

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



BÁO CÁO BÀI TẬP LỚN

MÔN: TÌM KIẾM THÔNG TIN VÀ TRÌNH DIỄN THÔNG TIN

**ĐỀ TÀI: ỨNG DỤNG SOLR
XÂY DỰNG HỆ THỐNG TÌM KIẾM ĐIỆN THOẠI**

Giáo viên hướng dẫn: TS. Nguyễn Bá Ngọc

Sinh viên thực hiện:

1. Nguyễn Đức Thắng
2. Nguyễn Thị Thuỳ Dương
3. Trịnh Khánh Linh

MSSV: 20163843

MSSV: 20160849

MSSV: 20162490

Hà nội, T6/2020

Mục lục

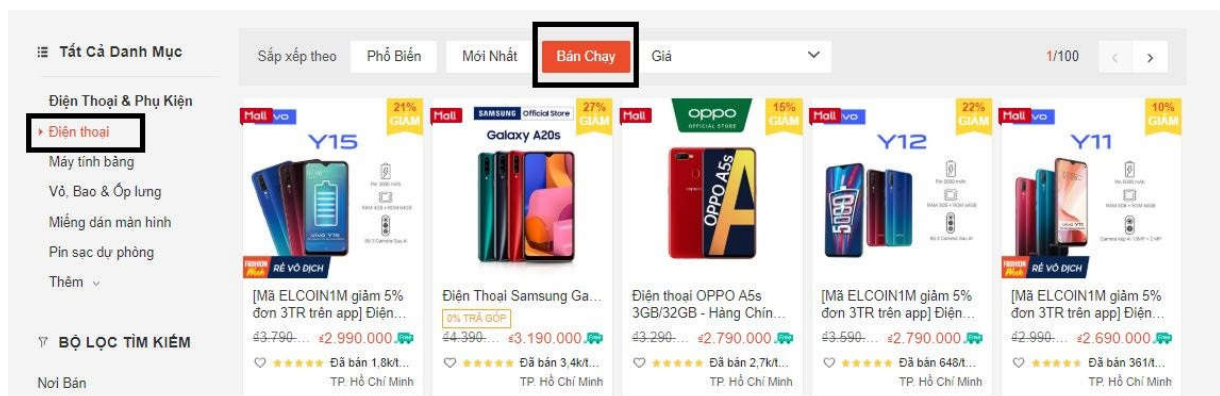
I. THU THẬP DỮ LIỆU	3
II. TÌM HIỂU SOLR	5
1 Tổng Quan Về Solr	5
2 Các Chức Năng Cơ Bản Của Solr	6
3 Cài Đặt Và Sử Dụng	7
4 Thêm Dữ Liệu Vào Solr	8
5 Truy Vấn	8
III. XÂY DỰNG GIAO DIỆN	11

I. THU THẬP DỮ LIỆU

Chúng em sử dụng dữ liệu danh sách điện thoại trên trang thương mại điện tử Shopee. Nhóm em lấy dữ liệu về thông qua việc gọi API đến Shopee.

API Shopee để lấy ra các sản phẩm điện thoại bán chạy nhất:

https://shopee.vn/api/v2/search_items/?by=sales&limit=50&match_id=1979&new_est=6000&order=desc&page_type=search&version=2



Ở đây chúng em gọi API GET 2 lần , lần 1 để lấy được danh sách ID các sản phẩm sau đó gọi tiếp lần 2 để lấy về chi tiết sản phẩm.Sau khi có các thuộc tính của sản phẩm chúng em lưu kết quả vào file json.

