

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH



ĐỒ ÁN KỸ THUẬT LẬP TRÌNH (CO1031)

Báo cáo Đồ án

BUILD A SIMPLE WEB CRAWLER

GVHD: Thầy Lưu Quang Huân

SV thực hiện:	Trần Phương Tĩnh	– 1927038
	Lê Tất Thiện	– 1920058
	Dương Quang Tuấn	– 1920069
	Nguyễn Đức Tuấn	– 1927040

Tp. Hồ Chí Minh, Tháng 12/2020

Mục lục

1	Giới thiệu đề tài	2
2	Giới thiệu công nghệ	4
2.1	Giới thiệu về Angular	4
2.2	Hướng dẫn triển khai lên Cloud	5
2.3	Giới thiệu về NodeJS	7
2.4	Thư viện dùng để crawl với NodeJS: Cherrio và Axios	7
2.5	Object Data Model: Mongoose	8
3	Kết quả đạt được	9
4	Hướng dẫn sử dụng	11
5	Triển khai	12
5.1	Backend	12
5.2	Frontend	12
6	Báo cáo hàng tuần	14
6.1	Báo cáo tuần 1	14
6.2	Báo cáo tuần 2	14
6.3	Báo cáo tuần 3	15
6.4	Báo cáo tuần 4	16
6.5	Báo cáo tuần 5	16
7	Tổng kết và hướng phát triển cho đề tài	17
7.1	Tổng kết	17
7.2	Hướng phát triển cho đề tài	17
8	Tài liệu tham khảo	17

1 Giới thiệu đề tài

• MỤC ĐÍCH - YÊU CẦU

- Deep learning, Machine Learning hay AI đã và đang là một chủ đề đang được quan tâm với nhiều ứng dụng trong thực tiễn và các nghiên cứu mang tính học thuật cao. Tuy nhiên, việc triển khai các ứng dụng, các mô hình cũng như việc tiến hành nghiên cứu gặp nhiều khó khăn. Một mặt, do các bộ dữ liệu huấn luyện hạn chế, về cả số lượng và không đảm bảo về chất lượng. Các bộ dữ liệu hiện có cũng hạn chế, đa phần bằng các ngôn ngữ không phải tiếng Việt.
- Xuất phát từ nhu cầu thực tế của việc thực hiện tác đề tài, đồ án môn học sau này hoặc khi làm LVTN liên quan đến dữ liệu, sinh viên gặp nhiều khó khăn trong việc thu thập dữ liệu mẫu (Dataset) để triển khai các mô hình Machine Learning, Deep Learning hoặc cơ bản hơn là thu thập dữ liệu làm dữ liệu mẫu cho một ứng dụng, hệ thống đã xây dựng. Trong bài tập lớn này, SV sẽ có cơ hội hiện thực việc thu thập các dữ liệu này, làm quen với các thao tác tiền xử lý, phân loại, lưu trữ và biểu diễn một tập dữ liệu (Data set).
- Mỗi nhóm sẽ được phân công thu thập dữ liệu về một chủ đề từ các website khác nhau.

• CÁCH THỨC TRIỂN KHAI

- Project này được thực hiện trong vòng 5 tuần với kế hoạch và kết quả dự kiến như sau:
 - * Task 2.0: Chốt danh sách đăng ký nhóm, khởi tạo các công cụ làm việc nhóm.
 - * Task 2.1: Tìm hiểu và hiện thực các công cụ Crawl dữ liệu theo yêu cầu. Thường mỗi nhóm sẽ được phân crawl từ nhiều website khác nhau, do đó các thành viên trong nhóm làm việc để tự phân chia công việc. Kết quả dự kiến tuần 1: source code, hướng dẫn sử dụng, sample dữ liệu đã thu được (1000 items). Trong quá trình crawl chắc chắn gặp nhiều khó khăn, nhóm cần tự tìm hiểu về những cách sử dụng VPN, proxy, sock để có thể hoàn thành yêu cầu.
 - * Task 2.2. Triển khai việc Crawl dữ liệu, xây dựng Bộ dataset (10.000.), cải tiến công cụ (source code đã xây dựng) để tạo thành công cụ hoàn chỉnh cho việc Crawl dữ liệu. Tìm hiểu cách làm sạch, chuẩn hóa dữ liệu, phân loại dữ liệu. Thiết kế CLDS, các định dạng file biểu diễn dataset. Kết quả dự kiến: dataset, bản thiết kế CSDL, báo cáo tìm hiểu về cách làm sạch, source code nếu có.
 - * Task 2.3. Kết quả dự kiến: Báo cáo + source code làm sạch dữ liệu, source code lưu trữ dữ liệu raw (dataset raw) vào database, source code truy vấn dữ liệu, model data...
 - * Task 2.4. Thực hành việc phân tích dữ liệu. Xây dựng một số tính năng phân tích, thống kê, tìm kiếm, export....
 - * Task 2.5. Báo cáo tổng kết.
- Trước 23h59 ngày chủ nhật hàng tuần, các nhóm nộp kết quả từng tuần lên BKEL (thông qua module Bài tập lớn). Trong đó, sv cần nêu báo cáo vấn tắt (dưới 1) trang theo format
 - * **Các công việc đã thực hiện**
 - Tìm hiểu scrapy (Hưng) đã có báo cái tại link (hoặc đã push git)
 - Implement vnexpresscrawler.py (Huân) đã push code lên git.
 - ...
 - * **Các khó khăn gặp phải**
 - Session limit, đã khắc phục bằng cách abc... đã viết lại báo cáo hướng dẫn crawl tại link
 - * **Dự kiến công việc tuần tới**
 - a
 - b
 - c
- Kèm theo đó là đường link chi tiết về báo cáo kĩ thuật, hoặc source code cần được push lên git. Những báo cáo này cần viết và quản lý để tuần thứ 5 tổng hợp thành báo cáo Đồ án môn học.

