How you crawled the corpus (e.g., source, keywords, API, library) and stored them (e.g., whether a record corresponds to a file or a line, meta information like publication date, author name, record ID)

Answer:

We aimed to crawel informative review data from Tripadvisor website and constructe a comprehensive format of datastore, in order to build up our own search engine that can satisfy users' needs. As for recommending the Attraction for users, we use Twitter API for crawling Twitter data.

TripAdvisor

Source & Keywords for crawling

The keywords for craweling are "**Singapore Attraction**" from "**http://www.tripadvisor.com.sg/Attractions-g294262-Activities-Singapore.html**".

This page link above refers to the summary of Attractions in Singapore (as a Country), which provides 4 subcategories of locations:

1. Singapore Attractions: http://www.tripadvisor.com.sg/Attractions-g294265-Activities Singapore.html
2. Sentosa Island Attractions: http://www.tripadvisor.com.sg/Attractions-g294264-Activities-Sentosa_Island.html
3. Pulau Ubin Attractions: http://www.tripadvisor.com.sg/Attractions-g1644875-Activities-Pulau_Ubin.html
4. Jurong Attractions: http://www.tripadvisor.com.sg/Attractions-g294263-Activities-Jurong.html

Since we do not have API, we start crawling links of needed information from the above 4 links. Thus, to be precise, "Singapore Attraction" is not the Keyword in the normal sense (when using API), but actually we have 4 specific starting points for retrieving data. We will crawl the reviews (in English) for all the attractions under each category for Singapore.

API & Library

As TripAdvisor does not offer an API for academic purpose, the information of the attractions and reviews are fetch by Java code implemented by ourselves. The following is the libraries/frameworks for crawling and storing the corpora (and log purpose).

1. Crawler4j: https://github.com/yasserg/crawler4j (for crawling)
2. Jsoup: http://jsoup.org/ (for parsing HTML)
3. Hibernate (ORM): http://hibernate.org/orm/

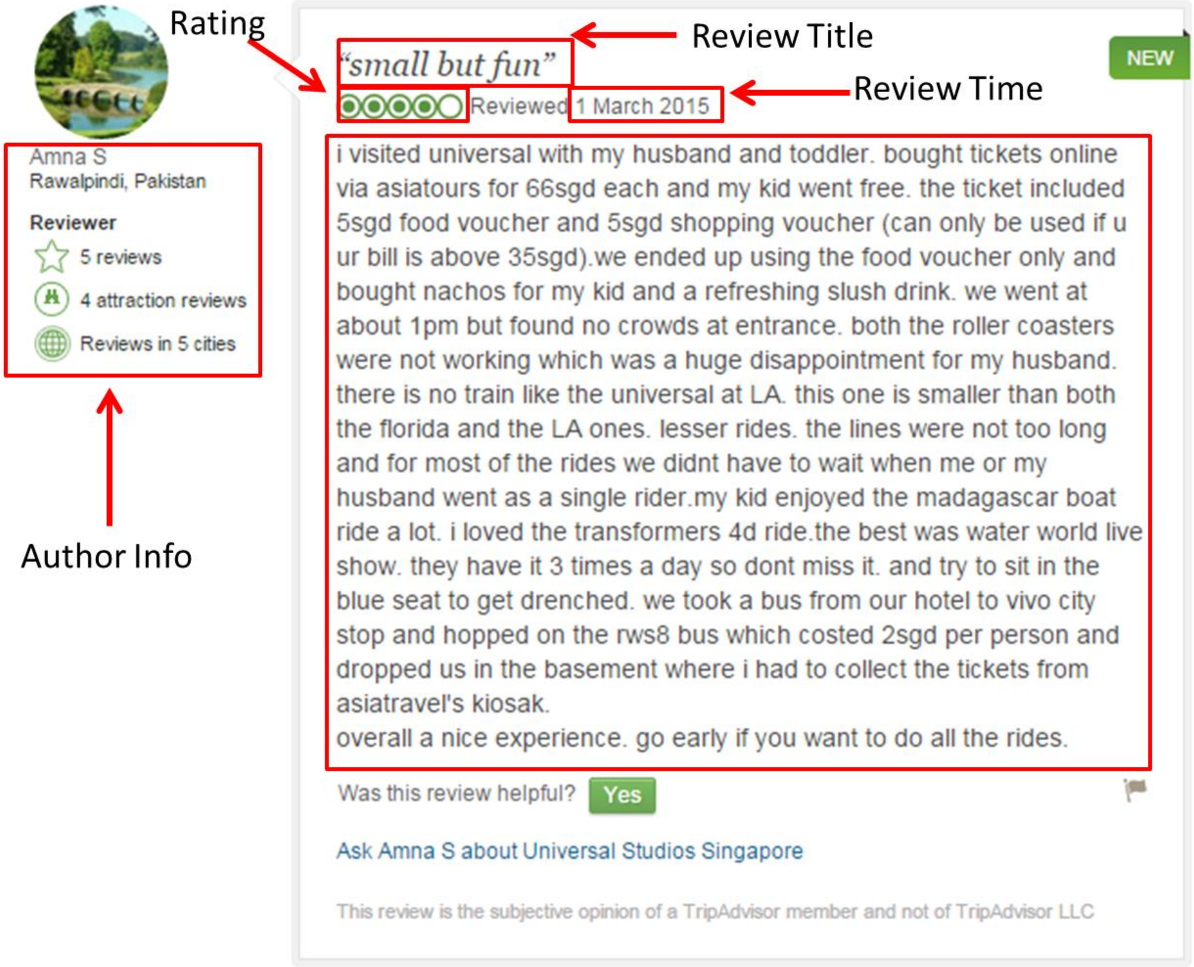
It is an Object/Relational Mapping framework, allowing mapping an object-oriented domain model to a traditional relational database. It relieves the developer from manual result set handling and object conversion.

