

BKWork: A System for Exporation and Integration of Job Portals in Vietnam

1

Nội dung

1. Giới thiệu đề tài
2. Cơ sở lý thuyết
3. Xây dựng hệ thống
4. Kết luận – Hướng phát triển



2

2

1. Giới thiệu đề tài



3

3

1. Giới thiệu đề tài

- ❖ Tích hợp dữ liệu tự động
- ❖ Khai thác dữ liệu tuyển dụng
 - ✓ Dự báo xu hướng tuyển dụng
 - ✓ Mối liên quan nhu cầu tuyển dụng và chương trình đào tạo



4

4

2. Cơ sở lý thuyết

2.1. Phương pháp kho dữ liệu

2.2. Dự báo chuỗi thời gian

2.3. Độ tương đồng WMD

5

5

2.1 Phương pháp kho dữ liệu



Quy trình phương pháp kho dữ liệu - ELT

- Extract: Truy cập dữ liệu nguồn để trích rút dữ liệu.
- Transform: Đối sánh, làm sạch và chuẩn hóa dữ liệu.
- Load: Lưu dữ liệu được chuyển đổi vào CSDL.

6

6

2.2 Dự báo chuỗi thời gian

- Một chuỗi thời gian là chuỗi các định lượng quan sát tại các thời điểm liên tiếp.
- Vượt trội cho dự báo ngắn hạn.

❖ Mô hình trung bình tích hợp tự hồi quy – ARIMA

Chuỗi thời gian không dừng $\xrightarrow{\Delta^d}$ Chuỗi dừng

$$\Delta^d y_t = c + \underbrace{\sum_{i=1}^p a_i \Delta^d y_{t-i}}_{\text{AR}(p)} + \underbrace{\varepsilon_t + \sum_{i=0}^q \beta_i \varepsilon_{t-i}}_{\text{MA}(q)}$$

\Rightarrow Mô hình ARIMA(p,d,q)

Biến đổi sai phân ngược \Rightarrow giá trị y_t

7

7

2.2 Dự báo chuỗi thời gian

ARIMA (p,d,q)

$$\Delta^d y_t = c + \sum_{i=1}^p a_i \Delta^d y_{t-i} + \varepsilon_t + \sum_{i=0}^q \beta_i \varepsilon_{t-i}$$

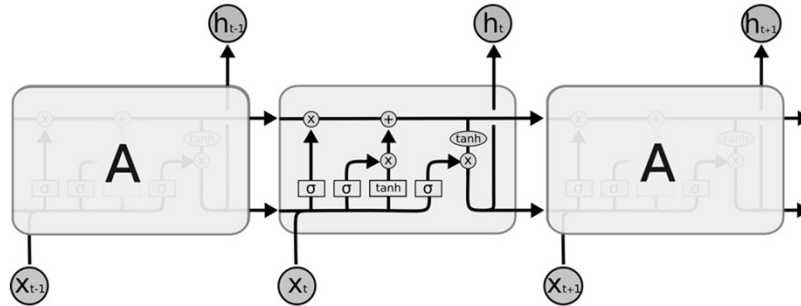
- Trong đó:
 - y_t : quan sát thứ t đối với biến phụ thuộc
 - c : hằng số
 - ε_t : nhiễu trắng
 - α_i, β_j : hệ số ước lượng
 - Δ^d : sai phân bậc d
- Biến đổi sai phân ngược \Rightarrow giá trị y_t

8

8

2.2 Dự báo chuỗi thời gian

❖ Mô hình LSTM



- Một dạng đặc biệt của RNN
- Có khả năng học được các phụ thuộc xa
- Gồm 4 tầng (3 tầng σ và 1 tầng \tanh)

9

9

2.2 Dự báo chuỗi thời gian

❖ Phương pháp đánh giá mô hình dự báo

Sai số RMSE

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

- Trong đó:
 - N : số lượng quan sát
 - y_i : giá trị quan sát trong thực tế
 - \hat{y}_i : giá trị dự đoán
- RMSE nhỏ => Mô hình dự báo cho kết quả tốt.

