

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

————— * —————



BÁO CÁO MÔN: TÌM KIẾM VÀ TRÌNH DIỄN
THÔNG TIN

ĐỀ TÀI: Xây Dựng Web Search Dựa Trên Nền Tảng
Apache Solr

GVHD: TS. Nguyễn Bá Ngọc

Nhóm sinh viên thực hiện:

Nguyễn Huy Phát	20143397
Trần Gia Nghĩa	20143180
Vũ Ngọc Hoàn	20141734
Chu Quang Tỉnh	20144517

Hà Nội ngày 18 tháng 05 năm 2018

Mục lục

Mô tả bài toán	3
1. Giới thiệu chung	3
2. Giới thiệu về Apache Solr	3
I. Triển Khai Hệ Thống	6
1. Thu thập dữ liệu:	6
2. Index cho dữ liệu sử dụng Apache Solr	7
2.1 Cài đặt Apache Solr:	7
2.2 Index dữ liệu theo cấu hình mặc định của solr:	9
2.3 Index dữ liệu theo cấu hình tùy chỉnh cho ngôn ngữ tiếng Việt:	16
2.4 Index dữ liệu sử dụng filter + stopword cho tiếng Việt:	16
3. Xây dựng Web Search	21
3.1 Back-end:	22
3.2 Front-end	22
3.3 Một số tính năng gợi ý cho người dùng:	22
4. Demo kết quả xây dựng hệ thống	23
5. Đánh giá kết quả tìm kiếm của hệ thống	32
6. Nhận xét và kết luận	34
7. Tài liệu tham khảo:	35

Mô tả bài toán

1. Giới thiệu chung

- Hiện nay lượng dữ liệu trên internet rất lớn, tổng số website ở mức 760 triệu (2013), tổng dữ liệu trên internet khoảng 670 tỷ GB. Trong khi đó lượng dữ liệu trùng lặp, dữ liệu rác cũng khá nhiều.
- Nhu cầu thông tin của con người này càng cao, nhu cầu về thông tin có chất lượng, không rác, không trùng lặp càng lớn. Đối mặt với những thách thức về lượng dữ liệu khổng lồ trên internet đó, cần phải có 1 hệ thống tìm kiếm để giúp người dùng có thể nhanh chóng tiếp cận với dữ liệu mình cần và dữ liệu cần phải có chất lượng cao: không rác, không trùng lặp.
- Các trang thương mại điện tử ngày càng phát triển với lượng dữ liệu vô cùng phong phú, tuy nhiên nó lại làm cho người dùng khó tìm kiếm lựa chọn sản phẩm phù hợp với mình
- Với một ứng dụng triển khai trên nền web site việc tìm kiếm trực tiếp vào cơ sở dữ liệu sẽ rất tốn kém thời gian, đặc biệt là khi dữ liệu tăng lên
- Có nhiều nền tảng opensource giúp triển khai 1 máy tìm kiếm như : Apache Solr, Elastic Search ...
- Trước nhu cầu đó thì nhóm em quyết định lựa chọn Apache Solr để xây dựng một trang web tìm kiếm giúp tìm kiếm:
 - Nội dung tìm kiếm : Tin tức văn bản hằng ngày
 - Ngôn ngữ tìm kiếm : Tiếng Việt.(có dấu hoặc không dấu)
 - Nguồn dữ liệu : Các website tin tức
 - Giao diện tương tác người sử dụng: Giao diện web.

2. Giới thiệu về Apache Solr

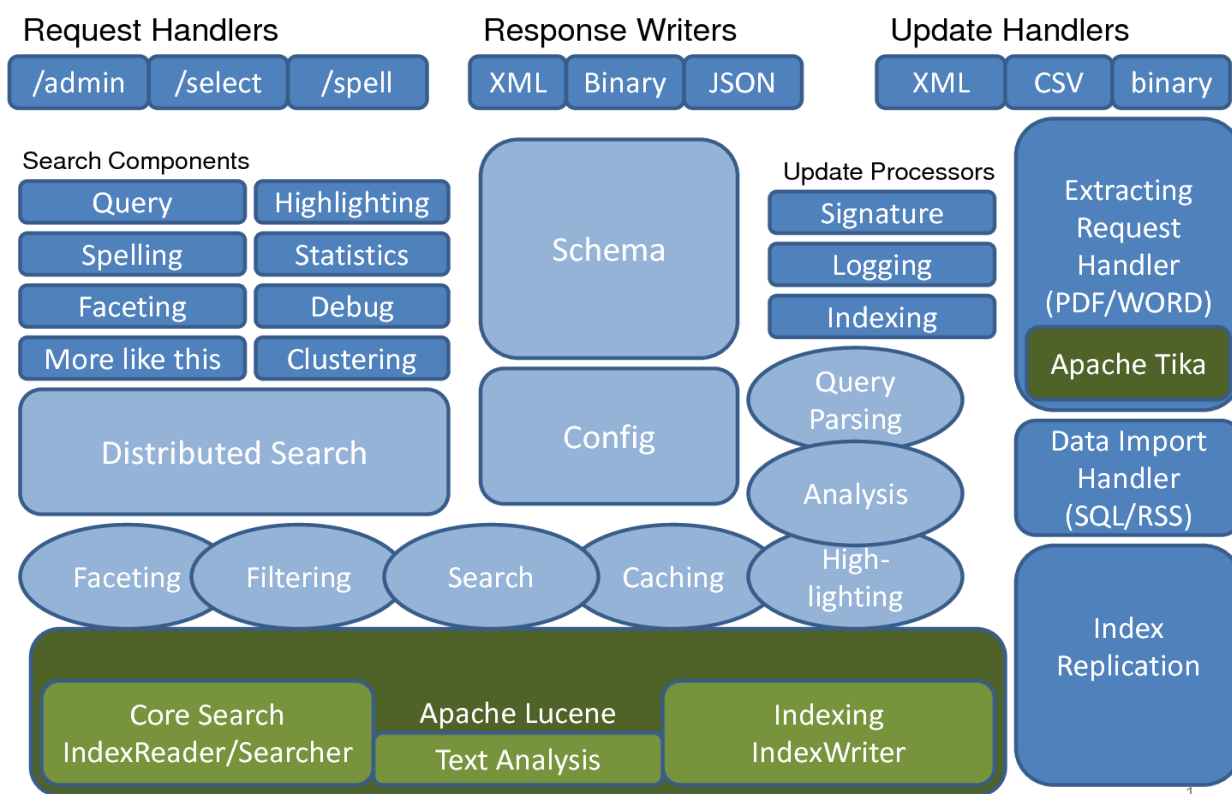
- Apache Solr là một open source full-text search platform dựa trên Apache Lucene. Lucene là một thư viện được viết bằng Java dùng để phân tích, đánh chỉ mục (indexing) và tìm kiếm thông tin được phát triển đầu tiên bởi Doug Cutting vào năm 2000. Cutting đồng thời cũng là tác giả của Hadoop lúc ông đang làm việc cho Yahoo vào năm 2005.

- Solr chạy như một máy chủ tìm full-text search độc lập. Nó sử dụng lõi thư viện tìm kiếm Lucene Java cho việc đánh chỉ mục full-text và tìm kiếm, và có REST (REpresentational State Transfer) như API HTTP/XML và JSON khiến nó có thể sử dụng bởi những ngôn ngữ lập trình phổ biến nhất. Cấu hình mở rộng của Solr cho phép nó có thể được thiết kế với nhiều loại ứng dụng mà không cần code Java, và nó có một kiến trúc plugin hỗ trợ chỉnh sửa nâng cao hơn.
- **Tổng quát thì Solr bao gồm có nhiều thành phần (components) khác nhau bao gồm:**
 - Apache Lucene để phân tích, đánh chỉ mục tìm kiếm dữ liệu.
 - Apache Tika dùng để trích xuất metadata, tìm kiếm và chỉ mục nhiều loại file document khác nhau như pdf, docx, mp3, jpg (hỗ trợ 66 file types khác nhau).
 - Apache UIMA (đọc như you-eee-mah, Unstructured Information Managemen Architecture), đây cũng là một project thuộc Apache Foudnation, nó được dùng để phân tích một lượng lớn dữ liệu không có cấu trúc nhằm tìm ra được những thông tin có ích cho người dùng. Ví dụ:
 - Phân tích các phim và trích xuất phụ đề rồi dựa vào đó để tìm ra diễn viên nào đóng trong phim đó.
 - Tìm các bài viết, video, hình ảnh có liên quan tới chủ đề của một bài viết cụ thể nào đó.
 - Apache Velocity là một Java-based template engine.
 - Carrot2 (search results clustering engine) dùng để phân loại và nhóm các kết quả tìm kiếm thành những danh mục có cùng chủ đề (thematic categories).
- **Một số đặc trưng của Apache Solr:**
 - Khả năng tìm kiếm văn bản toàn diện(Full-Text Search) giống kiểu Google.
 - Chỉnh sửa để hiệu năng tốt hơn.
 - Dựa trên các chuẩn mở trong giao tiếp với các hệ thống khác – XML, JSON và HTTP
 - Quản trị dưới dạng giao diện HTML đơn giản
 - Thống kê dưới dạng JMX
 - Cấu hình đơn giản dễ dàng với định dạng XML
 - Có khả năng bổ sung các phần mở rộng(plugin) mới. Ví dụ như phân tích mở rộng tiếng Việt: Bắt lỗi chính tả, bỏ dấu,...

- Cho phép highlighting kết quả tìm kiếm, như cách mà google hiện thị thông tin tóm tắt về kết quả mà ở đó câu truy vấn được in đậm
- Có thể xây dựng rất nhiều ứng dụng khác mà một trang tìm kiếm cần như: autosuggestion, spellchecking, xây dựng tagcloud, phân loại kết quả clustering (như Bing làm), trending keywords, category navigation, các kết quả liên quan, nhóm kết quả (field collapsed) ...
- Cho phép scale hệ thống một cách dễ dàng khi bạn có một lượng lớn dữ liệu mà không đủ chứa trên một máy chủ hay phải phục vụ rất nhiều người dùng đồng thời.
- Solr cũng có thể dùng như CSDL NoSQL hay như cache layer, dùng cho các listing cần performance tốt.
- Solr cũng sắp hỗ trợ realtime cho phép tìm kiếm ngay kết quả sau khi index. Điều này đặc biệt khó khi index rất lớn. Hiện tại Solr cho phép kết quả rất nhanh, nhưng phải hy sinh thời gian index. Với dữ liệu lớn có khi bạn phải mất 30 phút chỉ để cập nhật được một tài liệu.
- Solr hỗ trợ rất nhiều công cụ để tinh chỉnh kết quả tìm kiếm, bằng tất cả các thông tin mà bạn cung cấp làm sao để kết quả trả về là tốt nhất. Ví dụ như đánh trọng số các trường, click log, số lượt view, ...

- Kiến trúc

Lucene/Solr Architecture



Hình:

- Chức năng:
 - Phân tích, đánh chỉ mục dữ liệu tìm kiếm
 - Trích xuất metadata, tìm kiếm và chỉ mục nhiều loại tài liệu khác nhau : pdf, docx, mp3, jpg...
 - Hỗ trợ trên tìm kiếm trên ngôn ngữ: Anh, Pháp, Thái Lan....
 - Khả năng mở rộng, tùy biến cao.

