BÁO CÁO CÔNG VIỆC

Sinh viên thực hiện: Đinh Thanh Hải – 20151176 – Lớp CNTT 2.1 K60

Đến ngày 10/5:

1. **Công việc đã hoàn thành**

Bước 1:

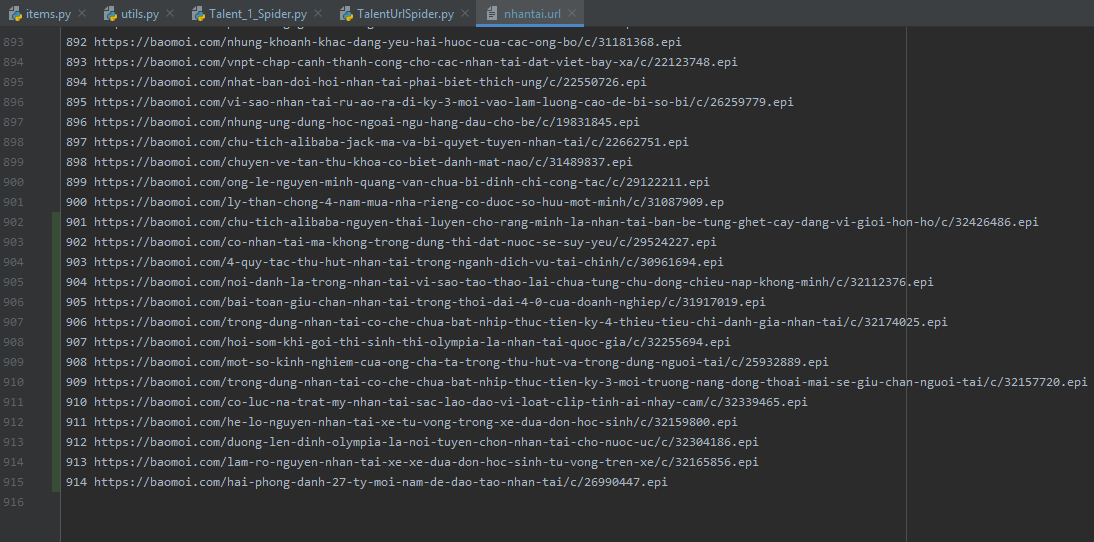
1. Crawl các bài báo trên intertet dựa trên 1 số query trên search engine như: "thu hút nhân tài", "đào tạo nhân tài", "giữ chân nhân tài", "nguồn lực chất lượng cao"

Công nghệ sử dụng: Framework Scrapy.

Các bước thực hiện:

* Crawl các link url dựa theo các từ khóa dựa trên Search Engine của Google..
* Lọc các url đã tồn tại và lưu vào file để làm đầu vào cho crawl nội dung bài báo.

Kết quả thực hiện:



Các url mới sẽ được lưu vào 4 file trong thư mục url trên Git: nguoitai.url, nhantai.url, nguonluc.url, nhanlucchatluong.url.

1. Lấy các dữ liệu thô từ các link bài báo: title, headline (abstract), content, source, published\_date, images và lưu vào Elastic Search ở local

Các bước thực hiện:

* Lấy 1 số URL cố định từ mỗi file trong thư mục URL. Mặc định: 50url / 1 File
* Thực hiện crawl và lưu tại Localhost.
* Lưu dữ liệu vào elasticsearch index dùng md5 hash của url làm id, thiết lập mapping các document

Kết quả thực hiện:

Dữ liệu crawl Demo:

{

"\_index": "newspapers",

"\_type": "talent",

"\_id": "efaa2c07a461eefc6f6d36d58d1960b7",

"\_version": 1,

"found": true,

"\_source": {

"url": "https://baomoi.com/bi-quyet-gin-giu-nhan-tai-o-nhung-nuoc-dang-phat-trien/c/4723132.epi",

"headline": "bí quyết gìn giữ nhân tài ở những nước đang phát triển",

"title": "(vtc news) - “chảy máu chất xám” vẫn luôn là một vấn đề nhức nhối đối với các quốc gia đang phát triển, khi mà hàng loạt nhân tài và những nhân công được rèn luyện có kỹ năng cao ở những nước nghèo đua nhau đi tìm “miền đất hứa” ở những cường quốc phát triển hơn.",