Các thuộc tính của sản phẩm :

- Name : Tên sản phẩm
- ID : ID của sản phẩm
- ShopID : Id của shop bán sản phẩm
- Location : Địa chỉ của shop
- Image : Link đến ảnh sản phẩm
- Rating : Số sao đánh giá của sản phẩm
- Price : Giá tiền

```

for page in range(MAX_PAGE):
    print(page)
    URL = "https://shopee.vn/api/v2/search_items/"
    PARAMS = { ...
    }
    HEADER = { ...
    }
    response = requests.get(url=URL, headers=HEADER, params=PARAMS)
    print(response)
    data = response.json()
    items = data['items']

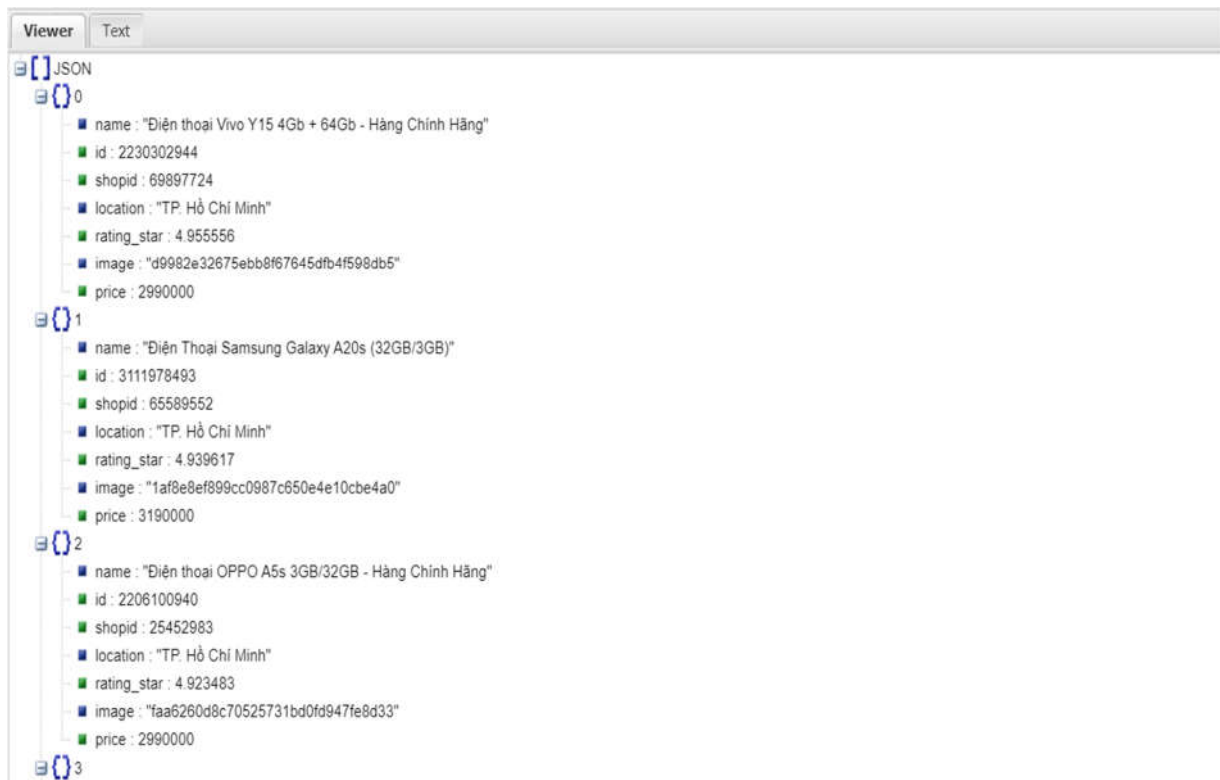
    for i in items:
        URL_ITEM = "https://shopee.vn/api/v2/item/get"
        PARAMS_ITEM = {
            'itemid': i['itemid'],
            'shopid': i['shopid']
        }
        response = requests.get(
            url=URL_ITEM, headers=HEADER, params=PARAMS_ITEM).json()
        item = response['item']

        json_load = {
            'name': item['name'],
            'id': item['itemid'],
            'shopid': item['shopid'],
            'location': item['shop_location'],
            'rating_star': item['item_rating']['rating_star'],
            'image': item['images'][0],
            'price': item['price']//100000
        }
        dataSolr.append(json_load)

with open('data.json', 'w') as outfile:
    json.dump(dataSolr, outfile)