10

10

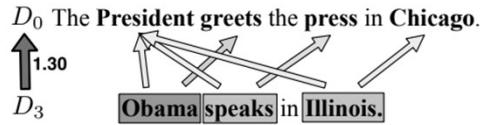
2.3 Độ tương đồng WMD

- WMD (Word Mover's Distance)
- Đo khoảng cách văn bản dựa trên ngữ nghĩa
- Tận dụng kết quả của kỹ thuật word embedding
- (Word2Vec, Glove...)
- $doc = [w_1, w_2, \dots, w_n]$
- Vec-tơ nBOW d :

$$d_i = \frac{c_i}{\sum_{j=1}^n c_j}$$

c_i : tần suất xuất hiện từ i trong văn bản

- $\forall i \in doc \longrightarrow j \in doc'$
- T_{ij} : số lần dịch chuyển từ i tới j



11

11

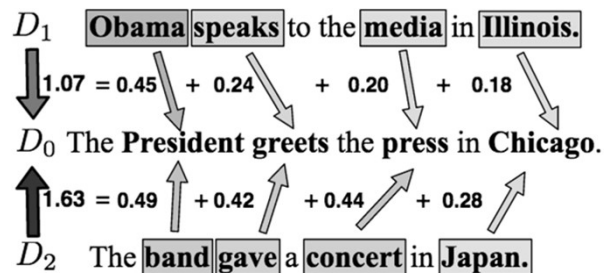
2.3 Độ tương đồng WMD

$$distance_{WMD} = \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} c(i,j)$$

