

# TripAdvisor Crawler

Development of a focused crawler in Java  
for reviews-extraction from TripAdvisor

*final project*  
*Web Information Management*

D. Grasso, M. Lombardi, A. Marani, P. Paris

# Analysis of TripAdvisor

- To adapt the crawler to the website:
  1. Which ones are the data of interest?
  2. On which pages are these data displayed?
  3. How do we reach all these pages?
- Data of interest are opinions, expressed in the reviews of the users.
- Each *review* can belong to a hotel, a restaurant or attractions.
- In general terms we can define them as *items*, each one having many *review*.

# Review



The screenshot shows a review from Hazel A. The review has a title "Best meal I have ever had", a rating of 5 stars (indicated by 5 green circles), and was posted on June 13, 2012. It includes a snippet of text: "We arrived at the bistro at about 9.30 and was told that it closed at 11 which wasn't a problem, although...". Annotations with arrows point to specific parts of the review:

- A red arrow labeled "rating" points to the 5-star rating icon.
- A red arrow labeled "date" points to the posting date "13 giugno 2012".
- A blue arrow labeled "author" points to the user profile picture and name "Hazel A".
- A blue arrow labeled "text" points to the main text snippet.
- A blue arrow labeled "title" points to the title of the review.

- We choose to focus only on **rating** and **date**
- Every review has also a unique identifier (reviewID) – like: “r133345545”

# Item

- Is an abstraction representing either:
  - Hotel
  - Restaurant
  - Attraction
- Also for items we focus only on essential data:
  - Name
  - Number of reviews (useful for crawling...)
  - Identifier (itemID) - like: “d1132860”

# Location

- Represents a geographic place of interest, where it is possible to find many items.
- The entities of type *location* are structured in a hierarchy, with a variable grain.

Home > Europa > Francia > Ile-de-France > Vacanze Parigi

- Each Location has a name and an ID – examples:
  - Europe - “g4”, France- “g187070”, Paris- “g187147”
  - Even if very useful to understand the structure of the website, these data are not going to be extracted.

# TripAdvisor - Fact sheet

- How many are the items we need to extract?
  - More than 60 million travel reviews
  - 108,000+ destinations
  - 1,600,000+ businesses
    - 600,000+ hotels ,
    - 198,000+ attractions,
    - 858,000+ restaurants

from: [http://www.tripadvisor.com/PressCenter-c4-Fact\\_Sheet.html](http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html)

- So, for a complete crawling, we'll have:
  - 60,000,000+ reviews
  - 1,600,000+ items

# Page Classes (1)

- Class: Home Page
    - A good starting point, containing many **many links** to pages of every kind, and specifically:
    - Most popular and newest Items
    - Main Locations
    - Other types of pages to ignore (good for testing)
      - like Flights, Forum, ...

**Plаниfica il viaggio perfetto**

Hotel    Voli    Ristoranti    Il meglio del 2012    I tuoi amici    IPO    Scrivimi una recensione

Inserire la città o il nome dell'hotel

Cerca hotel!

**Cosa dicono i viaggiatori italiani**

Ricerca Alberghi   Foto (97.380)   Forum (601)

 agostino p 10 recensioni 8 voti utili	 ciccio_bcn 15 recensioni 4 voti utili	 essencia 37 recensioni 18 voti utili
<b>Hotel Galileo</b> "Trovato come promesso"  07 luglio 2012	<b>For My Friend BAB</b> "comigliosissimo!"  07 luglio 2012	<b>Rome Marriott Park Hotel</b> "Comodo per l'aeroporto!"  07 luglio 2012

**Presentazione dei**

**Ad Hoc** (801 recensioni)  
**Ristorante D'Arborea** (369 recensioni)  
**Ristorante Mare - Artigiani del Pe...** (112 recensioni)  
Nostra (3.987 ristoranti)

**Colore: Città**

**1 Santa Prassede**  
dolobobol (30 recensioni)  
**2 Le Domande di Palazzo Valent...**  
domenico (236 recensioni)  
**3 San Giovanni in Laterano (Basilica...)**  
dolobobol (522 recensioni)  
Mostra le 988 attrazioni

**Foto dei viaggiatori da tutto il mondo**

  
Coco Plum Island...  
da cocacola

  
Fiji  
di Senza24

  
St. Maarten  
da helenm

  
Key West  
da helenm

**Destinazione in primo piano**

  
La Val di Fassa, nel Trentino nord-orientale, è la meta' ideale per una vacanza in montagna con la famiglia. In inverno offre le più belle piste sciistiche e le più belle viste sulle Dolomiti Superiori. In estate la valle propone molteplici attività da praticare all'aperto, tra le Dolomiti. Da non dimenticare poi i santi eventi, le tradizioni ladine e la buona cucina.

