## 2. Thu thập và mô tả dữ liệu

### 2.1. Thu thập dữ liệu

#### a. Nguồn dữ liệu

* Dữ liệu được thu thập trên trang: http://stackoverflow.itoa.vn.

#### b. Công cụ

* Môi trường: Node.js.
* Thư viện: Puppeteer.js.
* Trình soạn thảo văn bản (editor): Visual Studio Code.

#### c. Dữ liệu đầu vào

* Địa chỉ URL trang web chứa nguồn dữ liệu: http://stackoverflow.itoa.vn.

#### d. Dữ liệu đầu ra

* Tập tin chứa toàn bộ dữ liệu thu thập được trên trang web, có định dạng file comma-separated values (CSV).

#### e. Ví dụ minh họa

##### i) Mô tả quá trình cấu trúc JSON từ một URL đầu vào

##### 

URL

HTML

JSON

puppeteer.js

HTML DOM

**Hình 1. Sơ đồ khối quy trình cấu trúc ra JSON từ một URL.**

* Sử dụng thư viện puppeteer.js để lấy HTML từ một URL.
* Thông qua puppeteer.js, chạy javascript can thiệp vào HTML DOM để lấy dữ liệu.

##### ii) Mô tả cách thức tổng hợp dữ liệu

* Dùng puppeteer.js duyệt qua tất cả các URL. Theo bước (i) lấy kết quả và ghi JSON có được vào CSV.

URL

HTML

JSON

CSV

HTML

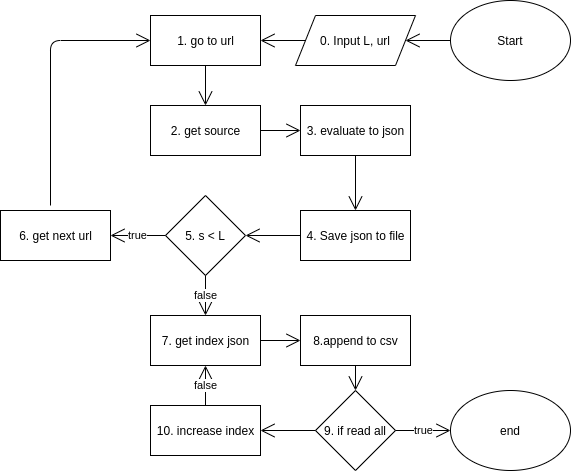
HTML

JSON

JSON

**Hình 2. Sơ đồ khối quy trình tổng hợp dữ liệu.**

##### iii) Tổng quan thuật toán



**Hình 3. Sơ đồ khối tổng hợp 2 quá tình trên.**

Mô tả thuật toán:

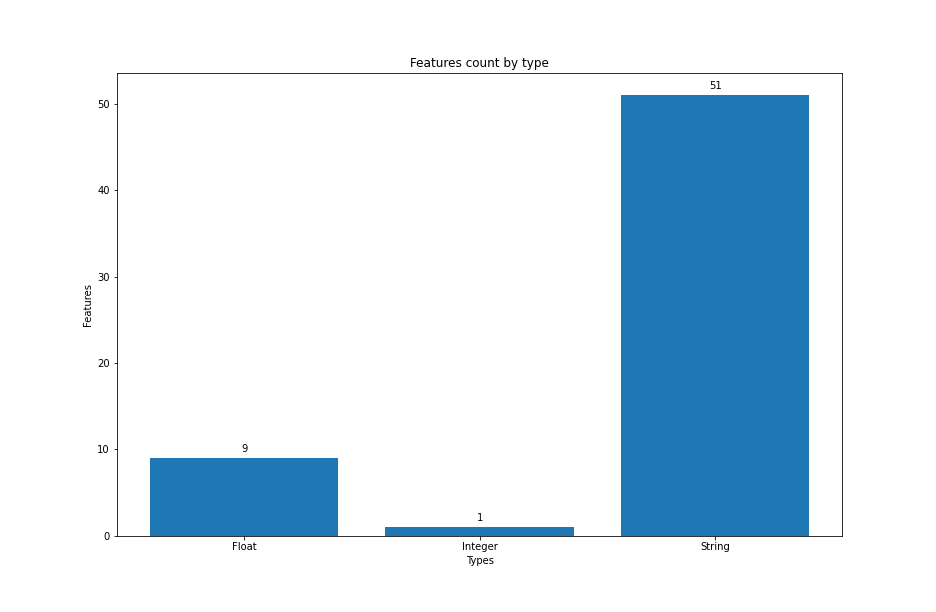
1. Mở kết nối đến trang (ban đầu là http://stackoverflow.itoa.vn).
2. Nhận source HTML từ phản hồi.
3. Nhận dạng, đánh giá HTML để đưa các thông tin cần thiết vào một JSON.
4. Lưu JSON vừa nhận được ra file.
5. Kiểm tra số lượng file (số mẫu) còn thiếu hay không?.
6. Nếu còn thiếu mẫu, tìm URL trang tiếp theo và quay lại bước 1.
7. Nếu đã đủ, đọc dữ liệu từ file JSON đang trỏ (con trỏ bắt đầu từ 0).
8. Điền dữ liệu từ file JSON thành một row của file CSV.
9. Kiểm tra file JSON đang trỏ có phải là file cuối hay không.
10. Nếu không, tăng con trỏ và quay lại bước 7.
11. Nếu đúng, kết thúc.

