỤ R

Chúng tôi tìm kiếm các công việc liên quan đến một từ khoá cụ thể trên trang <https://www.dice.com>, một danh sách các công việc liên quan được trả về (trang kết quả tìm kiếm). Trong đó, dữ liệu mà chúng tôi quan tâm: tên vị trí công việc, tên tổ chức, nơi làm việc, link. Đối với một thông báo việc làm cụ thể (trang chi tiết), dữ liệu chúng tôi quan tâm là phần mô tả công việc. Chúng tôi tạo các đối tượng để lưu giữ thông tin mà chúng tôi quan tâm và được mô tả trong bảng 4.3.

Bảng 4.3.1: Các đối tượng lưu trữ thông tin trong quá trình web scraping

Đối tượng	Kiểu	Nội dung
job_title	page	Lưu giữ thông tin tất cả các vị trí công việc
company_name	page	Lưu giữ tất cả thông tin về tên tổ chức
job_location	page	Lưu giữ tất cả thông tin nơi làm việc
links	page	Lưu trữ tất cả các link đến chi tiết việc làm
job_description	page	Lưu trữ phần tổng quan công việc
job_overview	page	Lưu trữ chi tiết công việc
full_df	data.frame	Chứa toàn bộ các đối tượng ở trên

Chúng tôi xây dựng vòng lặp để duyệt qua khoảng 15000 thông báo việc làm cho một từ khoá cụ thể. Giá trị khởi đầu là trang kết quả đầu tiên, giá trị kết thúc là trang kết quả cuối cùng và giá trị tăng biến đếm chính bằng số công việc có trong một trang kết quả. Bên trong vòng lặp chúng tôi thực hiện việc nạp dữ liệu cho các đối tượng. Kết thúc quá trình lặp, toàn bộ dữ liệu được xuất ra file CSV ở cuối quá trình. Chi tiết cài đặt (code) xin tham khảo phụ lục A. Ở đây chúng tôi xin trình bày mã giả để thực hiện công việc:

Duyệt lần lượt các trang kết quả tìm kiếm (page result) và thực hiện:

```
{
1.   Sử dụng thư viện Rvest để lấy text trong thẻ tương ứng gán cho các đối tượng
      job_title, company_name, job_location, links, job_description;
2.   Tạo khung dữ liệu df lưu giữ toàn bộ các đối tượng ở trên, khung dữ liệu
      full_df chứa kết quả cộng gộp các df;
}
```

Xuất dữ liệu sang CSV dựa trên dữ liệu từ full_df;

Chi tiết cấu hình máy để thực hiện web scraping và topic analysis được thể hiện chi tiết trong bảng 4.3.2.

Bảng 4.3.2: Chi tiết cấu hình máy thực hiện web scraping và topic analysis

Thành phần	Chỉ số
CPU	Core i5-7500, 2.5 GHz
RAM	8GB
HDD	1 TB
Hệ điều hành	Windows 10 Pro
Công cụ	Rwin 3.7.3, Rstudio 1.8.9

Quá trình thực hiện mất khoảng 20 giờ với chất lượng mạng ổn định. Dữ liệu thu được là dữ liệu thô chưa qua giai đoạn tiền xử lý gồm 15276 thông báo việc làm (dung lượng 36 MB). Trong phần sau chúng tôi tiếp tục trình bày quá trình tiền xử lý dữ liệu và tiến hành phân tích dữ liệu để đưa ra các kết luận.

CHƯƠNG 5: ỨNG DỤNG TOPIC MODEL PHÂN TÍCH THÔNG BÁO TUYÊN DỤNG

5.1 CÁC CÔNG CỤ VÀ THƯ VIỆN CẦN THIẾT

5.1.1 Thư viện Tm

Thư viện Tm nhằm hỗ trợ khai thác văn bản bằng phương pháp nhập dữ liệu, xử lý văn bản, tiền xử lý, quản lý siêu dữ liệu, tạo ma trận tài liệu... Một số hàm phổ biến của thư viện Tm được mô tả trong bảng 5.1.1.

Bảng 5.1.1: Một số hàm phổ biến trong thư viện Tm

STT	Các hàm phổ biến	Chức năng của hàm	Danh sách tham số	Kết quả trả về
1	corpus()	Tạo một kho văn bản từ tài liệu nguồn. Các nguồn tài liệu là: - Một vector ký tự của văn bản. - Một đối tượng corpusSource-class (lớp ảo), được xây dựng bằng cách sử dụng hàm textfile;	X,..., enc, docnames, docvars, source.notes, citation, c1, c2	Một đối tượng lớp corpus chứa các văn bản gốc, các biến cấp độ tài liệu, siêu dữ liệu cấp độ tài liệu, siêu dữ liệu cấp độ khối.
2	tm_map()	Giao diện để áp dụng các hàm biến đổi/ ánh xạ cho khối.	X,FUN,...,lazy = FALSE	Một kho văn bản với FUN- 1 hàm biến đổi, lấy 1 tài liệu văn bản đầu vào (vector ký tự khi x là SimpleCorpus) và trả về 1 tài liệu văn bản (vector ký tự có cùng độ dài với vector đầu vào cho SimpleCorpus) được áp dụng cho mỗi tài liệu trong x.

3	removeWords()	Xóa các từ chỉ định từ đoạn văn bản	X, words, ...	Đoạn văn bản , kí tự sau khi loại bỏ từ chỉ định.
---	---------------	-------------------------------------	---------------	---

5.1.2 Thư viện Quanteda

Quanteda là gói R để quản lý và phân tích dữ liệu văn bản. Gói áp dụng để xử lý ngôn ngữ tự nhiên cho văn bản, từ tài liệu đến phân tích cuối cùng.