Destinazione fornita da: Ufficio Turismo Val di Fassa   [Altre informazioni: Val di Fassa](#)

**Individua le tue esperienze**

[Scrivimi una recensione](#)  
[Come ottenere assistenza per recensioni e giudizi](#)  
[Fai una domanda o consulta i forum](#)  
[Referenze utente](#)  
[Aggiorna il tuo profilo o i tuoi abbonamenti](#)  
[Contattateli](#)  
[Invia un feedback o segnala un problema](#)  
[Per i proprietari](#)  
[Scopri le offerte Profilo Attivista](#)  
[Visita il Centro Gestione](#)  
[Total nelle destinazioni più richieste:](#)

Non ti sei ancora iscritto? Iscriviti oggi stesso su TripAdvisor e puoi trovare tante recensioni vere e

**Città in Italia** vedrete le ultime recensioni

Hotel Alghero    Hotel Lampedusa  
 Hotel Bologna    Hotel Livorno  
 Hotel Catania    Hotel Lombardia  
 Hotel Roma    Hotel Messina  
 Hotel Firenze    Hotel Milano  
 Hotel Genova    Hotel Napoli  
 Hotel Ischia    Hotel Palermo  
 Hotel Jesolo    Hotel Pisa  
 Hotel Venezia    Hotel Riccione  
 Hotel Ancona    Hotel Rimini  
 Hotel Barcellona    Hotel Riviera Liguria  
 Hotel Roma    Hotel Sardegna  
 Hotel Bruxelles    Hotel Sicilia  
 Hotel Parigi    Hotel Sizilie  
 Hotel Cagliari    Hotel Sirmione  
 Hotel Copenhagen    Hotel Sorrento  
 Hotel Milano    Hotel Taormina  
 Hotel Roma    Hotel Torino  
 Hotel Ascoli Piceno    Hotel Trieste  
 Hotel Barcellona    Hotel Venezia  
 Hotel Roma    Hotel Vittorio Veneto  
 Hotel Cagliari    Hotel Vittorio Veneto  
 Hotel Costa Azzurra    Hotel Zaragoza  
 Hotel Creta    Hotel Zaragoza  
 Hotel Dublino    Hotel Madrid  
 Hotel Formerra    Hotel Mafra  
 Hotel Gran Canaria    Hotel Monaco di Baviera  
 Hotel Granada    Hotel Montecatini Terme  
 Hotel Ibiza    Hotel Nizza  
 Hotel Stockholm    Hotel Oporto  
 Hotel Ischia di Malta    Hotel Palermo  
 Hotel Istanbul    Hotel Porto  
 Hotel Londra    Hotel Praga  
 Hotel Lourdes    Hotel Rodi  
 Hotel Madrid-Tropoz    Hotel Salzburg  
 Hotel Mafra    Hotel Salsburg  
 Hotel Roma    Hotel Savoia  
 Hotel Rio    Hotel St. Moritz  
 Hotel Monaco di Baviera    Hotel St. Moritz  
 Hotel Montecatini Terme    Hotel Tenere  
 Hotel Nizza    Hotel Valencia  
 Hotel Oporto    Hotel Vittorio Veneto  
 Hotel Porto    Hotel Zadar  
 Hotel Praga    Hotel Zinga  
 Hotel Rodi    Hotel Zuniga

**Nel mondo** vedrete le ultime recensioni

Hotel Amsterdam    Hotel Madrid  
 Hotel Atene    Hotel Mafra  
 Hotel Barcellona    Hotel Malesia  
 Hotel Roma    Hotel Monaco di Baviera  
 Hotel Bruxelles    Hotel Montecatini Terme  
 Hotel Parigi    Hotel Nizza  
 Hotel Cagliari    Hotel Oporto  
 Hotel Copenhagen    Hotel Palermo  
 Hotel Roma    Hotel Praga  
 Hotel Dublino    Hotel Rodi  
 Hotel Formerra    Hotel Roma  
 Hotel Gran Canaria    Hotel Salzburg  
 Hotel Granada    Hotel Salsburg  
 Hotel Ibiza    Hotel Savoia  
 Hotel Stockholm    Hotel St. Moritz  
 Hotel Ischia di Malta    Hotel Tenere  
 Hotel Istanbul    Hotel Valencia  
 Hotel Londra    Hotel Vittorio Veneto  
 Hotel Lourdes    Hotel Zadar  
 Hotel Madrid-Tropoz    Hotel Zinga  
 Hotel Mafra    Hotel Zuniga  
 Hotel Roma    Hotel Zuniga

**Mete preferite:**

Non ti sei ancora iscritto? Iscriviti oggi stesso su TripAdvisor e puoi trovare tante recensioni vere e