• ĐÁNH GIÁ

- Cách thức đánh giá chung: kết quả theo từng giai đoạn mà nhóm thu được thông qua source code, dataset và báo cáo. Kết quả đánh giá từng thành viên thông qua log git (do đó cần thường xuyên push code), các tài liệu cũng nên được commit lên git. Kết quả sẽ được GV đánh giá và update từng tuần trên file.
- Các nhóm làm việc với nhau dưới sự hướng dẫn của GV chủ yếu thông qua các công cụ online như email, bkel... nếu trong trường hợp chưa rõ, cần hỗ trợ hướng dẫn, SV cần chủ động liên hệ với GV qua email để setup meeting (online hoặc offline) để hiểu rõ yêu cầu và thực hiện.

2 Giới thiệu công nghệ

2.1 Giới thiệu về Angular



Angular 2 được biết đến tên rộng rãi như hiện tại là Angular. Nó là một framework cho frontend và là bản tiếp theo của AngularJS. Angular là mã nguồn mở giúp chúng ta xây dựng một Single Page Applications (SPAs). Angular là cung được xây dựng cả ứng dụng Mobile và Desktop. Nó được xây dựng sử dụng JavaScript. Bạn phải sử dụng nó để xây dựng ứng dụng hoàn chỉnh kết hợp với HTML, CSS và JavaScript. Angular có nhiều cải tiến thông so với AngularJS. Nó có nhiều cải tiến làm dễ học và phát triển ứng dụng cho doanh nghiệp. Bạn có thể xây dựng một ứng dụng dễ dàng mở rộng, bảo trì, test.

• Tính năng của Angular

Angular được load với tính năng Power-packaged. Một số tính năng được liệt kê ra đây như sau:

- Cơ chế Two-Way Data Binding: Đây là tính năng cool nhất của Angular. Data binding tự động và rất nhanh tức là bất cứ thay đổi nào trên view đều được tự động cập nhật vào component class và ngược lại
- Hỗ trợ cơ chế Routing mạnh mẽ: Angular có cơ chế routing tải trang một cách bất đồng bộ trên cùng một trang cho phép chúng ta tạo SPA.
- Mở rộng HTML: Angular cho phép chúng ta sử dụng cấu trúc lập trình giống như điều kiện if, vòng lặp for...để render các control.
- Thiết kế module hoá: Angular tiếp cận theo hướng thiết kế module hoá. Bạn phải tạo các Angular Module để tổ chức tốt hơn và quản lý source code.
- Hỗ trợ làm việc với hệ thống Backend: Angular được xây dựng hỗ trợ làm việc với backend server và thực thi bất cứ logic nào và nhận dữ liệu về.

- Chọn mục Hosting

```
=== Project Setup

First, let's associate this project directory with a Firebase project.
You can create multiple project aliases by running firebase use --add,
but for now we'll just set up a default project.

? Please select an option:
  Use an existing project
> Create a new project
  Add Firebase to an existing Google Cloud Platform project
  Don't set up a default project
```

Hình 2: Set up project 1

```
=== Project Setup

First, let's associate this project directory with a Firebase project.
You can create multiple project aliases by running firebase use --add,
but for now we'll just set up a default project.

? Please select an option: Create a new project
i If you want to create a project in a Google Cloud organization or folder,
  please use "firebase projects:create" instead, and return to this command w
  hen you've created the project.
? Please specify a unique project id (warning: cannot be modified afterward)
  [6-30 characters]:
  () ktlt-fe-angular
```

Hình 3: Set up project 2

```
Your public directory is the folder (relative to your project directory) tha
t
will contain Hosting assets to be uploaded with firebase deploy. If you
have a build process for your assets, use your build's output directory.

? What do you want to use as your public directory? public
? Configure as a single-page app (rewrite all urls to /index.html)? No
✓ Wrote public/404.html
✓ Wrote public/index.html

i Writing configuration info to firebase.json...
i Writing project information to .firebaserc...

✓ Firebase initialization complete!
```

Hình 4: Set up project 3

- Bước 7: Chạy lệnh "firebase deploy" để deploy lên firebase

```
$ firebase deploy

=== Deploying to 'ktlt-fe-angular'...

i deploying hosting
i hosting[ktlt-fe-angular]: beginning deploy...
i hosting[ktlt-fe-angular]: found 7 files in dist/web
✓ hosting[ktlt-fe-angular]: file upload complete
i hosting[ktlt-fe-angular]: finalizing version...
✓ hosting[ktlt-fe-angular]: version finalized
i hosting[ktlt-fe-angular]: releasing new version...
✓ hosting[ktlt-fe-angular]: release complete

✓ Deploy complete!

Project Console: https://console.firebase.google.com/project/ktlt-fe-angular/ov
erview
Hosting URL: https://ktlt-fe-angular.web.app
```