### 2.2. Mô tả dữ liệu

* Bộ dữ liệu thu thập được có số đặc trưng là 61 đặc trưng, số mẫu là 64,000 mẫu. Kiểu dữ liệu của mỗi đặc trưng được cho theo bảng dưới. Biểu đồ cột biểu diễn số các đặc trưng theo kiểu dữ liệu được cho theo sơ đồ ở hình 4.

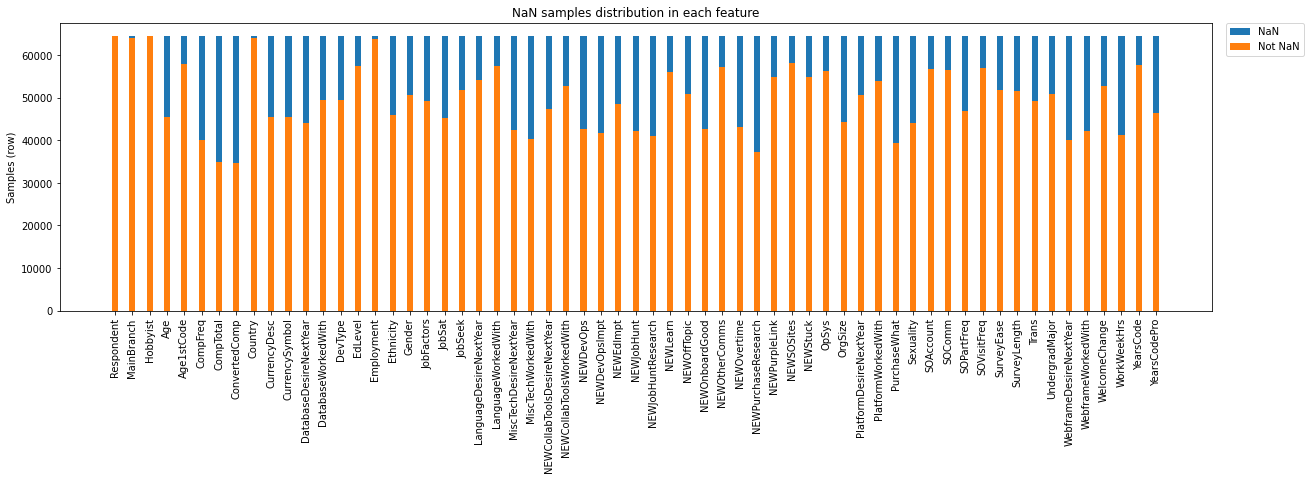
|  |  |
| --- | --- |
| Kiểu dữ liệu | Tên đặc trưng |
| Float | "Age", "CompTotal", "ConvertedComp", "NEWJobHunt", "NEWJobHuntResearch", "NEWOnboardGood", "PurchaseWhat", "SOPartFreq", "WorkWeekHrs". |
| Integer | "Respondent". |
| String | "MainBranch", "Hobbyist", "Age1stCode", "CompFreq", "Country", "CurrencyDesc", "CurrencySymbol", "DatabaseDesireNextYear", "DatabaseWorkedWith", "DevType", "EdLevel", "Employment", "Ethnicity", "Gender", "JobFactors", "JobSat", "JobSeek", "LanguageDesireNextYear", "LanguageWorkedWith", "MiscTechDesireNextYear", "MiscTechWorkedWith", "NEWCollabToolsDesireNextYear", "NEWCollabToolsWorkedWith", "NEWDevOps", "NEWDevOpsImpt", "NEWEdImpt", "NEWLearn", "NEWOffTopic", "NEWOtherComms", "NEWOvertime", "NEWPurchaseResearch", "NEWPurpleLink", "NEWSOSites", "NEWStuck", "OpSys", "OrgSize", "PlatformDesireNextYear", "PlatformWorkedWith", "Sexuality", "SOAccount", "SOComm", "SOVisitFreq", "SurveyEase", "SurveyLength", "Trans", "UndergradMajor", "WebframeDesireNextYear", "WebframeWorkedWith", "WelcomeChange", "YearsCode", "YearsCodePro". |