4. SLF4J & log4j: http://www.slf4j.org/index.html & http://logging.apache.org/log4j/1.2/ (Logging)
5. (Optional) Selenium: http://docs.seleniumhq.org/ (It is a browser testing automation tools, in the case of some pages that do not contain the information we need in the HTML but only show in the HTML modified by Javascript, We use this tool to execute the Javascript and return us the resultant HTML)

Analysis of TripAdvisor

As we choose “Singapore” & “Attractions” as the keywords, we have limited the information to a small portion of the whole data from TripAdvisor. We want to fetch the Attraction with its information along with all the Reviews (in English) written for each Attractions. We define in general the each attraction as an Object of Class Attraction and each review as an Object of Class Review. One Attraction has many reviews.

Review



Amna S
Rawalpindi, Pakistan
Reviewer
5 reviews
4 attraction reviews
Reviews in 5 cities

“small but fun”

●●●●○ Reviewed 1 March 2015

i visited universal with my husband and toddler. bought tickets online via asiatours for 66sgd each and my kid went free. the ticket included 5sgd food voucher and 5sgd shopping voucher (can only be used if u ur bill is above 35sgd).we ended up using the food voucher only and bought nachos for my kid and a refreshing slush drink. we went at about 1pm but found no crowds at entrance. both the roller coasters were not working which was a huge disappointment for my husband. there is no train like the universal at LA. this one is smaller than both the florida and the LA ones. lesser rides. the lines were not too long and for most of the rides we didnt have to wait when me or my husband went as a single rider.my kid enjoyed the madagascar boat ride a lot. i loved the transformers 4d ride.the best was water world live show. they have it 3 times a day so dont miss it. and try to sit in the blue seat to get drenched. we took a bus from our hotel to vivo city stop and hopped on the rws8 bus which costed 2sgd per person and dropped us in the basement where i had to collect the tickets from asiatravel's kiosak.
overall a nice experience. go early if you want to do all the rides.

Was this review helpful? Yes

Ask Amna S about Universal Studios Singapore

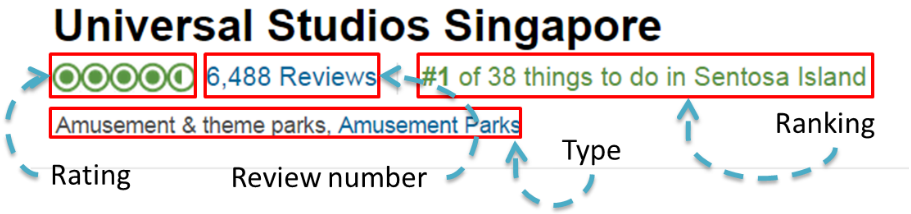
This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC

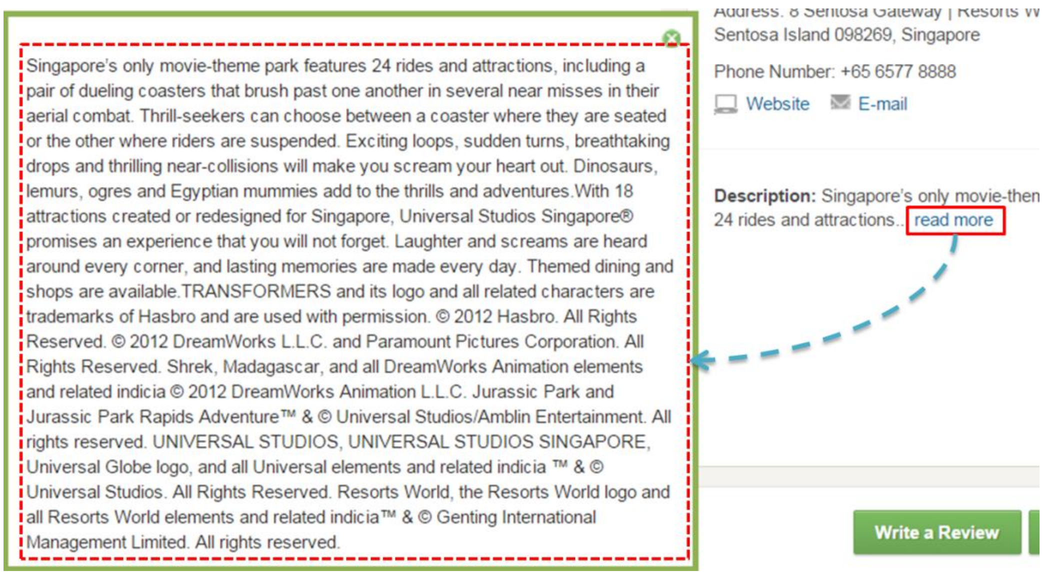
Page address: http://www.tripadvisor.com.sg/ShowUserReviews-g294264-d2439664-r257068853-Universal_Studios_Singapore-Sentosa_Island.html

Review ID:	**r257068853**
Attraction ID:	**d2439664**
Location ID:	**g294264**

We can get review title, review content, review time, rating and author information from the page (HTML)

Attraction


Universal Studios Singapore
6,488 Reviews
#1 of 38 things to do in Sentosa Island
Amusement & theme parks, Amusement Parks
Rating
Review number
Type
Ranking


Singapore's only movie-theme park features 24 rides and attractions, including a pair of dueling coasters that brush past one another in several near misses in their aerial combat. Thrill-seekers can choose between a coaster where they are seated or the other where riders are suspended. Exciting loops, sudden turns, breathtaking drops and thrilling near-collisions will make you scream your heart out. Dinosaurs, lemurs, ogres and Egyptian mummies add to the thrills and adventures.With 18 attractions created or redesigned for Singapore, Universal Studios Singapore® promises an experience that you will not forget. Laughter and screams are heard around every corner, and lasting memories are made every day. Themed dining and shops are available.TRANSFORMERS and its logo and all related characters are trademarks of Hasbro and are used with permission. © 2012 Hasbro. All Rights Reserved. © 2012 DreamWorks L.L.C. and Paramount Pictures Corporation. All Rights Reserved. Shrek, Madagascar, and all DreamWorks Animation elements and related indicia © 2012 DreamWorks Animation L.L.C. Jurassic Park and Jurassic Park Rapids Adventure™ & © Universal Studios/Amblin Entertainment. All rights reserved. UNIVERSAL STUDIOS, UNIVERSAL STUDIOS SINGAPORE, Universal Globe logo, and all Universal elements and related indicia ™ & © Universal Studios. All Rights Reserved. Resorts World, the Resorts World logo and all Resorts World elements and related indicia™ & © Genting International Management Limited. All rights reserved.

