

ONE LOVE. ONE FUTURE.



XÂY DỰNG HỆ THỐNG GOI Ý TIN TUYẾN DỤNG

TRONG LĨNH VỰC CÔNG NGHỆ THÔNG TIN

Sinh viên thực hiện: Đỗ Đức Thái

Giảng viên hướng dẫn: TS.Trần Việt Trung

HUST hust.edu.vn f fb.com/dhbkhn

Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển



Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển

Đặt vấn đề

Các web tuyển dụng trực tuyến là kênh chính trong lĩnh vực việc làm.

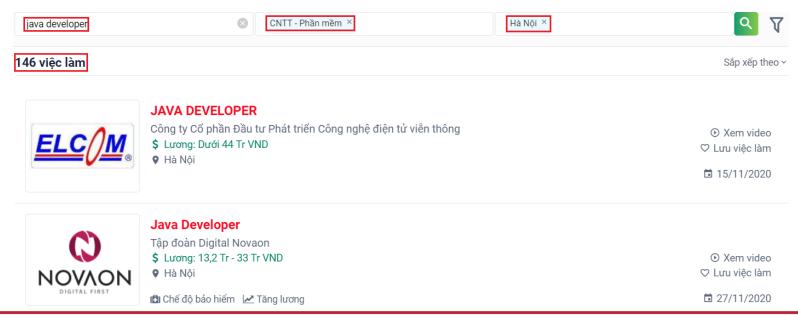






Nhược điểm:

- Lưu trữ nhưng không sử dụng thông tin người dùng.
- Tìm kiếm sử dụng từ khoá, không cá nhân hoá.





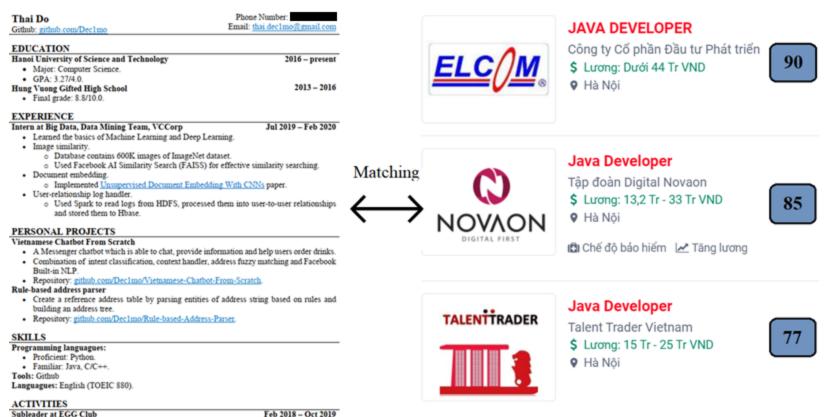
Định nghĩa bài toán

Learned Data Structures and Algorithms from older students and taught smaller ones.

Đầu vào:

- p: Thông tin tổng hợp của ứng viên = Sơ yếu lý lịch + hồ sơ ứng viên.
- $J = \{j_1, j_2, ... j_n\}$: Tập tin tuyển dụng.

Đầu ra: Một tập tin tuyển dụng $J' \subset J$ phù hợp nhất với p.







Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển

Phân tích khám phá dữ liệu

Dữ liệu sơ yếu lý lịch, hồ sơ ứng viên: 12.000.

Dữ liệu tin tuyển dụng: 6.000.

Hai tính chất của dữ liệu:

- Tính có sẵn: tỷ lệ số bản ghi khác rỗng trên tổng số bản ghi.
- Tính xác thực: độ phù hợp với thực tế của giá trị trong bản ghi.

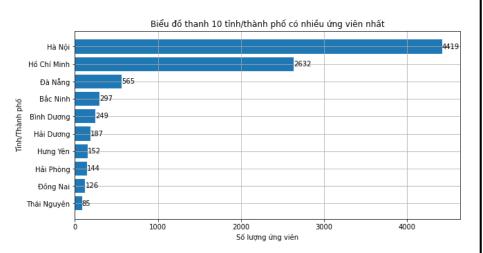
Dữ liệu hồ sơ ứng viên và tin tuyển dụng:

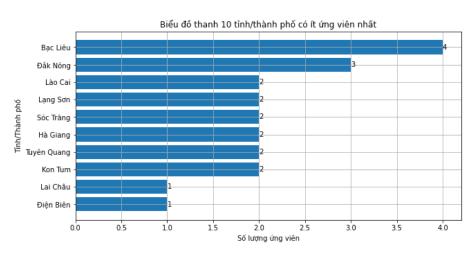
- Là dữ liệu thực tế, không nhãn, gặp nhiều vấn đề về chất lượng.
- Có sự tương đồng về ý nghĩa và phân phối.