**Destinazioni che hai visualizzato:**

  
Europa  
  
Stati Uniti  
  
Asia  
  
Africa  
  
Australia e Oceania  
  
Sud America  
  
Mezzogiorno Estremo Oriente  
  
Mezzogiorno America  
  
Africa  
  
Asia  
  
Australia e Oceania  
  
Sud America  
  
Mezzogiorno Estremo Oriente  
  
Mezzogiorno America

**Gli ospiti di hotel sono entusiasti di...**

  
Mostra prezzi  
O'Callaghan Elliott  
Grand Hotel Europe  
338 recensioni

  
Mostra prezzi  
Jaya Paris  
Parigi - Tre-de-France  
329 recensioni

  
Mostra prezzi  
Cor Doss  
Golfo di Guayaquil (Ecuador)  
17 recensioni

# Page Classes (2)

- Class: List-of-locations

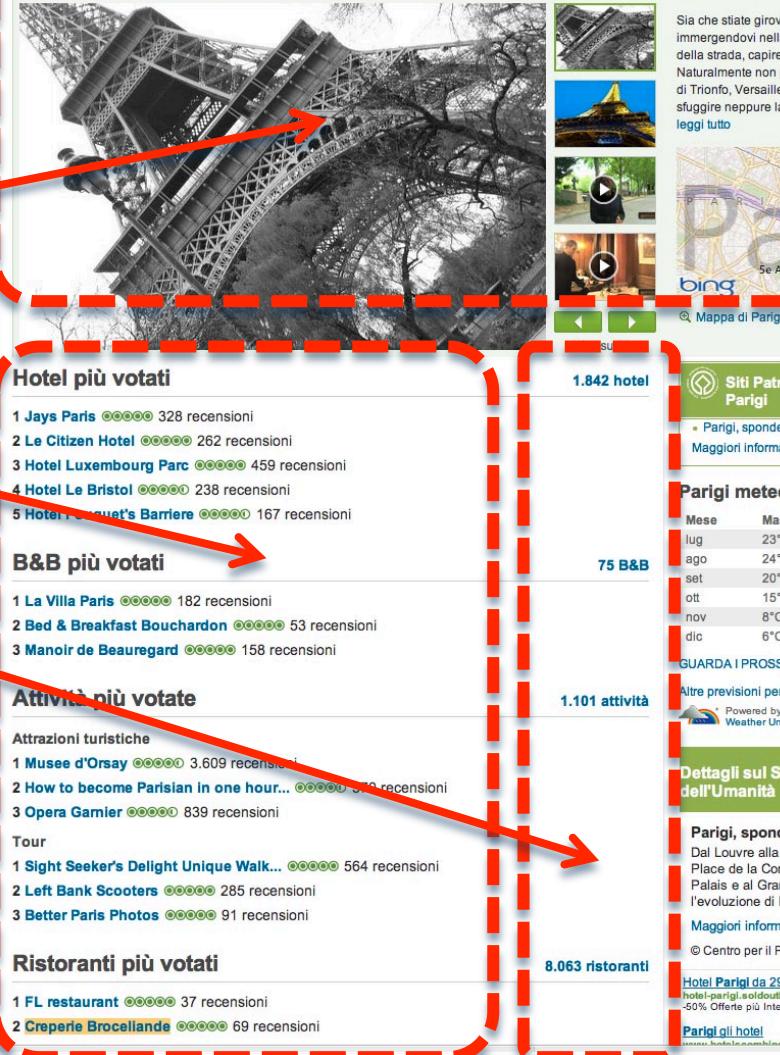
Allows navigation in the locations tree, in which:

- intermediate nodes are other list-of-locations
- leaves are pages of the class location.



## Vacanze Parigi, Francia

 **Vincitore dei premi Travellers'Choice™ 2012** Destinazioni | Enogastronomia (2011)



**Hotel più votati**

1	Jays Paris	★★★★★	328 recensioni
2	Le Citizen Hotel	★★★★★	262 recensioni
3	Hotel Luxembourg Parc	★★★★★	459 recensioni
4	Hotel Le Bristol	★★★★★	238 recensioni
5	Hotel Fouquet's Barrière	★★★★★	167 recensioni

**B&B più votati**

1	La Villa Paris	★★★★★	182 recensioni
2	Bed & Breakfast Bouchardon	★★★★★	53 recensioni
3	Manoir de Beauregard	★★★★★	158 recensioni

**Attività più votate**

1	Musee d'Orsay	★★★★★	3.609 recensioni
2	How to become Parisian in one hour...	★★★★★	372 recensioni
3	Opera Garnier	★★★★★	839 recensioni

**Tour**

1	Sight Seeker's Delight Unique Walk...	★★★★★	564 recensioni
2	Left Bank Scooters	★★★★★	285 recensioni
3	Better Paris Photos	★★★★★	91 recensioni

**Ristoranti più votati**

1	FL restaurant	★★★★★	37 recensioni
2	Creperie Broceliande	★★★★★	69 recensioni

# Page Classes (3)

- Class: Location

Given a locationID, shows:

- descriptive information
- a set of most popular items of any type.
- links to complete lists of item by type.

# Page Classes (4)

- Class: List-of-Items

Given a location (locationID) and an Item type, shows a list of all the local Items of that type and some search ui.

[Home > Europa > Francia > Île-de-France > Parigi](#)

## Hotel Parigi

Hotel (1.842)   B&B / Pensioni (75)   Altre sistemazioni (87)

Tutti gli hotel (1.842)   Qualità / prezzo (183)   Famiglia (442)

Vedi disponibilità

Arrivo dd/mm/aaaa Partenza dd/mm/aaaa Adulti 2 Ricerca

Perfeziona la ricerca

Prezzo a notte

Tutti (1.842)  
 €0 - €78 (382)  
 €78 - €156 (1.254)  
 €156 - €221 (928)  
 €221+ (592)

EUR

1842 su 1842 hotel visualizzati

In ordine di Classifica Vedi

**Jays Paris**  
 €492+ a notte\*  
 Classificato al n.1 di 1.842 hotel  
  
 \*Giusto il primo posto\* 02/11/2011  
 \*Scelta perfetta\* 10/01/2011

Classificato al n.1 di 1.842 hotel  
  
 328 recensioni

How to become Parisian in one hour?