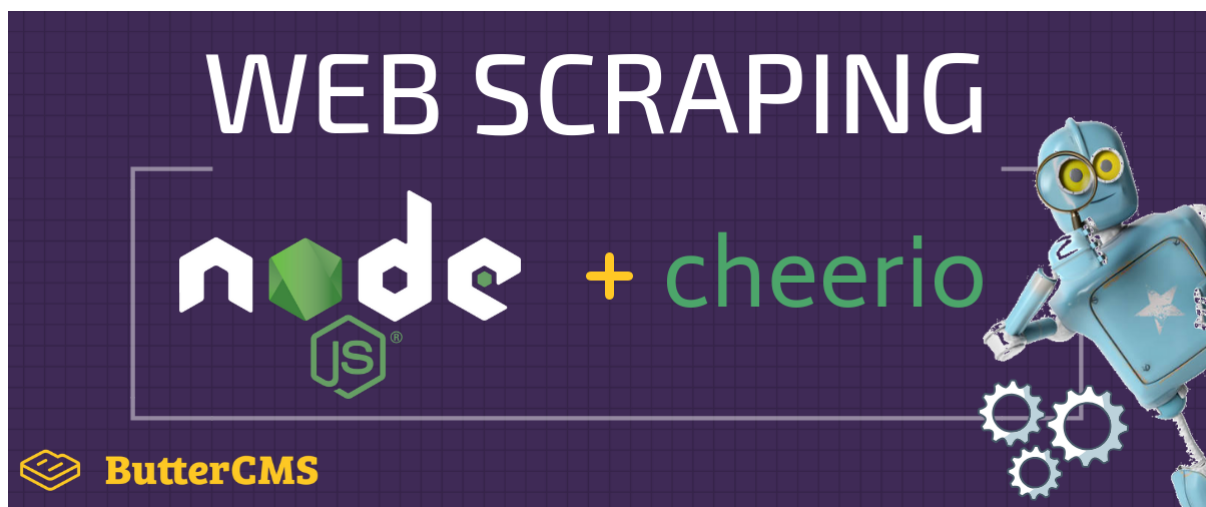
Hình 5: Firebase deploy

2.3 Giới thiệu về NodeJS



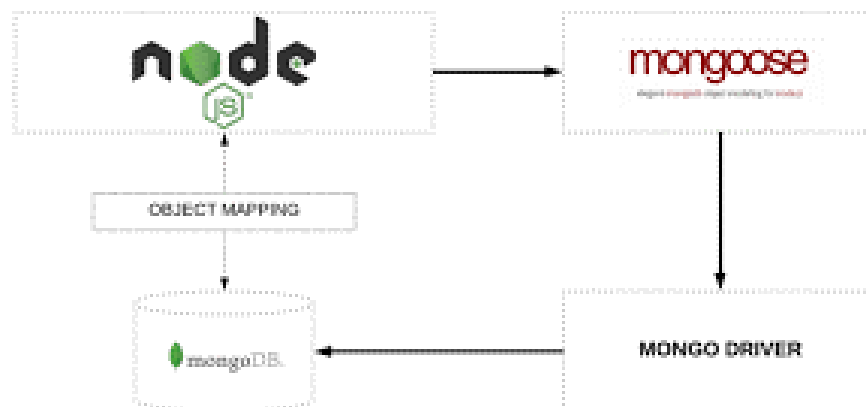
- Node.js là một hệ thống phần mềm được thiết kế để viết các ứng dụng internet có khả năng mở rộng, đặc biệt là máy chủ web. Chương trình được viết bằng JavaScript, sử dụng kỹ thuật điều khiển theo sự kiện, nhập/xuất không đồng bộ để tối thiểu tổng chi phí và tối đại khả năng mở rộng. Node.js bao gồm có V8 JavaScript engine của Google, libUV, và vài thư viện khác.
- Node.js được tạo bởi Ryan Dahl từ năm 2009, và phát triển dưới sự bảo trợ của Joyent.
- Node.js là một JavaScript runtime được build dựa trên engine JavaScript V8 của Chrome. Node.js sử dụng kiến trúc hướng sự kiện event-driven, mô hình non-blocking I/O làm cho nó nhẹ và hiệu quả hơn. Hệ thống nén của Node.js, npm, là hệ thống thư viện nguồn mở lớn nhất thế giới.
- Nodejs áp dụng cho các sản phẩm có lượng truy cập lớn, cần mở rộng nhanh, cần đổi mới công nghệ, hoặc tạo ra các dự án Startup nhanh nhất có thể.
- Nodejs tạo ra được các ứng dụng có tốc độ xử lý nhanh, realtime thời gian thực.
- Phần Core bên dưới của Nodejs được viết hầu hết bằng C++ nên cho tốc độ xử lý và hiệu năng khá cao.
- Trong đề tài lần này, nhóm chúng em sử dụng NodeJS và các thư viện liên quan để thiết kế web server cũng như kết nối với các thiết bị để thực hiện các chức năng đã đề ra.

2.4 Thư viện dùng để crawl với NodeJS: Cherrio và Axios



- Cherrio: là một thư viện hỗ trợ parse DOM giống như JQuery cung cấp nhiều công cụ để việc crawl được hiệu quả
- Axios là một thư viện HTTP Client dựa trên Promise. Cơ bản thì nó cung cấp một API cho việc xử lý XHR (XMLHttpRequests).