Trong đó: $\sum_{j=1}^n T_{ij} = d_i \quad \forall i \in \{1, 2, \dots, n\}$

$$\sum_{i=1}^n T_{ij} = d'_j \quad \forall j \in \{1, 2, \dots, n\}$$

$c(i,j)$: khoảng cách Euclide



12

12

3. Xây dựng hệ thống

3.1 Kiến trúc tổng quan

3.2 Tích hợp dữ liệu

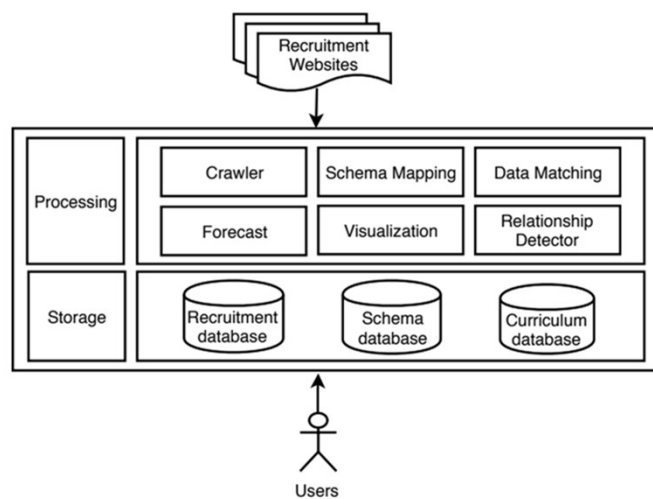
3.3 Dự báo

3.4 Mối liên quan tuyển dụng – CTĐT

13

13

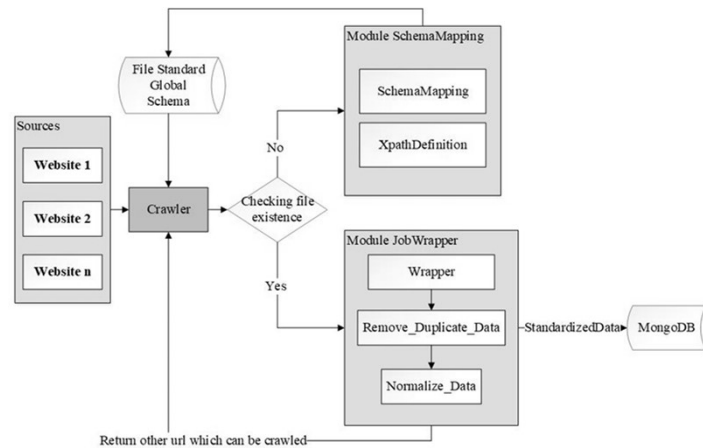
3.1 Kiến trúc tổng quan



14

14

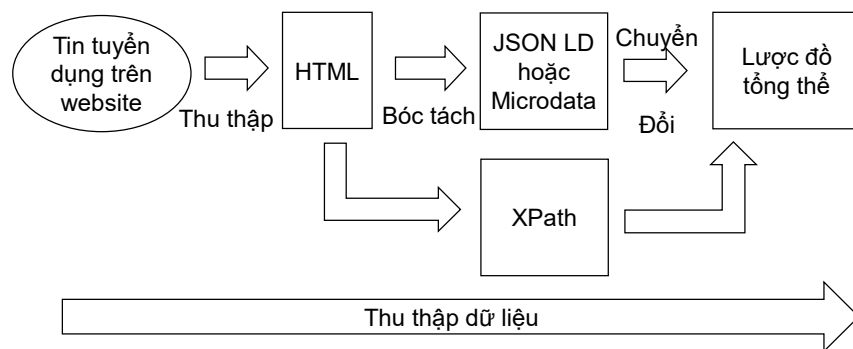
3.2. Tích hợp dữ liệu



15

15

3.2 Tích hợp dữ liệu



16

16

3.2 Tích hợp dữ liệu

Lược đồ tổng thể

```
{
  'title': "",
  'description': "",
  'jobBenefits': "",
  'skills': "",
  'qualifications': "",
  'experienceRequirements': "",
  'datePosted': datetime,
  'validThrough': datetime,
  'employmentType': "",
  'hiringOrganization_name': "",
  'jobLocation_address_addressRegion': "",
  'jobLocation_address_addressCountry': "",
  'jobLocation_address_addressLocality': "",
  'baseSalary_currency': "",
  'baseSalary_minValue': 0,
  'baseSalary_maxValue': 0,
  'baseSalary_unitText': "",
  'occupationalCategory': [],
  'totalJobOpenings': 0,
  'language': ""
}
```

17

17

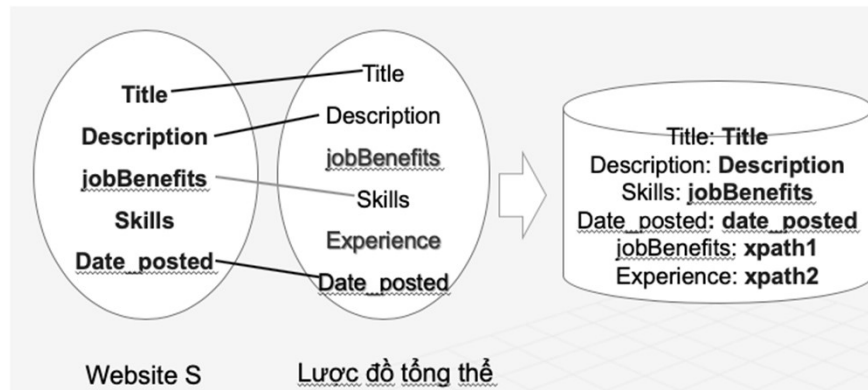
Global Schema của TopCV

```
"schema": {
  "datePosted": "datePosted",
  "baseSalary_value_unitText": "baseSalary_unitText",
  "employmentType": "employmentType",
  "validThrough": "validThrough",
  "baseSalary_currency": "baseSalary_currency",
  "skills": "occupationalCategory",
  "title": "title",
  "hiringOrganization_name": "hiringOrganization_name",
  "jobLocation_address_addressRegion": "jobLocation_address_addressRegion",
  "jobLocation_address_addressCountry": "jobLocation_address_addressCountry",
  "jobLocation_address_streetAddress": "jobLocation_address_addressLocality",
  "baseSalary_value_minValue": "baseSalary_minValue",
  "baseSalary_value_maxValue": "baseSalary_maxValue",
  "description": "description"
},
```

10

18

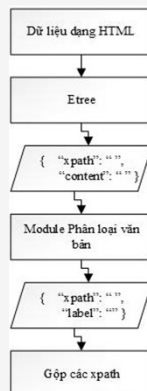
Schema mapping



19

19

XpathDefinition



✓ Đầu vào: Dữ liệu dạng HTML, N phần tử không có trong lược đồ kết quả

✓ Đầu ra: Đường dẫn xpath của N phần tử

✓ Hướng giải quyết bài toán:

1. Đưa về dạng bài toán phân loại văn bản, sử dụng SVM để nhận dạng xpath
2. Gộp xpath

11

20

TopCV

```
"selectors": {
  "job_url": "//*[@id='box-job-result']/div[1]/div/div/div[2]/h4/a",
  "next_page": "//*[@id='box-job-result']/div[2]/ul/li[last()]/a",
  "job_selectors": {
    "jobBenefits": "/html/body/div[3]/div[1]/div[4]/div[1]/div/div[1]/div[2]/p",
    "skills": "/html/body/div[3]/div[1]/div[4]/div[1]/div/div[1]/div[3]/p",
    "experienceRequirements": "/html/body/div[3]/div[1]/div[4]/div[1]/div/div[2]/div[2]/div[5]/span"
  }
}
```

12

21

Kết quả thử nghiệm

- Bảng 1. Kết quả thử nghiệm các mô hình phân ánh xạ lược đồ

Mô hình	NB	DT	LG	NB-DT	NB-LG
Kết quả(%)	98.27	98.87	98.17	99.01	98.21

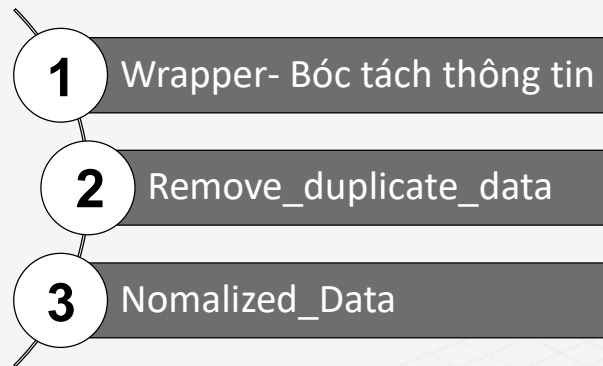
- Bảng 2. Kết quả thử nghiệm mô hình xác định xpath

Mô hình	SVM	Logistic	Naïve Bayes	Random Forest
Kết quả(%)	88.44	87.85	75.25	86.43

13

22

JobWrapper: Bộ bóc tách thông tin



14

23

Wrapper – Bóc tách dữ liệu

Website 1		Lược đồ kết quả	
Title	Nhân viên KD	Title	Title
Descriptions	Mô tả 1 Mô tả 2	Descriptions	Descriptions
jobBenefits	Yêu cầu công việc 1 Yêu cầu công việc 2	Skills	jobBenefits
Skills	Yêu cầu tuổi tác	...	
...		jobBenefits: xpath1	
		experience: xpath2	

15

24

Loại bỏ trùng lặp

Một tin tuyển dụng được coi là trùng lặp nếu:

$$\text{softTFIDF}(x, y) \geq 90\%$$

x: là bộ 3 phần tử (title, jobLocation, hiringOrganization) của tin tuyển dụng đầu vào

y: là bộ 3 phần tử (title, jobLocation, hiringOrganization) của tin tuyển dụng trong CSDL

Hướng giải quyết:

- ⇒ Đánh chỉ mục ngược
- ⇒ Tìm ra tập Y chứa các tin có khả năng trùng khớp (size filtering, prefix filtering, position filtering)
- ⇒ Tính độ tương tự giữa tin x và các tin trong Y bằng softTF-IDF

16

25

Chuẩn hóa dữ liệu

- Nguyên nhân: Sai khác cách biểu diễn dữ liệu trên các website
- Chuẩn hóa:
 - ✓ jobLocation (Tỉnh, thành phố):
 - Đưa về tên đầy đủ, ví dụ: Hà Nội, Hồ Chí Minh,...
 - ✓ occupationalCategory (Ngành nghề):
 - Xây dựng từ điển ngành nghề (Thủ công)
 - Ánh xạ dữ liệu thu thập được với từ điển