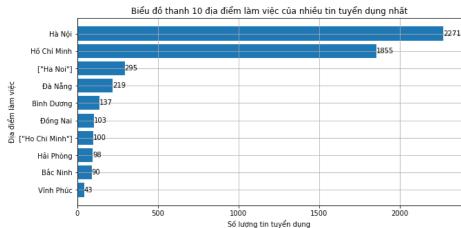


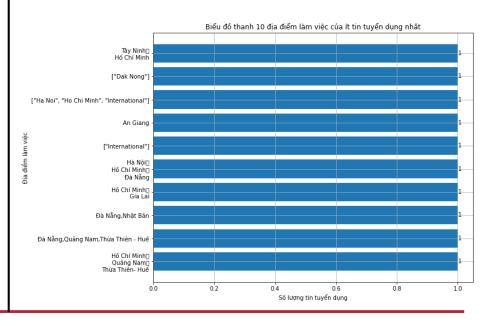
Phân tích khám phá dữ liệu – Địa điểm

Hồ sơ ứng viên





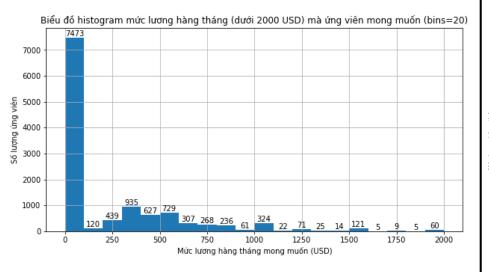




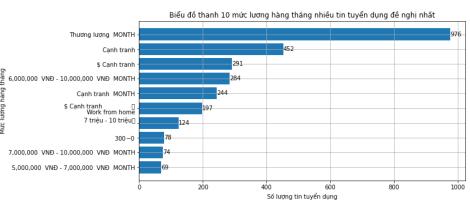


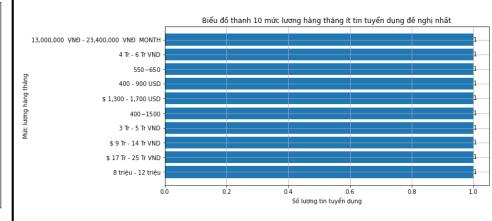
Phân tích khám phá dữ liệu – Mức lương hàng tháng

Hồ sơ ứng viên



Mức lương mong muốn (USD)





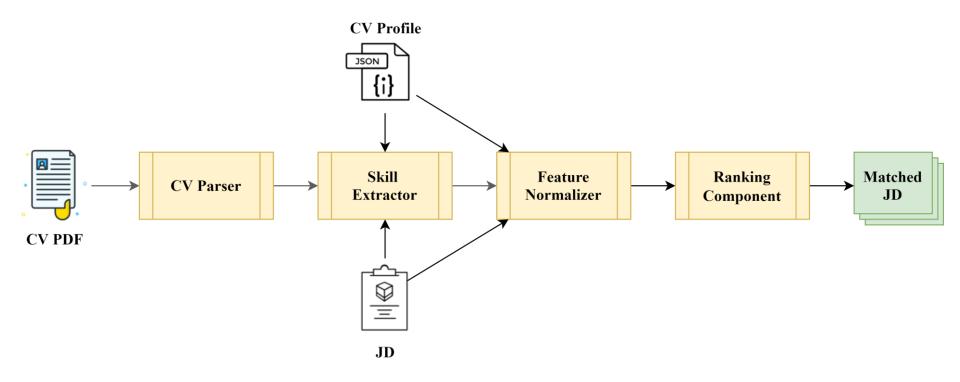




Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển

Tổng thể luồng xử lý của hệ thống

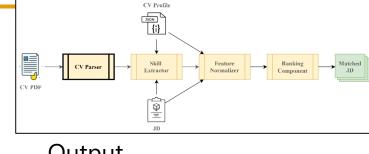




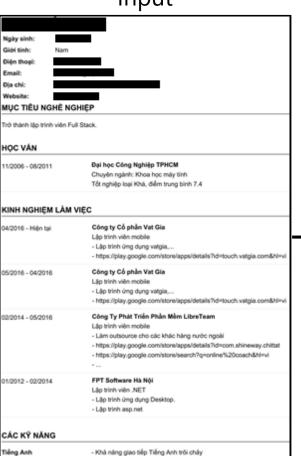
CV Parser

Input: CV định dạng PDF.

Output: JSON.