**Bảng 1. Bảng mô tả kiểu dữ liệu của các đặc trưng.**



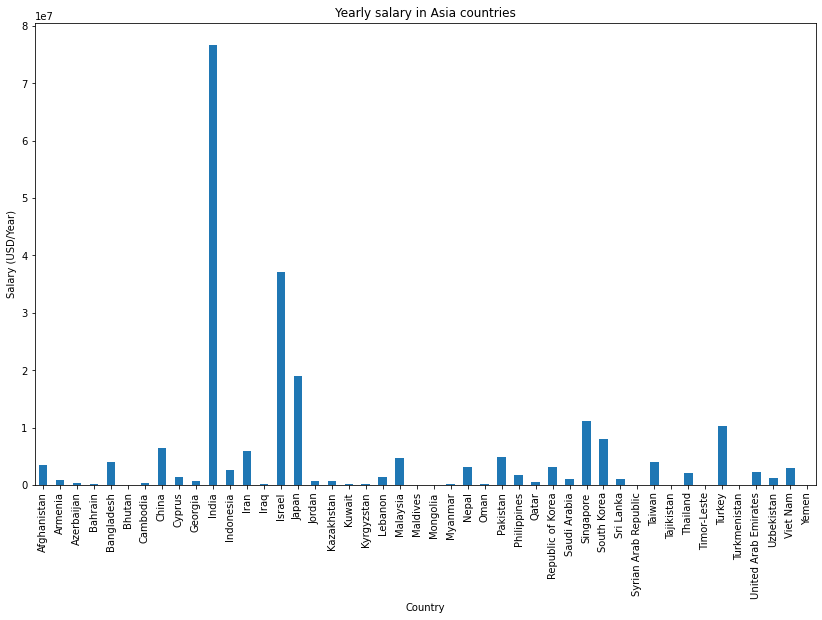
**Hình 4. Biểu đồ cột số đặc trưng theo kiểu dữ liệu.**

* Số mẫu dữ liệu trống của mỗi đặc trưng được mô tả bằng biểu đồ cột dưới đây.