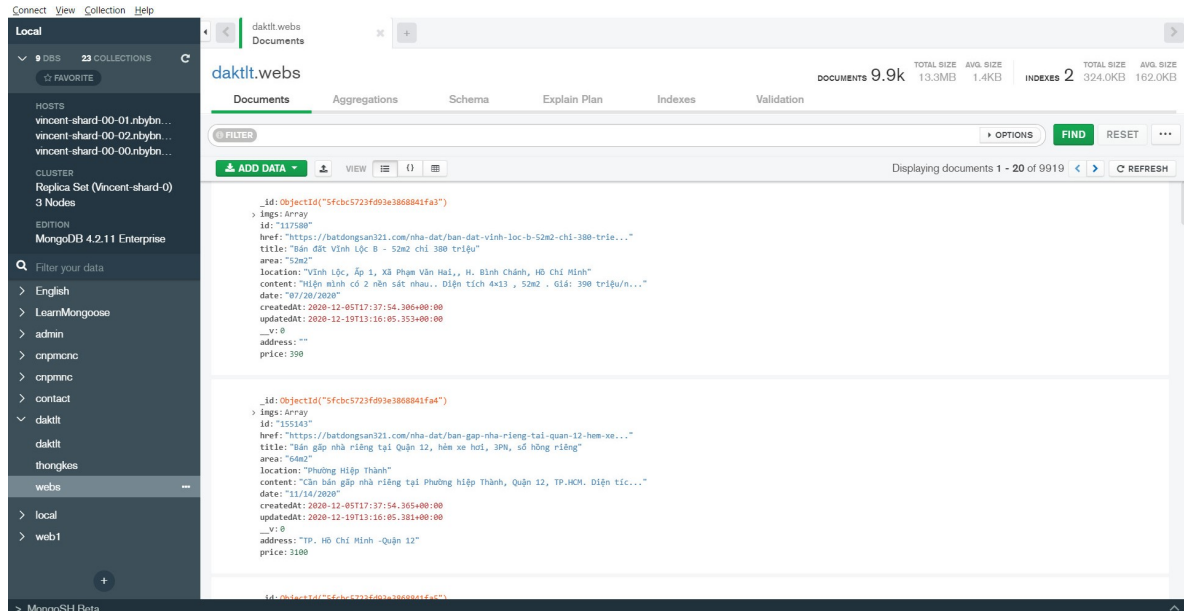
2.5 Object Data Model: Mongoose



Mongoose là một thư viện mô hình hóa đối tượng (Object Data Model - ODM) cho MongoDB và Node.js. Nó quản lý mối quan hệ giữa dữ liệu, cung cấp sự xác nhận giản đồ và được sử dụng để dịch giữa các đối tượng trong mã và biểu diễn các đối tượng trong MongoDB. Nhóm nhận thấy sử dụng Mongoose rất thuận tiện cho việc tạo model nên đã đưa vào sử dụng.

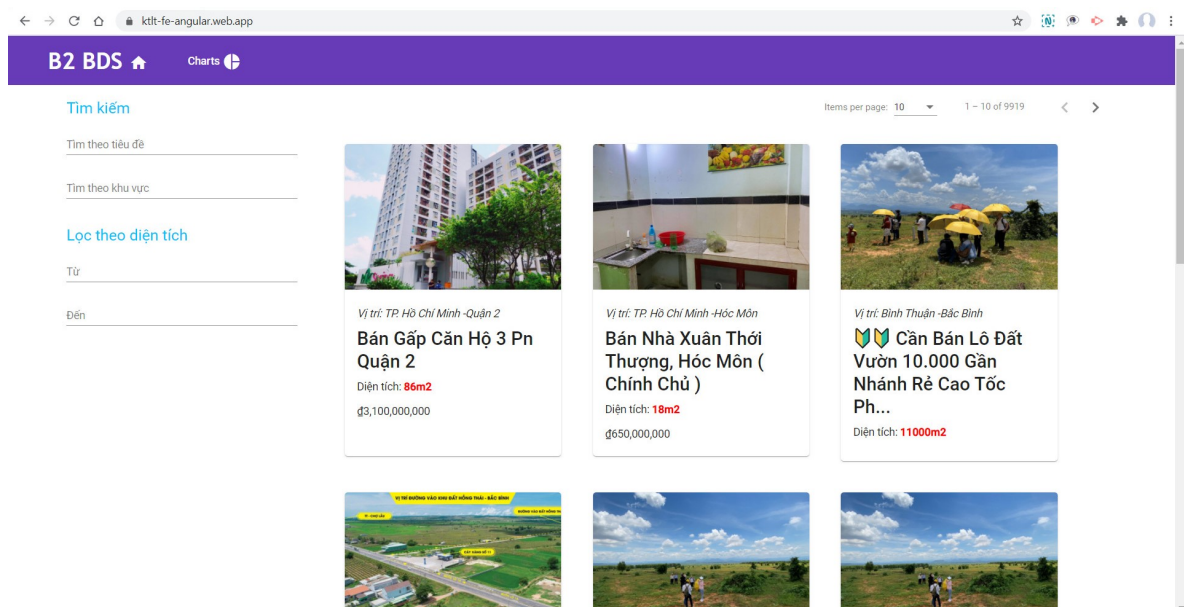
3 Kết quả đạt được

- Xây dựng được công cụ để crawl gần 10000 dữ liệu bất động sản từ trang batdongsan321.com và lưu trong MongoDB.