I. Triển Khai Hệ Thống

1. Thu thập dữ liệu:

- Nguồn thu thập: các bài báo trong trang <http://dantri.com.vn>
- Thư viện hỗ trợ : Jsoup của Java
- Làm sạch dữ liệu: Loại bỏ các ký tự đặc biệt, các thẻ html....
- Các trường thu thập :

- Title : Tiêu đề của tin tức.
 - Content : Nội dung của tin tức.
 - URL: Đường dẫn tới trang tin tức gốc.
 - Author : Tác giả bài viết
 - Time : Thời gian bài viết được đăng
- Đầu ra : file json gồm : 7700 tin tức. Thuộc các chủ đề : Ôtô-xe máy, giáo dục, thể giới, thể thao, sức mạnh số.

```

1 [
2   {
3     "Origin": "Nhom 4",
4     "Content": "Evan Blass, người nổi tiếng với việc đăng tải thông tin bị rò rỉ chính xác về các sản phẩm công nghệ sắp ra mắt, vừa đăng",
5     "Author": "T.Thùy",
6     "Title": "Lộ ảnh chính thức và cấu hình chi tiết smartphone "bom tấn" 4 camera của HTC",
7     "Time": "Thứ bảy, 19/05/2018 - 07:30",
8     "Url": "http://dantri.com.vn/suc-manh-so/lo-anh-chinh-thuc-va-cau-hinh-chi-tiet-smartphone-bom-tan-4-camera-cua-htc-2018051907292889",
9   },
10  {
11    "Origin": "Nhom 4",
12    "Content": "Ra mắt chưa đầy 1 tháng, phiên bản iPhone 8 và 8 Plus màu đỏ mới nhất của Apple đang dần mất sức hút tại Việt Nam. Đến",
13    "Author": "Già Hưng",
14    "Title": "iPhone 8 đỏ ế ẩm ở Việt Nam, người Việt thích iPhone X hơn",
15    "Time": "Thứ bảy, 19/05/2018 - 23:27",
16    "Url": "http://dantri.com.vn/suc-manh-so/iphone-8-do-e-am-o-viet-nam-nguoi-viet-thich-iphone-x-hon-20180519232540423.htm",
17  },
18  {
19    "Origin": "Nhom 4",
20    "Content": "Mới đây, tổ chức dân sự tự do mang tên Big Brother Watch (BBW) tại Anh đã công bố những phát hiện mới nhất của mình về",
21    "Author": "Nguyễn Nguyễn",
22    "Title": "Hệ thống nhận diện khuôn mặt ở Anh cho kết quả sai đến 98%",
23    "Time": "Thứ bảy, 19/05/2018 - 07:55",
24    "Url": "http://dantri.com.vn/suc-manh-so/he-thong-nhan-dien-khuon-mat-o-anh-cho-ket-qua-sai-den-98-20180519075308789.htm",
25  },
26  {
27    "Origin": "Nhom 4",
28    "Content": "Caviar, hãng sản xuất phụ kiện dành cho smartphone của Nga vừa trình làng tấm ốp lưng bảo vệ tích hợp pin năng lượng mặt",
29    "Author": "T.Thùy",
30    "Title": "Tấm ốp lưng giá 4.500 USD dành cho iPhone X có gì đặc biệt?",
31    "Time": "Thứ bảy, 19/05/2018 - 07:37",
32    "Url": "http://dantri.com.vn/suc-manh-so/tam-op-lung-gia-4500-usd-danh-cho-iphone-x-co-gi-dac-biet-20180519073344872.htm",
33  },
34  {
35    "Origin": "Nhom 4",
36    "Content": "Honor 10 là mẫu smartphone mới nhất của hãng này có sự thay đổi đáng kể trong thiết kế và camera. Máy được trang bị hệ

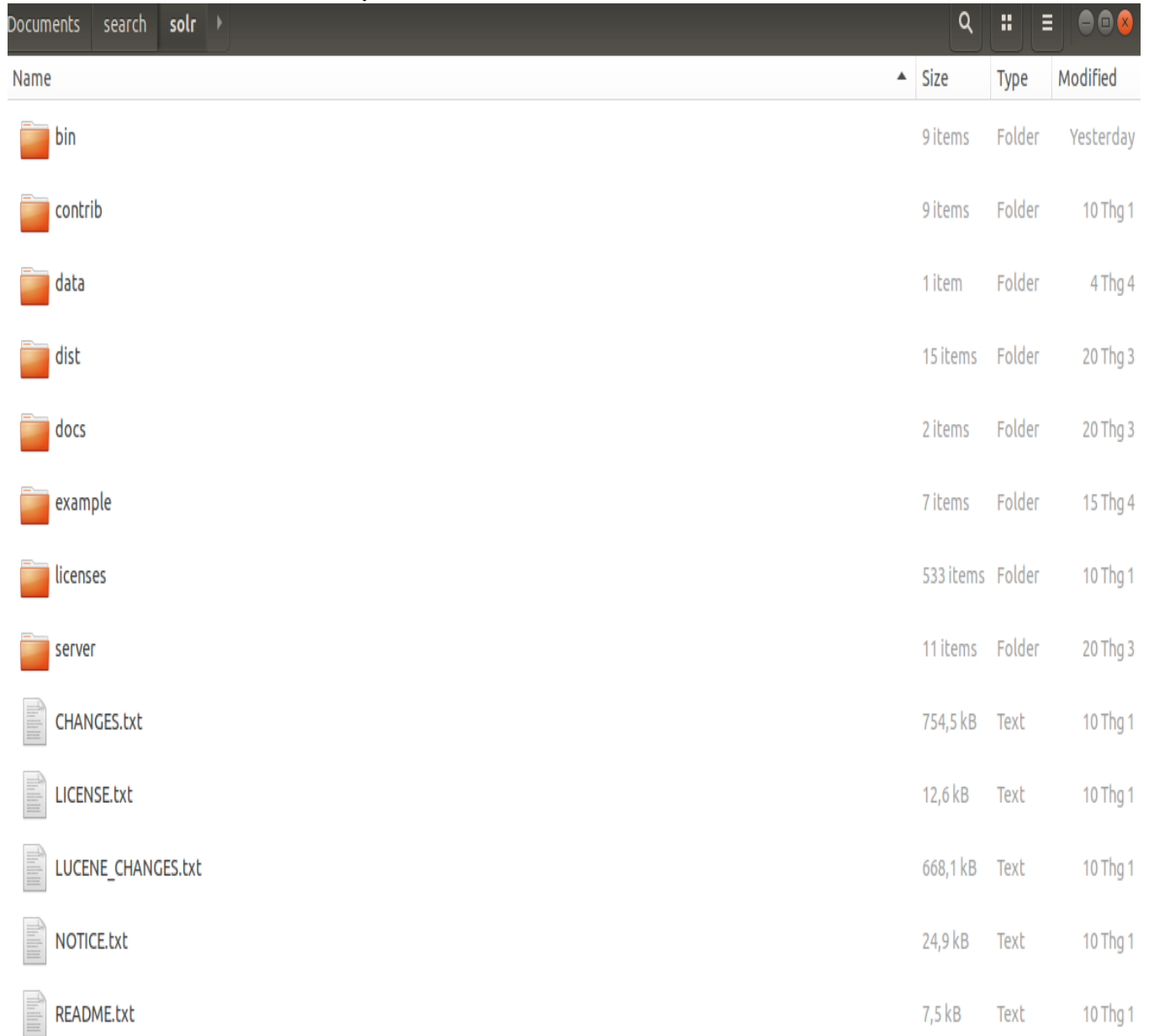
```

Hình:

2. Index cho dữ liệu sử dụng Apache Solr

2.1 Cài đặt Apache Solr:

- Môi trường cài đặt: Ubuntu 18.04, Java 1.9 (bước phải cài đặt môi trường java)
- Tải Solr tại : <http://archive.apache.org/dist/lucene/solr/7.2.1/>
- Cấu trúc thư mục của solr :

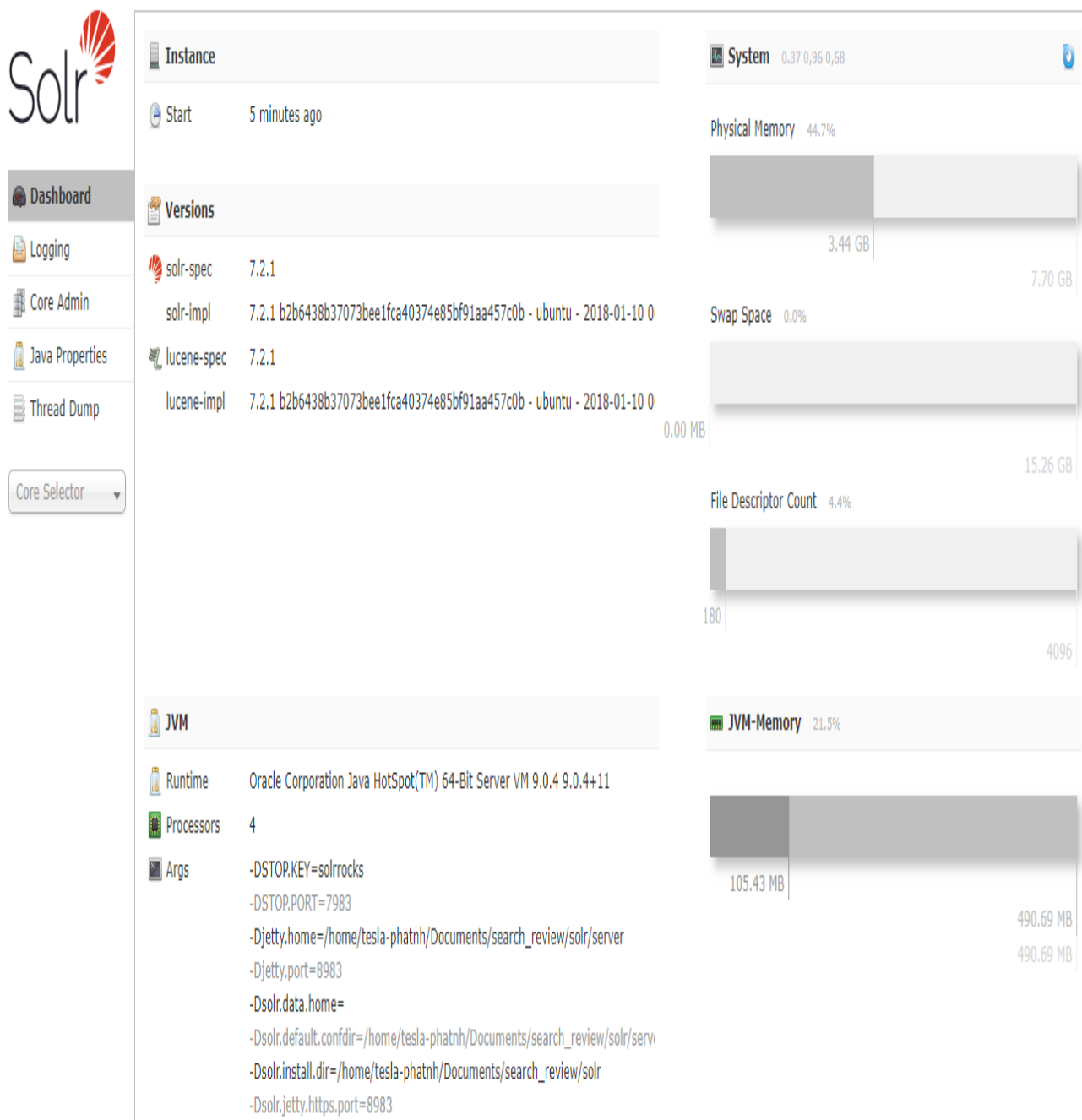


Name	Size	Type	Modified
bin	9 items	Folder	Yesterday
contrib	9 items	Folder	10 Thg 1
data	1 item	Folder	4 Thg 4
dist	15 items	Folder	20 Thg 3
docs	2 items	Folder	20 Thg 3
example	7 items	Folder	15 Thg 4
licenses	533 items	Folder	10 Thg 1
server	11 items	Folder	20 Thg 3
CHANGES.txt	754,5 kB	Text	10 Thg 1
LICENSE.txt	12,6 kB	Text	10 Thg 1
LUCENE_CHANGES.txt	668,1 kB	Text	10 Thg 1
NOTICE.txt	24,9 kB	Text	10 Thg 1
README.txt	7,5 kB	Text	10 Thg 1