Input



CV

Parser

Output

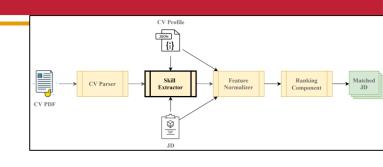
```
("name":
("email":
("phone": ""),
{"birthday": "
("company_time": "04 / 2016 - Hién_tai"),
("company": "Cong_ty Co_phan Vat_Gia"),
("position": "Lập_trình_viên mobile"),
("company_time": "05 / 2016 - 04 / 2016"),
("company": "Công ty Cổ phần Vat Gia"),
("position": "Lap_trinh_vien mobile"),
("company_time": "02 / 2014 - 05 / 2016"),
{"company": "Cong_Ty Phát_Triển Phần_Mêm_LibreTeam"},
("position": "Lập_trinh_viên mobile"),
{"company_time": "01 / 2012 - 02 / 2014"},
("company": "FPT Software Hà Nôi"),
("position": "Lập_trinh_viên . NET Lập_trinh ứng_dụng Desktop ."),
{"study_time": "11 / 2006 - 08 / 2011"},
("university": "Dai_hoc Cong_Nghiep TPHCM"),
"major": "Chuyên_ngành : Khoa_học máy_tính"},
("cpa": "Tot_nghiệp loại Khá")
("exp": [
            "Cong_ty Co_phan Vat_Gia",
            "04 / 2016 - Hiện_tại",
            "Lập trình viên mobile",
            "Lap_trinh ung_dung vatgia , ... https://play.google.com/
            "Cong_ty Co_phan Vat_Gia",
            "05 / 2016 - 04 / 2016",
            "Lập_trình_viên mobile",
            "Lap_trinh ung_dung vatgia , ... https://play.google.com/
            "Cong_Ty Phát_Triển Phần_Mêm_LibreTeam",
            "02 / 2014 - 05 / 2016",
            "Lập_trình_viên mobile",
            "Làm outsource cho các khác hàng nước_ngoài https://play.
             /search?q=online%20coach&hl=vi ...",
            "FPT Software Ha_Noi",
            "01 / 2012 - 02 / 2014",
            "Lập_trình_viên . NET Lập_trình ứng_dụng Desktop .",
            "Lap_trinh asp . net"
```

Skill Extractor

Input: Đoạn văn bản.

- CV: Kinh nghiệm làm việc.
- JD: Mô tả + Yêu cầu công việc.

Output: Danh sách kỹ năng.



Input

```
"Công ty Cổ phần Vat Gia",
"04 / 2016 - Hiện_tại",
"Lập trình viên mobile",
"Lâp trình ứng dụng vatgia , ... https://play.google.com/
"Công ty Cổ phần Vat Gia",
"05 / 2016 - 04 / 2016",
"Lâp trình viên mobile",
                                                          Extractor
"Lâp trình ứng dung vatgia , ... https://play.google.com/
"Công Ty Phát Triển Phần Mềm LibreTeam",
"02 / 2014 - 05 / 2016",
"Lập trình viên mobile",
"Làm outsource cho các khác hàng nước ngoài https://play.
 /search?q=online%20coach&hl=vi ...",
"FPT Software Hà Nội",
"01 / 2012 - 02 / 2014",
"Lập_trình_viên . NET Lập_trình ứng_dụng Desktop .",
"Lập trình asp . net"
```

Output

```
".net",
"asp",
"desktop".
"fpt software",
"lâp trình",
"lập_trình_viên_mobile",
"lập_trình_ứng_dụng",
"outsource",
"phát triển phần mềm",
"phần_mềm",
"software",
"ứng_dụng"
```

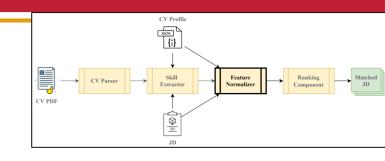
Skill

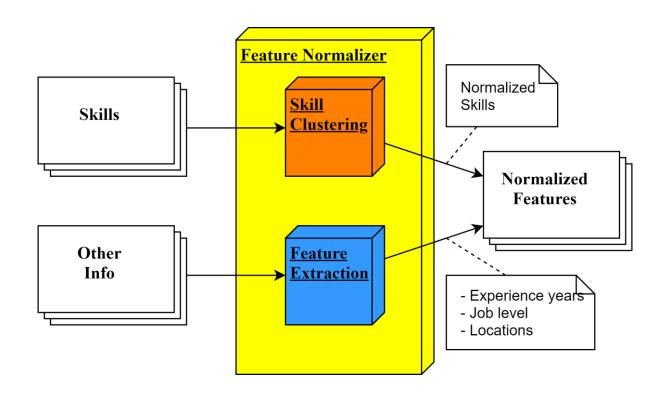


Feature Normalizer

Feature Normalizer: Bóc tách, xử lý, chuẩn hoá các đặc trưng.

- Skill Clustering: Bộ chuẩn hoá kỹ năng.
- Feature Extraction: bóc tách và xử lý các đặc trưng khác.