```

Dưới đây là file kết quả mà chúng em thu thập được :



II. TÌM HIỂU SOLR

1 Tổng Quan Về 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.

Apache Solr khởi đầu là một project nội bộ của CNET được tạo ra bởi Yonik Seeley, nhằm phục vụ chức năng tìm kiếm của website CNET vào năm 2004 và được đóng góp cho Apache Software Foundation năm 2006. Solr không hoàn toàn là một RESTful interface của Lucene mà là sử dụng Lucene như là một component trong toàn bộ hệ thống. Sau phiên bản Solr 1.4, từ version 3.1 (3/2011) thì Solr và Lucene dùng chung một codebase và version number.

Solr bao gồm nhiều thành phần (components) khác nhau:

Apache Lucene để phân tích, đánh chỉ mục tìm kiếm dữ liệu.

2 Các Chức Năng Cơ Bản Của Solr

- Khả năng tìm kiếm văn bản toàn diện (Full-Text Search) giống kiểu Google.
- 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
- Khả năng mở rộng ra nhiều server Solr
- 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, ...

3 Cài Đặt Và Sử Dụng

Các bước để chạy Solr:

- cd vào thư mục solr
- bin/solr start
- bin/solr status để kiểm tra trạng thái
- truy cập vào đường dẫn <http://localhost:8983/solr/#/> để sử dụng solr admin.
- bin/solr stop để ngắt kết nối với server solr

Các câu lệnh cơ bản

- Tạo core solr create -c name_core -p port_name -d conf_dir
- Xóa core solr delete -c name_core
- Start solr trên cổng khác bin/solr start -p 8984

Đơn vị thông tin cơ bản trong Solr là các tài liệu (các documents), mỗi tài liệu bao gồm các trường thông tin (các fields), mỗi trường thông tin có 1 kiểu định dạng (field type).

Người dùng có thể định nghĩa các trường thông tin, các kiểu định dạng trong file schema.xml

Các kiểu định dạng:

Quy định các thuộc tính của trường thông tin, các thao tác thực hiện khi tài liệu được đánh chỉ mục, cũng như khi xử lý câu truy vấn.

Thuộc tính	Ý nghĩa
indexed	Giá trị của trường có được đánh chỉ mục hay không.
stored	Giá trị của trường có được lưu trữ hay không.
multiValued	Trường trong tài liệu có thể nhận nhiều giá trị của kiểu hay không.

Các trường thông tin (field) :

Thuộc tính	Ý nghĩa
name	Tên của trường
type	Kiểu định dạng của trường
default	Giá trị mặc định của trường

4 Thêm Dữ Liệu Vào Solr

Để thêm dữ liệu vào solr nhóm sử dụng thư viện Pysolr :

```
1 import pysolr
2 import json
3
4 solr = pysolr.Solr('http://localhost:8983/solr/phone-data/')
5 with open('data.json', 'r') as json_file:
6     json_object = json.load(json_file)
7     solr.ping()
8     solr.add(json_object)
9     solr.commit()
```

5 Truy Vấn

Các tham số truy vấn chung

Tham số Chức năng Mô tả defType chọn trình phân tích cú pháp truy vấn mà solr nên sử dụng để xử lý tham số truy vấn chính (q) trong yêu cầu.

Ví dụ: defType=dismax, edismax....

- sort sắp xếp kết quả tìm kiếm tăng dần ASC hoặc giảm dần DESC. Solr có thể sắp xếp các câu trả lời truy vấn theo document score hoặc bất kì trường nào có một giá trị được lập chỉ mục hoặc sử dụng DocValues
- start chỉ định offset vào tập kết quả của truy vấn và hướng dẫn solr hiển thị kết quả từ offset này.
- Có thể dùng start để phân trang.