Hình 6: Gần 10000 dữ liệu được lưu trong MongoDB

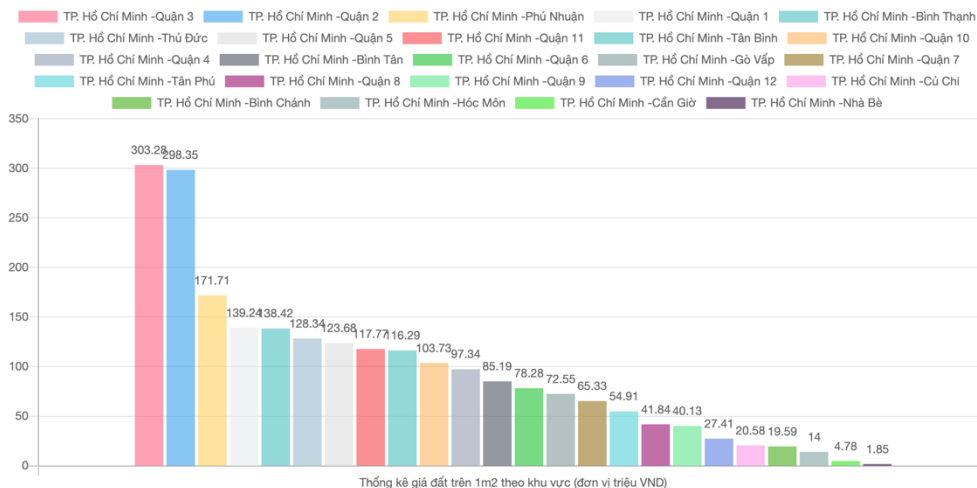
- Xây dựng được trang web hiển thị dữ liệu bất động sản đã crawl được, kèm theo công cụ tìm kiếm và bộ lọc:
 - Tìm kiếm theo tiêu đề.
 - Tìm kiếm theo khu vực.
 - Lọc theo diện tích.



Hình 7: Giao diện trang chủ với công cụ tìm kiếm và bộ lọc

- Xây dựng được trang web hiển thị biểu đồ thống kê giá nhà đất trong các khu vực trong tỉnh, kèm theo bộ lọc hiển thị theo tỉnh.

Chọn Tỉnh/Thành phố
TP. Hồ Chí Minh



Hình 8: Giao diện thống kê dữ liệu nhà đất các khu vực trong tỉnh

4 Hướng dẫn sử dụng

B2 BDS Charts 3

Tìm kiếm 1

Lọc theo diện tích

Items per page: 10
1 - 10 of 9919
2
<
>

Vị trí: TP. Hồ Chí Minh - Quận 2

Bán Gấp Căn Hộ 3 Pn Quận 2

Diện tích: **86m²**

₫3,100,000,000

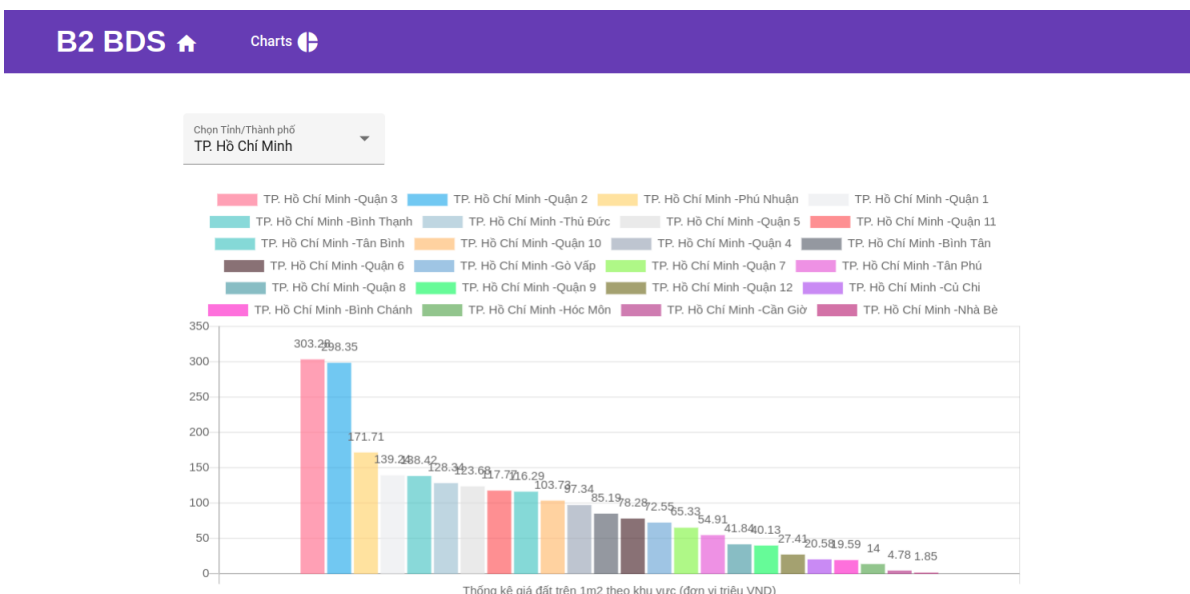
Vị trí: TP. Hồ Chí Minh - Hóc Môn

Bán Nhà Xuân Thới Thượng, Hóc Môn (Chính Chủ)

Diện tích: **18m²**

₫650,000,000

- Chú thích:
 - Số 1: Bộ lọc và tìm kiếm, dùng để tìm kiếm bất động sản theo tiêu đề, khu vực và trong khoảng diện tích nhất định.
 - Số 2: Paging dùng để xem các trang tiếp theo.
 - Số 3: Xem biểu đồ.