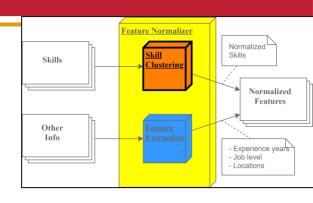


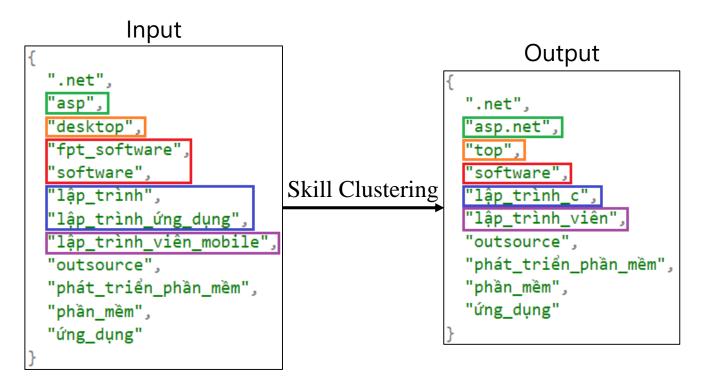


Skill Clustering

Input: Tập kỹ năng chưa chuẩn hoá.

Output: Tập kỹ năng đã chuẩn hoá.







Skill Clustering

Affinity Propagation:

- Độ tương đồng n-gram giữa các kỹ năng.
- Tự xác định số cụm.
- Tâm cụm là một thành phần trong cụm.

6760 kỹ năng → 1002 kỹ năng.

Các kỹ năng trước khi chuẩn hoá	Kỹ năng sau khi
	chuẩn hoá
.net, aps.net., c#.net, net, telnet, vb.net	.net
.net_mvc, aps.net_mvc, asp.net_mvc, net_mvc	.net_mvc
3d_max, 3d, 3ds_max	3d_max
3d_modeling, 3d_model, data_modeling, modeling	3d_modeling
ability, abc, ability_to, ability_to_work, capability, facility, usability,	ability
utility, vulnerability	
academy, academic, vtc_academy	academy
access, acc, acca, acceptance, less, ms_access	access
accounting, account, accountant, accounts, accouting, country,	accounting
fast_accounting	
activiti, activation, active, activity, cacti, tivi	activiti
actor, acb, editor, factor, motor, or	actor



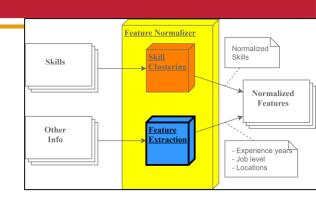
Feature Extraction

Input: JSON chứa các trường dữ liệu.

Output: Các đặc trưng ngoài kỹ năng.

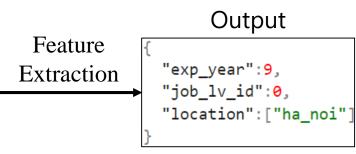
Kết hợp một số kỹ thuật nâng cao:

- Fuzzy Searching.
- Loại bỏ ngoại lai.



Input

```
{"company_time": "04 / 2016 - Hien_tai"},
{"company": "Công ty Cổ phần Vat Gia"},
{"position": "Lập_trình_viên mobile"},
{"company time": "05 / 2016 - 04 / 2016"}.
{"company": "Công ty Cổ phần Vat Gia"},
{"position": "Lập_trình_viên mobile"},
{"company_time": "02 / 2014 - 05 / 2016"},
{"company": "Công Ty Phát Triển Phần Mèm LibreTeam"},
{"position": "Lập_trình_viên mobile"},
{"company_time": "01 / 2012 - 02 / 2014"},
{"company": "FPT Software Ha Noi"},
{"position": "Lap trình viên . NET Lap trình ứng dụng Desktop ."}
{"study_time": "11 / 2006 - 08 / 2011"},
{"university": "Đại học Công Nghiệp TPHCM"},
{"major": "Chuyên_ngành : Khoa_học máy_tính"},
("cpa": "Tốt nghiệp loại Khá")
```

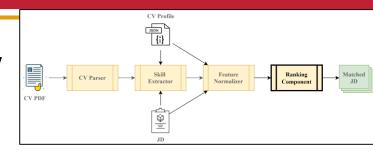


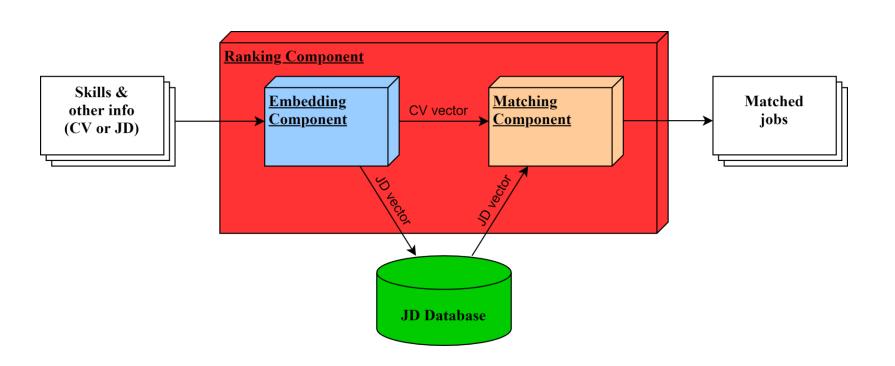


Ranking Component

Ranking Component: Nhúng đặc trưng, so sánh và xếp hạng kết quả.