Posizione n. 2 di 547 attrazioni a Parigi  
 579 recensioni

Tipo di attrazione Esibizioni  
 Distanza 2.0 km da Centro città  
 "Molto divertente ed utile" 26 mag 2012  
 "Divertente spettacolo...da n..." 6 mag 2012

Mappa | Foto

[Home > Europa > Francia > Île-de-France > Parigi > Attrazioni: Parigi](#)

## Attrazioni: Parigi

Attrazioni (547)   Tour (173)   Vita notturna (272)

Mostra filtri

**Attrazioni consigliate dai viaggiatori**

1-30 su 547   < 1 2 ... 19 >

Ordina 547 attrazioni per Classifica

**Musee d'Orsay**  
 Posizione n. 1 di 547 attrazioni a Parigi  
  
 3.695 recensioni

Tipo di attrazione Gallerie d'arte; Edifici architettonici; Musei  
 Distanza 1.9 km da Centro città  
 Mappa | Foto  
 "Uno dei più bei musei che ab..." 9 lug 2012  
 "Vacanza a Parigi" 8 lug 2012  
 "Imperdibile" 8 lug 2012

**How to become Parisian in one hour?**  
 Posizione n. 2 di 547 attrazioni a Parigi  
 579 recensioni

Tipo di attrazione Esibizioni  
 Distanza 2.0 km da Centro città  
 "Molto divertente ed utile" 26 mag 2012  
 "Divertente spettacolo...da n..." 6 mag 2012

Mappa | Foto

[Home > Europa > Francia > Île-de-France > Parigi > Ristoranti: Parigi](#)

## Ristoranti: Parigi

Cucina   Quartieri   Opzioni

Tutti (8.063)  
 Africana (81)  
 Americana (103)  
 Asiatica (186)  
 Bar (90)

Tutti (8.063)  
 Batignolles-Monceau (361)  
 Belleville (28)  
 Bercy (32)

Tutti (8.063)  
 Accetta prenotazioni (1.720)  
 Adatto ai bambini (263)  
 Adatto ai balli (203)

**Ristoranti consigliati dai viaggiatori**

1-30 su 8.063   < 1 2 ... 269 >

In ordine di: Classifica    Mostra prima i ristoranti prenotabili (296)

**FL restaurant**  
 N. 1 in classifica su 8.063 ristoranti a Parigi  
  
 43 recensioni  
 Prezzo: €18 - €45  
 Cucina: gastronomia, ristorante americano stile anni '50, francese, salutistica  
 Possibilità per mangiare: Accetta prenotazioni  
 "Ristorante favoloso dal cibo..." 5 lug 2012  
 "finalmente un ottimo ristorante..." 27 giu 2012  
 "un posticino davvero gradevo..." 1 nov 2011

**Creperie Broceliande**  
 N. 2 in classifica su 8.063 ristoranti a Parigi

# Page Classes (5)

- Item-Reviews

Given an itemID, shows:

- descriptive information and overall stats on the reviews.
- The complete list of all the reviews, divided into pages.  
Reviews are shortened!

The screenshot shows a TripAdvisor hotel review page for Hotel La Belle Juliette in Paris. At the top, it displays the hotel's name, address (92 rue du Cherche Midi, 75006 Parigi, Francia), and a summary rating of 4.5 stars. It also shows a 'Mi piace' button and a link to the hotel's website. Below this, there's a thumbnail image of a room, a star rating of 4.5 stars from 163 reviews, and a mention of being a 'Vincitore del premi Travellers'Choice™ 2012 Top 25'. A red arrow points from the 'Reviews' section of the slide to this star rating area. Further down, there's a search bar for prices, a list of travel partners (Booking.com, Prestigia.com, Expedia.it, agoda.it, Hotels.com), and a section for '163 recensioni dei viaggiatori'. This section includes a bar chart for review ratings (Excellent, Very Good, Average, Fair, Poor) and a list of traveler types (Family, Couple, Solo traveler, Work, Friends). A red arrow points from the 'Reviews' section of the slide to the start of this list. Below this, there are two examples of reviews with their titles and snippets. At the bottom, it shows '1-10 di 163 recensioni' and a page navigation bar with links for page 1, 2, ..., 17, and next. To the right of the main content, there are sidebar sections for 'Gli utenti hanno anche guardato', 'Cerca nei dintorni', 'Mappa interattiva', and a list of nearby restaurants.

# Page Classes (6)

- ReviewDetails

Given a reviewID, shows:

- Again descriptive information on the item to which the review belongs to.
- The complete text of the review, with many additional data.
- The other reviews of the item, in a detailed fashion

The screenshot shows a TripAdvisor page for Hotel La Belle Juliette in Paris. At the top, it displays the hotel's name, address (92 rue du Cherche Midi, 75008 Parigi, Francia), and various contact links. It also shows the hotel is ranked 22nd out of 1,842 hotels in Paris, has 163 reviews, and is a 'Travellers'Choice' award winner. Below this, there's a photo gallery, a price search bar, and a section for traveler reviews. A red arrow points from the 'ReviewDetails' section of the slide to this traveler review section. Another red arrow points from the 'other reviews' section of the slide to a specific review by 'Jennifer P' at the bottom of the page.