- rows dùng tham số này để chọn ra bao nhiêu bản ghi để phân trang kết quả truy vấn.
- fq (query filter) Bộ lọc fq có thể được chỉ định nhiều lần trong 1 truy vấn. hữu ích để tăng tốc độ các truy vấn phức tạp
- fq=id:[10 TO *]&fq=score:10 ---> chỉ những tài liệu có id lớn hơn 10 và có score 10 thì mới được đưa vào kết quả.
- wt kiểu dữ liệu phản hồi của truy vấn JSON,...

Các tham số:

- q: Xác định truy vấn bằng cú pháp chuẩn.
- q.op: chỉ định toán tử mặc định cho các biểu thức truy vấn, ghi đè toán tử mặc định và chỉ định trong lược đồ. Các gtri có thể là AND hoặc OR
- df: trường mặc định
- sow: Tách trên khoảng trắng: nếu đặt = false thì chuỗi phân tách bằng khoảng trắng sẽ được cung cấp cho phân tích tài liệu 1 lần (ý kiểu là nó sẽ phân tích cái dấu cách). Mặc định là true.

Cách thức thực hiện truy vấn với Solr

Để thực hiện truy vấn với Solr chúng ta có thể sử dụng trực tiếp giao diện mà solr cung cấp như hình dưới đây :

The screenshot shows the Solr Admin UI interface. On the left is a sidebar with navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'phone-data' containing Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (selected), Replication, Schema, and Segments info. The main panel is titled 'Request-Handler (qt)' and shows the path '/select'. The 'q' field contains 'hồ chí minh+'. The 'wt' dropdown is set to 'json'. The 'response' field shows a JSON object with headers and a list of documents.

```

{
  "responseHeader": {
    "status": 0,
    "QTime": 5,
    "params": {
      "q": "hồ chí minh+",
      "indent": "on",
      "start": "0",
      "wt": "json",
      "...": "1591882731263"
    }
  },
  "response": {
    "numFound": 1853,
    "start": 0,
    "docs": [
      {
        "name": ["Điện thoại Nokia 1110i giá rẻ"],
        "id": "5505175833",
        "shopid": [34285568],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [4.679012],
        "image": ["daff8fd255635e7b6db2206232eab51"],
        "price": [70000],
        "_version_": 1669210015519473666
      },
      {
        "name": ["Điện Thoại Mini BM10"],
        "id": "6124861013",
        "shopid": [197138574],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [4.736842],
        "image": ["3cc3278be687cdf371faf99d57f1c3de"],
        "price": [165000],
        "_version_": 1669210015520522240
      },
      {
        "name": ["Điện thoại Nokia hu"],
        "id": "4009512721",
        "shopid": [34285568],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [5.0],
        "image": ["7afe3c5dd8508191c49d6a265d16f8cc"],
        "price": [10000],
        "_version_": 1669210015540445184
      }
    ]
  }
}

```

Hoặc chúng ta có thể giao tiếp với solr thông qua HTTP API

The screenshot shows a web browser window with the address bar displaying the URL: `localhost:8983/solr/phone-data/select?indent=on&q=hồ%20chí%20minh+&start=0&wt=json`. The browser shows the JSON response from the Solr API, which is identical to the one shown in the Solr Admin UI screenshot.

```

{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "q": "hồ chí minh ",
      "indent": "on",
      "start": "0",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1853,
    "start": 0,
    "docs": [
      {
        "name": ["Điện thoại Nokia 1110i giá rẻ"],
        "id": "5505175833",
        "shopid": [34285568],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [4.679012],
        "image": ["daff8fd255635e7b6db2206232eab51"],
        "price": [70000],
        "_version_": 1669210015519473666
      },
      {
        "name": ["Điện Thoại Mini BM10"],
        "id": "6124861013",
        "shopid": [197138574],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [4.736842],
        "image": ["3cc3278be687cdf371faf99d57f1c3de"],
        "price": [165000],
        "_version_": 1669210015520522240
      },
      {
        "name": ["Điện thoại Nokia hu"],
        "id": "4009512721",
        "shopid": [34285568],
        "location": ["TP. Hồ Chí Minh"],
        "rating_star": [5.0],
        "image": ["7afe3c5dd8508191c49d6a265d16f8cc"],
        "price": [10000],
        "_version_": 1669210015540445184
      }
    ]
  }
}