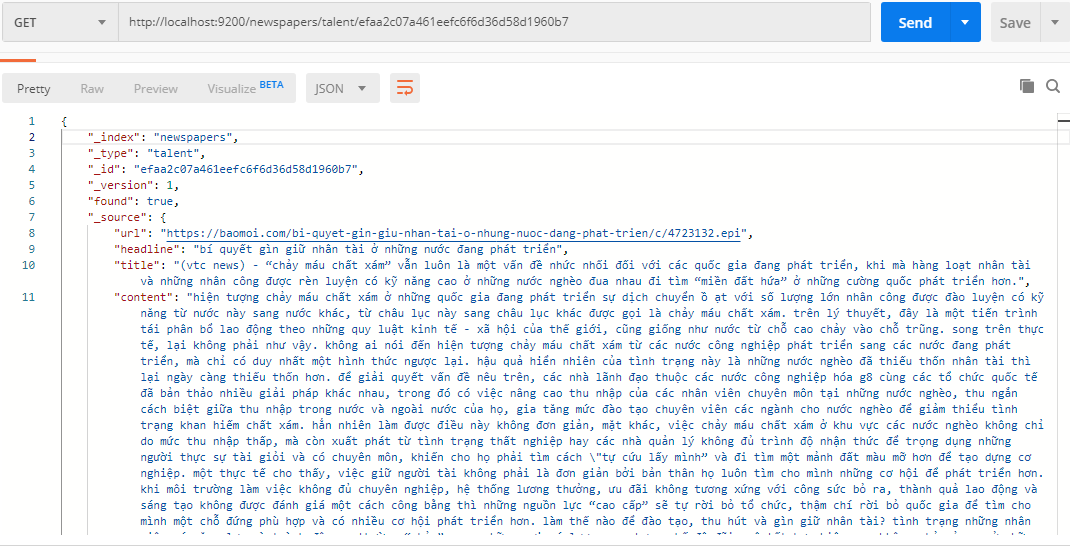
"content": "hiện tượng chảy máu chất xám ở những quốc gia đang phát triển sự dịch chuyển ồ ạt với số lượng lớn nhân công được đào luyện có kỹ năng từ nước này sang nước khác, từ châu lục này sang châu lục khác được gọi là chảy máu chất xám. trên lý thuyết, đây là một tiến trình tái phân bổ lao động theo những quy luật kinh tế - xã hội của thế giới, cũng giống như nước từ chỗ cao chảy vào chỗ trũng. song trên thực tế, lại không phải như vậy. không ai nói đến hiện tượng chảy máu chất xám từ các nước công nghiệp phát triển sang các nước đang phát triển, mà chỉ có duy nhất một hình thức ngược lại. hậu quả hiển nhiên của tình trạng này là những nước nghèo đã thiếu thốn nhân tài thì lại ngày càng thiếu thốn hơn. để giải quyết vấn đề nêu trên, các nhà lãnh đạo thuộc các nước công nghiệp hóa g8 cùng các tổ chức quốc tế đã bản thảo nhiều giải pháp khác nhau, trong đó có việc nâng cao thu nhập của các nhân viên chuyên môn tại những nước nghèo, thu ngắn cách biệt giữa thu nhập trong nước và ngoài nước của họ, gia tăng mức đào tạo chuyên viên các ngành cho nước nghèo để giảm thiểu tình trạng khan hiếm chất xám. hẳn nhiên làm được điều này không đơn giản, mặt khác, việc chảy máu chất xám ở khu vực các nước nghèo không chỉ do mức thu nhập thấp, mà còn xuất phát từ tình trạng thất nghiệp hay các nhà quản lý không đủ trình độ nhận thức để trọng dụng những người thực sự tài giỏi và có chuyên môn, khiến cho họ phải tìm cách \"tự cứu lấy mình” và đi tìm một mảnh đất màu mỡ hơn để tạo dựng cơ nghiệp. một thực tế cho thấy, việc giữ người tài không phải là đơn giản bởi bản thân họ luôn tìm cho mình những cơ hội để phát triển hơn. khi môi trường làm việc không đủ chuyên nghiệp, hệ thống lương thưởng, ưu đãi không tương xứng với công sức bỏ ra, thành quả lao động và sáng tạo không được đánh giá một cách công bằng thì những nguồn lực “cao cấp” sẽ tự rời bỏ tổ chức, thậm chí rời bỏ quốc gia để tìm cho mình một chỗ đứng phù hợp và có nhiều cơ hội phát triển hơn. làm thế nào để đào tạo, thu hút và gìn giữ nhân tài? tình trạng những nhân viên có năng lực và trình độ cao thường “nhảy” sang những