Bảng 5.1.2: Một số hàm phổ biến trong thư viện Quanteda

STT	Các hàm phổ biến	Chức năng của hàm	Danh sách tham số	Kết quả trả về
1	dfm()	Xây dựng một ma trận tính năng tài liệu từ một ký tự, kho văn bản, token hoặc thậm chí các đối tượng dfm khác.	X,..., verbose, clean, stem, ignoredFeatures, keptFeatures, matrixType, language, fromCorpus, bigrams, thesaurus, dictionary, dictionary_regex, addto, groups	
2	dfm_trim()	Cắt ma trận dfm bằng cách sử dụng lựa chọn tính năng dựa trên ngưỡng tần số	X,min_termfreq,max_termfreq,termfreq_type ,min_docfreq,max_docfreq,docfreq_type, sparsity,verbose, ...	Trả về tài liệu sau khi cắt giảm ma trận dfm (giảm kích thước ma trận)
3	topFeatures()	Xác định các đặc tính thường xuyên trong Dfm	X, n, decreasing, scheme, Groups	Một vector số được đặt tên của số đặc tính, trong đó tên là nhãn đặc tính hoặc danh sách các giá trị này nếu các nhóm được đưa ra.

5.1.3 Thư viện Topicmodels

Cung cấp các mô hình phân bố Dirichlet tiềm ẩn (LDA) và mô hình chủ đề tương quan (CTM). Mô hình thích hợp có thể được sử dụng để ước tính sự giống nhau giữa các tài

liệu cũng như giữa một tập hợp các từ khóa được chỉ định bằng cách sử dụng một lớp các biến tiềm ẩn bổ sung được gọi là chủ đề.

Bảng 5.1.2: Một số hàm phổ biến trong thư viện Quanteda

STT	Các hàm phổ biến	Chức năng của hàm	Danh sách tham số	Kết quả trả về
1	LDA()	Phân bố Dirichlet tiềm ẩn. Ước tính mô hình LDA bằng cách sử dụng ví dụ thuật toán VEM hoặc lấy mẫu Gibbs.	X, k, method, control, model	Trả về một đối tượng của lớp LDA.

5.2 DỮ LIỆU VÀ TIỀN XỬ LÝ

5.2.1 Mô tả dữ liệu

Dữ liệu cần phân tích chính là file Bigdata.xls thu thập được từ quá trình web scraping. Dữ liệu Bigdata bao gồm 15276 dòng với 6 thuộc tính:

- + job_title: tiêu đề của công việc
- + company_name: tên công ty tuyển nhân sự
- + job_location: vị trí nơi làm việc
- + job_description: mô tả tổng quan về công việc
- + job_overview: mô tả chi tiết về công việc
- + job_link: đường link tới trang web đăng thông tin tuyển công việc.

Trong đó, thuộc tính job_description là thuộc tính mang nhiều thông tin có giá trị cho khai phá nên chúng tôi sử dụng dữ liệu thuộc tính job_description để tiến hành phân tích để tìm ra chủ đề phổ biến. Trước khi bắt đầu quá trình phân tích, chúng ta cần tiến hành tiền xử lý dữ liệu.

5.2.2 Tiền xử lý dữ liệu

Các dữ liệu lưu trữ hoàn toàn dưới dạng thô chưa sẵn sàng cho việc phát hiện, khám phá thông tin ẩn chứa trong nó. Do vậy chúng ta cần phải qua giai đoạn tiền xử lý dữ liệu trước khi tiến hành bất kỳ một phân tích nào. Quá trình xử lý dữ liệu thô nhằm cải thiện chất lượng dữ liệu và do đó sẽ cải thiện chất lượng của kết quả khai phá. Các kỹ thuật tiền xử lý dữ liệu thường có:

- + Làm sạch dữ liệu: loại bỏ nhiễu, hiệu chỉnh những phần dữ liệu không nhất quán.
- + Tích hợp dữ liệu: trộn dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu.
- + Biến đổi dữ liệu: chuẩn hoá dữ liệu.
- + Thu giảm dữ liệu: thu giảm kích thước dữ liệu (giảm số phần tử) bằng kết hợp dữ liệu, loại bỏ các đặc điểm dư thừa (giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu.

Dựa theo đó, chúng tôi tiến hành tiền xử lý dữ liệu trong file BigData.csv:

- + Vì dữ liệu không có các thẻ html nên chúng ta không cần remove noise.
- + Xóa các chữ số như 1,2,3.. vì chúng không phù hợp với thuật toán.
- + Xóa đi các khoảng trắng thừa.
- + Chuyển các từ có chữ cái hoa thành chữ cái thường.
- + Loại bỏ các dấu chấm câu.
- + Loại bỏ các stopwords những từ không cần thiết hoặc thừa thãi.

Đây là các kỹ thuật phù hợp khi xử lý dữ liệu đầu vào để nâng cao hiệu suất của như tính chính xác của phân tích topic model sau này. Tuy nhiên đây chỉ là một phần của xử lý dữ liệu, ở các phần sau sẽ tiếp tục xử lý dữ liệu để ra kết quả tốt nhất.

5.3 THỰC HIỆN TOPIC ANALYSIS CÁC THÔNG BÁO TUYÊN DỤNG

Quá trình thực hiện topic analysis bao gồm việc nạp dữ liệu, xử lý, chọn tham số, áp dụng thuật toán và biểu diễn kết quả. Trong phần thực nghiệm này, chúng tôi áp dụng thuật toán LDA để tiến hành phân tích dữ liệu theo chủ đề. Các tham số và giá trị tham số áp dụng cho thuật toán LDA được mô tả trong bảng 5.3.

Bảng 5.3: Các tham số và giá trị tham số khi thực nghiệm với LDA

Tham số	Giá trị
α	5
β	0.1
k (số chủ đề)	10
niters (vòng lặp)	2000
savesteps (số vòng lặp mỗi lần lưu mô hình)	200
twords (số từ trong một chủ đề)	10

Lưu ý rằng tham số α, β được lấy theo mặc định của thư viện topicmodels LDA, trong đó: $\alpha = 50/k, \beta = 0.1$.

Chi tiết cài đặt(code) xin tham khảo phụ lục B. Ở đây chúng tôi xin trình bày mã giả thực hiện topic analysis:

1. Nạp dữ liệu, chỉ lấy dữ liệu thuộc tính Job_description
2. Tiền xử lý dữ liệu
3. Vector hóa dữ liệu
4. Chọn ra các từ có tần suất xuất hiện cao nhất và thể hiện qua biểu đồ
5. Áp dụng thuật toán và biểu diễn kết quả

CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ PHÂN TÍCH

6.1 BIỂU DIỄN KẾT QUẢ

Trước tiên chúng tôi thống kê các từ có tần suất xuất hiện cao nhất trong tập tài liệu và chúng có thể được xem như những đại diện quan trọng đóng góp vào chủ đề. Kết quả 50 từ có tần suất xuất hiện cao nhất được thể hiện trong hình 6.1.1.



