CS839 Data Science Spring 2018

<Group 9>

Ya-Chun Yang, yang364@wisc.edu
Han Wang, hwang729@wisc.edu
Yanghui Kang, ykang38@wisc.edu

Stage 2 Report - Restaurant Information Extraction from the Web

INTRODUCTION

In this project stage, we extracted restaurants information in Los Angeles Area from two popular crowd-sourced review website: Yelp and Tripadvisor. The data were summarized into two relational tables with common scheme that uniquely identifies each restaurant.

DATA SOURCE

The first web source is Yelp, where we searched all "Restaurants" in "Los Angeles, CA" . The web pages contain structured data of each restaurant, including name, address, phone number, price, rate. The second web source is TripAdvisor, and the search results contain structured data. Besides, we also scrapped individual restaurant page to extract more information about each restaurant.

WRAPPER CONSTRUCTION

Yelp data set

We extracted structured data directly from Yelp using Yelp Fusion API service, which provides access to the industry leading Yelp content and data. The Yelp API allows us to query information about restaurants based on city name and zip code. We requested at most 50 restaurants from each zip code and further filtered our data based on review count. The structured data could be retrieved in JSON format. We totally collected 3188 tuples from Yelp in Los Angeles area. The extracted schema includes category, open hour, phone address and area, review count, and rating.

TripAdvisor data set

We extracted data from TripAdvisor API with the corresponding url: "https://www.tripadvisor.com/RestaurantSearch?Action=PAGE&geo=32655&ajax=1&itags=10591&sortOrder=relevance&o=a**{30*(page-1)}**&availSearchEnabled=false", for **page** from 1 to 104. There are 30 restaurants listed per page, so we extracted 3120 restaurants data in total. From the above API, it returns a HTML template. We then manually construct the wrapper by using BeautifulSoup in Python to extract features from the web pages. From TripAdvisor data source, we extract the following schema: id, name, rate, n_reviews, money, type_1, type_2, type_3, address_street, address_locality, phone, sun_hours, mon_hours, tue_hours, wed_hours, thur_hours, fri_hours, sat_hours.

TABLE DESCRIPTION

The yelp table (Table A) contains 3188 restaurants, while the tripadvisor table (Table b) contains 3120 restaurants. Both tables share the following schema. Name [text]

- Category_1 [text]
- Category_2 [text]
- Address [text]
- City [text]
- Zipcode [5-digit]
- Phone [10-digit]
- Price [number of dollar signs]
- Rating [1 to 5, allow half]
- Review_count [number]
- A set of attributes for open hours
    - hours_day_<open/close>: [HHMM] in 24 hour format
    - Examples:
        - hours_sun_open = 1100
        - hours_sun_close = 200

Note that not all attributes are contextually the same in both tables. For example, price, rating, and review_count are specific to website data and might not be useful in entity matching stage, but are of potential interest for analysis in later stages.

TOOLS USED

ConvertCSV: This tool is a fast web controver that parse JSON format into comma separated CSV format.

Beautiful Soup: This is a python library designed for quick turnaround projects like screen-scraping.