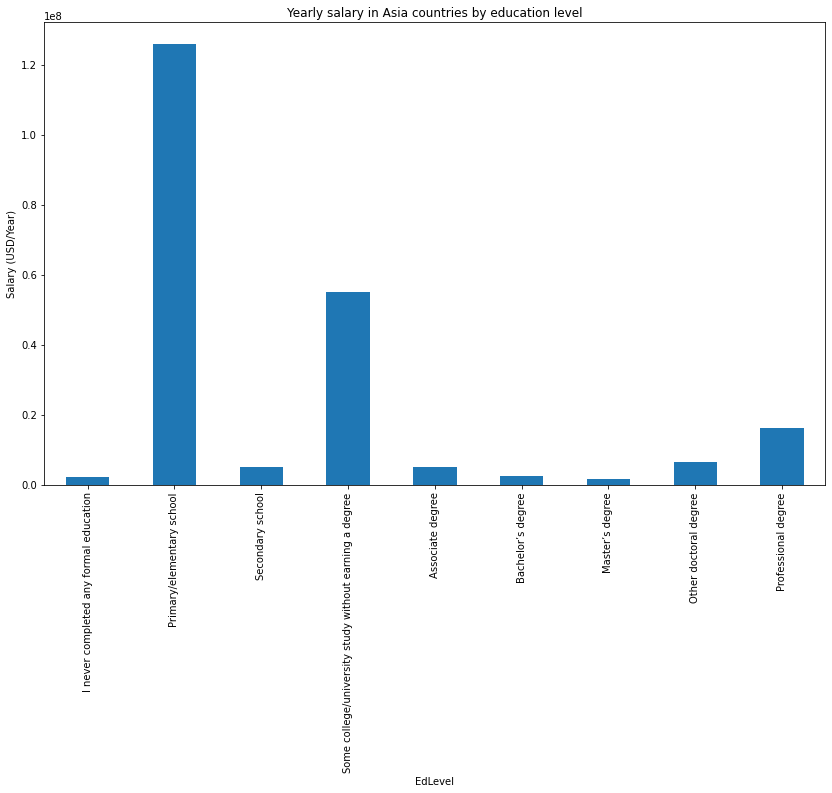


**Hình 5. Biểu đồ cột số mẫu dữ liệu trống trên tổng số mẫu của các đặc trưng.**

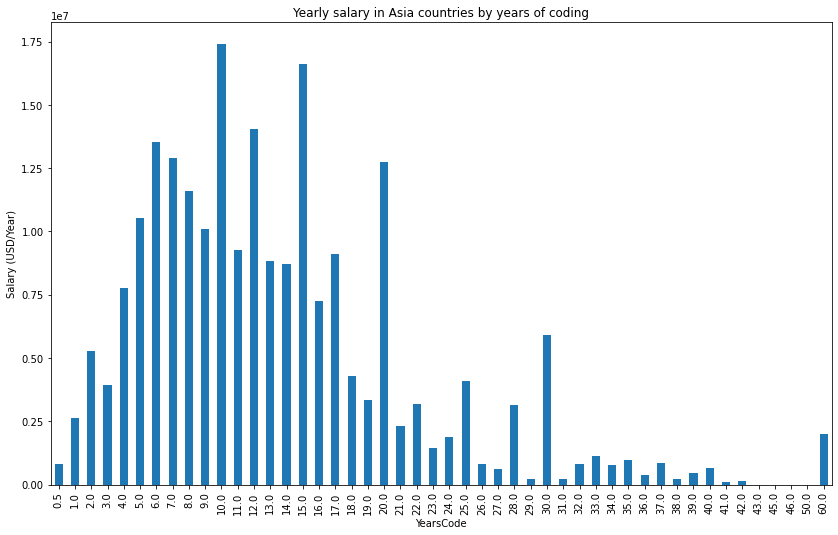
* Các thông kê mô tả trực quan về một vài đặc trưng được mô tả bằng cách biểu đồ dưới đây.



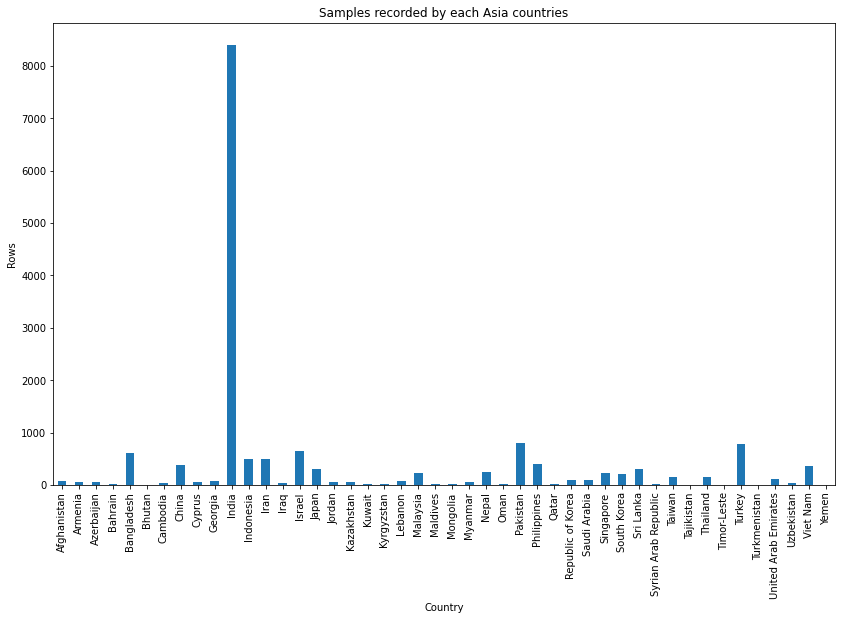
Hình . Biểu đồ cột thể hiện mức lương hằng năm giữa các quốc gia ở Châu Á.

****

**Hình 6. Biểu đồ cột thể hiện quan hệ giữa mức lương và mức học vấn ở Châu Á.**

****

Hình . Biểu đồ cột thể hiện quan hệ giữa mức lương và số năm kinh nghiệm ở Châu Á.



Hình . Biểu đồ cột thể hiện số lượng mẫu theo các quốc gia ở Châu Á.