- Embedding Component: Biểu diễn đặc trưng về dạng vector.
- Matching Component: So ghép và xếp hạng các vector.





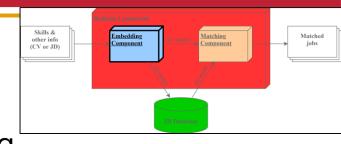


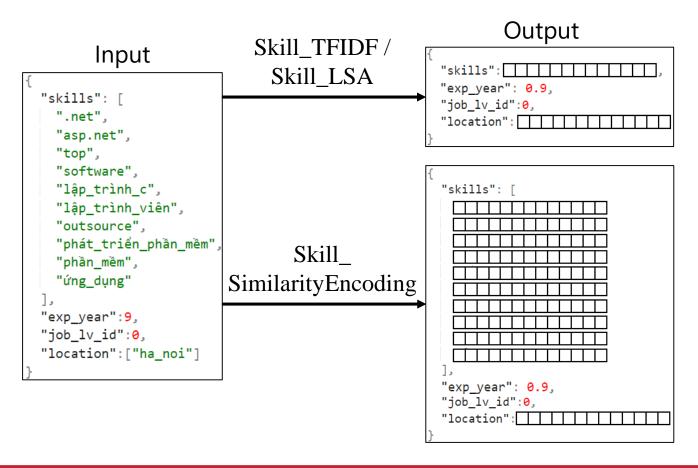
Embedding Component

Input: Các đặc trưng.

Output: Vector biểu diễn.

Ba mô hình: TFIDF, LSA, SimilarityEncoding







Matching Component

$$Score = \begin{bmatrix} s_{skills}(skills_{CV_PF}, skills_{JD}) \\ s_{expyear}(expyear_{CV_PF}, expyear_{JD}) \\ s_{joblv}(joblv_{CV_PF}, joblv_{JD}) \\ s_{location}(location_{CV_PF}, location_{JD}) \end{bmatrix} \begin{bmatrix} w_{skills} \\ w_{expyear} \\ w_{joblv} \\ w_{location} \end{bmatrix}$$

Trong đó: s_{feat} , w_{feat} là lần lượt độ tương đồng giữa CV_PF với JD và trọng số tương ứng xét trên (các) đặc trưng feat.

- s_{skills} : Độ tương đồng cosine, Word Mover's Distance (WMD).
- $s_{expyear}$, s_{joblv} : 1 $max(0, t_j t_c)$, trong đó: t_j và t_c là expyear hoặc joblv của lần lượt tin tuyển dụng và thông tin ứng viên.
- $s_{location}$: 0 nếu ($location_{CV-PF}$ AND $location_{JD}$) = 0; 1 nếu ngược lại.
- Nếu dữ liệu có nhãn, để chọn các trọng số w phù hợp nhất, sử dụng mô hình phân loại.
- $(w_{skills}, w_{expyear}, w_{joblv}, w_{location}) = (0.4, 0.2, 0.2, 0.2)$





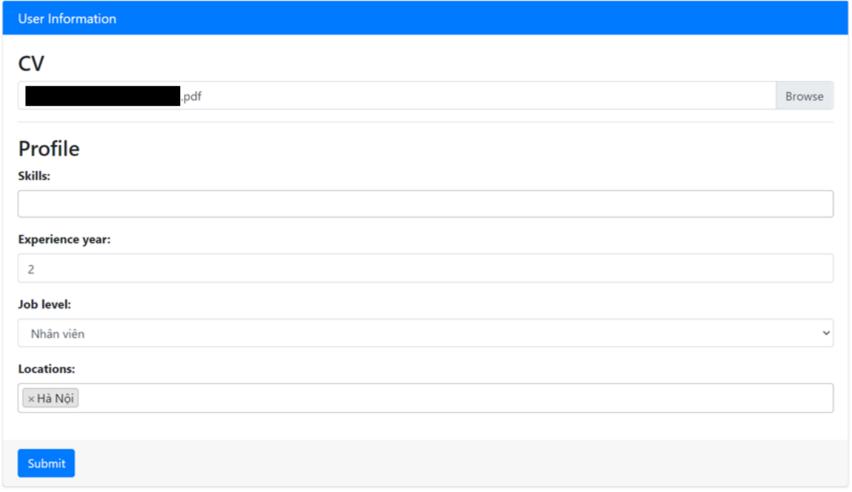
Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển

Trang người dùng cung cấp thông tin

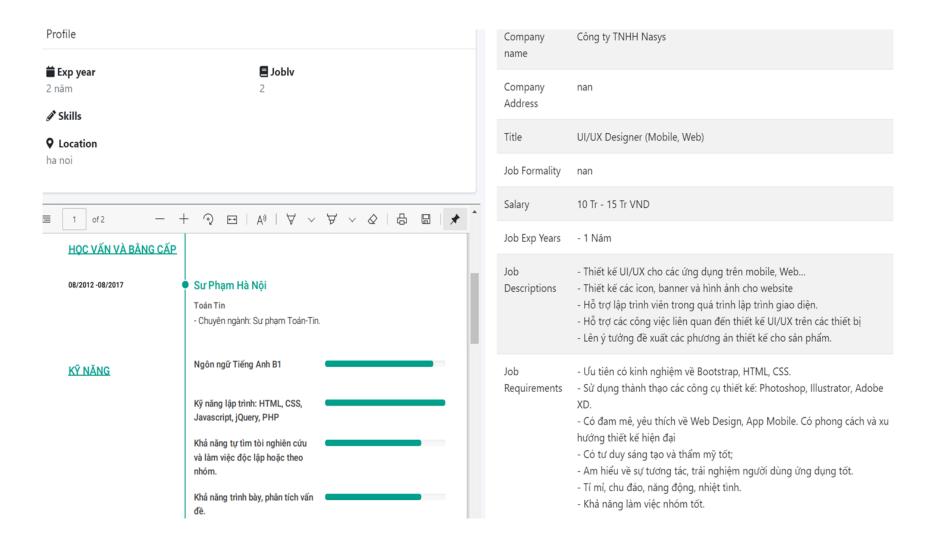


HỆ THỐNG GỢI Ý TIN TUYỂN DỤNG TRONG LĨNH VỰC CNTT



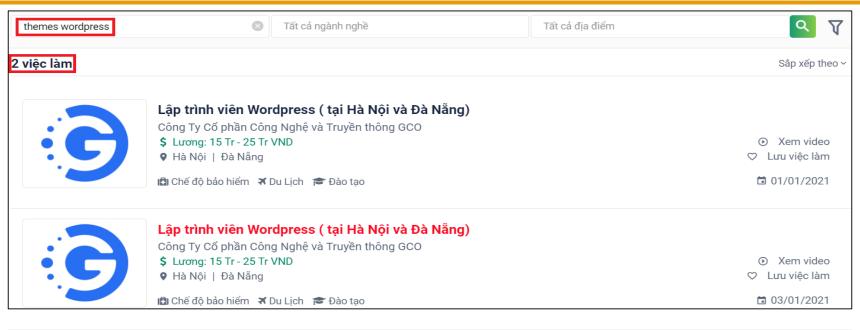


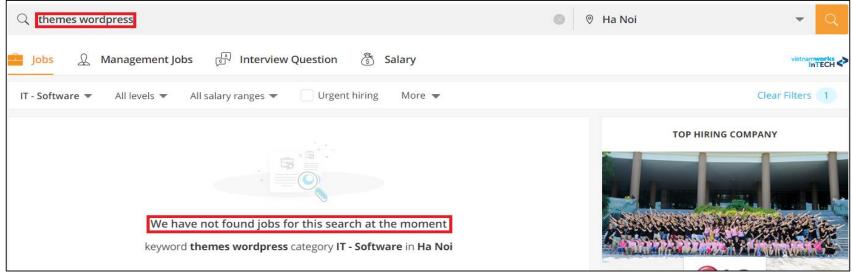
Trang kết quả





Kết quả tìm kiếm từ trang web khác







Đánh giá kết quả

Do kết quả định tính, sử dụng Mean Opinion Score (MOS).

Đánh giá	Ý nghĩa
5	Rất phù hợp
4	Phù hợp
3	Bình thường
2	Không phù hợp
1	Rất không phù hợp

Do kinh phí không cho phép, đánh giá chưa thể hoàn toàn chính xác và khách quan:

- Số cặp (thông tin ứng viên, tin tuyển dụng) đánh giá: 675 cặp.
- Số người đánh giá: 2 người. Thời gian đánh giá: 6h/người.