Hotel La Belle Juliette: Recensioni ★★★★★

Mi piace 2

92 rue du Cherche Midi, 75008 Parigi, Francia

Sito internet hotel Invia un'e-mail all'hotel +33 142229740 Servizi dell'hotel

Classificato al n.22 di 1.842 hotel in Parigi

163 Recensioni

Vincitore del premio Travellers'Choice™

Top 25

Mostra la tariffa più bassa per quest'hotel\*

Arrivo 21/7/2012 Parenza 22/7/2012 Adulti 2

Foto professionali

45 foto dei viaggiatori

Mostra prezzi

Booking.com  agoda.it

Prestigia.com  Hotels.com

Expedia.it

\*dai nostri partner

Scrivi una recen...

**163 recensioni dei viaggiatori**

Valutazione

Eccellente	123
Molto buono	31
Nella media	6
Scarsa	3
Pessimo	0

Tipo di viaggio

- In famiglia (19)
- In coppia (103)
- Di viaggiatori solitari (5)
- Di lavoro (16)
- Con amici (10)

Le camere preferite dai viaggiatori: 36 consigli dei viaggiatori

163 recensioni filtrate per Data Punaggio Qualsiasi

1 di 163 recensioni

Traduzione automatica Originale in Inglese Il tuo giudizio su questa traduzione: Cattiva Buona

Jennifer P 1 recensione

**"Questo hotel è adorabile!"**

Recensito il 2 luglio 2012

Siamo appena tornati dal nostro primo viaggio a Parigi. La Belle Ju era un hotel francese elegante e speciale. Sono rimasto molto colpito dalla pulizia delle camere e le strutture erano. L'hotel è arredato in modo incantevole. Il nostro servizio è stato sempre impeccabile e sono molto gentili con noi come americani che non parlava una parola di francese! La colazione è molto caro, ma ci sono opzioni nelle vicinanze che vendono frutta e generi alimentari. Consigliamo questo hotel e ritorniamo sicuramente per La Belle Ju.

# URL and page classes

- To distinguish different page classes it has been enough to look at the url:

[http://www.tripadvisor.it/Tourism-g187147-Paris\\_Ile\\_de\\_France-Vacations.html](http://www.tripadvisor.it/Tourism-g187147-Paris_Ile_de_France-Vacations.html)

[http://www.tripadvisor.it/Restaurants-g187147-Paris\\_Ile\\_de\\_France.html](http://www.tripadvisor.it/Restaurants-g187147-Paris_Ile_de_France.html)

[http://www.tripadvisor.it/Restaurant\\_Review-g187147-d1991708-Reviews...](http://www.tripadvisor.it/Restaurant_Review-g187147-d1991708-Reviews...)



- This section identifies the kind of page
  - Accordingly, the next sections can refer to the location and the item identifiers.