Hình :

- Giải nén file tải về, trong thư mục chứa source code mở một Terminal chạy lệnh : **bin/solr start**

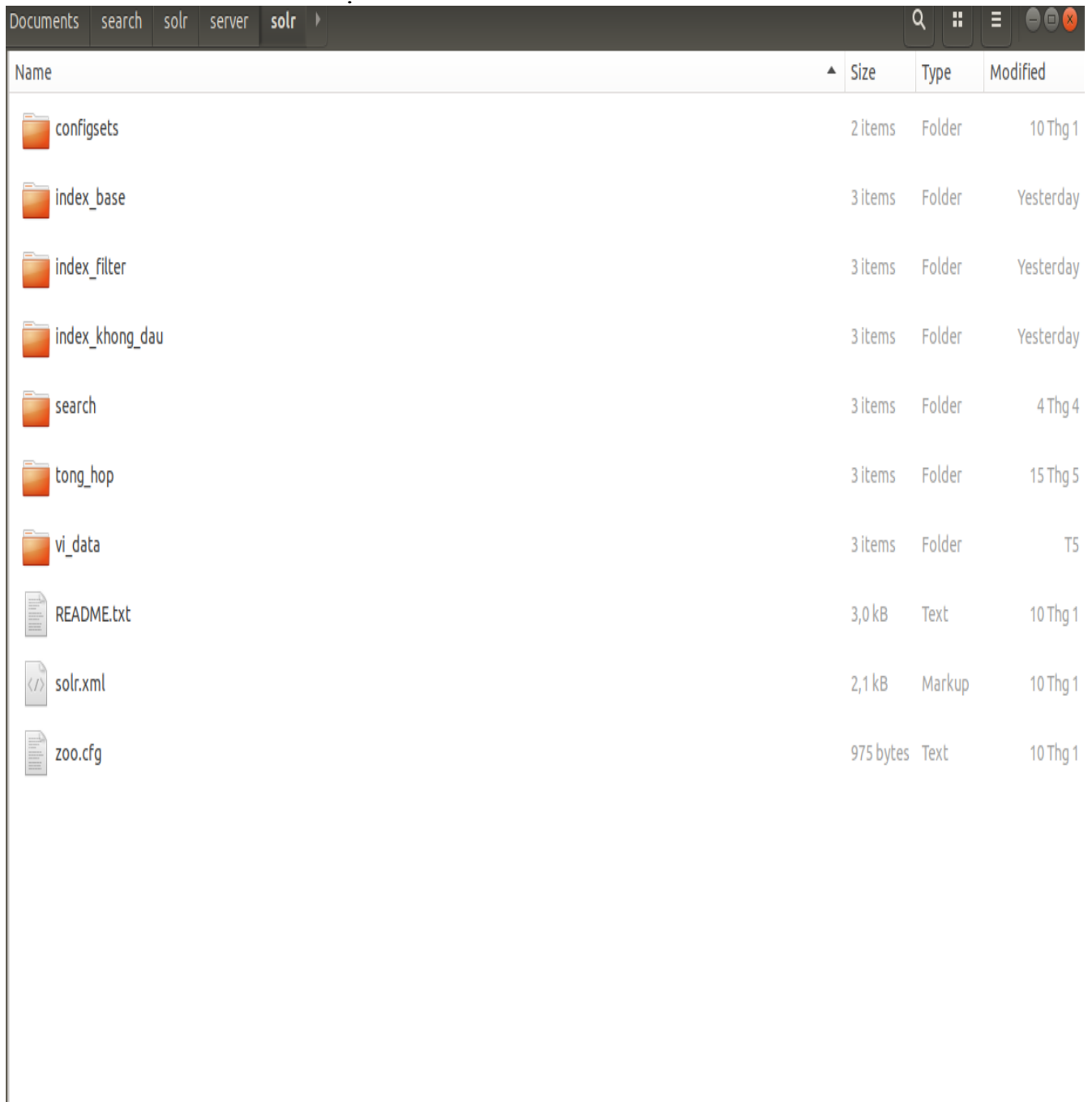
- Vào giao diện trang quản trị của solr :



Hình :




2.2 Index dữ liệu theo cấu hình mặc định của solr:

- Tạo một core mới:
bin/solr create -c core_name
- Index dữ liệu:
bin/post -c path_data/data_name.json
- Cấu trúc thư mục chứa các core:











Hình:

- Cấu trúc một core :

Documents search solr server solr index_base conf ▶				Q	≡	≡	⌵	⌵	✖
Name ▲			Size	Type	Modified				
	conf		8 items	Folder	18 Thg 4				
	data		3 items	Folder	Yesterday				
	core.properties		80 bytes	Text	Yesterday				

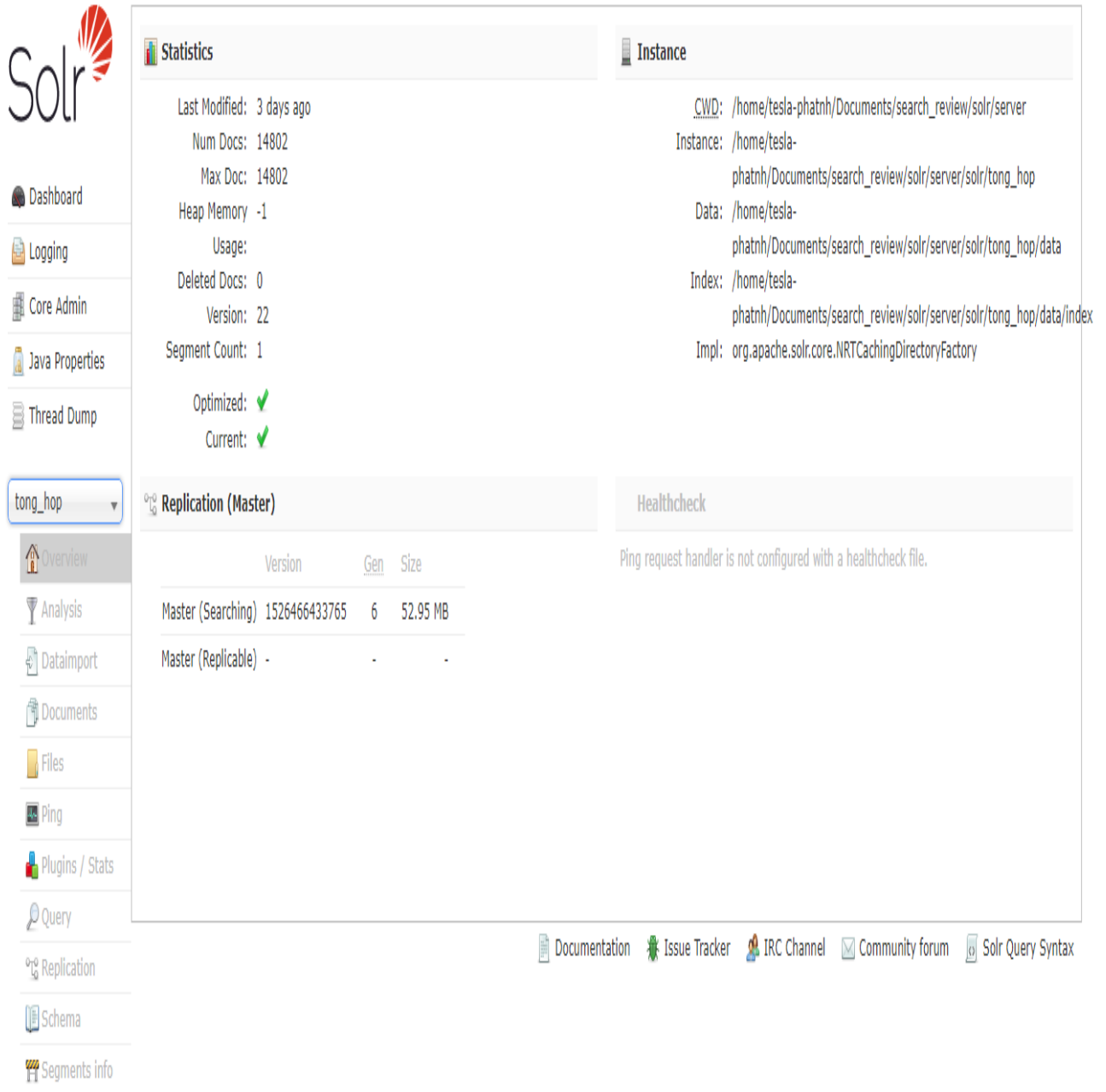
Hình :

- Thư mục cấu hình một core :

Documents	search	solr	server	solr	index_base	conf		Q	⌵	☰	⌵	⌵	⌵	⌵
Name	Size	Type	Modified											
 lang	38 items	Folder	10 Thg 1											
 managed-schema	28,8 kB	Markup	Yesterday											
 managed-schema_backup	50,7 kB	Markup	10 Thg 1											
 params.json	308 bytes	Program	10 Thg 1											
 protwords.txt	873 bytes	Text	10 Thg 1											
 solrconfig.xml	54,0 kB	Markup	10 Thg 1											
 stopwords.txt	781 bytes	Text	10 Thg 1											
 synonyms.txt	1,1 kB	Text	10 Thg 1											

Hình :

- Vào trang quản trị core mới tạo:



The screenshot displays the Apache Solr Admin UI. On the left is a sidebar with navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu currently showing 'tong_hop'. Below the dropdown are links for Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query, Replication, Schema, and Segments info.