	MOS	
k	k=1	k=3
Skills_TFIDF + cosine similarity	2.2467	2.3467
Skills_LSA+ cosine similarity	2.9467	3.3067
Skills_SimilarityEncoder + WMD	3.0533	3.3667





Nội dung

- 1. Giới thiệu đề tài
- 2. Phân tích khám phá dữ liệu
- 3. Luồng xử lý
- 4. Demo và đánh giá kết quả
- 5. Kết luận và hướng phát triển

Kết luận

Đóng góp:

- Phân tích và xử lý dữ liệu thực tế.
- Thiết kế, xây dựng và thử nghiệm đánh giá một hệ thống gợi ý tin tuyển dụng hoàn chỉnh:
 - Xây dựng một bộ chuẩn hoá kỹ năng.
 - Xây dựng thành phần Ranking Component.

Hạn chế:

- Chưa đánh giá được bộ chuẩn hoá kỹ năng.
- Độ tương đồng WMD cần nhiều thời gian tính toán.
- Số lượng tin tuyển dụng ít.



Hướng phát triển

- Thu thập thêm dữ liệu.
- Gán nhãn dữ liệu.
- Áp dụng online learning cho hệ thống.
- Cải thiện tốc độ WMD bằng Word Centroid Distance, Relaxed Word Moving Distance,...





THANK YOU!



Phụ lục

Phân tích khám phá dữ liệu (Tiếp)

Sơ yếu lý lịch trong demo



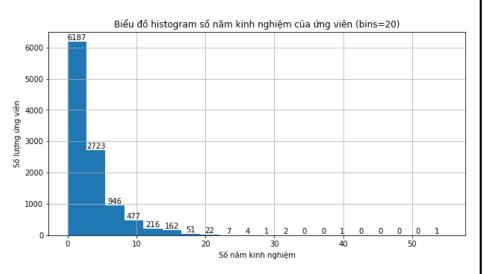
Phụ lục

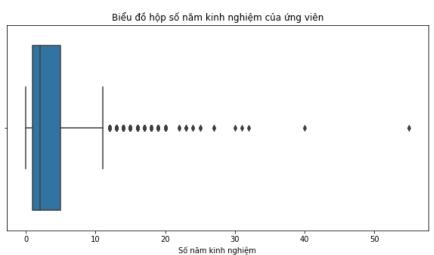
Phân tích khám phá dữ liệu (Tiếp)

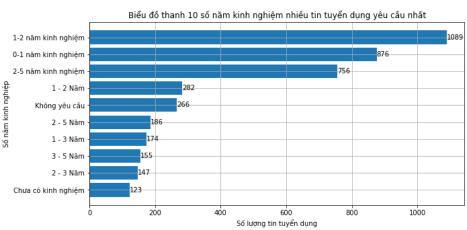
Sơ yếu lý lịch trong demo

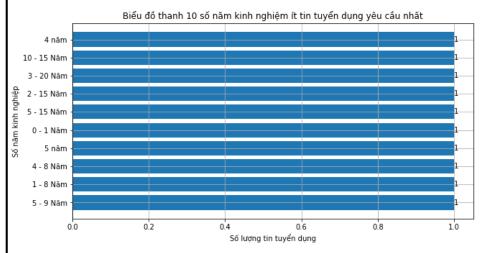
Phân tích khám phá dữ liệu – Số năm kinh nghiệm

