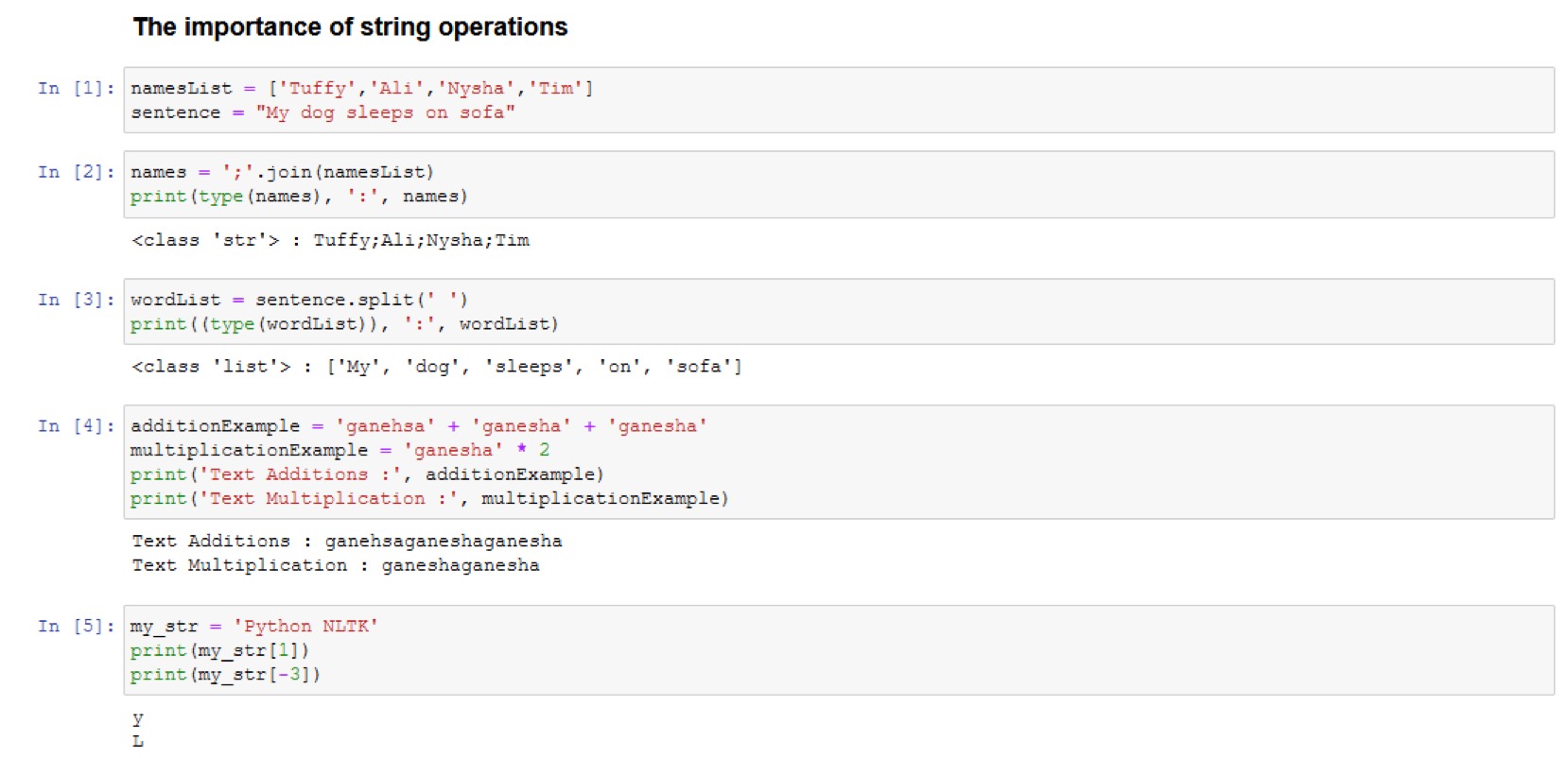
**THỰC HÀNH BUỔI 2**

**TIỀN XỬ LÝ VĂN BẢN - THAO TÁC VỚI TỆP**

|  |