17

26

3.2 Tích hợp dữ liệu

- Ánh xạ lược đồ: Naïve Bayes và cây quyết định
- Xác định Xpath cho thuộc tính thiếu: SVM
- Kiểm tra trùng lặp: 5 thuộc tính
 - title
 - jobLocation
 - hiringOrganization
 - datePosted
 - validThrough
$$\left. \begin{array}{l} \text{title} \\ \text{jobLocation} \\ \text{hiringOrganization} \end{array} \right\} \text{sim}(x, y) \geq 0.9$$

và

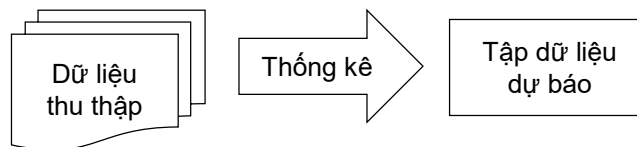
$$\left. \begin{array}{l} \text{datePosted} \\ \text{validThrough} \end{array} \right\} \text{month_range}(x, y) < 1$$
- Kết quả thu thập dữ liệu:
 - 14 websites
 - Hơn 700,000 tin tuyển dụng (2015 – 2020)

27

27

3.3 Dự báo

- 2 nguồn dữ liệu: timviecnhanh.com và topcv.vn



- 10 tập dữ liệu dự báo (8 ngành/ngành + 2 thành phố)

Công nghệ thông tin	Kế toán – Kiểm toán
Dệt may – Da giày – Thời trang	Dịch vụ – Khách sạn – Nhà hàng – Du lịch
Kinh doanh – Tư vấn – CSKH	Bảo hiểm – Chứng khoán
Bất động sản	Thành phố Hà Nội
Tài chính – Ngân hàng	Thành phố Hồ Chí Minh

28

28

3.3 Dự báo

- 59 điểm quan sát (1/2015 – 11/2019)
- Tập huấn luyện 70%
- Tập kiểm thử 30%

❖ Xây dựng mô hình dự báo ARIMA

B1. Tiền xử lý

$$\hat{y}_i = \ln y_i$$

B2. Xác định p, q, d

B3. Huấn luyện mô hình ARIMA(p,d,q)

Sử dụng thuật toán Rolling ARIMA

B4. Đánh giá mô hình trên tập kiểm thử

B5. Lựa chọn mô hình phù hợp

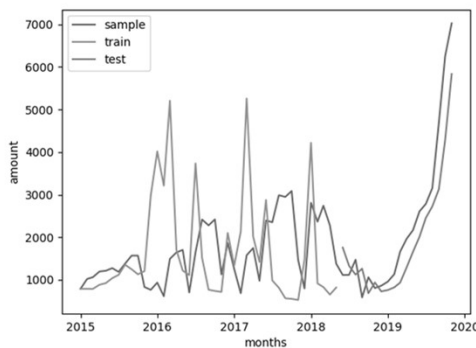
29

29

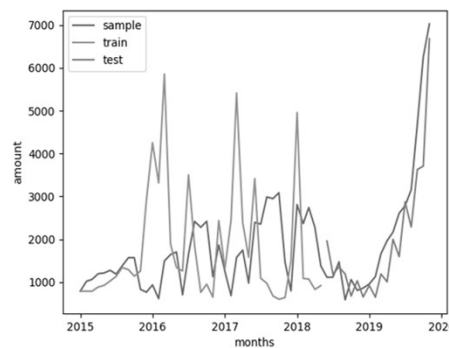
3.3 Dự báo

❖ Thử nghiệm mô hình ARIMA(p,d,q)

Tập dữ liệu ngành Công nghệ thông tin



Mô hình ARIMA(3,2,1)



Mô hình ARIMA(4,2,1)

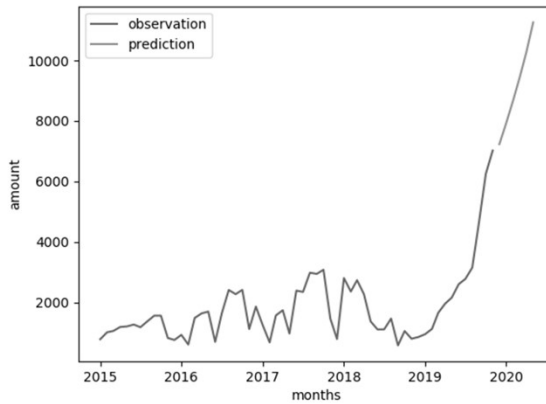
Mô hình	RMSE
ARIMA(3,2,1)	780.45
ARIMA(4,2,1)	839.45