Hình 6.1.1: Biểu đồ thể hiện 50 từ có tần suất xuất hiện cao nhất khi phân tích

Chi tiết 50 từ và tần suất xuất hiện (TS) của chúng được thể hiện trong bảng 6.1.1.

Bảng 6.1.1: Chi tiết tần suất của 50 từ có tần suất xuất hiện cao nhất

Mục từ	TS	Mục từ	TS	Mục từ	TS	Mục từ	TS	Mục từ	TS
experience	46441	project	7076	communication	4804	network	3544	analytics	2955
data	37408	tools	6972	database	4577	product	3518	able	2952
business	15261	analysis	6215	performance	4562	maintain	3460	procedures	2949
management	13166	sql	6020	engineering	4278	enterprise	3425	standards	2922
development	12980	system	5984	projects	4214	practices	3413	service	2889
design	10703	develop	5524	architecture	4186	customer	3378	web	2841
technical	9757	information	5269	computer	4149	big	3190	level	2828
systems	9709	technology	5227	cloud	4135	developing	3102	infrastructure	2673
software	8028	applications	4874	science	3827	manage	3102	java	2668
security	7271	services	4816	quality	3739	server	3062	user	2647

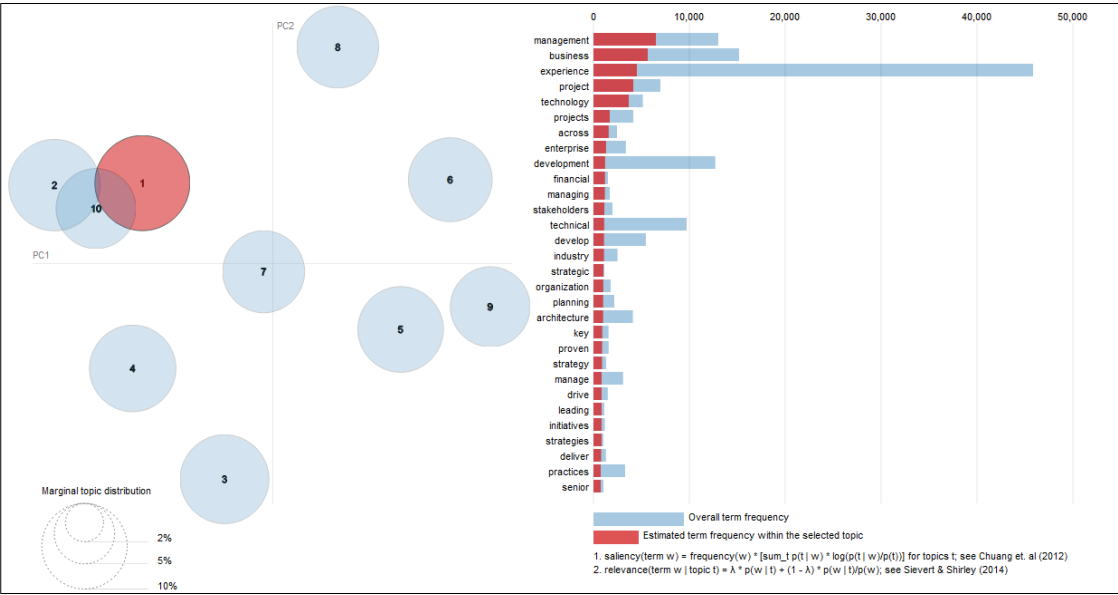
Áp dụng thuật toán LDA phân tích chủ đề cho tập tài liệu, chúng tôi xác định được các chủ đề quan tâm. Kết quả 10 chủ đề (topic) được thể hiện trong bảng 6.1.2. Mỗi chủ đề bao gồm tất cả các từ trong tập từ vựng với các xác **xuất** khác nhau. Từ có xác **xuất** cao hơn được đưa lên trên và dùng để suy luận ra tên của chủ đề. Trong bảng 6.1.2 chỉ đưa ra 10 từ có xác suất cao nhất ứng với mỗi chủ đề.

Bảng 6.1.2: Kết quả 10 chủ đề của của tập tài liệu khi phân tích với LDA

Topic1	Topic2	Topic3	Topic4	Topic5
management	able	security	systems	experience
business	maintain	experience	experience	development
experience	communication	network	technical	software
project	projects	systems	system	design
technology	project	management	engineering	web
projects	quality	system	information	applications
across	effectively	infrastructure	software	code
enterprise	manage	administration	analysis	java
development	tasks	monitoring	provides	programming
financial	management	configuration	development	developing
Topic6	Topic7	Topic8	Topic9	Topic10
data	business	data	experience	product
experience	experience	analysis	data	company
sql	user	experience	cloud	experience
database	design	analytics	big	customer
design	systems	business	aws	services
etl	functional	science	architecture	business
oracle	management	tools	python	communication
development	system	research	hadoop	products
server	sap	techniques	tools	software
databases	technical	analytical	spark	make

Ta có thể xem chi tiết số lần xuất hiện của các từ trong một chủ đề dựa vào biểu đồ R cung cấp (textplot wordcloud). Khi chọn vào các hình tròn chứa con số, thông tin chi tiết các từ trong topic tương ứng với số đó được hiển thị bên phải màn hình. Hình 6.1.2

cung cấp chi tiết thông tin chủ đề 1 (topic 1) bao gồm các từ và số lần xuất hiện của chúng.