| --- |
| Họ tên và MSSV:  Lớp: Ngày học: |

1. **Thực hành code**
2. **Thực hành thao tác với chuỗi (String operations)**



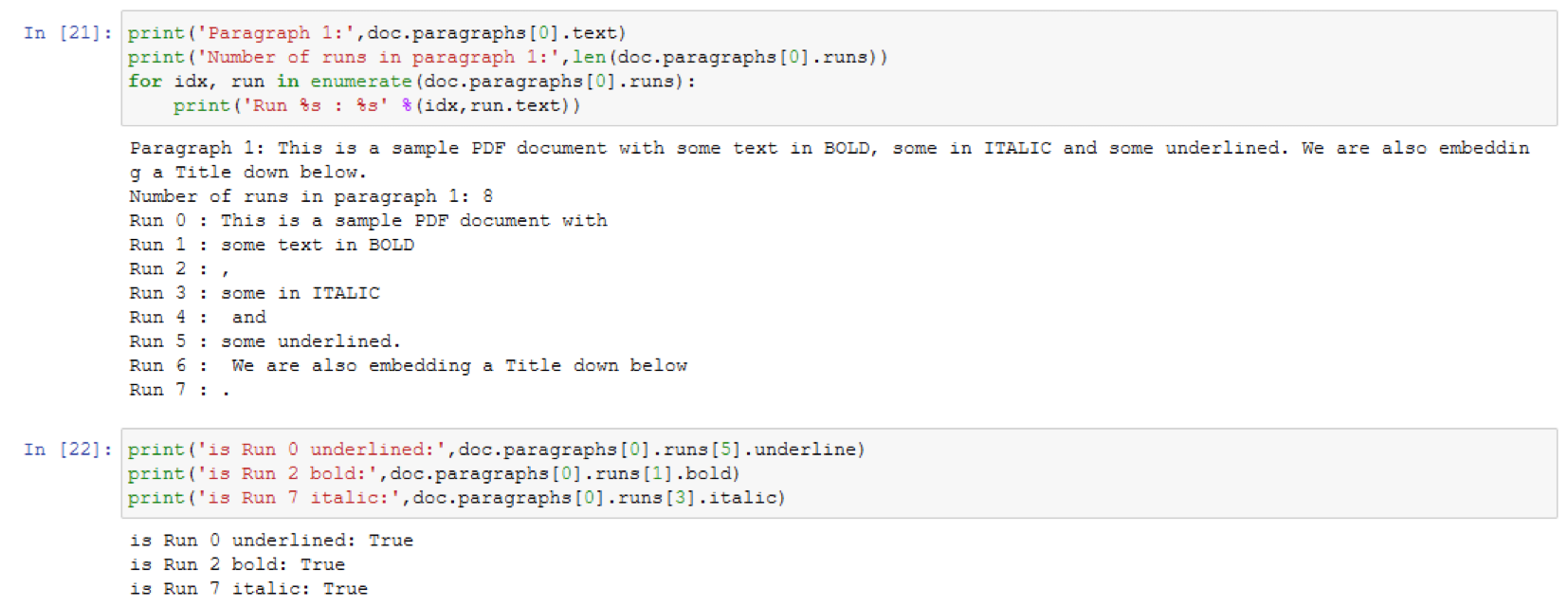
1. **Xử lý văn bản**



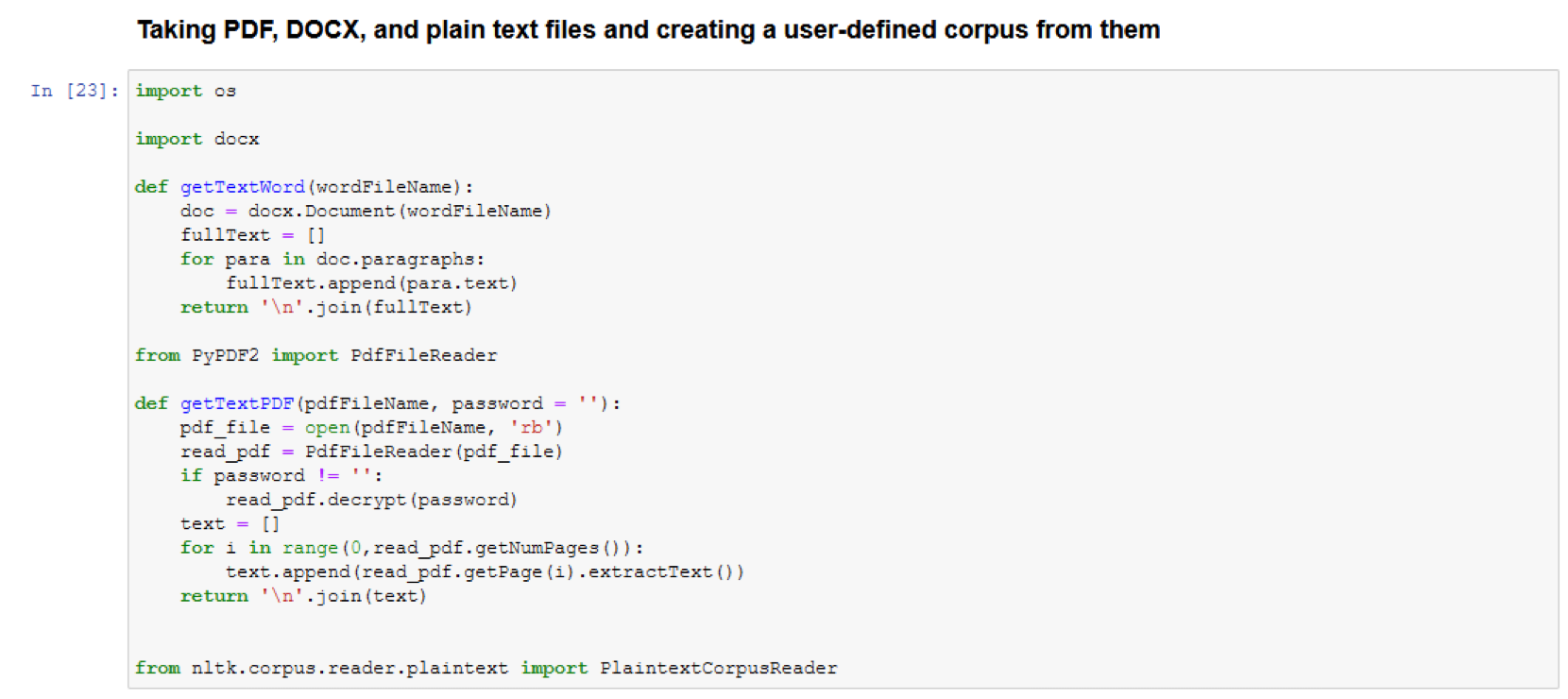