30

30

3.3 Dự báo

❖ Kết quả dự báo ARIMA(p,d,q)



Tháng	Số lượng
12 – 2019	7,235
1 – 2020	7,938
2 – 2020	8,684
3 – 2020	9,430
4 – 2020	10,293
5 – 2020	11,259

Nhu cầu tuyển dụng ngành CNTT (6 tháng)

31

31

3.3 Dự báo

❖ Xây dựng mô hình dự báo LSTM

B1. Tiền xử lý

Chuẩn hóa giá trị trong phạm vi 0 - 1

B2. Xác định p

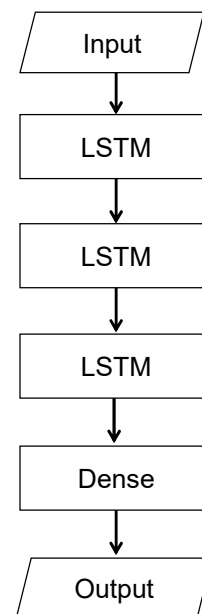
B4. Huấn luyện mô hình LSTM

B5. Đánh giá mô hình trên tập kiểm thử

B6. Lựa chọn mô hình phù hợp

❖ Mô hình LSTM

- Hàm kích hoạt ReLU
- Hàm lỗi mse
- Thuật toán tối ưu Adam



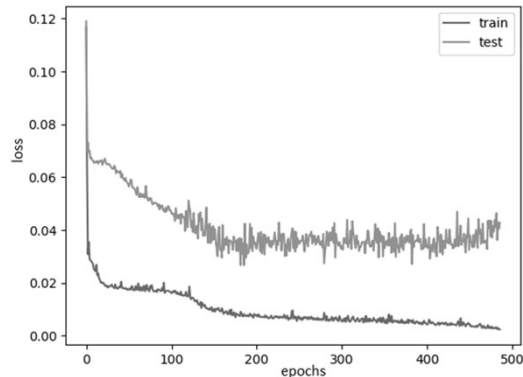
32

32

3.3 Dự báo

❖ Tối ưu tham số

- Phương pháp Early Stopping
- Phương pháp Model Checkpoint



Lỗi trong quá trình học mô hình

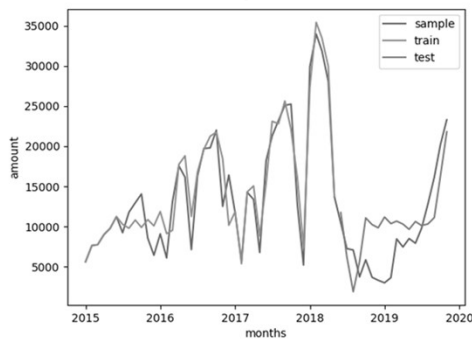
33

33

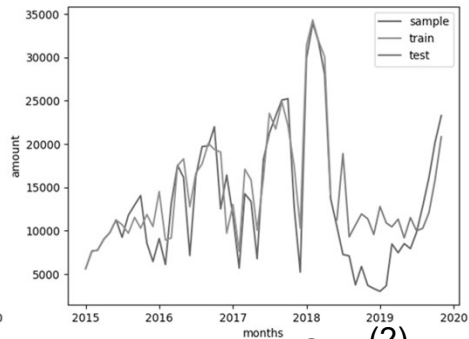
3.3 Dự báo

❖ Thử nghiệm mô hình LSTM

Tập dữ liệu thành phố Hồ Chí Minh



Mô hình LSTM(1)



Mô hình LSTM(2)

Mô hình	RMSE
(1) $p=6$, $\eta = 2e-3$, epochs = 401	780.45
(2) $p=3$, $\eta = 1e-3$, epochs = 295	839.45