Hình 6.1.2: Biểu đồ thể hiện thông tin chi tiết chủ đề 1 (topic 1)

Mỗi tài liệu là một phân bố xác suất các chủ đề, tức là mỗi tài liệu đều bao gồm tất cả các chủ đề nhưng mức độ đóng góp của các chủ đề vào tài liệu là khác nhau. Chi tiết các xác suất này xin tham khảo phụ lục D. Hình 6.1.3 thể hiện mức độ đóng góp của 10 chủ đề đầu tiên vào 37 tài liệu đầu tiên trong tập tài liệu.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
text1	0.07321226	0.01872872	0.04143019	0.04597049	0.05051078	0.24120318	0.03688990	0.06867196	0.40011351	0.02326901
text2	0.09201141	0.02924394	0.01783167	0.03495007	0.04065621	0.24607703	0.04065621	0.14907275	0.33166904	0.01783167
text3	0.41617770	0.05741827	0.01383068	0.01383068	0.06077117	0.05071249	0.06077117	0.04735960	0.22506287	0.05406538
text4	0.48573975	0.02228164	0.03654189	0.05080214	0.03654189	0.05080214	0.08645276	0.04367201	0.13636364	0.05080214
text5	0.20808203	0.03437877	0.01990350	0.04402895	0.01507841	0.24185766	0.07297949	0.02472859	0.29975875	0.03920386
text6	0.10341880	0.02820513	0.02136752	0.04871795	0.14444444	0.14444444	0.02820513	0.10341880	0.24700855	0.13076923
text7	0.17837627	0.04410617	0.02224824	0.08157689	0.03473849	0.08157689	0.12217018	0.05347385	0.33138173	0.05035129
text8	0.14563758	0.04362416	0.03288591	0.24765101	0.08120805	0.09731544	0.05973154	0.05436242	0.21543624	0.02214765
text10	0.20701619	0.04047803	0.02814187	0.01580571	0.03739399	0.06823439	0.11141095	0.03430995	0.36738628	0.08982267
text11	0.03882353	0.02941176	0.05764706	0.03882353	0.15176471	0.10470588	0.02941176	0.06705882	0.45294118	0.02941176
text14	0.11323155	0.04198473	0.04198473	0.05216285	0.29643766	0.04198473	0.04198473	0.07251908	0.24554707	0.05216285
text16	0.16793409	0.05133080	0.03105196	0.25411914	0.29974651	0.02091255	0.04119138	0.03105196	0.04119138	0.06147022
text17	0.02886836	0.02886836	0.04734411	0.07505774	0.28752887	0.07505774	0.02886836	0.11200924	0.27829099	0.03810624
text19	0.42521110	0.13088058	0.01990350	0.03920386	0.04885404	0.02955368	0.08745476	0.04885404	0.07297949	0.09710495
text20	0.11502347	0.05868545	0.13380282	0.05868545	0.09624413	0.15258216	0.05868545	0.05868545	0.20892019	0.05868545
text21	0.02646240	0.02274838	0.01160631	0.04131848	0.05617456	0.16388115	0.03389044	0.08960074	0.52785515	0.02646240
text25	0.06871036	0.05179704	0.03488372	0.03488372	0.12790698	0.14482030	0.02642706	0.04334038	0.44080338	0.02642706
text26	0.14383562	0.20958904	0.04520548	0.05616438	0.18767123	0.04520548	0.07808219	0.05616438	0.12191781	0.05616438
text28	0.18490153	0.08862144	0.19365427	0.10612691	0.02735230	0.18490153	0.02735230	0.07111597	0.07111597	0.04485777
text29	0.06846473	0.06846473	0.16804979	0.10165975	0.05186722	0.10165975	0.08506224	0.05186722	0.25103734	0.05186722
text30	0.04462659	0.04462659	0.08834244	0.03734062	0.14663024	0.05191257	0.04462659	0.06648452	0.09562842	0.37978142
text31	0.16766169	0.03233831	0.02835821	0.03631841	0.03631841	0.10000000	0.04427861	0.08805970	0.27114428	0.19552239
text35	0.03202614	0.04248366	0.06862745	0.03202614	0.05816993	0.13137255	0.02679739	0.07385621	0.51307190	0.02156863
text36	0.06648452	0.04462659	0.04462659	0.19763206	0.09562842	0.22677596	0.03734062	0.03005464	0.19034608	0.06648452
text37	0.06862745	0.07983193	0.04621849	0.07983193	0.14705882	0.14705882	0.05742297	0.04621849	0.24789916	0.07983193

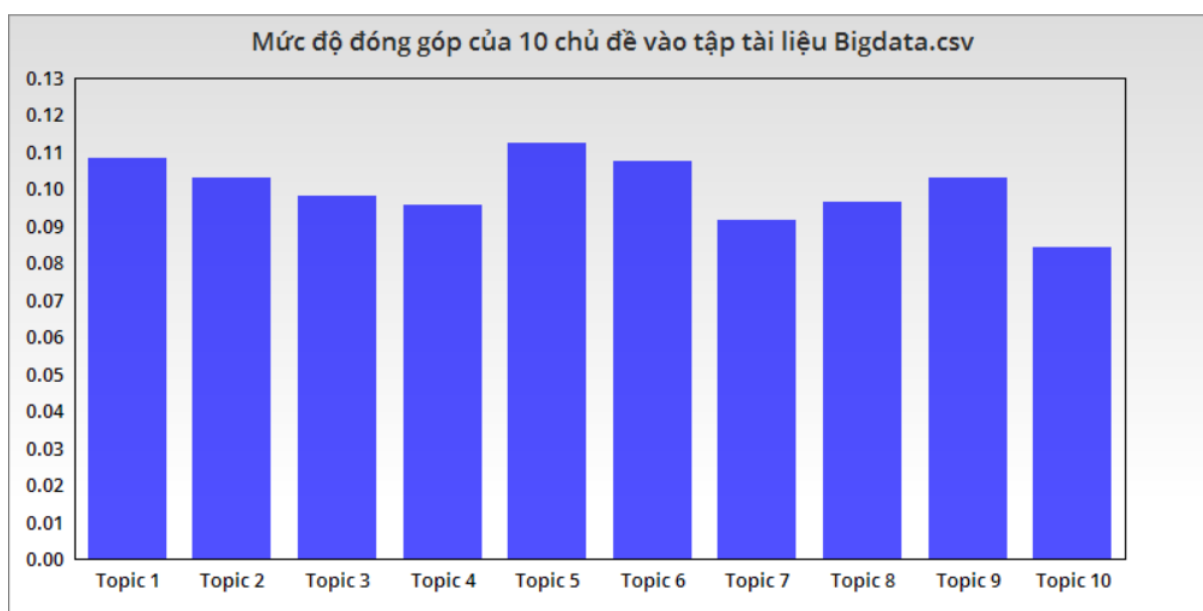
Hình 6.1.3: Mức độ đóng góp của các chủ đề vào các tài liệu

Mỗi đối tượng text trong hình 6.1.3 tương ứng với một thông báo việc làm trong tập các thông báo việc làm thu thập được. Tại mỗi dòng, tổng xác suất các chủ đề đóng góp vào tài liệu đúng bằng một. Đối với một tài liệu, xác suất chủ đề x (topic x) càng cao cho biết thông báo việc làm này đề cập nhiều đến chủ đề đó. Bảng 6.1.3 đề cập đến mức độ đóng góp của các chủ đề vào tập tài liệu Bigdata.csv.