The main content area is divided into two columns:

- Statistics:**
 - Last Modified: 3 days ago
 - Num Docs: 14802
 - Max Doc: 14802
 - Heap Memory: -1
 - Usage:
 - Deleted Docs: 0
 - Version: 22
 - Segment Count: 1
 - Optimized: ✔
 - Current: ✔
- Instance:**
 - CWD: /home/tesla-phatnh/Documents/search_review/solr/server
 - Instance: /home/tesla-phatnh/Documents/search_review/solr/server/solr/tong_hop
 - Data: /home/tesla-phatnh/Documents/search_review/solr/server/solr/tong_hop/data
 - Index: /home/tesla-phatnh/Documents/search_review/solr/server/solr/tong_hop/data/index
 - Impl: org.apache.solr.core.NRTCachingDirectoryFactory

Below these columns are two more sections:


- Replication (Master):** A table showing replication status.

	Version	Gen	Size
Master (Searching)	1526466433765	6	52.95 MB
Master (Replicable)	-	-	-
- Healthcheck:** A message stating "Ping request handler is not configured with a healthcheck file."

At the bottom of the page, there are links for Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

Hình:

- Tiến hành tìm kiếm thử trên giao diện quản lý của solr:



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

tong_hop

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema

Segments info

Request-Handler (qt)

/select

common

q

Kinh tế thế giới

fq

sort

start, rows

010

fl

df

Content

Raw Query Parameters

key1=val1&key2=val2

wt

.....

☐ indent off
☐ debugQuery
☐ dismax
☐ edismax

http://192.168.1.17:8983/solr/tong_hop/select?df=Content&q=Kinh tế thế giới

```

{
  "responseHeader":{
    "status":0,
    "QTime":8,
    "params":{
      "q":"Kinh tế thế giới",
      "df":"Content",
      "_":"1526736023954"}},
  "response":{"numFound":12732,"start":0,"docs":[
    {
      "Origin":["Nhóm 4"],
      "Content":["Bloomberg ngày 26/12 dẫn báo cáo mới nhất của Trung tâm nghiên cứu Kinh tế và Kinh doanh (CEBR) tại London, A
      "Author":["Thành Đạt"],
      "Title":["Trung Quốc sẽ \"vượt mặt\" Mỹ trong 15 năm tới?\"],
      "Time":["Thứ ba, 26/12/2017 - 22:12\"],
      "Url":["http://dantri.com.vn/the-gioi/trung-quoc-se-vuot-mat-my-trong-15-nam-toi-201712262211358.htm\"],
      "id":["5a2e123a-3078-4815-acd9-551306f92cda\"],
      "Time_str":["Thứ ba, 26/12/2017 - 22:12\"],
      "_version_":["1600616065883176961\",
      "Author_str":["Thành Đạt"]}],
    {
      "Origin":["Nhóm 4"],
      "Author":["Nguyễn Nhân"],
      "Title":["Bức tranh kinh tế toàn cầu 2018: Lạc quan với nhiều gam màu sáng\"],
      "Time":["Thứ sáu, 16/02/2018 - 20:07\"],
      "Url":["http://dantri.com.vn/the-gioi/buc-tranh-kinh-te-toan-cau-2018-lac-quan-voi-nhieu-gam-mau-sang-20180216200745893.ht
      "id":["29c37402-1654-4986-879f-ce0d4eb41b59\"],
      "Content":["Những \"đột phá\" tăng trưởng Năm 2018 được dự báo, kinh tế Mỹ khởi sắc với mức tăng trưởng 3% đến 3,1%, kinh t
      "Time_str":["Thứ sáu, 16/02/2018 - 20:07\"],
      "_version_":["1600616065625227264\",
      "Author_str":["Nguyễn Nhân"]}]
  ]}
}

```

Hình:

- Xem bộ Analyzer của solr:



Field Value (Index)

Kính tế thể giới có nhiều biến động

Field Value (Query)

kính tế

Dashboard

Logging

Core Admin

Java Properties

Thread Dump

base_index

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema

Segments info

Analyse Fieldname / FieldType: Content

Schema Browser

Verbose Output

Analyse Values

ST	text	Kính	tế	thể	giới	có	nhiều	biến	động
	raw_bytes	[4b 69 6e 68]	[74 e1 ba bf]	[74 68 e1 ba bf]	[67 69 e1 bb 9b 69]	[63 c3 b3]	[6e 68 69 e1 bb 81 75]	[62 69 e1 ba bf 6e]	[c4 91 e1 bb 99 6e 67]
	start	0	5	8	12	17	20	26	31
	end	4	7	11	16	19	25	30	35
	positionLength	1	1	1	1	1	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	termFrequency	1	1	1	1	1	1	1	1
	position	1	2	3	4	5	6	7	8
SF	text	Kính	tế	thể	giới	có	nhiều	biến	động
	raw_bytes	[4b 69 6e 68]	[74 e1 ba bf]	[74 68 e1 ba bf]	[67 69 e1 bb 9b 69]	[63 c3 b3]	[6e 68 69 e1 bb 81 75]	[62 69 e1 ba bf 6e]	[c4 91 e1 bb 99 6e 67]
	start	0	5	8	12	17	20	26	31
	end	4	7	11	16	19	25	30	35
	positionLength	1	1	1	1	1	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	termFrequency	1	1	1	1	1	1	1	1
	position	1	2	3	4	5	6	7	8
LCF	text	kính	tế	thể	giới	có	nhiều	biến	động
	raw_bytes	[6b 69 6e 68]	[74 e1 ba bf]	[74 68 e1 ba bf]	[67 69 e1 bb 9b 69]	[63 c3 b3]	[6e 68 69 e1 bb 81 75]	[62 69 e1 ba bf 6e]	[c4 91 e1 bb 99 6e 67]
	start	0	5	8	12	17	20	26	31
	end	4	7	11	16	19	25	30	35
	positionLength	1	1	1	1	1	1	1	1

Hình:

⇒ Nếu sử dụng cách index dữ liệu theo mặc định của solr thì dữ liệu sẽ được tắc bởi dấu: “ “ sau đó chúng được đánh index. Điều này sẽ có lợi cho việc tìm kiếm đầy đủ dấu. Tuy nhiên nếu người dùng nhập vào một tìm kiếm không đầy đủ dấu thì kết quả của nó sẽ không chính xác như mong muốn.

2.3 Index dữ liệu theo cấu hình tùy chỉnh cho ngôn ngữ tiếng Việt:

- Solr quản lý việc đánh index và query bằng file managed-schema
Để solr thực hiện việc bỏ dấu trước khi đánh index, làm những bước sau:
 - Xác định kiểu dữ liệu cho các trường cần bỏ dấu trước khi đánh index là kiểu text_general
 - Đối với kiểu dữ liệu text_general, quy định trước khi đánh index phải đi qua 1 class MappingCharFilterFactory, class này có nhiệm vụ map các ký tự sử dụng file mapping.txt.
- file mapping.txt có dạng như sau:

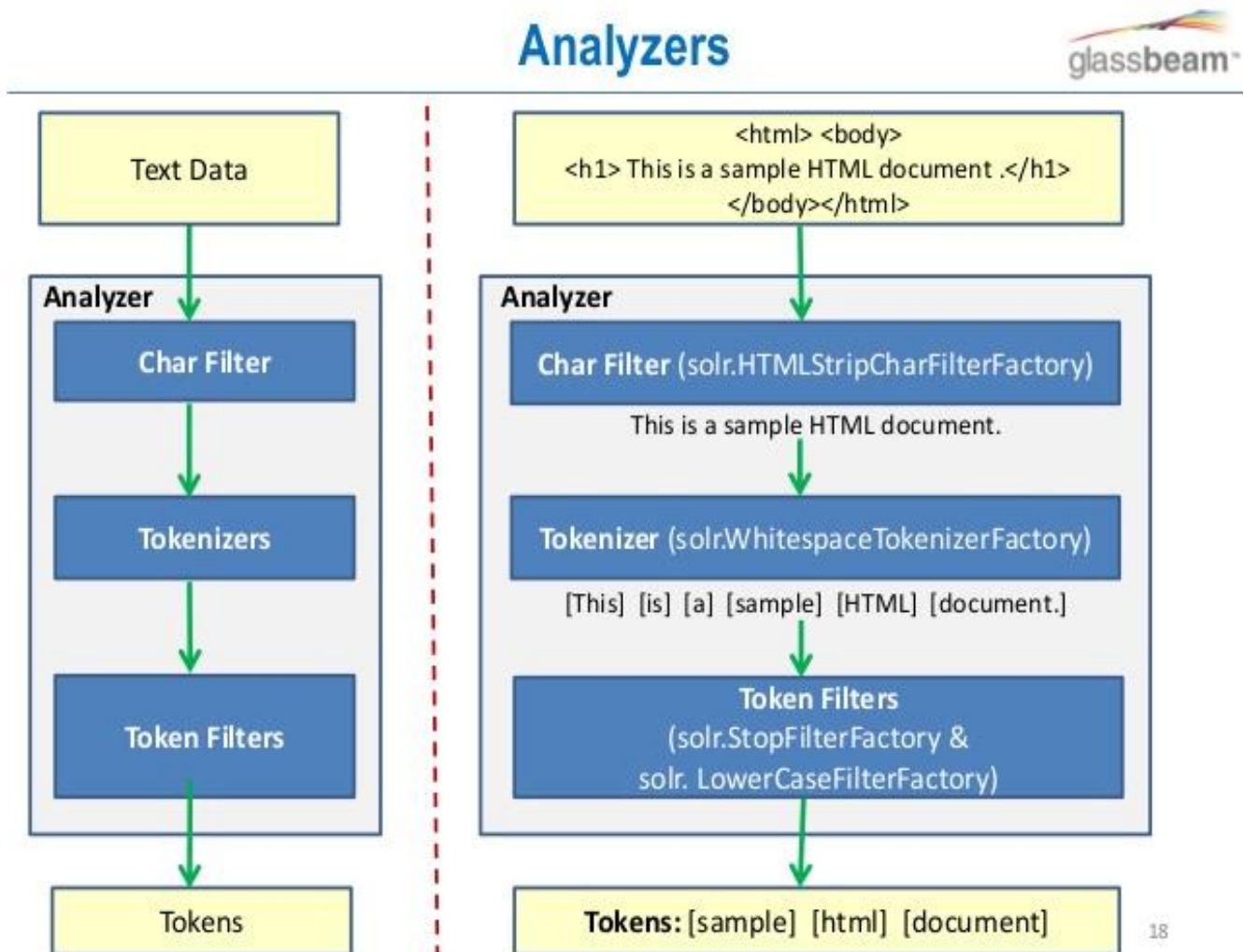
```
"À" => "A"  
"Á" => "A"  
"Â" => "A"  
"Ã" => "A"  
"È" => "E"  
"É" => "E"  
"Ê" => "E"  
"Ì" => "I"  
"Í" => "I"  
"Ò" => "O"  
"Ó" => "O"  
"Ô" => "O"  
"Õ" => "O"  
"Ù" => "U"  
"Ú" => "U"  
"Ý" => "Y"  
"à" => "a"  
"á" => "a"  
"â" => "a"  
"ã" => "a"  
"è" => "e"
```

.....

2.4 Index dữ liệu sử dụng filter + stopwords cho tiếng Việt:

- Solr hỗ trợ việc cấu hình dễ dàng việc thêm các filter và tokenizer nên chúng ta có thể xử lý các vấn đề của nhiều ngôn ngữ khác nhau.