34

34

3.3 Dự báo

❖ So sánh ARIMA và LSTM

Tập dữ liệu	RMSE		Δ RMSE (%)
	LSTM	ARIMA	
Công nghệ thông tin	806.88	780.45	3.28
Dệt may – Da giày – Thời trang	396.92	405.57	-2.13
Kế toán – Kiểm toán	658.23	642.55	2.38
Bất động sản	925.86	740.39	20.03
Tài chính – Ngân hàng	1545.55	982.91	36.40
Dịch vụ - Khách sạn – Nhà hàng – Du lịch	1933.62	1401.39	27.52
Bảo hiểm – Chứng khoán	706.49	450.20	36.28
Kinh doanh – Tư vấn – CSKH	4183.96	2294.60	28.42
Thành phố Hồ Chí Minh	6350.14	3729.18	41.27
Thành phố Hà Nội	5068.08	2343.31	53.76

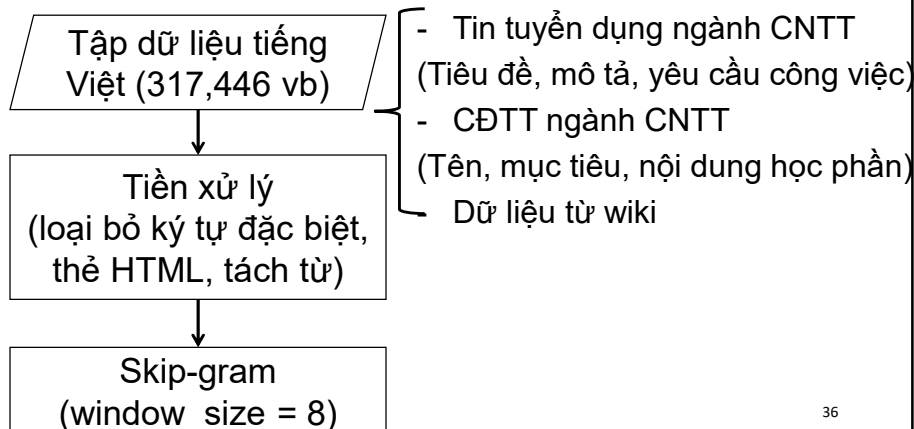
35

35

3.3 Mối liên quan tuyển dụng - CTĐT

- Thống kê mức độ phù hợp học phần – yêu cầu tuyển dụng (1)
- Top 5 học phần liên quan tới công việc cụ thể (2)

❖ Xây dựng mô hình Word2Vec



36

36

3.3 Mối liên quan tuyển dụng - CTĐT

❖ Xác định độ tương đồng WMD

B1. Loại bỏ từ dừng

B2. Tính độ tương đồng giữa các văn bản

```
instance = WmdSimilarity(wmd_corpus,model,num_best)
sims = instance[query]
```

	Vấn đề 1	Vấn đề 2
wmd_corpus	tập tin tuyển dụng CNTT (c_1)	CTĐT ngành CNTT (c_2)
num_best	kích thước c_1	kích thước c_2
query	học phần	tin tuyển dụng

(1) $sim(job, course) \geq \theta$

=> Học phần phù hợp với công việc tuyển dụng.

37

37

3.3 Mối liên quan tuyển dụng - CTĐT

Top 10 học phần đáp ứng công việc ngành CNTT
(10,272 vị trí tuyển dụng, $\theta = 0.525$)

Học phần	Số lượng công việc	Tỷ lệ (%)
Nhập môn công nghệ phần mềm	7,845	76.37
Phát triển hệ thống Web an toàn	7,711	75.07
Công nghệ Web và dịch vụ trực tuyến	7,606	74.05
Quản lý dự án phần mềm	7,372	71.77
Đồ án: Các công nghệ xây dựng hệ thống thông tin	7,311	71.17
Linux và phần mềm mã nguồn mở	7,303	71.10
Đồ án môn học: Phát triển phần mềm chuyên nghiệp	6,704	65.26
Phát triển ứng dụng cho thiết bị di động	6,662	64.86
Lập trình kịch bản với Javascript	6,400	62.31
Lập trình hệ thống	6,328	61.60

38

38

Dự báo xu hướng



41

41

Chương trình đào tạo

Mối liên hệ tin tuyển dụng - CTĐT

Search for...