```

III. XÂY DỰNG GIAO DIỆN

Các công nghệ nhóm sử dụng để hoàn thiện bài tập:

- Python để thực hiện thu thập dữ liệu
- Solr
- Vuejs

Nhóm 5 - Tìm Kiếm Điện Thoại

Search Filter

Tắt cả mức giá

< 2.000.000

2.000.000 - 5.000.000

5.000.000 - 10.000.000

> 10.000.000

Tắt cả các vùng










Hà Nội

Hồ Chí Minh

Đà Nẵng

Cần Thơ

Kết quả Tìm kiếm Với samsung - Giá:5.000.000 - 10.000.000 - Địa chỉ: Hồ Chí Minh

 <p>Điện thoại Samsung Galaxy A71 (8/128Gb) Chính hãng, mới 100%</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 8.090.000 Đánh giá: 0</p>	 <p>Điện thoại Samsung Galaxy Note 10 Lite 8GB 128GB - Hàng chính hãng</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 9.840.000 Đánh giá: 5</p>	 <p>Điện thoại Samsung Galaxy S9 Plus 2sim mới Fullbox ram 6G/64G mới zin</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 6.190.000 Đánh giá: 5</p>	 <p>Điện Thoại Samsung Galaxy Note 8 2 Sim mới / chính hãng chơi game tốt</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 5.799.000 Đánh giá: 0</p>	 <p>Điện thoại samsung Galaxy Note 8 64gb mới fullbox uy tín giá tốt nhất tphcm</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 7.000.000 Đánh giá: 0</p>
 <p>Điện thoại Samsung Galaxy S9 64GB (Bản Mỹ) Nguyên zin 99% giá tốt nhất hcm</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 5.600.000 Đánh giá: 0</p>	 <p>Điện Thoại Samsung Galaxy S8+ Quốc Tế 4GB/64GB Chuẩn Kháng Nước</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 8.900.000 Đánh giá: 5</p>	 <p>Điện thoại samsung Galaxy S9 Plus 64GB (Bản Mỹ) nguyên zin 99% giá rẻ.Ship COD toàn Quốc</p> <p>Địa chỉ: TP. Hồ Chí Minh</p> <p>Giá: 6.500.000 Đánh giá: 0</p>	 <p>Điện thoại Samsung Galaxy A31 - Hàng chính hãng</p> <p>Địa chỉ: Bến Tre</p> <p>Giá: 5.090.000 Đánh giá: 4.833333</p>	 <p>Điện thoại Samsung Galaxy A51 - Hàng chính hãng</p> <p>Địa chỉ: Bến Tre</p> <p>Giá: 6.079.000 Đánh giá: 4.909091</p>

« 1 2 3 4 5 6 »

Hệ thống nhóm em có các chức năng:

- Phân trang các kết quả trả về
- Lọc kết quả về theo mức giá
- Lọc theo trường Địa chỉ
- Hiển thị thông tin các kết quả trả về và có link dẫn đến trang mua hàng trên shopee.

Quản lý code trong nhóm:

GitHub: <https://github.com/thuyduongbka/IT4853-BTL-SearchEngine>

Tài liệu tham khảo

Slide thầy Nguyễn Bá Ngọc

Trang Tài liệu hướng dẫn của Solr

https://lucene.apache.org/solr/guide/6_6/common-query-parameters.html

Cài đặt solr

https://www.tutorialspoint.com/apache_solr/apache_solr_basic_commands.htm