- Analyzer : đóng vai trò khảo sát các trường văn bản để tạo ra một token stream.



Hình:

- Tokenizer: chia nhỏ các stream đó thành những tokens (đơn vị nhỏ nhất để index, có thể là từ hay ký tự). Các ký tự trong input stream có thể bị bỏ qua như các ký tự không nhìn thấy được (whitespace như khoảng trắng, tab) hay các dấu phân cách (delimiter như dấu phẩy, dấu chấm).
- Filter: đọc input stream và tạo ra các token, nhưng dữ liệu đầu vào sẽ được xử lý thêm (ví dụ chuyển chữ hoa thành chữ thường, chuyển từ viết tắt (tên bang, tên thành phố) thành kiểu viết đầy đủ hoặc ngược lại).

- Cấu hình lại filter + Stopword tiếng việt:

```
<!-- Thêm một fieldType để index cho tiếng việt -->
<fieldType name="text_general_vn" class="solr.TextField" positionIncrementGap="100" multiValued="true">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_vn.txt" ignoreCase="true"/>
    <filter class="solr.ASCIIFoldingFilterFactory" preserveOriginal="true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_vn.txt" ignoreCase="true"/>
    <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <!-- <filter class="solr.ASCIIFoldingFilterFactory" preserveOriginal="true"/> -->
  </analyzer>
</fieldType>
<fieldType name="text_gl" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_gl.txt" ignoreCase="true"/>
    <filter class="solr.GalicianStemFilterFactory"/>
  </analyzer>
</fieldType>
```


Hình:

- Thêm một số field name:

```
<field name="Author" type="text_general_vi" indexed="true" stored="true"/>
<field name="Content" type="text_general_vi" indexed="true" stored="true"/>
<field name="Origin" type="text_general_vi" indexed="true" stored="true"/>
<field name="Time" type="text_general"/>
<field name="Title" type="text_general_vi" indexed="true" stored="true"/>
<field name="Url" type="text_general_vi" indexed="true" stored="true"/>
<field name="_root_" type="string" docValues="false" indexed="true" stored="false"/>
<field name="_text_" type="text_general" multiValued="true" indexed="true" stored="false"/>
<field name="_version_" type="plong" indexed="false" stored="false"/>
<field name="id" type="string" multiValued="false" indexed="true" required="true" stored="true"/>
```

Hình:

- Sau khi cấu hình Index mới cho solr, bộ Analysis mới hoạt động:



Field Value (Index)

Kính tế giới có nhiều biến động

Field Value (Query)

kính tế

Dashboard

Logging

Core Admin

Java Properties

Thread Dump

vi_data

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema

Segments info

Analyse Fieldname / FieldType: Content

Schema Browser

Verbose Output

Analyse Values

ST	text	Kính	tế	thế	giới	có	nhiều	biến	động
	raw_bytes	[4b 69 6e 68]	[74 e1 ba bf]	[74 68 e1 ba bf]	[67 69 e1 bb 9b 69]	[63 c3 b3]	[6e 68 69 e1 bb 81 75]	[62 69 e1 ba bf 6e]	[c4 91 e1 bb 99 6e 67]
	start	0	5	8	12	17	20	26	31
	end	4	7	11	16	19	25	30	35
	positionLength	1	1	1	1	1	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	termFrequency	1	1	1	1	1	1	1	1
	position	1	2	3	4	5	6	7	8
SF	text	Kính	tế		giới			biến	động
	raw_bytes	[4b 69 6e 68]	[74 e1 ba bf]		[67 69 e1 bb 9b 69]			[62 69 e1 ba bf 6e]	[c4 91 e1 bb 99 6e 67]
	start	0	5		12			26	31
	end	4	7		16			30	35
	positionLength	1	1		1			1	1
	type	<ALPHANUM>	<ALPHANUM>		<ALPHANUM>			<ALPHANUM>	<ALPHANUM>
	termFrequency	1	1		1			1	1
	position	1	2		4			7	8
ASCIIFF	text	Kính	te	tế		giới	giới		
	raw_bytes	[4b 69 6e 68]	[74 65]	[74 e1 ba bf]		[67 69 6f 69]	[67 69 e1 bb 9b 69]		
	start	0	5	5		12	12		
	end	4	7	7		16	16		
	positionLength	1	1	1		1	1		

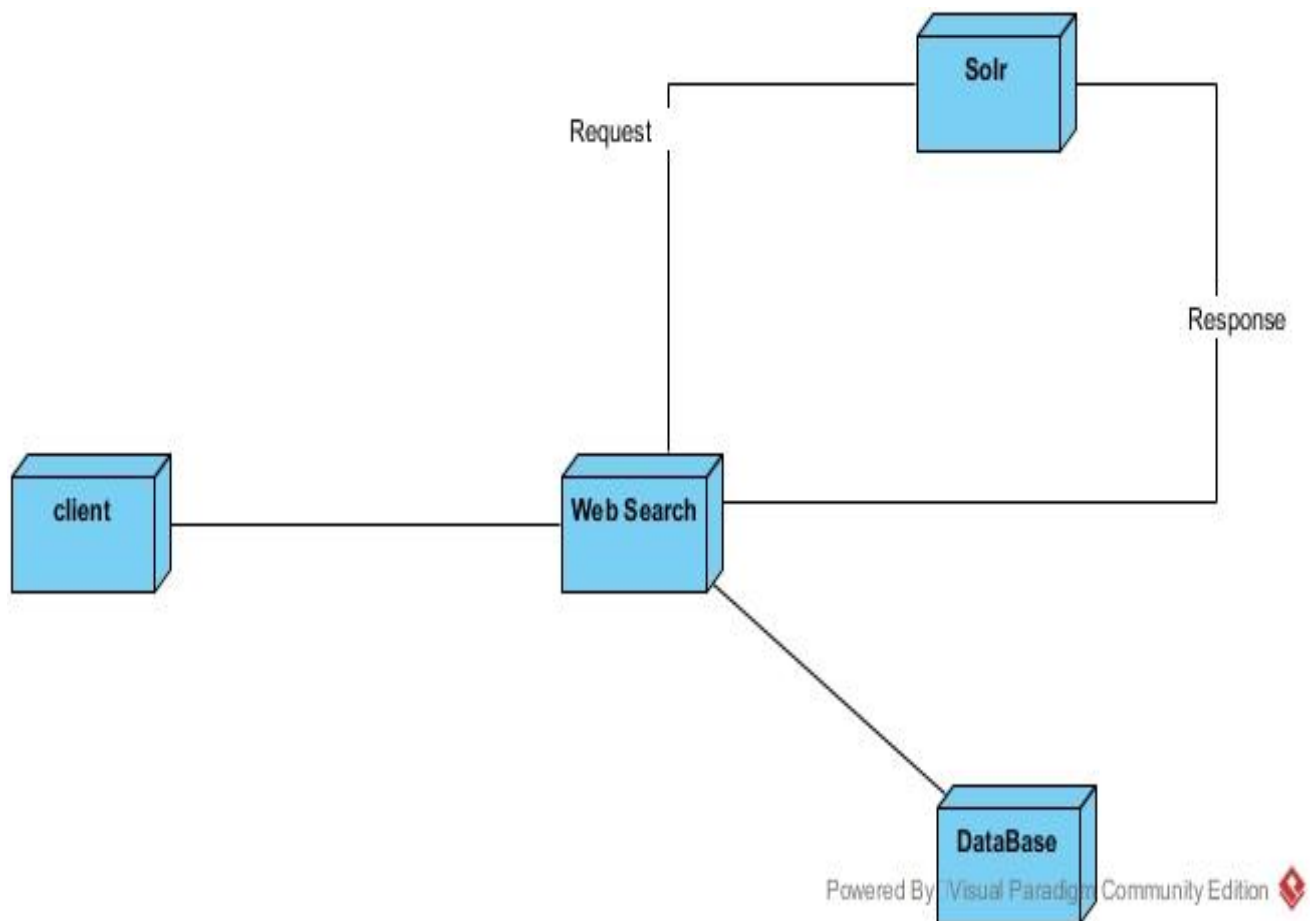
Hình:

⇒ Khi index dữ liệu theo cách cấu hình lại bộ Filter + Stopword của solr thì bộ Analysis của solr sẽ hoạt động khác với việc sử dụng index mặc định một chút. Đầu tiên dữ liệu sẽ được loại bỏ từ dừng, sau đó chúng được tách thành hai trường độc lập : có dấu và không có dấu. Bộ index của solr sẽ đánh index cho cả hai trường này của dữ liệu. Vì vậy mà việc sử dụng cách index mới này sẽ giúp việc index nhanh hơn, thời gian chuyển vấn nhanh hơn. Việc index theo cách mới này còn hỗ trợ việc tìm kiếm có dấu và không có

dấu đối với tiếng việt. Tuy nhiên do phải index trên cả hai trường có dấu và không dấu nên sẽ tốn bộ nhớ hơn cách index thông thường.

3. Xây dựng Web Search

Sơ đồ hoạt động của hệ thống:



Hình:

- Solr hỗ trợ trả lại API cho các hệ thống khác:

Hình:

3.1 Back-end:

- Laravel Framework 5.6
- Call API: Guzzle
- Mysql 5.7

- Luồng hoạt động:

- Tại giao diện tìm kiếm, người dùng gõ một câu tìm kiếm bất kỳ nào đó
- Từ khóa được gửi về server web search, tại đây nó được loại bỏ các ký tự đặc biệt
- Sử dụng: Guzzle để gửi một request tới server solr cùng với truy vấn xây dựng sẵn + câu truy vấn của người dùng
- Solr trả lại API chứa kết quả tìm kiếm tới server web search
- Từ khóa tìm kiếm được lưu lại vào cơ sở dữ liệu
- Server search sử lý API do solr trả lại để lấy ra các thông tin cần thiết
- Server search gửi kết quả lại cho giao diện tìm kiếm
- Giao diện hiển thị kết quả cho người dùng

3.2 Front-end

- Bootstrap 3
- JQuery
- Ajax

3.3 Một số tính năng gợi ý cho người dùng:

- Gợi ý kết quả tìm kiếm real-time khi người dùng gõ tìm kiếm
 - Khi người dùng gõ từ khóa vào ô tìm kiếm, bắt sự kiện thay đổi ký tự trong ô tìm kiếm
 - Sử dụng Ajax để gửi từ khóa đang gõ của người dùng đến server search theo thời gian thực
 - Bắt đầu một luồng tìm kiếm khi nhận được từ khóa
 - Server search gửi lại kết quả qua Ajax để hiển thị kết quả tìm kiếm theo thời gian thực
- Gợi ý hoàn thiện tìm kiếm của người dùng dựa trên thống kê từ khóa phổ biến của hệ thống
 - Lấy ra 10 từ khóa được tìm kiếm nhiều nhất
 - Server search tiến hành gửi 10 từ khóa phổ biến nhất về giao diện người dùng