Bảng 6.1.3: Mức độ đóng góp của 10 chủ đề vào tập tài liệu Bigdata.csv

Topic1	Topic2	Topic3	Topic4	Topic5
0.108313	0.103153	0.098025	0.09566	0.112291
Topic6	Topic7	Topic8	Topic9	Topic10
0.107571	0.091626	0.096462	0.10284	0.084059

Từ bảng 6.1.3 ta suy ra kết quả xác suất của 10 chủ đề trong cả tập tài liệu Bigdata.csv bằng cách tính trung bình xác suất của tất cả đối tượng text. Xác suất càng cao, nội dung của chủ đề đó trong tập tài liệu Bigdata càng được nhắc tới nhiều lần. Kết quả này được minh họa bằng đồ thị trong hình 6.1.4.



Hình 6.1.4: Biểu đồ thể hiện mức độ đóng góp của 10 chủ đề vào tập tài liệu Bigdata.csv

6.2 GIẢI THÍCH KẾT QUẢ

Từ các chủ đề (topic) tìm được, ta sẽ dự đoán tên các chủ đề. Chúng tôi trình bày một số tên dự đoán cho các chủ đề: topic 1, topic 3, topic 5, topic 6, topic 7, topic 8, topic 9. Chi tiết được thể hiện trong bảng 6.2.1. Dòng đầu tiên là tên dự đoán ứng với topic đó.

Bảng 6.2.1: Kết quả dự đoán tên chủ đề của một số topic

" Business manager"	"Network security"	"Developer"	" Database administration"
TOPIC 1	TOPIC 3	TOPIC 5	TOPIC 6
management	security	experience	data
business	experience	development	experience
experience	network	software	sql
project	systems	design	database
technology	management	web	design
projects	system	applications	etl
across	infrastructure	code	oracle
enterprise	administration	java	development
development	monitoring	programming	server
financial	configuration	developing	databases
management	security	experience	data

"Business design"	"Business analytics"	"Cloud infrastructure"
TOPIC 7	TOPIC 8	TOPIC 9
business	data	experience
experience	analysis	data
user	experience	cloud
design	analytics	big
systems	business	aws
functional	science	architecture
management	tools	python
system	research	hadoop
sap	techniques	tools
technical	analytical	spark
business	"business analytics"	"Cloud infrastructure"

Bây giờ, chúng tôi có một số các keyword và muốn biết chúng có mức độ quan tâm như thế nào trong các thông báo tuyển dụng. Để thực hiện điều này, chúng tôi viết một hàm thống kê số lần xuất hiện của keyword đó trong các chủ đề. Ý tưởng đơn giản là đếm sự xuất hiện của các keyword trong kết quả chủ đề. Chi tiết cài đặt (code) tham khảo phụ lục B. Kết quả thực hiện được thể hiện trong bảng 6.2.2.

Bảng 6.2.2: Thống kê số lần xuất hiện của một số keyword trong các chủ đề

Keyword	Count
sql	3371
java	2498
python	1952
aws	1469
excel	1332
hadoop	1095
spark	1052
sas	861
azure	752
tableau	747
nosql	635
r	625

Dựa vào kết quả thống kê đơn giản trong bảng 6.2.2, chúng tôi nhận thấy nhà tuyển dụng có quan tâm nhiều hơn tới sql so với nosql. Cũng theo thống kê này, dễ thấy số lượng việc làm liên quan đến java và python nhiều hơn hẳn so với lượng việc làm liên quan đến r. Ngoài ra, việc thành thạo Excel cũng được quan tâm không kém so với việc nắm vững framework liên quan đến xử lý dữ liệu lớn như aws, hadoop, spark.

KẾT LUẬN

1. Kết quả đạt được

1.1. Về ý nghĩa khoa học

Báo cáo đã trình bày khái niệm web scarping, quá trình thu thập dữ liệu từ website và giới thiệu cấu trúc HTML cơ bản. Nội dung chính của đề tài trình bày các thuật toán để phân tích chủ đề là PLSA, LDA và áp dụng chúng để phân tích các thông báo tuyển dụng. Các thư viện liên quan, các khảo sát về các trang web tuyển dụng cũng được chúng tôi giới thiệu sơ lược. Thông qua đề tài, chúng tôi biết được cách đánh giá và biểu diễn kết quả của bài toán phân tích theo chủ đề cũng như các bài toán khai phá dữ liệu nói chung, nâng cao hiểu biết và kỹ năng sử dụng R cùng các thư viện hỗ trợ để phân tích dữ liệu. Bên cạnh đó, chúng tôi còn nâng cao khả năng đọc hiểu tài liệu, khả năng làm việc nhóm và khả năng trình bày báo cáo khoa học.

1.2. Về ý nghĩa thực tiễn

Chúng tôi biết được nhiều hơn các trang web việc làm phổ biến tại Việt **nam** và trên thế giới. Áp dụng được kiến thức đã tìm hiểu để thu thập và phân tích cho bài toán phát hiện các mối quan tâm trong các thông báo tuyển dụng. Kết quả đạt được là các chủ đề nổi bật liên quan đến 15276 thông báo, trong đó một số chủ đề nổi bật liên quan đến: data, developer, software,... cho thấy mức độ quan tâm của nhà tuyển dụng vào các vị trí liên quan đến dữ liệu và phát triển phần mềm. Thông qua việc thực hiện đề tài, chúng tôi biết được phân tích theo chủ đề đang trở thành một trong những hướng nghiên cứu rất phát triển, đặc biệt là tại các doanh nghiệp. Chẳng hạn như các mạng xã hội Facebook, Youtube, các trang bán hàng điện tử Amazon,... đang thu thập và phân tích các bình luận, các nội dung do người dùng đăng tải để phát hiện mối quan tâm nhằm đưa ra các giải pháp kinh doanh hiệu quả. Báo cáo cũng **đã** **đã** cung cấp lý thuyết về cách xử lý bài toán phân tích chủ đề, làm tài liệu để người dùng áp dụng và cải tiến để giải quyết nhiều bài toán tại doanh nghiệp và trường học.