Go!

Thống kê học phần (ngưỡng=0.525)











STT	Mã học phần	Học phần	Số lượng tin	Tỷ lệ (%)
1	IT4080	Nhập môn công nghệ phần mềm	7845	76.37266355140187
2	IT4403	Phát triển hệ thống Web an toàn	7711	75.06814641744549
3	IT4409	Công nghệ web và dịch vụ trực tuyến	7606	74.04595015576324
4	IT4541	Quản lý dự án phần mềm	7372	71.76791277258567
5	IT4421	Đồ án: Các công nghệ xây dựng hệ thống thông tin	7311	71.17406542056075
6	IT3110	LINUX và phần mềm nguồn mở	7303	71.09618380062304
7	IT4551	Đồ án môn học: Phát triển phần mềm chuyên nghiệp	6704	65.26479750778816
8	IT4785	Phát triển ứng dụng cho thiết bị di động	6662	64.85591900311525
9	IT4766	Lập trình kịch bản với JavaScript	6400	62.305295950155774
10	IT4786	Lập trình hệ thống	6328	61.604361370716504

First Previous 1 2 3 4 5 Next Last

42

42

Top học phần đáp ứng công việc

Danh sách học phần		Search for...	Go!
Thực tập sinh Mobile (Android/iOS)			
 Phát triển ứng dụng cho thiết bị di động 0.540374973429056 IT4785	1	 Đồ án tốt nghiệp cử nhân 0.533576320325819 IT4995	2
 Multimedia, trò chơi và các Hệ thống giải trí 0.5203020060243719 IT4898	3	 Công nghệ JAVA 0.5202248526090135 IT4784	4
 Đồ án môn học: Phát triển phần mềm chuyên nghiệp 0.5200659077048594 IT4551	5	 Thực tập kỹ thuật 0.5184867647837098 IT4991	6
 Project II 0.5168945427510953 IT3920	7	 Project I 0.5167649480670902 IT3910	8
 Văn phòng kỹ thuật 0.5158060347405133 IT4888	9	 Lập trình hướng đối tượng 0.5153698346519424 IT3100	10
First Previous 1 2 3 4 5 Next Last			

43

4. Kết luận

- ❖ Kết quả đạt được
 - Phát triển module tích hợp dữ liệu
 - Thu thập hơn 700,000 tin tuyển dụng từ 2 nguồn chính: topcv.vn và timviecnhanh.com
 - Cải thiện thuật toán kiểm tra trùng lặp
 - Tìm hiểu và xây dựng các mô hình dự báo
 - Áp dụng dự báo xu hướng nghề nghiệp 10 tập dữ liệu
 - Tìm hiểu độ tương đồng WMD
 - Xác định danh sách học phần phù hợp với công việc dựa trên độ tương đồng WMD

44

44

4. Kết luận

- ❖ Hạn chế:
 - Tiền xử lý đơn giản => Chưa loại bỏ nhiễu
 - Chưa chứng minh kết quả dự báo là phù hợp
 - Chọn giá trị ngưỡng mang tính chủ quan
- ❖ Hướng phát triển
 - Thu thập dữ liệu tuyển dụng 2015 trở về trước
 - Kết hợp các yếu tố ảnh hưởng cho bài toán dự báo
 - Áp dụng mô hình ARIMA online
 - Thu thập chi tiết nội dung học phần
 - Kết hợp mô hình word embedding BERT và WMD

45

45

4. Tài liệu tham khảo

- [1] Đ. T. T. Nga, "BKWorks - Hệ thống tích hợp và khai thác dữ liệu việc làm," 2019.
- [2] P. H. Quang, "Github," [Online]. Available: https://github.com/quangph-1686/FramgiaBlog/tree/master/Blog01_Word_embedding.
- [3] Kusner, Matt and Sun, Yu and Kolkin, Nicholas and Weinberger, Kilian, "From word embeddings to document distances," in *International conference on machine learning*, 2015.
- [4] Thực Đoan và Hào Thi, Nhập môn kinh tế lượng với các ứng dụng, Hồ Chí Minh: Trường chính sách công và quản lý Fulbright, 2013.
- [5] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

46

46