**REPORT DATA CRAWLER**

**Topic: Địa điểm du lịch và giải trí**

**Web crawler: tripadvisor.com**

1. **Hướng dẫn cài đặt**

- Cài đặt scrapy framework: pip install scrapy

1. **Hướng dẫn crawler data**

- Tạo project bằng comand line:

*scrapy startproject crawlerdata*

- Tạo mới file *tripadvisor.py* trong thư mục spiders

- Định nghĩa UTF-8 trong file setting.py

FEED\_EXPORT\_ENCODING = "utf-8"

- Khai báo các thuộc tính item cần crawler về:

category = scrapy.Field()

name = scrapy.Field()

rating = scrapy.Field()

reviews\_number = scrapy.Field()

rank = scrapy.Field()

address = scrapy.Field()

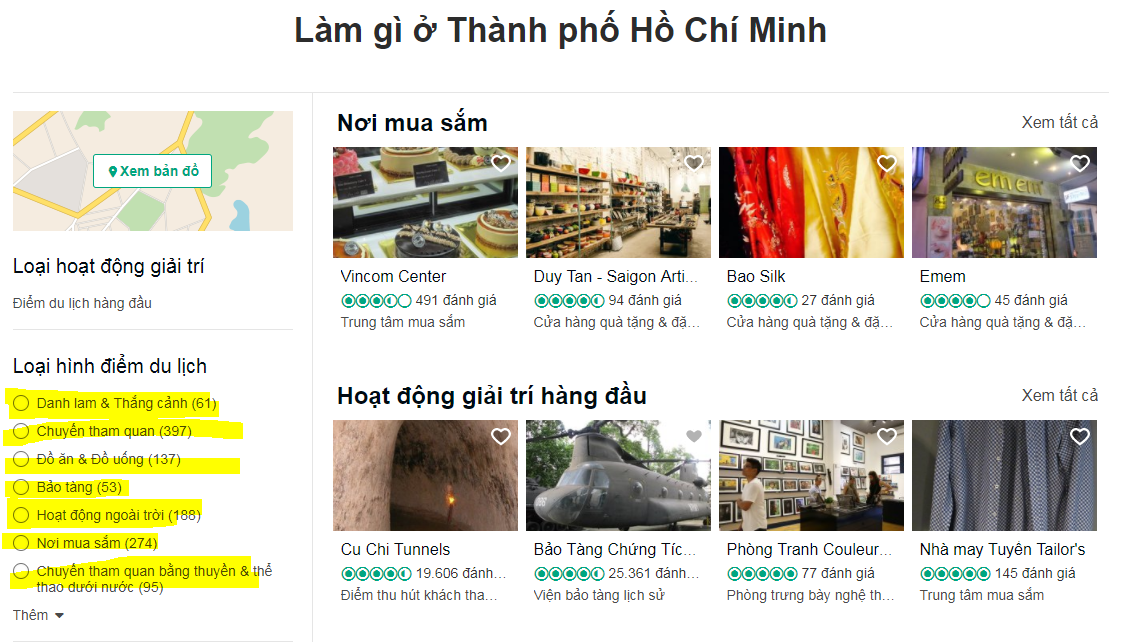
avatar = scrapy.Field()

attraction = scrapy.Field()

- Trong file *tripadvisor.py,* ta crawler về theo địa danh của từng tỉnh, với mỗi tỉnh chunsg lấy link theo từng Loại hình du lịch để chạy code

- Vào link: <https://www.tripadvisor.com.vn/Attractions>

- Tìm kiếm địa danh theo tỉnh, sau đó lấy link của từng loại hình điểm du lịch



- Parse link vào file *tripadvisor.py* và đặt lại tên Category trong mục item[‘category’] = “ ”

- Chạy lệnh để scrapy crawler data về dạng json

*scrapy crawl tripadvisor –o hochiminh.json*

- Lần lượt làm tương tự các loại hình điểm du lịch khác trong tỉnh. Và tiếp tục crawl với tỉnh khác tương tự