2. Hạn chế và đề xuất cải thiện

Do sự giới hạn nguồn lực và về thời gian, chúng tôi chỉ thu thập cho các website cấu trúc HTML tương đối đồng nhất. Việc xử lý loại bỏ dữ liệu trùng lặp được thực hiện thủ công. Phân tích chỉ dừng lại ở một thuộc tính. Nhóm đề xuất thực hiện thu thập thông báo việc làm đối với nhiều website khác nhau và cấu trúc khác nhau, sử dụng

công cụ hoặc thuật toán để loại bỏ các thông báo trùng lặp và mở rộng phân tích cho hai hoặc nhiều thuộc tính sẽ cho kết quả phản ánh chính xác hơn về mối quan tâm của nhà tuyển dụng.

3. Hướng phát triển

Lý thuyết được trình bày trong báo cáo có thể được áp dụng để giải quyết các bài toán phân tích trong doanh nghiệp hoặc trong trường học như: phân tích quan tâm của sinh viên về học phí trường Đại học Sư Phạm Kỹ Thuật TP.HCM trong giai đoạn bước sang tự chủ tài chính, phân tích phản hồi người dùng về dòng sản phẩm SamsungGalaxy Note 9 tại Thế Giới Di Động,... Đề tài này còn có thể phát triển tiếp tục theo hướng nghiên cứu cải thiện chất lượng phân tích thông qua:

- + Lựa chọn số k (số chủ đề) tối ưu nhất với dữ liệu. Ý tưởng chính là sử dụng phương pháp cross validation (hold out, k fold, leave one out) để đánh giá các kết quả sinh ra bởi các lần lựa chọn k khác nhau. Từ đó giúp tìm ra số k phù hợp nhất với dữ liệu [6].
- + Giải quyết vấn đề về sự khác biệt chủ đề khi thay đổi thứ tự tập dữ liệu đào tạo.
- + Bổ sung mối quan hệ tương đồng và quan hệ cú pháp giữa các từ trong câu để dự đoán chủ đề phù hợp hơn. Cụ thể những từ có cùng ngữ cảnh xuất hiện hoặc có quan hệ cú pháp gần với nhau thì thường cùng xuất hiện trong cùng một chủ đề.
- + Quá trình tiền xử lý dữ liệu đưa ra các cụm từ thay vì một từ thì khi đưa vào phân tích sẽ cho ra kết quả có ý nghĩa phù hợp hơn với dữ liệu.

TÀI LIỆU THAM KHẢO

Tiếng Việt

- [1] Nguyễn Trung Kiên, Lê Quang Minh, *Xây dựng ứng dụng thu thập dữ liệu tự động từ các Website*, Luận văn tốt nghiệp, Đại học Bách Khoa TP.HCM, 2009, trang 13-15.
- [2] Đào Minh Tùng, *Phân cụm đa mức web bằng k-means dựa trên chủ đề ẩn và thực nghiệm đánh giá*, Luận văn tốt nghiệp, Đại học Công nghệ- ĐHQGHN, 2011, trang 10-20-21.

Tiếng Anh

- [3] Simon Munzert, *Automated Data Collection with R- A Practical Guide to Web Scraping and Text Mining*, John Wiley & Sons Ltd, 2015, trang 292.
- [4] ChengXiang Zhai, Sean Massung, *Text Data Management and Analysis - A Practical Introduction to Information Retrieval and Text Mining*, ACM Books, 2016, trang 331.
- [5] Foster Provost, Tom Fawcett, *Data Science For Business- What You Need To Know About Data Mining And Data-Analytic Thinking*, O'Reilly Media, 2013, trang 252.

Nguồn khác

- [6] *Cross-validation of topic modelling*, links <http://freerangestats.info/blog/2017/01/05/topic-model-cv>, 06/2019.

PHỤ LỤC A

Code thực hiện thu thập dữ liệu thông báo việc làm:

```
library(tidyverse)
library(rvest)
library(wordcloud)

full_df <- data.frame()

for(i in 0:1000) {

  first_page_url <- "https://www.dice.com/jobs?q=Big+data"
  url <- paste0(first_page_url, "&p=", i)
  page <- xml2::read_html(url)
  # Sys.sleep pauses R for two seconds before it resumes
  # Putting it there avoids error messages such as "Error in open.connection(con, "rb") :
  Timeout was reached"
  Sys.sleep(5)

  #get the job title
  job_title <- page %>%
    rvest::html_nodes('.complete-serp-result-div') %>%
    rvest::html_nodes("h3") %>%
    rvest::html_nodes("a") %>%
    rvest::html_attr("title")

  #get the company name
  company_name <- page %>%
    rvest::html_nodes('.complete-serp-result-div') %>%
    rvest::html_nodes('.hidden-xs') %>%
    rvest::html_nodes('.compName') %>%
    rvest::html_text() %>%
    stringi::stri_trim_both() -> company.name

  #get job location
  job_location <- page %>%
    rvest::html_nodes('.complete-serp-result-div') %>%
    rvest::html_nodes('.jobLoc') %>%
    rvest::html_text() %>%
    stringi::stri_trim_both()

  # get links
  links <- page %>%
    rvest::html_nodes('.complete-serp-result-div') %>%
    rvest::html_nodes("h3") %>%
    rvest::html_nodes("a") %>%
    rvest::html_attr("href")
}
```

```

job_description <- c()
Job_Overview <- c()
for(j in seq_along(links)) {

  url <- paste0("https://www.dice.com", links[j])
  page <- xml2::read_html(url)

  text <- page %>%
    html_nodes(".highlight-black") %>%
    html_nodes("li") %>%
    rvest::html_text() %>%
    stringi::stri_trim_both()

  for (k in 1:length(text)){
    job_description[j] <- paste(text[k], job_description[j], sep = " ")
  }

  text <- page %>%
    html_nodes(".highlight-black") %>%
    html_nodes("p") %>%
    rvest::html_text() %>%
    stringi::stri_trim_both()
  for (k in 1:length(text)){
    Job_Overview[j] <- paste(text[k], Job_Overview[j], sep = " ")
  }
}
df <- data.frame(job_title, company_name, job_location, job_description, Job_Overview,
links)
full_df <- rbind(full_df, df)
}

write.csv(full_df, "BigDAta.csv")