- Sử dụng thuộc tính suggestions của thẻ input trong html5 để gợi ý từ khóa cho người dùng khi họ gõ tìm kiếm
- Gợi ý tìm kiếm cho người dùng khi người dùng gõ truy vấn bị sai một vài ký tự dựa trên thống kê từ khóa tìm kiếm + độ tương đồng Jacard:
 - Người dùng gõ một từ khóa tìm kiếm
 - Từ khóa được gửi về server search
 - Lấy tất cả các từ khóa của trong cơ sở dữ liệu
 - Sử dụng độ đo tương đồng : để so sánh từ khóa của người dùng với các từ khóa trong hệ thống để trả về từ khóa gần giống nhất mà có nhiều người tìm kiếm
- Bôi đậm tìm kiếm của người dùng trong kết quả trả về:
 - Tại giao diện tìm kiếm, sau khi nhận được kết quả từ server search, tiến hành phân tích ra các trường dữ liệu để hiển thị
 - Đối với trường: Title và trường Content: từ khóa tìm kiếm của người dùng được loại bỏ các ký tự đặc biệt
 - Viết biểu thức chính quy để tìm từ khóa trong trường Title và trường Content và thay thế nó bởi các ký tự in đậm
 - Nếu cả từ khóa không xuất hiện thì tiến hành tách từ khóa thành các từ đơn, sau đó chuyển chúng thành các từ viết hoa chữ cái đầu, các từ viết thường rồi tiến hành tìm kiếm các từ đó, nếu chúng tồn tại trong Title và Content thì thay chúng bằng các ký tự in đậm
- Phân trang kết quả tìm kiếm (mỗi trang chứa 10 kết quả)

4. Demo kết quả xây dựng hệ thống

- Giao diện tìm kiếm:



Index dữ liệu theo mặc định của Solr	Q	Tìm Với Google
--------------------------------------	---	----------------

Index dữ liệu loại bỏ dấu tiếng việt	Index dữ liệu sử dụng Filter+Stopword
--------------------------------------	---------------------------------------

Hình :

- Kết quả tìm kiếm với index mặc định:



tại sao khởi nghiệp lại hay thất bại



Tìm Với Google

Khoảng: 14643 kết quả (khoảng: 1 giây)

Mọi người cũng hay tìm kiếm: tại sao khởi nghiệp lại hay thất bại ?

« 1 2 3 4 5 6 7 8 9 10 »

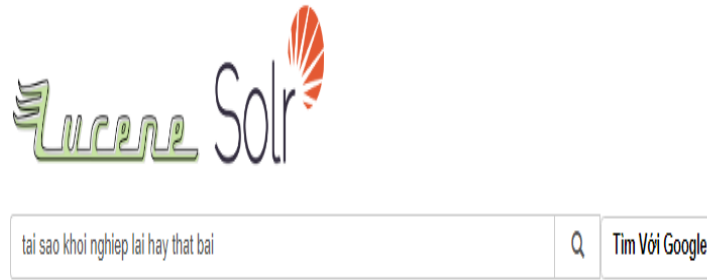
Doanh nghiệp khởi nghiệp Việt nên dám làm, dám mạo hiểm hơn

Link: <http://dantri.com.vn/kinh-doanh/doanh-nghiep-khoi-nghiep-viet-nen-dam-lam-dam-mao-hiem-hon-20171114202345469.htm>

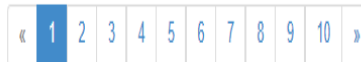
Cụ thể, trả lời câu hỏi của phóng viên Dân trí về cộng đồng khởi nghiệp (startup) Việt có điểm giống và khác gì so với các startup trên thế giới, ông Thuận cho biết: "Tôi thấy cộng đồng khởi nghiệp ở Việt Nam rất hứa hẹn, rất trẻ trung, tài năng và có học. Quan trọng là họ có khao khát làm nên những thay đổi lớn lao, được xây dựng công ty của riêng họ". Bên cạnh đó, ông Thuận chia sẻ, cộng đồng khởi nghiệp ở những nơi khác nhau trên thế giới ví dụ như tại Mỹ, Thung lũng Silicon, Anh, Canada,... tất nhiên sẽ có những điểm khác nhau nhưng cơ bản là giống. Cộng đồng khởi nghiệp ở Việt Nam cũng không ngoại lệ. Cụ thể, ông cho rằng, cộng đồng khởi nghiệp ở Việt Nam rất trẻ, năng động, và được giáo dục bài bản. Chính sự năng động, khao khát làm nên điều mới như vậy tạo nên sức trẻ rất đặc biệt ở cộng đồng khởi nghiệp Việt Nam, CEO Công nghệ Toàn cầu của Uber nói. "Có lẽ, tôi chỉ khuyên cộng đồng khởi nghiệp Việt Nam rằng các bạn nên dám làm và dám mạo hiểm hơn. Cũng không nên sợ thất bại vì khi vượt qua nỗi sợ các bạn mới đứng cảm hơn, và chỉ khi đứng cảm hơn thì các bạn mới làm được điều to lớn hơn", ông Thuận cho hay. Bên cạnh đó, ông Thuận cũng chia sẻ những kiến thức mà ông tích lũy được dành cho các startup Việt. Theo ông, đầu tiên, các DN khởi nghiệp cần biết vấn đề chính mình cần giải quyết là gì. "Tôi nghĩ họ cần có sự tập trung cao độ, để có thể hiểu được mình muốn làm gì và muốn giải quyết vấn đề gì. Sau đó họ cần tập trung hết sức để giải quyết vấn đề đó", ông Thuận chia sẻ và cũng lưu ý rằng, khi khởi nghiệp mà DN ôm đồm quá nhiều vấn đề thì có thể thất bại. Vì vậy, ngay từ đầu DN cần tập trung vào 1 vấn đề cần giải quyết và luôn giữ được sự tập trung đó. Sau đó, DN startup nên biết điều gì là quan trọng và tập trung vào đó. Theo ông Thuận, mấu chốt của khởi nghiệp không phải là xây dựng DN khởi nghiệp mà là giải quyết vấn đề, gia tăng giá trị. Khi đó, có thể một ai đó trên thị trường thấy đó là giải pháp hữu hiệu và sẵn sàng chi trả cho giải pháp đó. Do đó, trọng tâm của khởi nghiệp không phải là sự thành công, mà là việc giải quyết vấn đề, khi đó tự khắc thành công sẽ đến với bạn, vị CEO Công nghệ Toàn cầu của Uber nhấn mạnh. Cuối cùng, startup cần trụ vững vì con đường khởi nghiệp chắc chắn sẽ gặp nhiều khó khăn, bạn cần tồn tại mạnh mẽ được đến

Hình:

- Kết quả tìm kiếm một truy vấn không dấu khi sử dụng index mặc định:



Khoảng: 7179 kết quả (khoảng: 2 giây)



Học tiếng Anh mỗi ngày: Cách dùng của Present Continuous là gì?

Link: <http://dantri.com.vn/giao-duc-khuyen-hoc/hoc-tieng-anh-moi-ngay-cach-dung-cua-present-continuous-la-gi-20180131081112403.htm>

2. Usage: Ta dùng Present Continuous (Hiện tại tiếp diễn) để diễn đạt - Một hành động đang xảy ra tại thời điểm đó hoặc xung quanh thời điểm đó. Cách dùng này thường đi với các trạng từ now, at the moment, at present, presently, currently, ... Ex: We are learning English now. (Hiện giờ chúng tôi đang học tiếng Anh.) - Một hành động, một dự định **hay** kế hoạch chắc chắn sẽ xảy ra trong tương **lai** gần Ex: We are visiting my grandparents this weekend. (Chúng tôi chắc chắn sẽ đi thăm ông bà vào cuối tuần này.) - Diễn đạt sự thay đổi **hay** chuyển biến. Cách dùng này thường **hay** đi với các động từ to get, to become, to change, to turn + tính từ cấp so sánh hơn Ex: It's getting hotter. (Trời đang ngày càng nóng dần.) - Diễn đạt sự phàn nàn, luôn đi với always Ex: He is always talking in class. (Anh ấy luôn nói chuyện trong giờ.) Bây giờ chúng ta hãy cùng vận dụng kiến thức đã học để hiểu rõ hơn cách dùng qua bài tập sau nhé! Bài tập: Hãy điền dạng đúng của từ vào đoạn hội thoại sau: An: What (1) (you / do) Bình: (2) (I / write) a letter to a friend. He's a disc jockey. Vicky and I (3) (try) to organize a party. An: That sounds a lot of work. How (4) (you / find) time for your studies? Bình: Well, as I said, Vicky (5) (help) me. An: That's great. Good luck with the party! I have to go now. See you later. Bình: Goodbye, An. Answer key: 1. Are you doing? 2. I'm writing/ I am writing 3. Are trying 4. Are you finding 5. Is helping

Tác Giả: Nhật Hồng

Ngày Đăng: Thứ tư, 31/01/2018 - 08:07

Học phí của con: Gánh nặng tài chính không của riêng ai

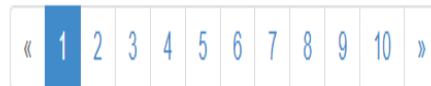
Hình:

- Kết quả tìm kiếm không dấu với dữ liệu được index sau khi loại bỏ dấu của các từ trong tiếng việt:



Khoảng: 14257 kết quả (khoảng: 6 giây)

Mọi người cũng hay tìm kiếm: tai sao khoi nghiep lai hay that bai



Bạn thất nghiệp vì bạn muốn thế!

Link: <http://dantri.com.vn/giao-duc-khuyen-hoc/ban-that-nghiep-vi-ban-muon-the-2017061417311553.htm>

Mỗi khi có thông tin số liệu về con số cử nhân thất nghiệp, dư luận lại rộ lên sôi nổi bàn về nguyên nhân thất nghiệp, nào là do chất lượng đào tạo kém, lý thuyết không đi đôi với thực hành; do bằng đại học mất giá; do cung quá nhiều so với cầu... Trong khi đó, nhân tố chủ thể trong "vấn nạn" thất nghiệp là bản thân người lao động lại ít được đưa ra bàn luận. Tôi nghĩ rằng, cử nhân sẽ không thất nghiệp nếu họ thực sự muốn có việc làm! Sở dĩ cử nhân thất nghiệp là bởi vì họ kén việc. Theo tôi, có một số khía cạnh liên quan đến tâm lý kén việc của cử nhân. Thứ nhất, nhiều người vẫn theo xu hướng thích vào biên chế. Trong buổi tiếp xúc với hơn 200 cử tri của phường Hưng Thạnh (quận Cái Răng, TP Cần Thơ) ngày 27/4/2017, Chủ tịch Quốc hội Nguyễn Thị Kim Ngân nói, hiện cả nước có khoảng 200.000 sinh viên sau khi ra trường chưa tìm được việc làm. Trong khi đó, quan điểm của người dân là con em mình phải làm việc ở cơ quan nhà nước, phải vào biên chế. Chọn vào biên chế, tức là bạn chọn một "cuộc chơi" khó cho mình, vì chỉ tiêu biên chế vô cùng hạn chế, quá ít ỏi so với những chỉ tiêu việc làm khác ngoài xã hội. Vì cạnh tranh cho một số lượng rất ít chỉ tiêu, nên sự khó khăn, khắc nghiệt trong cuộc đua giành suất biên chế là điều dễ hiểu. Và thất nghiệp là kết cục của đa số người tham gia vào cuộc chơi này. Thứ hai, nhiều người nghĩ phải làm một công việc xứng tầm với mình - vị trí tốt, lương cao. Nhưng thế nào là công việc xứng tầm? Trong cuốn sách "Học cách tiêu tiền" (NXB Lao động), tác giả Larry Winget (diễn giả, người dẫn chương trình truyền hình nổi tiếng tại Mỹ) khẳng định: "Không có công việc lương thấp nào lại dưới tầm đối với tôi". Ông Trần Anh Tuấn, Phó Giám đốc Trung