1. **Đọc tệp PDF**



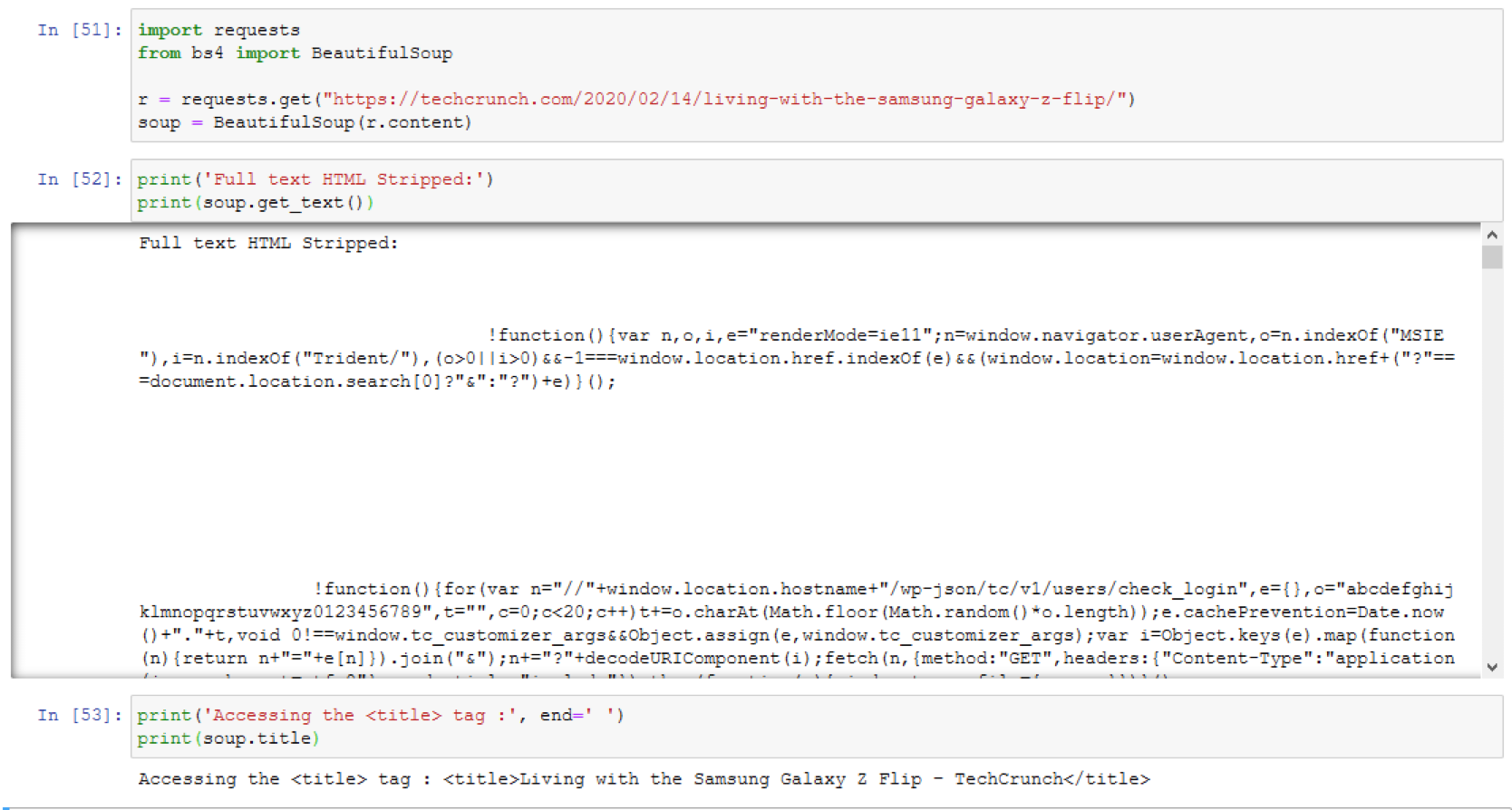
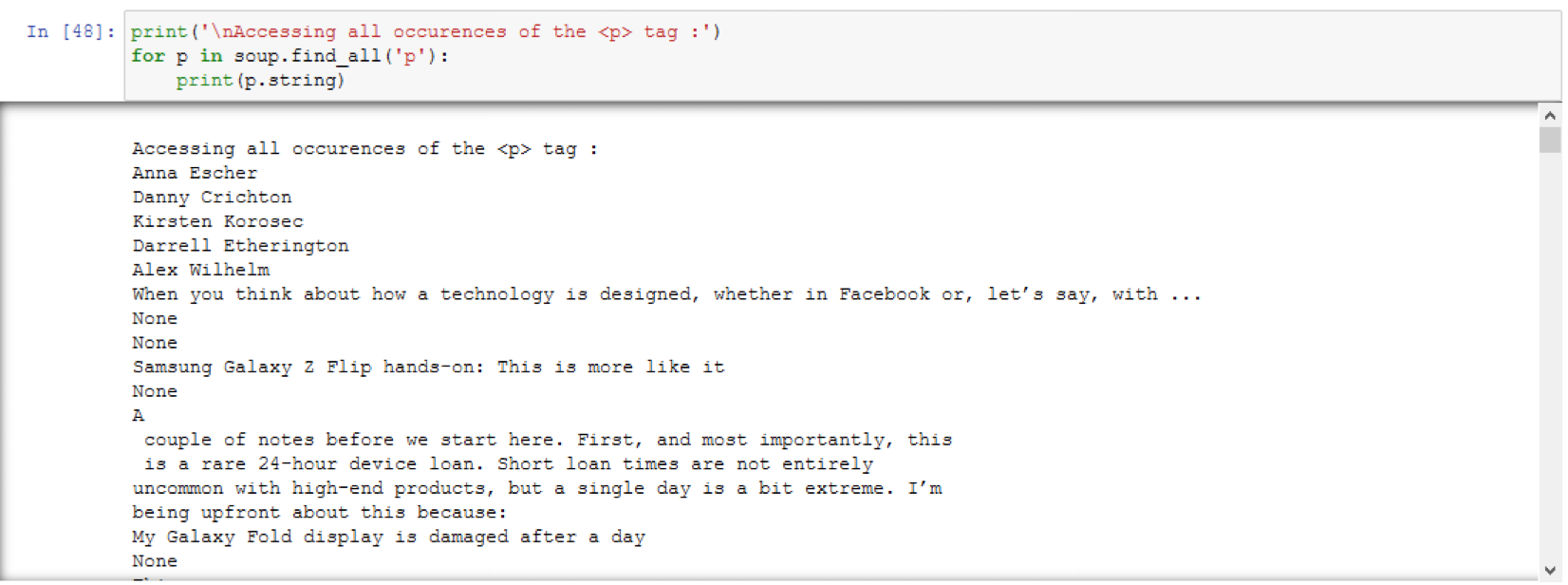