```


PHỤ LỤC B

Code phân tích theo chủ đề áp dụng LDA:

```
library(quantda)
library(tidyverse)
library(RColorBrewer)
library(servr)
library(topicmodels)
library(LDAvis)
library(tm)
library(SnowballC)
library(MASS)
# đọc file
dataset <- read_csv("BigData.csv")
# xóa các kí tự
dataset <- str_replace_all(dataset$job_description,"[^[:graph:]]", " ")
#dataset <- stemDocument(dataset, language = "english")

corpus<-Corpus(VectorSource(dataset))
corpus<-tm_map(corpus, removeNumbers)
corpus<-tm_map(corpus, content_transformer(tolower))
corpus<-tm_map(corpus, removeWords, stopwords('english'))
corpus<-tm_map(corpus, removePunctuation)
corpus<-tm_map(corpus, stripWhitespace)
#corpus<-tm_map(corpus, stemDocument)
corpus<-tm_map(corpus, removeWords,
c("years","solution","skills","ability","work","knowledge","strong","working","team","including",
"solutions","including",
"using","understanding","related","etc","environment","new","required","teams","provide","testing",
"life","full",
"complex","must","s","integration","plus","preferred","multiple","excellent","ensure","plan","clients",
"review",
"identify","build","will","implementation","within","create","building","large","needs","reports",
"hands","implement","agile","lead","demonstrated",
"various","best","like","time","sales","can","help","risk","marketing","technologies",
"high","use","understand","well","least","good","relevant","minimum","test","degree","compliance",
"eg","delivery","internal","intelligence","languages","one","meet","process","client","reporting",
"application","written",
"assist","verbal","results","sets","needed","plans","requirements","support","issues","processes",
"leadership"))

answer<-unlist(as.list(corpus))
answer <- removeWords(answer, "na")
#create corpus
myCorpus <- corpus(answer)

dfm <- dfm(myCorpus,
ngrams=1L,
```

```

    stem = F,
    remove_symbols = TRUE)

vdfm <- dfm_trim(dfm, min_termfreq = 10, min_docfreq = 5)
topfeatures(vdfm, n = 50)
write.csv(topfeatures(vdfm, n = 50), "50tu.csv")

textplot_wordcloud(vdfm, scale=c(3.5, .75), colors=brewer.pal(8, "Dark2"),
    random.order = F, rot.per=0.1, max.words=150, main = "Raw Counts")

# we now export to a format that we can run the topic model with
dtm <- convert(vdfm, to="topicmodels")

# estimate LDA with K topics
K <- 10
lda <- LDA(dtm, k = K, method = "Gibbs",
    list(alpha = 50/k, estimate.beta = TRUE,
        verbose = 0, prefix = tempfile(), save = 0, keep = 0,
        seed = as.integer(Sys.time()), nstart = 1, best = TRUE,
        delta = 0.1,
        iter = 2000, burnin = 0, thin = 2000))

#Create Json for LDavis
topicmodels_json_ldavis <- function(fitted, dfm, dtm){
  # Required packages
  library(topicmodels)
  library(dplyr)
  library(stringi)
  library(quanteda)
  library(LDAvis)
  library(tm)

  # Find required quantities
  phi <- posterior(fitted)$terms %>% as.matrix
  theta <- posterior(fitted)$topics %>% as.matrix
  vocab <- colnames(phi)

  doc_length <- ntoken(dfm[rownames(dtm)])

  temp_frequency <- as.matrix(dtm)
  freq_matrix <- data.frame(ST = colnames(temp_frequency),
    Freq = colSums(temp_frequency),
    stringsAsFactors = F)
  rm(temp_frequency)

  # Convert to json
  json_lda <- LDAvis::createJSON(phi = phi, theta = theta,
    vocab = vocab,
    doc.length = doc_length,
    term.frequency = freq_matrix$Freq)

  return(json_lda)

```

```

}

json <- topicmodels_json_ldavis(lda,vdfm,dtm)
new.order <- RJSONIO::fromJSON(json)$topic.order

# change open.browser = TRUE to automatically open result in browser
serVis(json, out.dir = 'unccResearch', open.browser = T)

term <- terms(lda, 10)

# Topic #'s reordered!!
term <- term[,new.order]
colnames(term) <- paste("Topic",1:K)
term
write.csv(term, "10topic.csv")

# to get topic probabilities per document
postlist <- posterior(lda)
probttopics <- data.frame(postlist$topics)
probttopics <- probttopics[,new.order]
colnames(probttopics) <- paste("Topic",1:K)
probttopics

```

Hàm thực hiện chọn tham số cho LDA:

```

topicmodels_json_ldavis <- function(fitted, dfm, dtm){
  # Required packages
  library(topicmodels)
  library(dplyr)
  library(stringi)
  library(quanteda)
  library(LDAvis)
  library(tm)

  # Find required quantities
  phi <- posterior(fitted)$terms %>% as.matrix
  theta <- posterior(fitted)$topics %>% as.matrix
  vocab <- colnames(phi)

  doc_length <- ntoken(dfm[rownames(dtm)])

  temp_frequency <- as.matrix(dtm)
  freq_matrix <- data.frame(ST = colnames(temp_frequency),
                           Freq = colSums(temp_frequency),
                           stringsAsFactors = F)
  rm(temp_frequency)

  # Convert to json
  json_lda <- LDAvis::createJSON(phi = phi, theta = theta,
                                vocab = vocab,

```

```

doc.length = doc_length,
term.frequency = freq_matrix$Freq)

return(json_lda)
}