Hồ sơ ứng viên





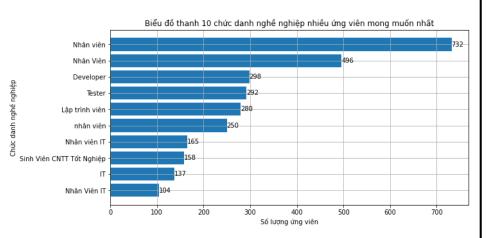


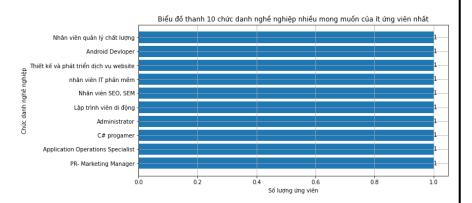


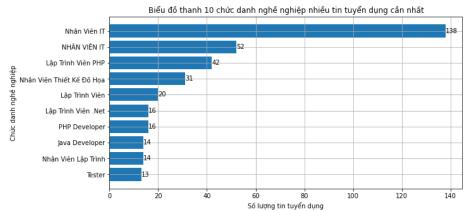


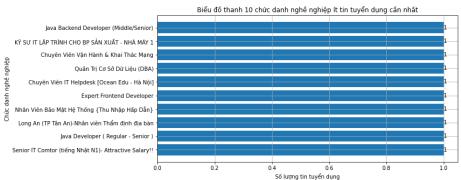
Phân tích khám phá dữ liệu – Chức danh

Hồ sơ ứng viên





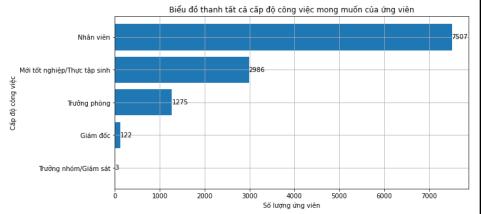


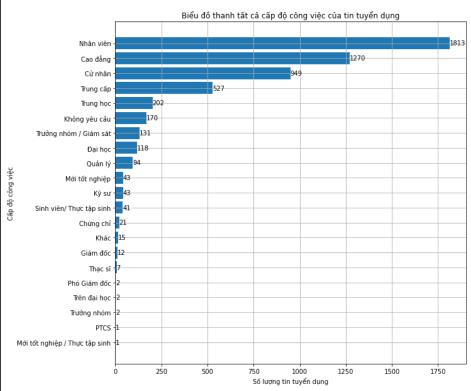




Phân tích khám phá dữ liệu – Cấp độ công việc

Hồ sơ ứng viên

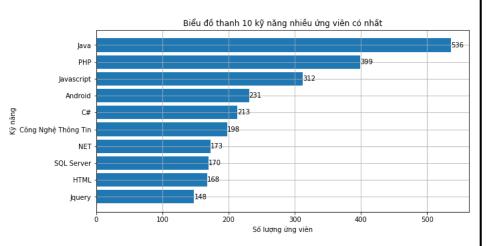


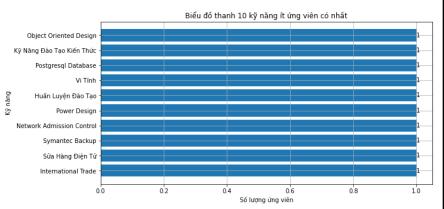


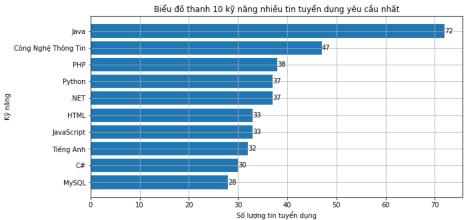


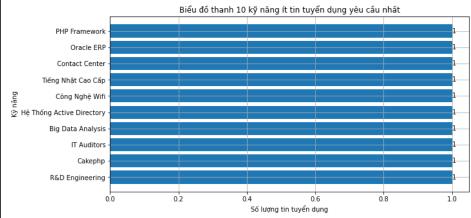
Phân tích khám phá dữ liệu – Kỹ năng

Hồ sơ ứng viên













Phụ lục

Phân tích khám phá dữ liệu (Tiếp)

Sơ yếu lý lịch trong demo

Sơ yếu lý lịch trong demo



MUC TIÊU NGHỀ NGHIỆP

- Mong muốn được học tập và làm việc trong lĩnh vực IT, cụ thể là lập trình viết themes wordpress, để có thể trau đồi, vận dụng và phát huy những kiến thức đã học.
- Mục tiêu trong 3 năm tới có thể trở thành một Developer Magento chuyển nghiệp, góp phần xây dựng nên những app có thể áp dụng vào nhiều mặt của đời sống một cách hiệu quả.

HỌC VẤN VÀ BẰNG CẮP Sư Phạm Hà Nội Toán Tin - Chuyển ngành: Sư phạm Toán-Tin. KỸ NĂNG Ngôn ngữ Tiếng Anh B1 Kỹ năng lập trình: HTML, CSS, Javascript, jQuery, PHP Khá năng tự tìm tòi nghiên cứu và làm việc độc lập hoặc theo nhóm. Khá năng trình bày, phân tích vấn đề.

KINH NGHIỆM LÀM VIỆC

08/2016 - 12/2016

Đã có 2 năm kinh nghiệm

PVESER

- Lập trình viên
- Đọc, nghiên cứu tài liệu .
- Làm việc nhóm với đội lập trình.

DỰ ÁN ĐÃ THAM GIA

06/2016 -08/2017

Viết themes bán hàng và plugins

PVESER

Vị trí công việc: Nhân viên IT

Quy mô dự án: Vừa

 - Được phân công làm về các mảng: cắt ảnh photoshop css, xử lí multiple screens, custom wigdet, custom view, fix bug.

Úng viên	Tin tuyển dụng
- Kinh nghiệm: 2 năm	- Kinh nghiệm yêu cầu: 1
- Có kỹ năng về HTML,	năm
CSS, Javascript, jQuery	- Ưu tiên có kinh nghiệm về
- Đã từng làm về các	Bootstrap, HTML, CSS
mång cắt ảnh Photoshop	- Sử dụng thành thạo công cụ
CSS, xử lý multiple	thiết kế: Photoshop,
screens, custom widget,	- Có đam mê yêu thích về
custom view	Web Design,