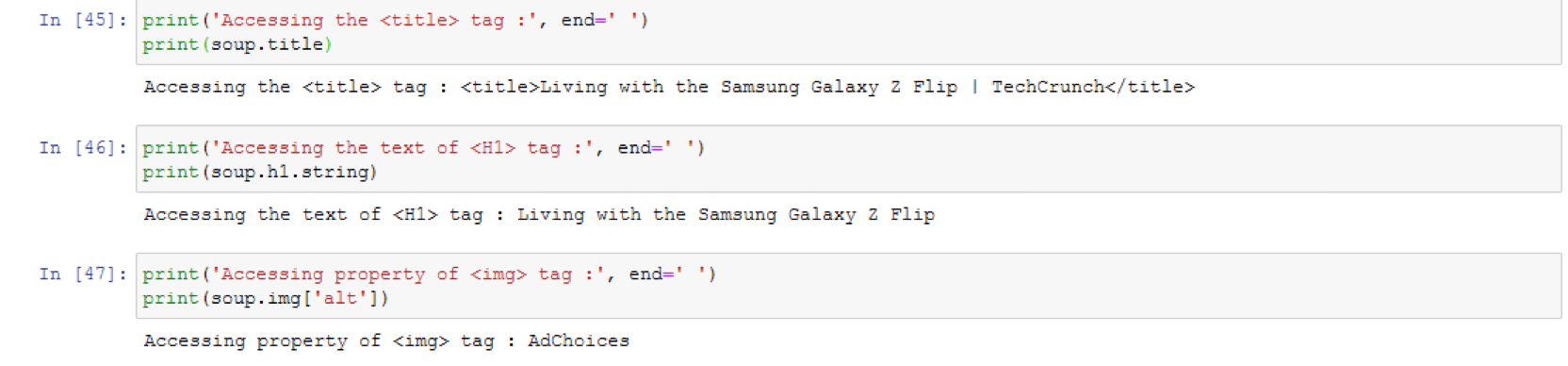
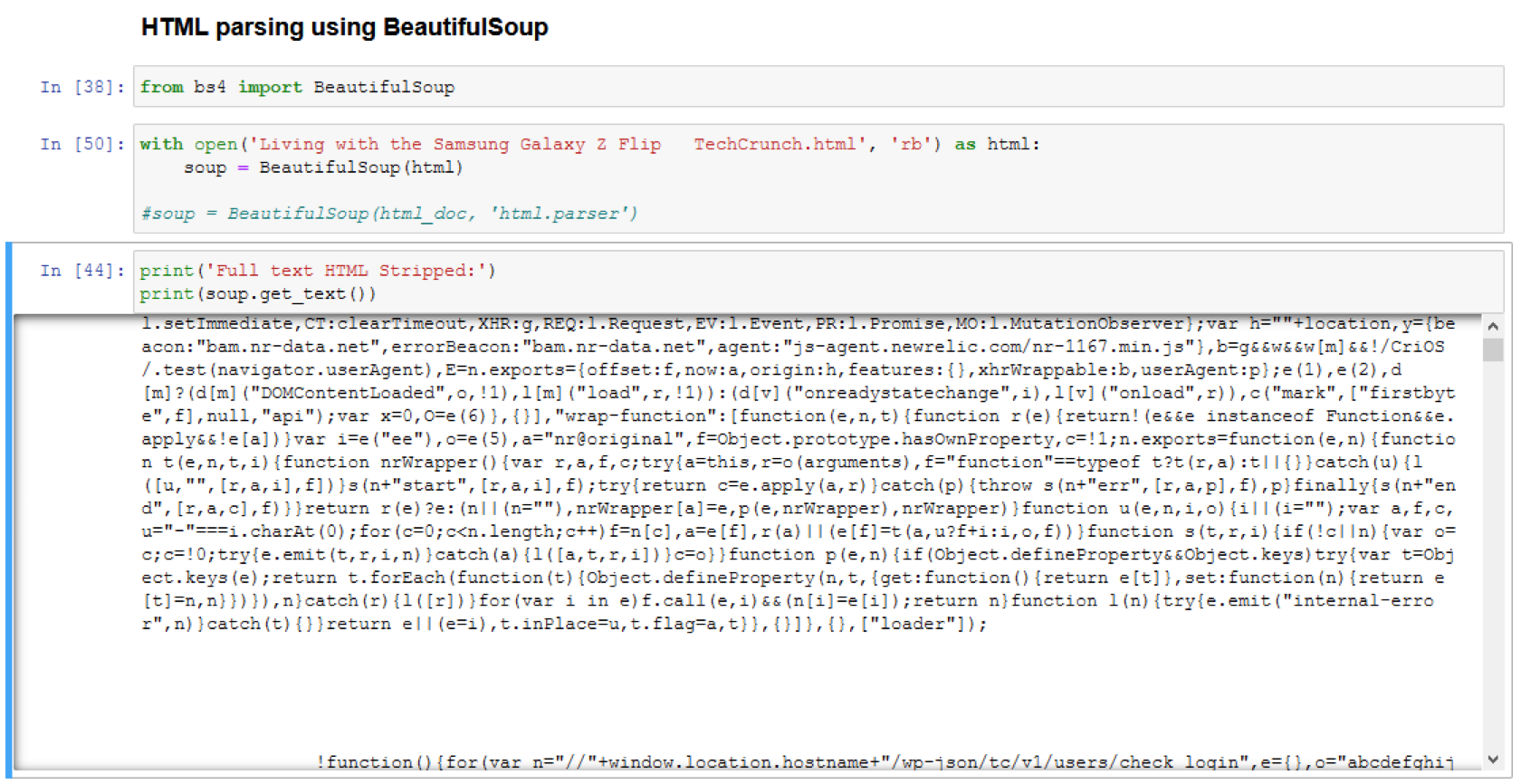
1. **Đọc tệp Word**