Address: 8 Sentosa Gateway | Resorts W
Sentosa Island 098269, Singapore

Phone Number: +65 6577 8888

Description: Singapore's only movie-then
24 rides and attractions... read more

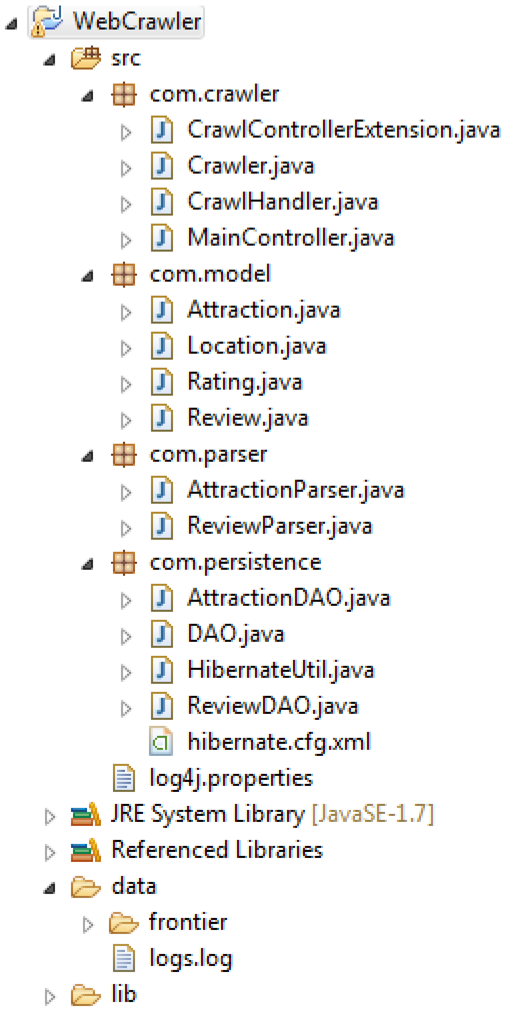
Page address: http://www.tripadvisor.com.sg/**Attraction_Review**-g294264-d2439664-Reviews-Universal_Studios_Singapore-Sentosa_Island.html

Attraction ID:	**d2439664**
Location ID:	**g294264**
Attraction Name:	**Universal_Studios_Singapore**

We can get rating, number of reviews, and detailed description from the web page (HTML)

We will also save the link address of the attraction imagine in the database for user query.

Code Implementation



```
WebCrawler
  src
    com.crawler
      CrawlControllerExtension.java
      Crawler.java
      CrawlHandler.java
      MainController.java
    com.model
      Attraction.java
      Location.java
      Rating.java
      Review.java
    com.parser
      AttractionParser.java
      ReviewParser.java
    com.persistence
      AttractionDAO.java
      DAO.java
      HibernateUtil.java
      ReviewDAO.java
      hibernate.cfg.xml
    log4j.properties
  JRE System Library [JavaSE-1.7]
  Referenced Libraries
  data
    frontier
    logs.log
  lib
```

com.crawler

Classes using the **crawler4j** library, for initiating and controlling the program and doing the craweling.

- **CrawlControllerExtension**: it extends the original crawler4j.crawler.CrawlController class for transfer a CrawlHandler into the Crawler class, which is not provided in the original class
- **Crawler**: it extends the original crawler4j.crawler.WebCrawler class for implement customized crawler. It overrides the *shouldVisit* and *visit* class for parsing the attraction and review pages.
- **CrawlHandler**: it is an intermediate class for calling parsers from the crawler when craweling pages needed and record information for each attraction and review.
- **MainController**: it contains the main function. It adds the seeds to the crawler and start craweling, and it also works Hibernate framework for saving data into the database.

com.model

Model Classes for different types of data, and **Hibernate** annotations in classes needed for storing. The Attraction, Review, classes will stored into database as Tables; one-to-many relationship will be established between Attraction-to-Review.

com.parser

Html parsers for Attraction page and Review pages, making use of **Jsoup**. It downloads the html of each page and parse the information we need and saves the data as Attraction/Review objects.

com.persistence

Data Access Object (DAO) classes for linking the database using **Hibernate** framework and Hibernate configuration files (*HibernateUtil.java* and *hbernate.cfg.xml*).

log4j.properties

Configurations for the log4j library for the log file

Data Storage

Hibernate (ORM) is used here for transferring the information into the database.

Three tables are created here: **Attraction**, **Review**, & **AttractionReviews**. The table ***AttractionReviews*** is specified in the *Attraction.java* code, as “Set<Review> _reviews” is instantiated in the class.