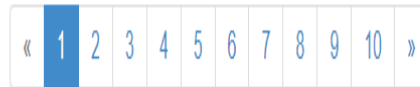
Hình:

- Kết quả tìm kiếm không dấu với dữ liệu được index bằng cách cấu hình Filter + StopWord trong tiếng việt:



Khoảng: 9910 kết quả (khoảng: 0 giây)

Mọi người cũng hay tìm kiếm: vi sao khoi nghiep hay that bai



Doanh nghiệp khởi nghiệp Việt nên dám làm, dám mạo hiểm hơn

Link: <http://dantri.com.vn/kinh-doanh/doanh-nghiep-khoi-nghiep-viet-nen-dam-lam-dam-mao-hiem-hon-20171114202345469.htm>

Cụ thể, trả lời câu hỏi của phóng viên Dân trí về cộng đồng khởi nghiệp (startup) Việt có điểm giống và khác gì so với các startup trên thế giới, ông Thuận cho biết: "Tôi thấy cộng đồng khởi nghiệp ở Việt Nam rất hứa hẹn, rất trẻ trung, tài năng và có học. Quan trọng là họ có khao khát làm nên những t **hay** đổi lớn lao, được xây dựng công ty của riêng họ". Bên cạnh đó, ông Thuận chia sẻ, cộng đồng khởi nghiệp ở những nơi khác nhau trên thế giới ví dụ như tại Mỹ, Thung lũng Silicon, Anh, Canada,... tất nhiên sẽ có những điểm khác nhau nhưng cơ bản là giống. Cộng đồng khởi nghiệp ở Việt Nam cũng không ngoại lệ. Cụ thể, ông cho rằng, cộng đồng khởi nghiệp ở Việt Nam rất hứa hẹn và có nhiều điểm thú vị. Tất nhiên sẽ khó so sánh Việt Nam với những nước khác vì điều kiện phát triển mỗi nơi đều khác nhau nhưng cộng đồng khởi nghiệp ở Việt Nam rất trẻ, năng động, và được giáo dục bài bản. Chính sự năng động, khát khao làm nên điều mới như vậy tạo nên sức trẻ rất đặc biệt ở cộng đồng khởi nghiệp Việt Nam, CEO Công nghệ Toàn cầu của Uber nói. "Có lẽ, tôi chỉ khuyên cộng đồng khởi nghiệp Việt Nam rằng các bạn nên dám làm và dám mạo hiểm hơn. Cũng không nên sợ thất bại vì khi vượt qua nỗi sợ các bạn mới đứng cảm hơn, và chỉ khi đứng cảm hơn thì các bạn mới làm được điều to lớn hơn", ông Thuận cho **hay**. Bên cạnh đó, ông Thuận cũng chia sẻ những kiến thức mà ông tích lũy được dành cho các startup Việt. Theo ông, đầu tiên, các DN khởi nghiệp cần biết vấn đề chính mình cần giải quyết là gì. "Tôi nghĩ họ cần có tư tưởng trung tâm để có thể hiểu được mình muốn làm gì và muốn giải quyết vấn đề gì. Sau đó họ cần tập trung hết sức để giải quyết vấn đề đó" ông

- Gợi ý kết quả thời gian thực:



Điểm	Q	Tìm Với Google
<ul style="list-style-type: none">• Trường Đại học Đà Lạt công bố điểm trúng tuyển năm 2017• Cụm thi số 34 - ĐH Hồng Đức: Thí sinh điểm cao nhất đạt 28,85 điểm khối A• Điểm chuẩn vào Học viện Công nghệ Bưu chính viễn thông từ 19 - 25 điểm• TPHCM: Có 2 thí sinh đạt 3 điểm 10• Nữ sinh dân tộc Thái lọt top 10 thí sinh điểm khối C cao nhất toàn quốc		

Khoảng: 2176 kết quả (khoảng: 0 giây)

«	1	2	3	4	5	6	7	8	9	10	»
---	---	---	---	---	---	---	---	---	---	----	---

Yêu hoa, thương hoa cũng cần có văn hóa

Link: <http://dantri.com.vn/nhip-song-tre/yeu-hoa-thuong-hoa-cung-can-co-van-hoa-20151206222428972.htm>

Tan nát những cánh đồng hoa vì "tự sướng" quá đà Cách đây không lâu, ngay trước khi Lễ hội hoa tam giác mạch Hà Giang được tổ chức, cả nghìn người miền xuôi đổ xô về cao nguyên đá để tận mắt nhìn thấy những loài hoa biểu trưng cho vẻ đẹp vùng cao. Nhưng khi lễ hội còn chưa khai màn, rất nhiều cánh đồng hoa đã tan hoang cùng với đó là la liệt rác rưởi, chai nước du khách bỏ lại... Những đồng cỏ trắng xóa Mộc Châu cũng chịu chung số phận, bởi đây đã biến thành nơi để cho những người mê chụp ảnh thì nhau nằm, ngồi tạo dáng. Còn tại Hà Nội, thung lũng hoa hồ Tây - địa điểm "thượng hoa" có tiếng Hà thành phải đóng cửa chỉ sau 3 ngày mở cửa miễn phí vì du khách giẫm đạp hoa không thương tiếc. Ngay sau khi thung lũng hoa mở cửa, lượng người đổ về đây ngày một đông. Theo thống kê, chỉ trong ngày thứ sáu, có 1.500 người vào cửa, thì sang ngày chủ nhật đã có 7.000 người chen chúc để vào được ruộng hoa. "Thật kinh hoàng, lẽ ra đã có một vườn hoa tam giác mạch đẹp thì nay mất hết. Cả vườn hoa bướm bướm chưa kịp ra hoa cũng bị giẫm nát" - anh Bùi Mạnh Hiếu, chủ thung lũng hoa hồ Tây chia sẻ với PV sau 3 ngày "vỡ trận". Giờ đây, 5ha hoa của anh, trong đó có 2ha hoa tam giác mạch, còn lại là vô số các loại hoa bướm bướm, bách nhật, cải, cúc họa mi... đây công vun trồng, nay lại phải chờ trồng lại. Biền cảm có đẹp được thói quen xấu? Hoa tam giác mạch, hoa cải, hay hướng dương vốn không phải được trồng cho đẹp. Phần lớn những loài hoa này được người nông dân trồng để làm nguồn lương thực cho gia súc. Nhưng chính nhờ "phát hiện" của du khách, mà những cánh đồng hoa đã trở thành dịch vụ để khai thác, phát triển du lịch. Người người đổ xô đến, chen chúc lội xuống ruộng hoa, cúi lấy vài phút tạo dáng, chụp ảnh. "Cú hích" du lịch thì chưa thấy đâu, sau mỗi mùa hoa nở chỉ thấy nổi bức xúc, thất vọng của những người dân bản địa, những người trồng hoa. Ngẫm ra, việc người ta chen nhau giẫm nát cả ruộng hoa chỉ để chụp ảnh không khác cảnh tượng hàng trăm người giẫm đạp, tranh giành nhau một vài suất sushi miễn phí, hay những ông bố, bà mẹ đồ con vượt rào leo vào Công viên nước hồ Tây trong ngày mở cửa miễn phí. Tất cả chỉ vì: Miễn phí! Sẽ chẳng có ai nài tính toán cụ thể thiệt hại cho những người nông dân khi những du khách giẫm đạp lên những vườn hoa ở Hà Giang. Sơn

Hình:

- Gợi ý hoàn thành từ khóa dựa trên thông kê từ khóa phổ biến trong hệ thống:



Điểm thi đại học bị	▼	Q	Tim Với Google
điểm thi đại học bách khoa			
Điểm thi đại học Bách Khoa hà nội			

Khoảng: 7179 kết quả (khoảng: 2 giây)

«	1	2	3	4	5	6	7	8	9	10	»
---	---	---	---	---	---	---	---	---	---	----	---

Học tiếng Anh mỗi ngày: Cách dùng của Present Continuous là gì?

Link: <http://dantri.com.vn/giao-duc-khuyen-hoc/hoc-tieng-anh-moi-ngay-cach-dung-cua-present-continuous-la-gi-20180131081112403.htm>

2. Usage: Ta dùng Present Continuous (Hiện tại tiếp diễn) để diễn đạt - Một hành động đang xảy ra tại thời điểm đó hoặc xung quanh thời điểm đó. Cách dùng này thường đi với các trạng từ now, at the moment, at present, presently, currently, ... Ex: We are learning English now. (Hiện giờ chúng tôi đang học tiếng Anh.) - Một hành động, một dự định **hay** kế hoạch chắc chắn sẽ xảy ra trong tương **lại** gần Ex: We are visiting my grandparents this weekend. (Chúng tôi chắc chắn sẽ đi thăm ông bà vào cuối tuần này.) - Diễn đạt sự thay đổi **hay** chuyển biến. Cách dùng này thường **hay** đi với các động từ to get, to become, to change, to turn + tính từ cấp so sánh hơn Ex: It's getting hotter. (Trời đang ngày càng nóng dần.) - Diễn đạt sự phàn nàn, luôn đi với always Ex: He is always talking in class. (Anh ấy luôn nói chuyện trong giờ.) Bây giờ chúng ta hãy cùng vận dụng kiến thức đã học để hiểu rõ hơn cách dùng qua bài tập sau nhé! Bài tập: Hãy điền dạng đúng của từ vào đoạn hội thoại sau: An: What (1) (you / do) Binh: (2) (I / write) a letter to a friend. He's a disc jockey. Vicky and I (3) (try) to organize a party. An: That sounds a lot of work. How (4) (you / find) time for your studies? Binh: Well, as I said, Vicky (5) (help) me. An: That's great. Good luck with the party! I have to go now. See you later. Binh: Goodbye, An. Answer key: 1. Are you doing? 2. I'm writing/ I am writing 3. Are trying 4. Are you finding 5. Is helping

Tác Giả: Nhật Hồng

Ngày Đăng: Thứ tư, 31/01/2018 - 08:07

Học phí của con: Gánh nặng tài chính không của riêng ai

Hình:

- Gợi ý từ khóa phổ biến giống từ khóa bạn đang tìm kiếm nhất:



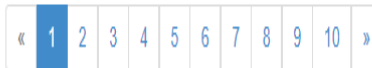
Điểm thi đại học Bách khoa năm 2018



Tìm Với Google

Khoảng: 12947 kết quả (khoảng: 0 giây)

