Yelp vs Zomato

Goal of the project

To compare data from two popular restaurant review websites - Yelp and Zomato, and identify which website is more preferred by users for rating restaurants and for checking out the menus and prices.

Set of questions to answer

1) Which website contains more reviews for a particular restaurant?
2) Based on the answer to the above question, which site do users prefer to rate and review restaurants?
3) How closely related are the data from each website?

Data source

We collected data from two popular restaurant review sites - Yelp and Zomato. We scraped the reviews of different restaurants and created 310 text documents, each containing one review.

How we extracted structured data

We identified the popular restaurants from four cities in the US and extracted the following details of each restaurant: **Name**, **Address**, **Price range**, **Rating** and **Number of reviews**. We then structured all the details in a Comma Separated Values (CSV) format and stored them in two files - one for Yelp and the other for Zomato.

What we want to extract from the text documents

We would like to extract the following details from the text documents:
(i) The dishes / cuisines.
(ii) Identify the polarity of the review using adjectives.

Tools we used to scrape data

We used an open-source python library called **Scrapy** to extract data from the two aforementioned websites. Scrapy provides functions to send the Uniform Resource Locators (URL) of pages from which to extract data through a **Request** object and populates the data from those URLs in a **Response** object which we can then write to a file. It also provides a web crawler which can identify sub-links in a page and then scrape data from them recursively. In addition to extracting all the data from a page, It provides two options to extract specific attributes from it -

(i) **CSS -** Accepts the HTML node of the attribute to be extracted as input and provides its value.
(ii) **XPath -** Accepts the XPath of the attribute to be extracted as input and provides its value.