- Ở trang biểu đồ, ta có thể chọn Tỉnh/Thành phố tương ứng để xem các thông tin về bất động sản của các quận/huyện thuộc tỉnh đó.



5 Triển khai

5.1 Backend

- Nền tảng sử dụng: Heroku
- Link api backend sau khi deploy: <http://daktlt.herokuapp.com/>
- Truy cập api: getAll



- Truy cập api getThongKe



5.2 Frontend

- Nền tảng sử dụng: Firebase
- Url sau khi deploy: <https://ktlt-fe-angular.web.app/>
- Giao diện trang chủ:

B2 BDS Charts

Tìm kiếm

Tìm theo tiêu đề


Tìm theo khu vực

Lọc theo diện tích

Từ

Đến

Items per page: 10 1 - 10 of 9919




Vị trí: TP. Hồ Chí Minh - Quận 2

Bán Gấp Căn Hộ 3 Pn Quận 2

Diện tích: **86m²**

₫3,100,000,000




Vị trí: TP. Hồ Chí Minh - Hóc Môn

Bán Nhà Xuân Thới Thượng, Hóc Môn (Chính Chủ)

Diện tích: **18m²**

₫650,000,000



Vị trí: Bình Thuận - Bắc Bình

Cần Bán Lô Đất Vườn 10.000 Gàn Nhánh Rẻ Cao Tốc Ph...

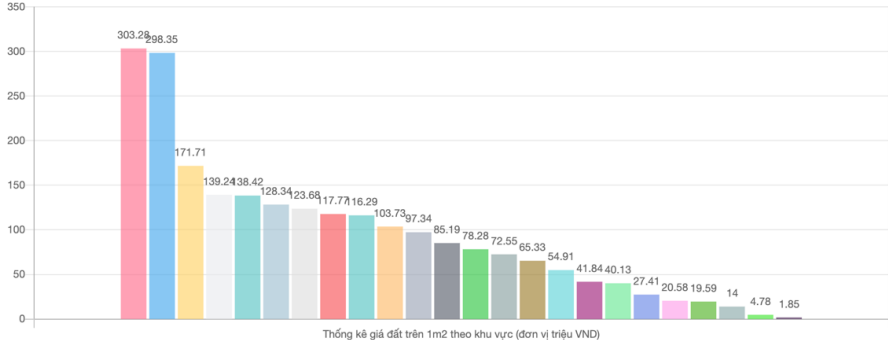
Diện tích: **11000m²**

- Giao diện trang thống kê:

B2 BDS Charts

Chọn Tỉnh/Thành phố
TP. Hồ Chí Minh

■ TP. Hồ Chí Minh - Quận 3
 ■ TP. Hồ Chí Minh - Quận 2
 ■ TP. Hồ Chí Minh - Phú Nhuận
 ■ TP. Hồ Chí Minh - Quận 1
 ■ TP. Hồ Chí Minh - Bình Thạnh
 ■ TP. Hồ Chí Minh - Thủ Đức
 ■ TP. Hồ Chí Minh - Quận 5
 ■ TP. Hồ Chí Minh - Quận 11
 ■ TP. Hồ Chí Minh - Tân Bình
 ■ TP. Hồ Chí Minh - Quận 10
 ■ TP. Hồ Chí Minh - Quận 4
 ■ TP. Hồ Chí Minh - Bình Tân
 ■ TP. Hồ Chí Minh - Quận 6
 ■ TP. Hồ Chí Minh - Gò Vấp
 ■ TP. Hồ Chí Minh - Quận 7
 ■ TP. Hồ Chí Minh - Tân Phú
 ■ TP. Hồ Chí Minh - Quận 8
 ■ TP. Hồ Chí Minh - Quận 9
 ■ TP. Hồ Chí Minh - Quận 12
 ■ TP. Hồ Chí Minh - Củ Chi
 ■ TP. Hồ Chí Minh - Bình Chánh
 ■ TP. Hồ Chí Minh - Hóc Môn
 ■ TP. Hồ Chí Minh - Cần Giờ
 ■ TP. Hồ Chí Minh - Nhà Bè



Thống kê giá đất trên 1m² theo khu vực (đơn vị triệu VND)

Khu vực	Giá đất (triệu VND)
TP. Hồ Chí Minh - Quận 3	303.28
TP. Hồ Chí Minh - Quận 2	298.35
TP. Hồ Chí Minh - Phú Nhuận	171.71
TP. Hồ Chí Minh - Quận 1	139.24
TP. Hồ Chí Minh - Bình Thạnh	138.42
TP. Hồ Chí Minh - Thủ Đức	128.34
TP. Hồ Chí Minh - Quận 5	23.68
TP. Hồ Chí Minh - Quận 11	17.77
TP. Hồ Chí Minh - Tân Bình	16.29
TP. Hồ Chí Minh - Quận 6	103.73
TP. Hồ Chí Minh - Gò Vấp	97.34
TP. Hồ Chí Minh - Quận 7	85.19
TP. Hồ Chí Minh - Tân Phú	78.28
TP. Hồ Chí Minh - Quận 8	72.55
TP. Hồ Chí Minh - Quận 9	65.33
TP. Hồ Chí Minh - Quận 12	54.91
TP. Hồ Chí Minh - Củ Chi	41.84
TP. Hồ Chí Minh - Bình Chánh	40.13
TP. Hồ Chí Minh - Hóc Môn	27.41
TP. Hồ Chí Minh - Cần Giờ	20.58
TP. Hồ Chí Minh - Nhà Bè	19.59
TP. Hồ Chí Minh - Quận 4	14
TP. Hồ Chí Minh - Quận 10	4.78
TP. Hồ Chí Minh - Quận 1	1.85