Mọi người cũng hay tìm kiếm: Điểm thi đại học năm 2018



Năm 2018 , điểm sàn do các trường đại học quyết định

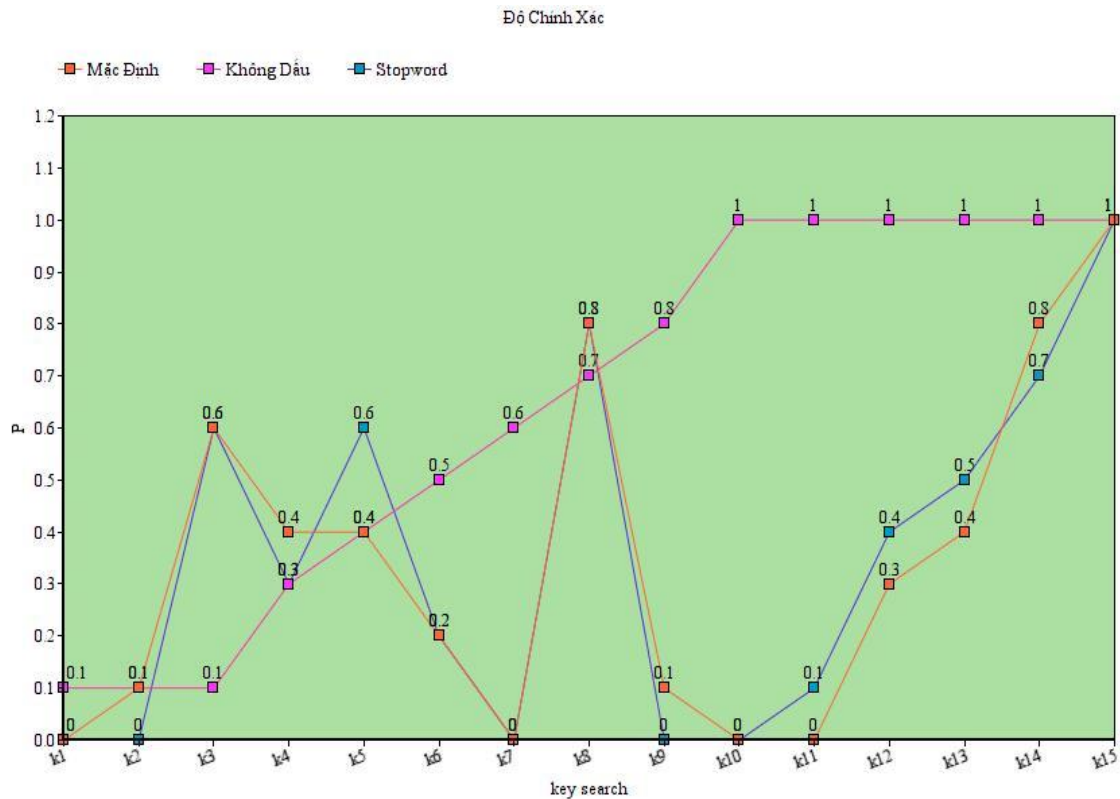
Link: <http://dantri.com.vn/giao-duc-khuyen-hoc/nam-2018-diem-san-do-cac-truong-dai-hoc-quyet-dinh-20170624183921341.htm>

Kỳ thi THPT Quốc gia 2017 kết thúc, một số ý kiến cho rằng, với đề thi như năm nay có khả năng sẽ nhiều điểm cao và do vậy, việc xét tuyển vào các trường đại học sẽ khó khăn hơn. Vụ trưởng Vụ giáo dục Đại học Nguyễn Thị Kim Phụng cho biết, về điểm sàn theo quy định tuyển sinh hiện nay đã quy định rõ, các quy định về điểm sàn áp dụng trong năm 2017. Từ năm 2018, có một số điều kiện được bổ sung và khi thực hiện điều kiện được bổ sung đó, các trường cũng sẽ phải thay đổi. Cụ thể, các trường đại học sẽ phải xây dựng đề án tuyển sinh đầy đủ, hoàn chỉnh. Trong đó, đặc biệt là những quy định về công khai các điều kiện đảm bảo chất lượng của trường, công khai tỷ lệ việc làm của sinh viên theo từng ngành đào tạo trong 2 năm gần nhất và công khai tỷ suất đầu tư đào tạo từng sinh viên trong năm học... Theo bà Phụng, tất cả những nội dung đó sẽ cung cấp cho xã hội thí sinh điều kiện đảm bảo chất lượng của các trường thế nào, tỷ lệ việc làm... Sự đầu tư của nhà trường cũng sẽ tương đương với học phí thu. Đó là cơ sở để xã hội giám sát chất lượng và thí sinh lựa chọn được trường phù hợp với ngành học, trình độ năng lực, mức điểm thi của mình... "Khi đã cung cấp cho thí sinh, xã hội tất cả điều kiện lựa chọn rồi thì Bộ GD&ĐT có thể không cần thiết phải quy định điểm sàn nữa mà trao quyền đó cho từng trường và xã hội sẽ có quyền lựa chọn", bà Phụng nói. Vụ trưởng Vụ giáo dục Đại học khẳng định, từ năm 2018 trở đi, điểm sàn do các trường đại học quyết định phù hợp với điều kiện mà Bộ đã quy định. Và đó cũng là quá trình đảm bảo quyền tự chủ của trường đại học, vừa cung cấp đầy đủ thông tin cho xã hội để thí sinh tự lựa chọn ngôi trường đại học phù hợp với bản thân. Về bản khoản của phòng viên cho rằng, điểm đại học cao sẽ khó xét tuyển đại học, bà Nguyễn Thị Kim Phụng lưu ý, khi chưa chấm thi chúng ta không thể vì cảm nhận mà nói điều đó. Tuy nhiên, qua theo dõi dư luận và trao đổi với giáo viên hầu hết các ý kiến đều đánh giá đề năm nay có tính phân loại cao. "Khối đại học cũng vui mừng và phản hồi tốt về đề thi năm nay. Nếu tính phân loại của đề cao thì đó sẽ là điều kiện để xét tuyển vào các trường đại học", bà Phụng nói.

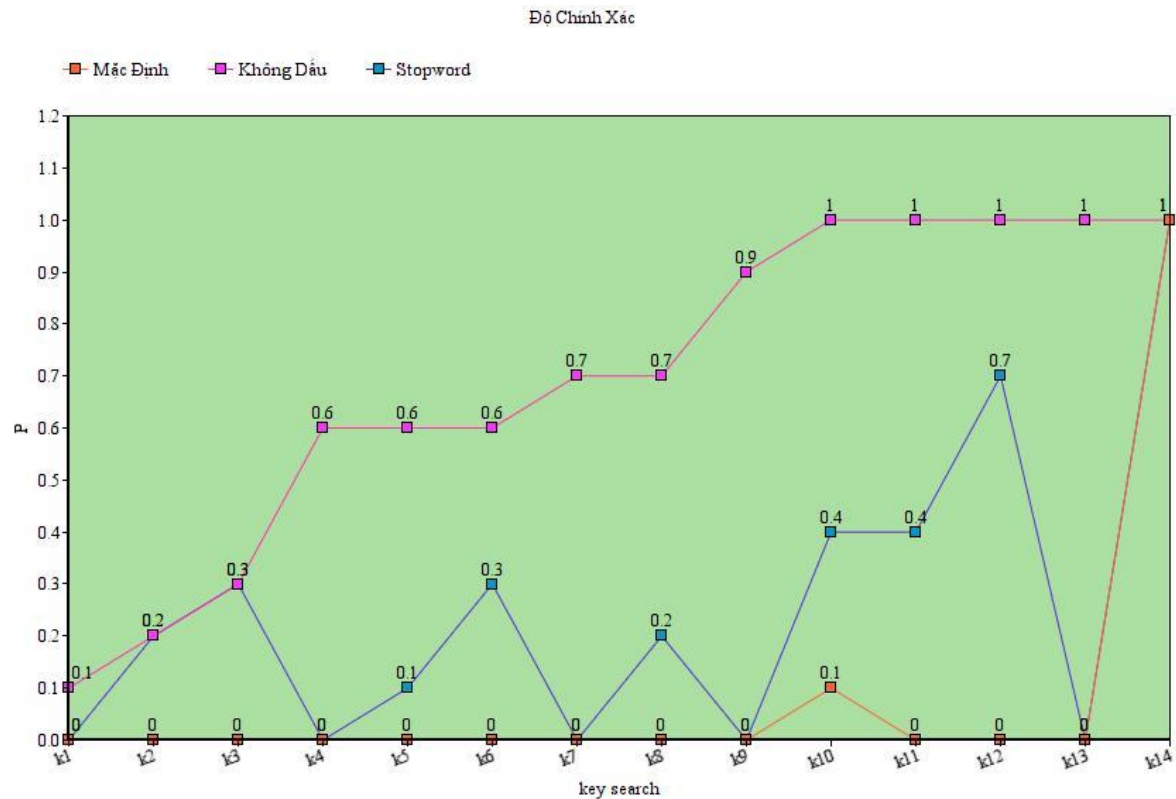
Hình:

5. Đánh giá kết quả tìm kiếm của hệ thống

Truy vấn đầy đủ dấu	Core_1	Core_2	Core_3	Số KQ
tuyển sinh	8	10	7	10
giáo dục đại học	4	10	5	10
siêu xe	3	10	4	10
khởi nghiệp là gì vậy	4	3	3	10
năng suất lao động	0	1	0	10
hội thánh của đức chúa trời	0	6	0	10
hà nội có gì nổi bật	6	1	6	10
giá vàng	1	8	0	10
môi trường	8	7	8	10
giá xăng	0	10	0	10
Điện thoại smartphone	0	10	1	10
giá ô tô năm nay thế nào	4	4	6	10
smartphone	10	10	10	10
bách khoa hà nội	1	1	0	10
công nghệ 4.0 là gì	2	5	2	10



Truy thiếu dấu hoặc không dấu	Core_1	Core_2	Core_3	Số KQ
tuyen sinh	0	10	7	10
giao duc dai hoc	1	10	4	10
sieu xe	0	10	4	10
khoi nghiep la gi	0	3	3	10
nang suat lao dong	0	1	0	10
hoi thanh duc chua troi	0	6	0	10
gia vang	0	7	0	10
moi truong	0	6	1	10
gia xang	0	9	0	10
dien thoai smartphone	0	10	0	10
gia o to nam nay the nao	0	6	3	10
smartphone	10	10	10	10
bach khoa ha noi	0	2	2	10
cong nghe 4.0 la gi	0	7	2	10



6. Nhận xét và kết luận

Ưu điểm:

- Có khả năng search theo kiểu full text search
- Kết quả trả về khá chính xác
- Thời gian trả về kết quả ngắn
- Có khả năng scale tốt

Nhược điểm:

- Dữ liệu còn ít nên không bao quát được các trường hợp tìm kiếm
- Gợi ý từ còn hạn chế do dữ liệu trong cơ sở dữ liệu còn ít
- Hỗ trợ tiếng việt chưa được tốt

7. Tài liệu tham khảo:

- http://lucene.apache.org/solr/guide/7_2/solr-tutorial.html
- <https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt>
- <https://jqueryui.com/autocomplete/>
- <https://laravel.com/docs/5.6>
- <http://butchiso.com/2013/08/tim-hieu-ve-apache-solr.html>
- <https://issues.apache.org/jira/browse/LUCENE-2507>
- <https://wiki.apache.org/solr/LanguageAnalysis>
- <https://wiki.apache.org/solr/AnalyzersTokenizersTokenFilters>