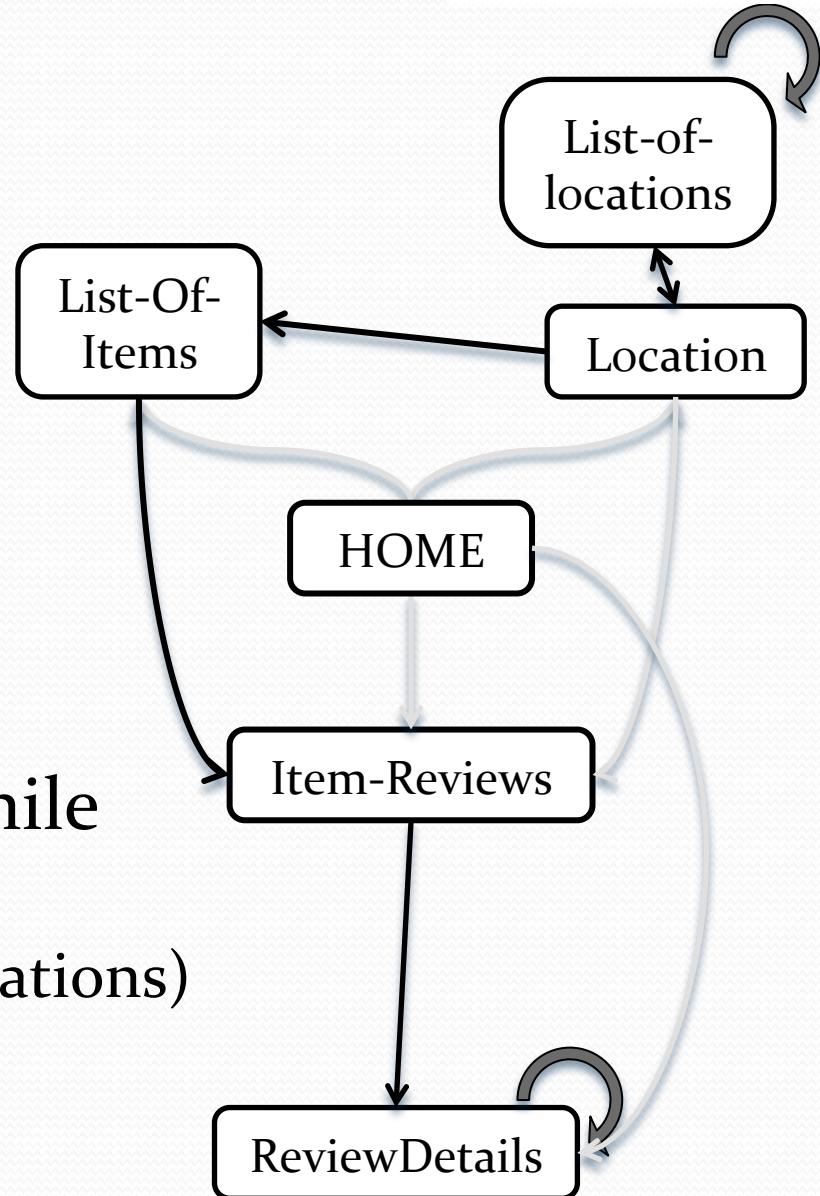
# Mapping URL - Classes

URL MAPPING	Class	#estimated pages
/AllLocations	List-of-Locations	< 100,000
/Tourism	Location	~108,000
/Hotels /Restaurants /Attractions	List-of-Items	< 1,000,000
/Hotel_Review /Restaurant_Review /Attraction_Review	Item-Reviews	6,000,000
/ShowUserReview	ReviewDetails	60,000,000
...	...	?

# SiteMap

Linking page-classes according to the contained links, we can obtain a SiteMap.

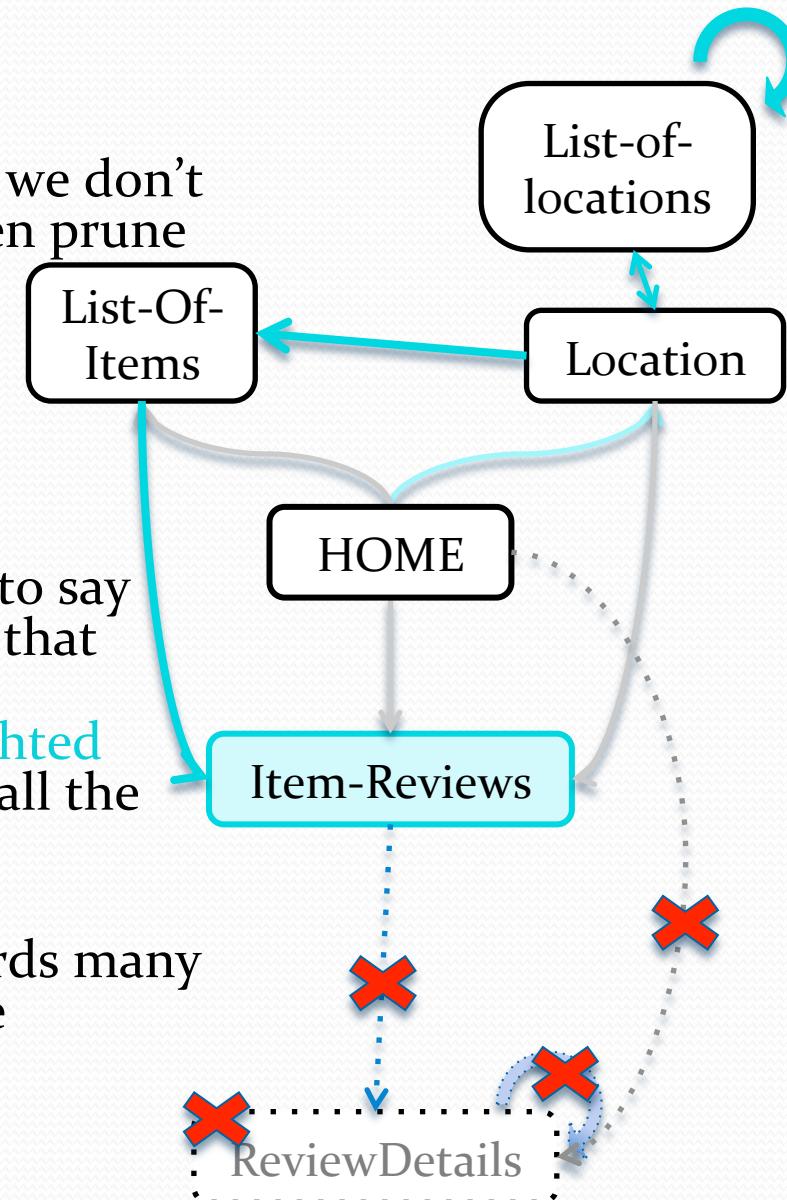
**black** links are exhaustive, while gray ones are partials.  
(e.g.: home contains only some locations)



# Analysis

- If we only have to extract date and rating, we don't need to reach ReviewDetails. We can then prune part of the graph.
- All the data that we need to extract are contained in Item-Reviews pages
- The structure of list-of-location allows us to say that it's enough to visit one single page of that class to be able to reach every location.  
From each Location, following the **highlighted path**, it is possible to reach the reviews of all the items.

