TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**BÀI TẬP GIỮA KỲ**

**MÔN XỬ LÝ DỮ LIỆU LỚN**

*Người thực hiện:* **NGUYỄN QUỐC CƯỜNG – 518H0003**

*Lớp*: **18H50203**

*Khoá*: **22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**BÀI TẬP GIỮA KỲ**

**MÔN XỬ LÝ DỮ LIỆU LỚN**

*Người thực hiện:* **NGUYỄN QUỐC CƯỜNG – 518H0003**

*Lớp*: **18H50203**

*Khoá*: **22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**MỤC LỤC**

[I. THU THẬP DỮ LIỆU 5](#_Toc87256699)

[1. Cào dữ liệu từ trang web, lưu kết quả vào file và mô tả cấu trúc trang web: 5](#_Toc87256700)

[- Cấu trúc trang web: 6](#_Toc87256701)

[2. Thực hiện các yêu cầu sau khi cào dữ liệu: 7](#_Toc87256702)

[1. Hãy đọc tất cả các thẻ html (div) với lớp là "quote" và lưu nó trong biến 'result’, hiển thị giá trị biến 'result’ ra màn hình 7](#_Toc87256703)

[2. Hãy tìm trong biến 'result’ vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" và in kết quả ra màn hình 7](#_Toc87256704)

[3. Hãy viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung 7](#_Toc87256705)

[4. Hãy lưu kết quả ở câu c vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu. Bạn được yêu cầu thu thập ít nhất 40 câu nói nổi tiếng từ trang web trên một cách tự động theo code của các ý trên 7](#_Toc87256706)

[II. TỰ ĐÁNH GIÁ 8](#_Toc87256707)

[III. TÀI LIỆU THAM KHẢO 9](#_Toc87256708)

# THU THẬP DỮ LIỆU

## Cào dữ liệu từ trang web, lưu kết quả vào file và mô tả cấu trúc trang web:

* Code cào dữ liệu:

Text, letter

Description automatically generated

Text

Description automatically generated with medium confidence

## Cấu trúc trang web:

## Thực hiện các yêu cầu sau khi cào dữ liệu:

### Text Description automatically generatedHãy đọc tất cả các thẻ html (div) với lớp là "quote" và lưu nó trong biến 'result’, hiển thị giá trị biến 'result’ ra màn hình

### Text Description automatically generatedHãy tìm trong biến 'result’ vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" và in kết quả ra màn hình

### Text Description automatically generatedHãy viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung

Text, letter

Description automatically generated

### Hãy lưu kết quả ở câu c vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu. Bạn được yêu cầu thu thập ít nhất 40 câu nói nổi tiếng từ trang web trên một cách tự động theo code của các ý trên

# TỰ ĐÁNH GIÁ

|  |  |  |  |
| --- | --- | --- | --- |
| **Bài tập** | **Nội dung** | **Điểm** | **Ghi chú** |
| **Bài 1**  **(5 điểm)** | **Mô tả cấu trúc dữ liệu (1 đ)** | 1 đ |  |
| **Giải thuật BFS (2 đ)** | 1.75 đ |  |
| Sơ đồ giải thuật (0.5 đ) | 0.5 đ |  |
| Hiện thực (1 đ) | 1 đ |  |
| Kết quả và thảo luận (0.5 đ) | 0.25 đ |  |
| **Giải thuật UCS (2 đ)** | 1.75 đ |  |
| Sơ đồ giải thuật (0.5 đ) | 0.5 đ |  |
| Hiện thực (1 đ) | 1 đ |  |
| Kết quả và thảo luận (0.5 đ) | 0.25 đ |  |
| **Bài 2**  **(5 điểm)** | **Mô tả cấu trúc dữ liệu (1 đ)** | 1 đ |  |
| **Giải thuật Greedy Best First Search (Heuristic) (2 đ)** | 1.75 đ |  |
| Sơ đồ giải thuật (0.5 đ) | 0.5 đ |  |
| Hiện thực (1 đ) | 1 đ |  |
| Kết quả và thảo luận (0.5 đ) | 0.25 đ |  |
| **Giải thuật A\* (2 đ)** | 1.75 đ |  |
| Sơ đồ giải thuật (0.5 đ) | 0.5 đ |  |
| Hiện thực (1 đ) | 1 đ |  |
| Kết quả và thảo luận (0.5 đ) | 0.25 đ |  |

# TÀI LIỆU THAM KHẢO

1. <https://www.javatpoint.com/ai-informed-search-algorithms>
2. <https://www.javatpoint.com/ai-uninformed-search-algorithms>
3. <https://www.geeksforgeeks.org/uniform-cost-search-dijkstra-for-large-graphs/>
4. <https://en.wikipedia.org/wiki/Breadth-first_search>
5. <https://www.geeksforgeeks.org/find-number-of-islands/>
6. <https://www.geeksforgeeks.org/best-first-search-informed-search/>
7. <https://www.geeksforgeeks.org/a-search-algorithm/>
8. Slide bài giảng:
   1. Lesson 2 - Solving problems by searching
   2. Lesson 3 - Informed Search