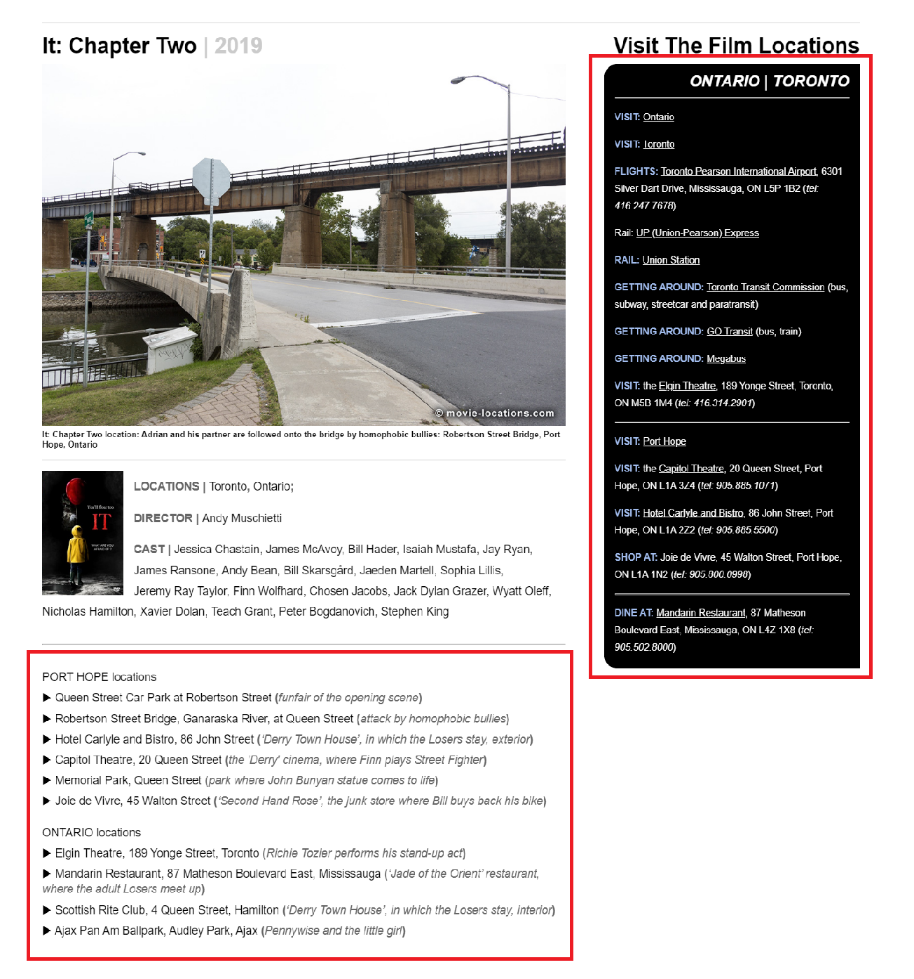
Task: Many travellers are often interested in visiting famous spots where movies were shot. How can we build a database of such famous spots?

# Data source

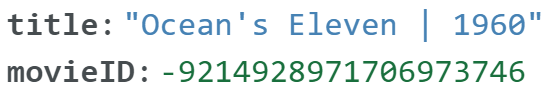
* Site: <http://www.movie-locations.com/>
* Reason: the site has detailed and expansive documentation of locations where movies are shot.

# Data acquisition

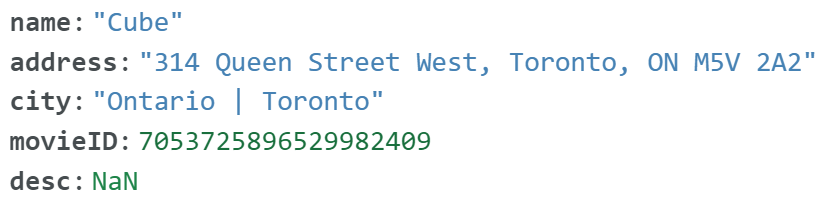
* Target:
  + the two regions in the webpage indicated by the red boxes below



* + URLs of the pages are stored in ‘urls.txt’
* Database structure:
  + Table 1: each entry corresponds to a movie. Included are:
    - title: title of the movie
    - movieID: obtained from applying Python’s hash() function on the title
    - pop (for future work): popularity of the movie



* + Table 2: each entry corresponds to a location where a movie is shot
    - name: name of the location
    - desc: description of the movie scene at the location
    - address: address of the location
    - city: the city where the location is located
    - movieID: corresponds to the title in table 1
    - pop (for future work): popularity of the location



* Scraping:
  + code written in python
  + Beautiful Soup package to parse the page’s HTML
  + pandas library to merge the two sets of data obtained from the two regions indicated by the red boxes
* Storage:
  + offline: as .csv files
  + online: mongoDB online databased named ‘movies’
    - Table 1 is named ‘movieIDs’
    - Table 2 is named ‘locations’

# Future work

* Write code to scrape the URLs of all the film location pages and store in ‘urls.txt’
* Fill in popularity data. The rationale is to have the ability to decide which location is more appealing to potential tourists. Methods:
  + Scrape movie scores from IMDB, Rotten Tomatoes and Metacritic
  + Use Google’s Custom Search JSON API to obtain the number of search results as a measure of popularity
* Construct another table or add another column for Table 2 for a Google Maps location