This approach, unlike *casual crawling*, discards many pages improving Efficiency, at the same time ensuring Completeness.

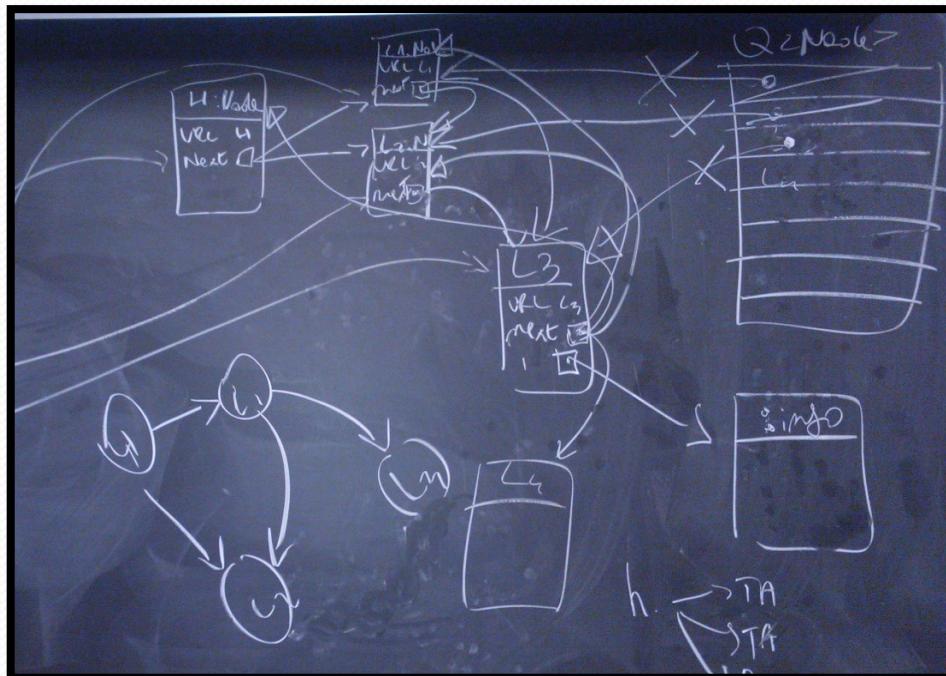


# Estimates

- The total number of pages we want to visit lies around 6,500,000, of which:
- 6,000,000 are the pages to contain all the reviews (10 per page) to be extracted.
- 500,000 are pages to be “traversed” extracting only links to the items

# Development of the System

- First steps...
  - Wanting to create a crawler from scratch, we built a first working version.
  - It allowed us to study the structure of a crawling engine and to face the typical involved problems and tradeoffs.



## Problems

- Slow! (multithreading?)
- Manage many server response types
- Manage the delay between 2 requests
- Do not re-invent the wheel!

# Development of the System(2)

- Crawler4j
  - A very useful library, clean and with all the tools
  - It allowed us to configure the crawling over the whole website with n crawlers at the same times, the extraction of some basic data and controlled minimum inter-request delay.  
(to configure everything: 10 minutes ☺)
  - We found in the structure of crawler4j the main ideas of our first crawler
  - We avoided to rewrite tons of code

# Development of the System(3)

- Optimization based on the structure of the website
  - discarded useless links
  - parser created to extract data of interest from the pages of class: *item-reviews*
  - Extension of some of the classes of crawler4j
  - Fix of some navigation problems between the reviews
- Development of the persistency layer

# Crawling: management & execution

CrawlHandler

1. CrawlHandler prepares the environment for the execution and management of crawling and the domain model, and starts crawler4j through CrawlController. In addition instantiates and interacts with the parser ItemReviewsPageParser.

CrawlController

2. CrawlController start threads of TACrawler instances, equal in number as indicated in the setup. Each TACrawler is created with a reference to CrawlHandler.

TACrawler

3. Each thread of TACrawler takes care of crawling TripAdvisor. It calls CrawlHandler to invoke the parser in case it needs to parse the reviews of a page of type Item.

# CrawlHandler

- Requires three setup parameters
  - number of crawlers (threads)
  - Path to temporary folder for the frontier data
  - Minimum delay between two requests (in ms)
- Instanciates the specific parser object for the pages of TripAdvisor
  - Low coupling between crawler and parser
- Initializes the “seed” links
- Creates a Map<String,Item>, with the Item and the related reviews as soon as they are collected by the many crawlers.
  - Each crawler, knowing the CrawlHandler, can access that map to insert and update items.
- Starts the crawling using CrawlController, with the setup parameters and a reference to itself, for the creation of the crawlers.
- At the end of the crawling the domain model, contained in the map, will be saved in the database using the persistence layer.

# CrawlController

- Crawler4j has a own CrawlController
  - Launch and management of the crawling algorithm of crawler4j
  - MultiThreading management
    - Launch of the requested number of crawlers, one per thread.
- It has been necessary to extend CrawlController from crawler4j so that threads are instantiated with a reference to the CrawlHandler (not existing in crawler4j)
- Our CrawlControllerOwn extends CrawlController allowing to pass the Handler to every new thread of generic type (T)
  - So this is not dependent on the specific crawler, given that in the constructor a reference to CrawlHandler is requested.