nơi có lương cao hơn, chế độ đãi ngộ tốt hơn hiện nay không chỉ xảy ra ở những ngành, lĩnh vực thiếu nhân lực như ngân hàng, chứng khoán, điện lực... mà đang trở thành “vấn nạn”, làm đau đầu hầu hết các doanh nghiệp, đặc biệt là doanh nghiệp nhà nước. đây là một hiện tượng bình thường và ngày càng xảy ra thường xuyên hơn trong thị trường lao động. nền kinh tế càng phát triển, hiện tượng này càng phổ biến và thực ra là một điều kiện không thể thiếu của sự phát triển lành mạnh. hiện tượng này tất yếu sẽ xảy ra đối với các doanh nghiệp không đủ bản lĩnh và chiến lược để giữ những nhân tài của mình. đội ngũ lao động tri thức là những người có tính năng động rất cao. họ làm gì, ở đâu, làm thế nào, phần lớn là do quyết định của họ. người sử dụng lao động cần họ hơn là ngược lại. theo nghiên cứu của viện quản lý nguồn nhân lực quốc gia singapore (shri), một đất nước sống chủ yếu bằng nền kinh tế tri thức và là một trong những nơi có tỉ lệ nhân sự cao cấp (bao gồm các thạc sỹ, tiến sỹ, giáo sư, chuyên gia giỏi trên nhiều lĩnh vực) cao nhất thế giới, thì việc đào tạo, thu hút và gìn giữ nhân tài chính là một trong những “chìa khóa thành công” quan trọng nhất đưa singapore sớm trở thành một “con rồng châu á”. các chuyên gia quản lý nhân sự của shri cho biết, có 3 bí quyết gìn giữ nhân tài cơ bản mà đất nước này từng áp dụng thành công. bí quyết 1: giữ chân người giỏi là chiến lược, không phải là biện pháp đối phó nhất thời. vì vậy, chiến lược giữ người giỏi phải tiến hành song song 4 yếu tố: thu hút, tuyển dụng, hội nhập và cộng tác. bí quyết 2: cần có những tiêu chí định tính và định lượng giúp doanh nghiệp nhận diện ra nhân viên giỏi cần giữ. đó là những nhân viên luôn hoàn thành xuất sắc công việc, đảm trách những công việc đòi hỏi kỹ năng, kiến thức hiếm trên thị trường lao động, tâm huyết với sự phát triển của doanh nghiệp … bí quyết 3: doanh nghiệp phải tìm ra được những yếu tố quyết định để giữ nhân viên giỏi. nếu muốn nhân viên giỏi hài lòng và ở lại lâu dài với doanh nghiệp cần giảm yếu tố bất mãn bao gồm lương bổng và chế độ làm việc, yếu tố tạo nguồn, bao gồm thu hút và tuyển dụng… hoài thư (tổng hợp)",