1. **Tạo corpus tùy chỉnh**



1. **Phân tích HTML với BeautifulSoup**



1. **Bài tập**

Sau khi thực hiện các đoạn code trên, sinh viên trả lời các câu hỏi sau:

### ****Thao tác với chuỗi (String operations)****

* Lệnh strip() có tác dụng gì trên chuỗi?
* Làm thế nào để tách một câu thành danh sách các từ?
* Nếu muốn thay thế dấu phẩy (,) trong chuỗi bằng dấu chấmxs (.), bạn sẽ làm thế nào?

1. **Đọc tệp PDF**

* Python có thể dùng thư viện nào để đọc file PDF?
* Đoạn code sau giúp đọc nội dung file PDF, nhưng có sai sót gì không?



* Nếu muốn trích xuất nội dung của trang đầu tiên trong file PDF, bạn sẽ làm gì?

### ****Đọc tệp Word****

* Thư viện nào giúp đọc file .docx trong Python?
* Nếu muốn in ra tất cả đoạn văn trong file Word, bạn sẽ dùng vòng lặp nào?
* Đoạn code sau giúp đọc file .docx, nhưng thiếu gì?



1. ****Tạo corpus tùy chỉnh****

* Corpus là gì?
* Một corpus có thể chứa những gì?
* Nếu muốn lưu danh sách các câu vào một file .txt, bạn sẽ làm như thế nào?

### ****5. Phân tích HTML với BeautifulSoup****

* Thư viện BeautifulSoup giúp làm gì?
* Lệnh soup.find\_all("a") dùng để làm gì?
* Nếu muốn lấy nội dung của thẻ <p>, bạn sẽ làm thế nào?

Sinh viên nhập câu trả lời tại đây

|  |
| --- |
| **FILE PHẢI NỘP CHO BÀI THỰC HÀNH NÀY** |
| **Một file Word** chứa đáp án các yêu cầu |

*--- Hết ---*