Đồ án Kỹ thuật lập trình - Niên khóa 2020-2021

Trang 13/17

6 Báo cáo hàng tuần

6.1 Báo cáo tuần 1

- Thời gian: 22/11 – 29/11/2020
- Danh sách nhóm:
 - 1920058 Lê Tất Thiện
 - 1927038 Trần Phương Tĩnh
 - 1920068 Dương Quang Tuấn
 - 1927040 Nguyễn Đức Tuấn
- Các công việc đã thực hiện
 - Setup backend, hiện thực crawl dữ liệu bằng selenium và cheerio, làm module xuất dữ liệu ra file csv, crawl 400 record từ batdongsan.com (Thiện), đã push code lên repository DAKTLT-Backend.
 - Crawl dữ liệu trên trang homedy.com, viết hàm convert thời gian sử dụng cho các trang không ghi rõ định dạng ngày (Tĩnh), đã push code lên repository DAKTLT-Backend.
 - Crawl dữ liệu trên trang batdongsan321.com (Tuấn Dương), đã push code lên repository DAKTLT-Backend.
 - Crawl dữ liệu trên trang 123nhadat.vn (Tuấn Nguyễn), đã push code lên repository DAKTLT-Backend.
- Các khó khăn gặp phải
 - Ban đầu nhóm sử dụng Selenium để tiến hành crawl dữ liệu, nhưng khi sử dụng vòng lặp với selenium thì có vấn đề với bất đồng bộ, kết quả chạy không được như mong muốn.
 - Nhóm chuyển qua sử dụng Cheerio và nhận thấy thời gian crawl tăng đáng kể và khắc phục được vấn đề trên.
- Dự kiến công việc tuần tới
 - Triển khai Frontend để thuận tiện cho việc crawl các dữ liệu cần thiết.
 - Sử dụng MongoDB để lưu trữ và GraphQL để thao tác và truy vấn API.
 - Hoàn thiện module xuất file csv.

6.2 Báo cáo tuần 2

- Thời gian: 30/11 – 6/12/2020
- Danh sách nhóm:
 - 1920058 Lê Tất Thiện
 - 1927038 Trần Phương Tĩnh
 - 1920068 Dương Quang Tuấn
 - 1927040 Nguyễn Đức Tuấn
- Các công việc đã thực hiện
 - Viết prototype function crawlWeb (Thiện), đã update vào file module/tinh.js trong repository DAKTLT-Backend.
 - Hoàn thiện function crawlWeb và chạy thử (Tĩnh), đã update vào file module/tinh.js trong repository DAKTLT-Backend.
 - Tìm hiểu format chuẩn, cách làm sạch data set (Tuấn Nguyễn).
 - Tuần 1 đã crawl 10k record và gửi cho Tĩnh (Tuấn Dương), tuần 2 làm sạch và đưa về format chuẩn mong muốn (Tĩnh, Thiện), đã push file 10kRecords.json lên folder samplerecord trong repository DAKTLT-Backend.

- Tìm hiểu và setup frontend (Tuấn Dương, Tuấn Nguyễn), đã push lên repository DAKTLT-Frontend.
- Kết nối backend với mongoDB, insert 10k record vào mongoDB và tạo các model cho web (Thiện), đã push lên repository DAKTLT-Backend.
- Viết báo cáo tuần 2 (Tĩnh), đã push báo cáo lên folder report/tuan2 trong repository DAKTLT-Backend.

- **Các khó khăn gặp phải**

- Trang web <https://www.muabannhadat.com.vn/> hiện đang bị lỗi nên không thể crawl dữ liệu về.
- 10k record được crawl về ở tuần 1 chưa đủ thuộc tính, chưa sạch và chưa đúng format chuẩn mong muốn, ở tuần 2 này nhóm đã viết thêm function crawlWeb để crawl trực tiếp từng mẫu tin của trang chính, địa chỉ để crawl lấy từ 10k record cũ. Từ đây crawl được 10k record mới đủ thuộc tính, sạch, và đúng format chuẩn mong muốn, lưu vào file json, sau đó insert vào mongoDB..

- **Dự kiến công việc tuần tới**

- Sử dụng MongoDB để lưu trữ và graphql để thao tác và truy vấn API.
- Backend: Viết API trả về dữ liệu.
- Frontend: Hiển thị dữ liệu.

6.3 Báo cáo tuần 3

- **Thời gian: 7/12 – 13/12/2020**

- **Danh sách nhóm:**

- 1920058 Lê Tất Thiện
- 1927038 Trần Phương Tĩnh
- 1920068 Dương Quang Tuấn
- 1927040 Nguyễn Đức Tuấn

- **Các công việc đã thực hiện**

- Hiện thực function searchLocation dùng để tìm kiếm theo địa điểm (Thiện), đã push code lên repository DAKTLT-Backend, branch master.
- Hiện thực function searchArea dùng để tìm kiếm diện tích trong đoạn mong muốn (Thiện), đã push code lên repository DAKTLT-Backend, branch master.
- Hiện thực function getall dùng để lấy tất cả tin (Tĩnh), đã push code lên repository DAKTLT-Backend, branch tinh.
- Hiện thực function searchTitle dùng để tìm kiếm theo tựa đề (Tĩnh), đã push code lên repository DAKTLT-Backend, branch tinh.
- Dùng api của Thiện và Tĩnh viết để render data ra Frontend (Tuấn Dương + Tuấn Nguyễn), đã push code lên repository DAKTLT-Frontend.
- Viết báo cáo tuần 3 (Tĩnh), đã push báo cáo lên folder report/tuan3 trong repository DAKTLT-Backend, branch master.