"source": "/bi-quyet-gin-giu-nhan-tai-o-nhung-nuoc-dang-phat-trien/r/4723132.epi",

"published\_date": "18/08/10 06:36"

}

}



1. Lưu dữ liệu vào elasticsearch index (dùng md5 hash của url làm id). Cần học cách index, query, thiết lập mapping các document trên nền tảng elasticsearch để phục vụ việc dễ dàng truy query, aggregate data.

Mapping của index:

{

"newspapers": {

"mappings": {

"talent": {

"properties": {

"content": {

"type": "text"

},

"headline": {

"type": "text"

},

"images": {

"type": "text",

"fields": {

"keyword": {

"type": "keyword",

"ignore\_above": 256

}

}

},

"published\_date": {

"type": "date",

"format": "dd/MM/yy HH:mm"

},

"source": {

"type": "keyword"

},

"title": {

"type": "text"

},

"url": {

"type": "keyword"

}

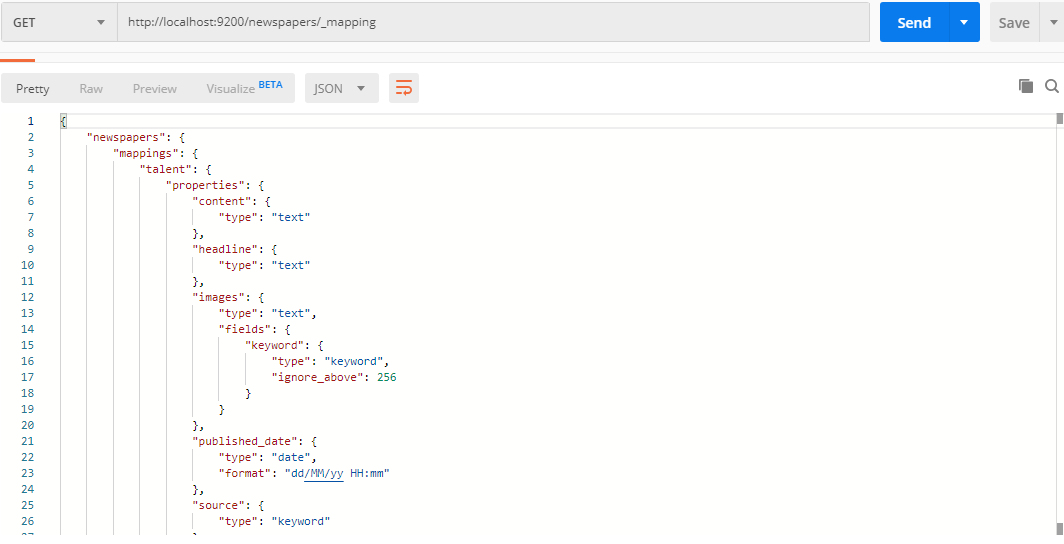
}

}

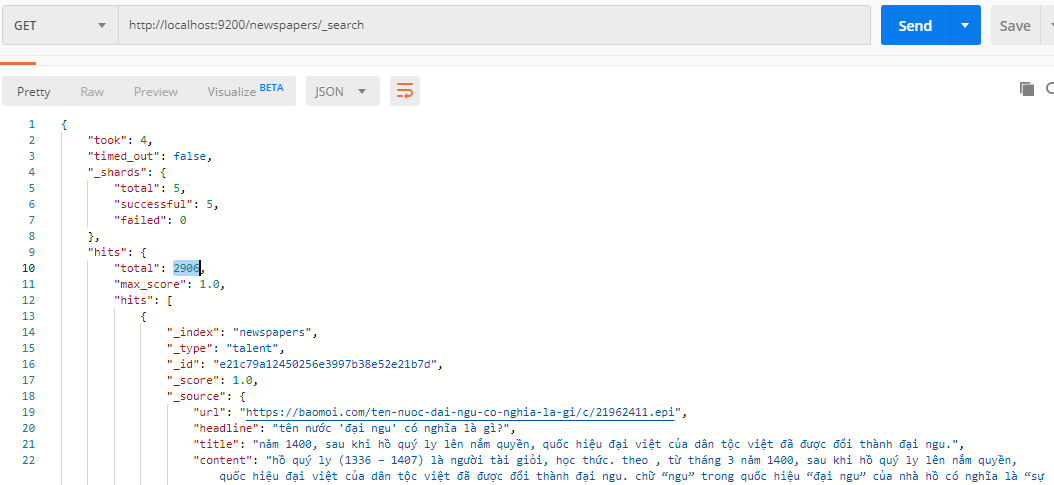
}

}

}



Tổng số lượng dữ liệu đã crawl : 2906 documents.

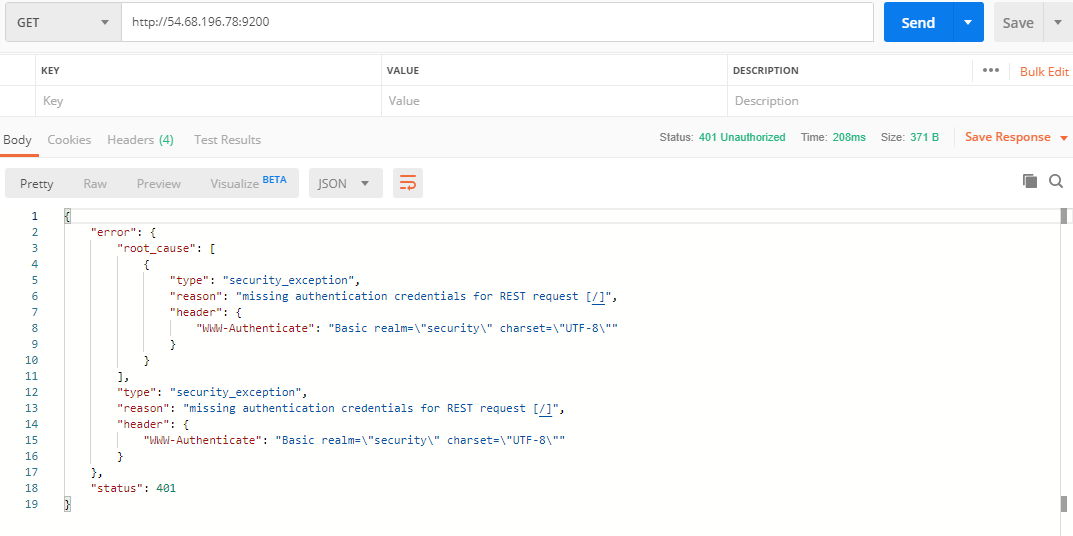


1. **Công việc đang thực hiện**

Bước 1:

Lưu dữ liệu lên Server:

Khó khăn: Lỗi Authentication



Bước 2:

Xây dựng các modul xử lý dữ liệu. Các bài báo "thô" sẽ được xử lý qua 1 data pipleline gồm các bước:

1. Phát hiện bài báo không liên quan (Filter Processor): nếu bài báo không liên quan gì đến vấn đề "nhân tài", "nhân lực", quá ít nội dung (video) thì lọc bỏ.

Hướng thực hiện:

1. Tìm hiểu Spacy.
2. Tìm hiểu thuật toán.

Khó khăn: tìm hiểu thuật toán hiệu quả để áp dụng