Attraction

The following is the columns (with data types)of the Attraction table (in *com.model.Attraction.java*), stores meta information for each attraction. The Hibernate Annotation is also showed in the following.

```
@Id
@Column(name="attractionId")
String _attractionId;         // Attraction ID: e.g. d2439664

@Column(name="attractionname")
String _name;                 // Attraction Name: e.g. Universal Studio Singapore

@Column(name="rating")
float _rating;                // Attraction Rating: e.g. 4.5, 5.0, etc.

@Column(name="numofreviews")
int _numOfReviews;            // Number of Reviews it has: e.g. 6488

@OneToMany(fetch = FetchType.EAGER)
@JoinTable(name="AttractionReviews",  // One-to-Many relationship:AttractionReviews
            joinColumns=@JoinColumn(name="AttractionId"),
            inverseJoinColumns=@JoinColumn(name="ReviewId"))
Set<Review> _reviews;         // Set of all the reviews this Attraction has

@Column(name="description")
@Type(type="text")
String _description;          // Attraction Description: text

@Column(name="imgsource")
String _imgSource;            // Link of Imagine of this Attraction

@Column (name="url")
String _url;                  // Attraction URL

@Column (name="location")
String _location;             // Attraction location

@Column (name="latitude")
```

```
String _gpsLatitude;          // Attraction latitude;             1.281566

@Column (name="longtitude")
String _gpsLongtitude;        // Attraction longitude;            103.86361
```

Review

The following is the columns (with data types)of the Review table (in *com.model.Review.java*), stores meta information for each review. The Hibernate Annotation is also showed in the following.

```
@Id
@Column(name="ReviewId")
String _reviewId;             // Review ID: e.g. r257068853

@Column(name="content")
@Type(type="text")
String _content = "";         // Review content: text

@Column(name="rating")
Rating _rating;               // Enum: terrible/poor/average/verygood/excellent

@Column(name="date")
@Temporal(TemporalType.DATE)
Date _date;                   // Review data: time stamp

@Column(name="title")
String _title;                // Review title: e.g. 'small but fun'

@Column(name="author")
String _author;               // Review Author: e.g. Amna S

@Column(name="wordcount")
int _wordCount = 0;           // Review word count
```