- **Các khó khăn gặp phải**

- Khó khăn trong việc xử lý hơn 10k dữ liệu một cách hợp lý và trả về data nhanh; khó khăn khi làm phân trang. Đã khắc phục bằng cách lấy hết về rồi đếm phần tử.

- **Dự kiến công việc tuần tới**

- Backend: Hoàn thiện các api.
- Frontend: Hoàn thiện render data.
- Thống kê dữ liệu.

6.4 Báo cáo tuần 4

- **Thời gian:** 14/12 – 20/12/2020
- **Danh sách nhóm:**
 - 1920058 Lê Tất Thiện
 - 1927038 Trần Phương Tĩnh
 - 1920068 Dương Quang Tuấn
 - 1927040 Nguyễn Đức Tuấn
- **Các công việc đã thực hiện**
 - Hiện thực api cleanAddress, fixAddress, findnotadd, batdongsan321, batdongsan (Tĩnh) để clean address, đã push code lên repository DAKTLT-Backend, branch tinh.
 - Hiện thực api fixpricebds, thongketheoquan, fixthongke (Thiện) để clean price và thống kê theo quận, đã push code lên repository DAKTLT-Backend, branch master.
 - Render thống kê ra Frontend (Tuấn Dương), đã push code lên repository DAKTLT-Frontend.
 - Viết báo cáo tuần 4 (Tĩnh), đã push báo cáo lên folder report/tuan4 trong repository DAKTLT-Backend, branch master.
- **Các khó khăn gặp phải**
 - Seller ghi địa chỉ không thống nhất (ví dụ lúc thì Hồ Chí Minh lúc thì TP. Hồ Chí Minh,...), đã xử lý cho thống nhất.
 - Seller ghi giá ảo (ví dụ 2333333333 tỷ đồng hoặc 1000 đồng,...), giá không thống nhất (ví dụ lúc thì giá theo mét vuông, lúc thì giá tổng,...), đã xử lý lại cho đúng.
- **Dự kiến công việc tuần tới**
 - Hoàn thiện cũng như sửa các lỗi nhỏ.
 - Bổ sung tính năng cần thiết.

6.5 Báo cáo tuần 5

- **Thời gian:** 21/12 – 27/12/2020
- **Danh sách nhóm:**
 - 1920058 Lê Tất Thiện
 - 1927038 Trần Phương Tĩnh
 - 1920068 Dương Quang Tuấn
 - 1927040 Nguyễn Đức Tuấn
- **Các công việc đã thực hiện**
 - Gộp api get all + tìm theo title, theo location, theo area vào làm một (Thiện, Tĩnh), đã push code lên repository DAKTLT-Backend, branch master.
 - Fix paging (Tuấn Dương), đã push code lên repository DAKTLT-Backend, branch master.
 - Display products (Tuấn Dương), đã push code lên repository DAKTLT-Frontend.
 - Add search to side bar, Done filter (Tuấn Nguyễn), đã push code lên repository DAKTLT-Frontend.
 - Viết báo cáo tuần 5 (Tĩnh), đã push báo cáo lên folder report/tuan5 trong repository DAKTLT-Backend, branch master.
- **Các khó khăn gặp phải**
 - Các khó khăn đã được giải quyết.
- **Dự kiến công việc tuần tới**
 - Đã hoàn thành project.

7 Tổng kết và hướng phát triển cho đề tài

7.1 Tổng kết

Với đề tài: Build a simple web crawler, nhóm đã xây dựng được một công cụ có thể crawl được gần 10000 dữ liệu từ một trang bất động sản, tuy dữ liệu thu được vẫn còn ít và chưa được xử lý tốt, song nếu có thời gian phát triển và hoàn thiện hơn thì công cụ này sẽ có ích trong việc thu thập dữ liệu mẫu (Dataset) để triển khai các mô hình Machine Learning, Deep Learning hoặc cơ bản hơn là thu thập dữ liệu làm dữ liệu mẫu cho một ứng dụng, hệ thống đã xây dựng.

7.2 Hướng phát triển cho đề tài

- Tiếp tục cải tiến để công cụ có khả năng crawl thêm dữ liệu từ nhiều website, từ đó sẽ có thêm nhiều thông tin phục vụ cho việc so sánh giá giữa các khu vực, đồng thời tìm ra được các khu vực có giá tốt.
- Thông báo cho người dùng khi giá nhà đất trong một khu vực nào đó tăng hoặc giảm đột biến.
- Cải tiến bộ lọc và công cụ tìm kiếm để lọc ra những thông tin phù hợp nhất với nhu cầu của người dùng.

8 Tài liệu tham khảo

- Slide bài giảng Kỹ thuật lập trình trên BKeL.
- <https://towardsdatascience.com/how-to-build-a-simple-web-crawler-66082fc82470>
- <https://scrapy.org/>
- <https://viblo.asia/p/lan-dau-tien-crawl-du-lieu-cua-toi-nhu-the-nao-RnB5p78JlPG>
- https://www.youtube.com/watch?v=hkF_oIm3lU4&t=2197s