```

Hàm thực thống kê các keyword quan tâm:

```

library(quanteda)
library(tidyverse)
library(RColorBrewer)
library(servr)
library(topicmodels)
library(LDAvis)
library(tm)
library(SnowballC)
library(stringi)
library(stringr)
# đọc file
dataset <- read_csv("BigData.csv")
# xóa các kí tự
dataset <- str_replace_all(dataset$job_description,"^[[:graph:]]", " ")
#dataset <- stemDocument(dataset, language = "english")

corpus<-Corpus(VectorSource(dataset))
corpus<-tm_map(corpus, removeNumbers)
corpus<-tm_map(corpus, content_transformer(tolower))
corpus<-tm_map(corpus, removeWords, stopwords('english'))
corpus<-tm_map(corpus, removePunctuation)
corpus<-tm_map(corpus, stripWhitespace)

corpus<-tm_map(corpus, removeWords,
c("years","skills","work","knowledge","strong","working","team","including","using","under
standing","related","etc","environment","new","required","teams","provide","complex","must
","s","integration","plus","preferred","multiple","excellent","ensure","identify","build","will",
"implementation","within","create","building","large","needs",
"hands","implement","agile","lead","demonstrated",
,"various","best","like","time","high","use","understand","well","least","good","relevant","mi
nimum","eg","delivery","internal","intelligence","languages",
"one","meet","assist","verbal","results","sets","needed","plans"))
#corpus<-tm_map(corpus, stemDocument)
answer<-unlist(as.list(corpus))
answer <- removeWords(answer, "na")
#create corpus
myCorpus <- corpus(answer)

texts <- myCorpus$documents$texts
head(texts)
KEYWORDS <- c('hadoop','python','\\bsql\\b', 'nosql','\\br\\b', 'spark', 'sas', 'excel', 'aws',
'azure', 'java', 'tableau')
KEYWORDS_DISPLAY <- c('hadoop','python','sql', 'nosql','r', 'spark', 'sas', 'excel', 'aws',
'azure', 'java', 'tableau')

```

```
running <- data.frame(skill = KEYWORDS_DISPLAY, count = rep(0,
length(KEYWORDS_DISPLAY)))
num_jobs <- 0
res <- list("running" = running, "num_jobs" = num_jobs)

for(i in 1:length(texts)){
  df <- data.frame(skill = KEYWORDS, count = ifelse(str_detect(texts[i], KEYWORDS), 1,
0))
  res$running$count <- res$running$count + df$count
  res$num_jobs <- res$num_jobs + 1
}

res
```

PHỤ LỤC C

Các biến đổi toán học:

* Trong các biến đổi bên dưới đây, chủ đề được ký hiệu là z (thay vì θ).

Các quy tắc biến đổi logarit:

$$\log x^n = n \log x; \log(xy) = \log x + \log y; \log\left(\frac{x}{y}\right) = \log x - \log y;$$

Các xác suất:

$$P(d_i, w_j) = \sum_{k=1}^K P(d_i)P(z_k|d_i)P(w_j|z_k)$$

Hay: $P(d_i, w_j) = P(d_i)P(w_j|d_i)$ trong đó:

$$P(w_j|d_i) = \sum_{k=1}^K P(w_j, z_k|d_i) = \sum_{k=1}^K P(w_j|z_k)P(z_k|d_i)$$

Biến đổi likelihood và log-likelihood:

$$L = \prod_{i=1}^N \prod_{j=0}^M P(d_i, w_j)^{n(d_i, w_j)}$$

$$\log L = \sum_{i=1}^N \sum_{j=0}^M n(d_i, w_j) \log P(d_i, w_j)$$

$$= \sum_{i=1}^N \sum_{j=0}^M n(d_i, w_j) \log \left[\sum_{k=1}^K P(d_i) P(z_k|d_i) P(w_j|z_k) \right]$$

$$= \sum_{i=1}^N \sum_{j=0}^M n(d_i, w_j) \log \left[P(d_i) \sum_{k=1}^K P(z_k|d_i) P(w_j|z_k) \right]$$

$$= \sum_{i=1}^N \sum_{j=0}^M n(d_i, w_j) \left[\log P(d_i) + \log \sum_{k=1}^K P(z_k|d_i) P(w_j|z_k) \right]$$

$$= \sum_{i=1}^N \sum_{j=0}^M n(d_i, w_j) \log P(d_i) + n(d_i, w_j) \log \left[\sum_{k=1}^K P(z_k|d_i) P(w_j|z_k) \right]$$

$$= \sum_{i=1}^N n(d_i) \log P(d_i) + \frac{n(d_i)}{n(d_i)} \sum_{j=0}^M n(d_i, w_j) \log \left[\sum_{k=1}^K P(z_k|d_i) P(w_j|z_k) \right]$$

$$= \sum_{i=1}^N n(d_i) \left[\log P(d_i) + \sum_{j=0}^M \frac{n(d_i, w_j)}{n(d_i)} \log \left[\sum_{k=1}^K P(z_k|d_i) P(w_j|z_k) \right] \right]$$

Biến đổi bước E trong thuật toán EM:

$$P(z_k | d_i, w_j) = \frac{P(w_j, z_k | d_i)}{P(w_j, d_i)} = \frac{P(w_j | z_k, d_i) P(z_k | d_i)}{P(w_j | d_i)} = \frac{P(w_j | z_k) P(z_k | d_i)}{\sum_{l=1}^K P(w_j | z_l) P(z_l | d_i)}$$

PHỤ LỤC D

Chi tiết mức độ đóng góp 10 chủ đề vào 10 tài liệu đầu của tập tài liệu:

	text1	text2	text3	text4	text5	text6	text7	text8	text10
Topic 1									
Topic 2									
Topic 3									
Topic 4									
Topic 5									
Topic 6									
Topic 7									
Topic 8									
Topic 9									
Topic 10									
Topic 11									
Topic 12									
Topic 13									
Topic 14									
Topic 15									
Topic 16									
Topic 17									
Topic 18									
Topic 19									
Topic 20									
Topic 21									
Topic 22									
Topic 23									
Topic 24									
Topic 25									
Topic 26									
Topic 27									
Topic 28									
Topic 29									
Topic 30									
Topic 31									
Topic 32									
Topic 33									
Topic 34									
Topic 35									
Topic 36									
Topic 37									
Topic 38									
Topic 39									
Topic 40									