# TACrawler

- This is the specific Crawler, of which many instances are launched, each one knowing CrawlHandler
- Extends the class WebCrawler of Crawler4j, overriding its main methods:
  - boolean **shouldVisit**(WebUrl url)
    - Given a url, extracted from the outgoing links, determines if this has to be loaded by the fetcher in a page and put in the frontier.
  - void **visit**(Page page)
    - Is invoked on a page extracted from the frontier.
    - if the page belongs to the class: Item-Reviews it calls ItemReviewParser for data-extraction. The data are then saved by CrawlHandler.

# Parser (ItemReviewPageParser)

- While link extraction from the HTML is performed by crawler4j using Apache Tika, this parser works on the already loaded HTML.
- Has the goal of extracting data with a higher level of abstraction, only on some of the pages.
- Keeps a good level of low coupling from the crawler, while being the part of the system that is more specifically bounded to the internal structure of the pages of TripAdvisor.

# Parser (ItemReviewPageParser)

- Based on the library Jericho HTML parser
  - Documentation defines it as:  
*“... neither an event nor tree based parser...”*
  - *Compared to a tree based parser such as DOM, the memory and resource requirements can be far better if only small sections of the document need to be parsed or modified. Incorrect or badly formatted HTML can easily be ignored, unlike tree based parsers which must identify every node in the document from top to bottom.*
  - *Compared to an event based parser such as SAX, the interface is on a much higher level and more intuitive, and a tree representation of the document element hierarchy is easily created if required.*

# The collected data

- Collected data were saved into a relational DB
  - Easy to query
  - Few updates, indexes are “for free”
  - Allows Stored Procedures
- Persistence provided by the Hibernate framework
  - Independency from the specific DBMS
  - Only needs a few lines in the config file
  - adapts to the variations in the models

# DB Schema

- The schema of the database looks like a tiny datamart
  - Item (*itemid, crawlDate, description, type*)
  - Review (*reviewid, date, review, value*)
  - ItemReviews (*itemid, reviewid*)
    - With obvious foreign keys and constraints
- This schema allows, if needed, to assign more informations to the correlation between item and review

# ...extra tips for the DB

- Typical queries will be mainly based upon the date of the reviews for a given item:
  - Indexes are created to speed up the queries
    - they are convenient also since updates are not frequent
- Use of Stored Procedures for the queries
  - `showAverageReviewPerDate ('item_name', 'start_date', 'end_date')`
  - `showReviewsPerDate ('item_name', 'start_date', 'end_date')`

# Execution times

- Execution times were a little too long for our laptops:
  - 16.000 items/hour -average
  - 100 hours, given by a total of 1.600.000 item (Restaurants, Hotel, Attractions)
- Too much, so we decided to execute our crawler on Amazon EC2:
  - 45.000 items/hour (about 3x faster)
  - About 18 hours for all the items of TripAdvisor

# Test: effectiveness

- Useful to understand if:
  - only pages of type Item are parsed.
  - all reviews are extracted for each Item
- The test has been conducted limiting the frontier expansion to a subset of the items, extracting outgoing links only from their pages and completing the extraction of all their reviews.
- As the seed page we set the page of a single Item
  - Different for every test.
- An aim was to find corner cases and possible intervention measures.

# Test: effectiveness - results

- Setup parameters:
  - 500 crawlers
  - 2 ms delay, so a maximum of 500 requests/second
- Only Items have been parsed.
  - No false positive
- Test results were divided by Item type
- Note: Every test had a total duration of about 10 minutes

# Results - Hotels

- Hotel
  - I Test: missed 2 reviews out of 42.625, for a total of 155 different hotels. Accuracy: 99,99%
  - II Test: missed 3 reviews out of 33.125, for a total of 79 different hotels. Accuracy: 99,99%
  - III Test: perse 6 recensioni su 40.847, for a total of 188 different hotels. Accuracy: 99,98%
- Average: 99,99% of extracted reviews, for a total of 422 different hotels.

# Results - Restaurants

- Restaurant
  - I Test: missed 20 reviews out of 15.339, for a total of 118 different restaurants. Accuracy: 99,87%
  - II Test: missed 0 reviews out of 9.012, for a total of 98 different restaurants. Accuracy: 100%
  - III Test: missed 20 reviews out of 15.313, for a total of 121 different restaurants. Accuracy: 99,87%
- Average: 99,91% of extracted reviews, for a total of 337 restaurants

# Results - Attractions

- Attraction
  - I Test: missed 20 reviews out of 14.309, for a total of 87 attractions. Accuracy: 99,86%
  - II Test: missed 140 reviews out of 33.212, for a total of 103 attractions. Accuracy: 99,58%
  - III Test: missed 100 reviews out of 35.193, for a total of 130 attractions. Accuracy: 99,72%
- Average: 99,72% of extracted reviews, for a total of 320 attractions

# Effectiveness - Conclusions

- Hotels were the type of item which was more in line with our original study
- From these tests some bugs were found in the extraction logic, due to inconsistencies in the pages of TripAdvisor.
  - A problem was due to the presence of some reviews without any rating (which is currently not allowed by TripAdvisor, but maybe was possible before)