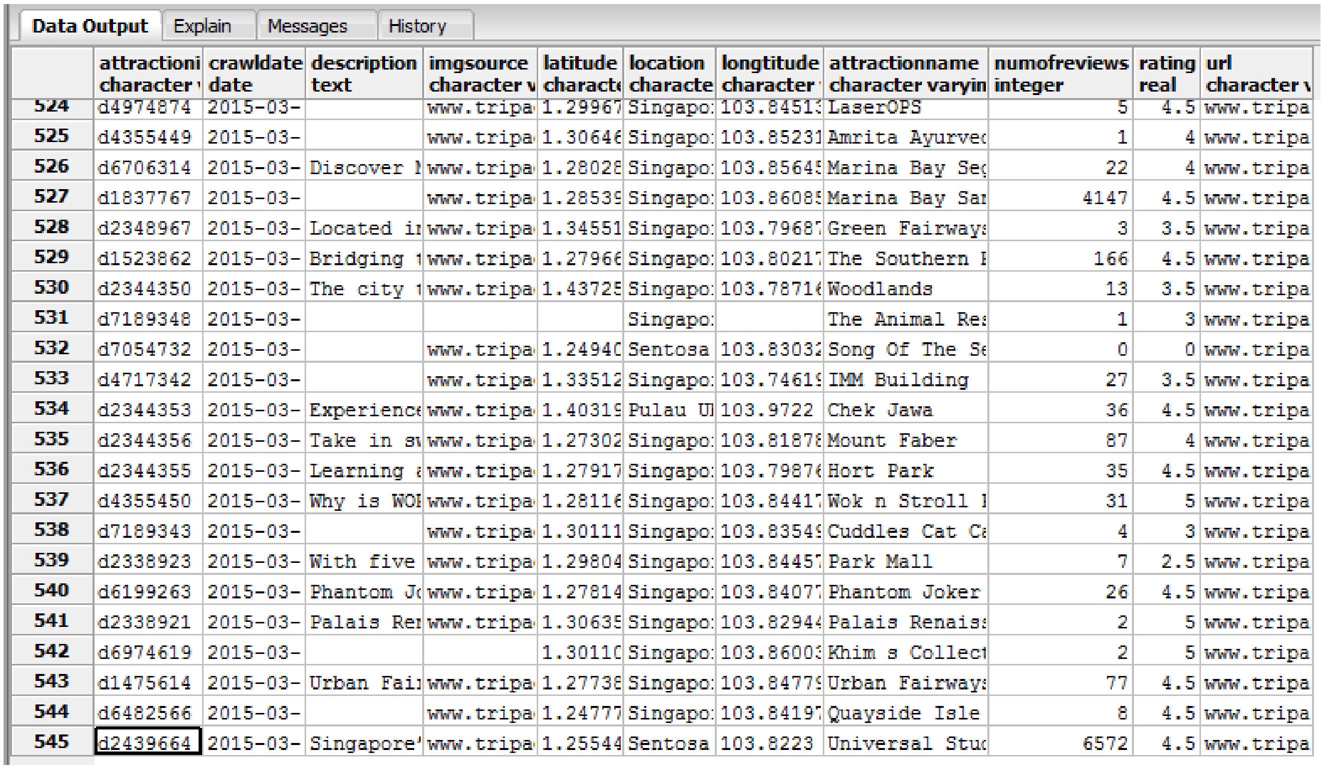
Hibernate Configuration file

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE hibernate-configuration PUBLIC
    "-//Hibernate/Hibernate Configuration DTD 3.0//EN"
    "http://hibernate.sourceforge.net/hibernate-configuration-3.0.dtd">
<hibernate-configuration>
```

```xml
    <session-factory>
        <property
name="hibernate.connection.driver_class">org.postgresql.Driver</property>
        <property
name="connection.url">jdbc:postgresql://localhost/tripadvisor</property>
        <property name="connection.username">postgres</property>
        <property name="connection.password"><!--Password-->
</property>
        <property
name="hibernate.dialect">org.hibernate.dialect.PostgreSQLDialect</property>
        <property
name="cache.provider_class">org.hibernate.cache.NoCacheProvider</property>
        <property name="hbm2ddl.auto">create</property>
        <property name="show_sql">true</property>
       <!-- Mapping the classes in com.model to tables in database -->
        <mapping class="com.model.Attraction"/>
         <mapping class="com.model.Review"/>
    </session-factory>
</hibernate-configuration>
```

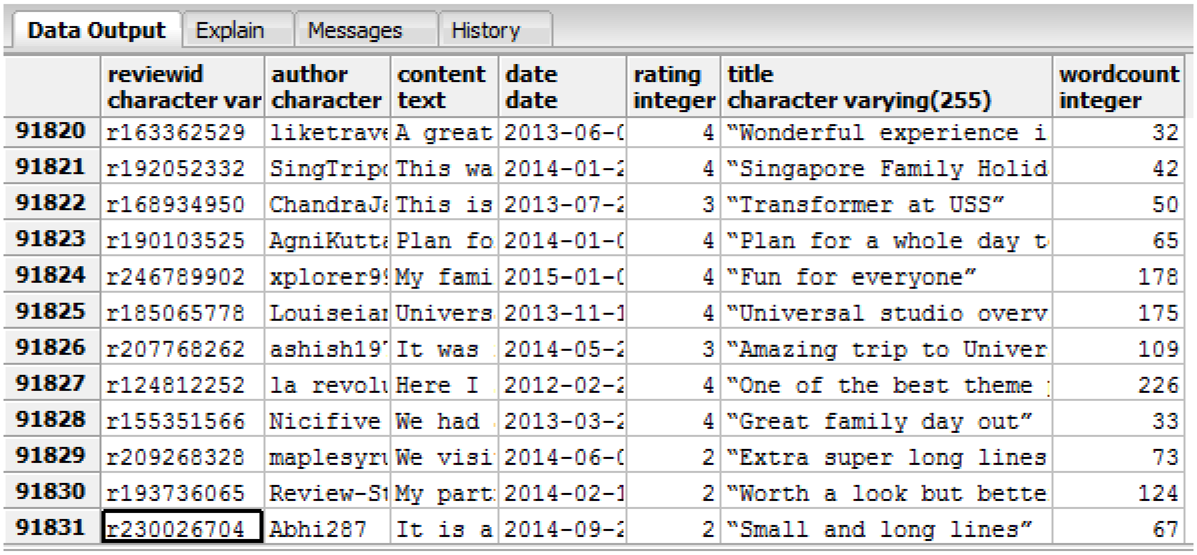
ScreenShot of Database

Attraction Table



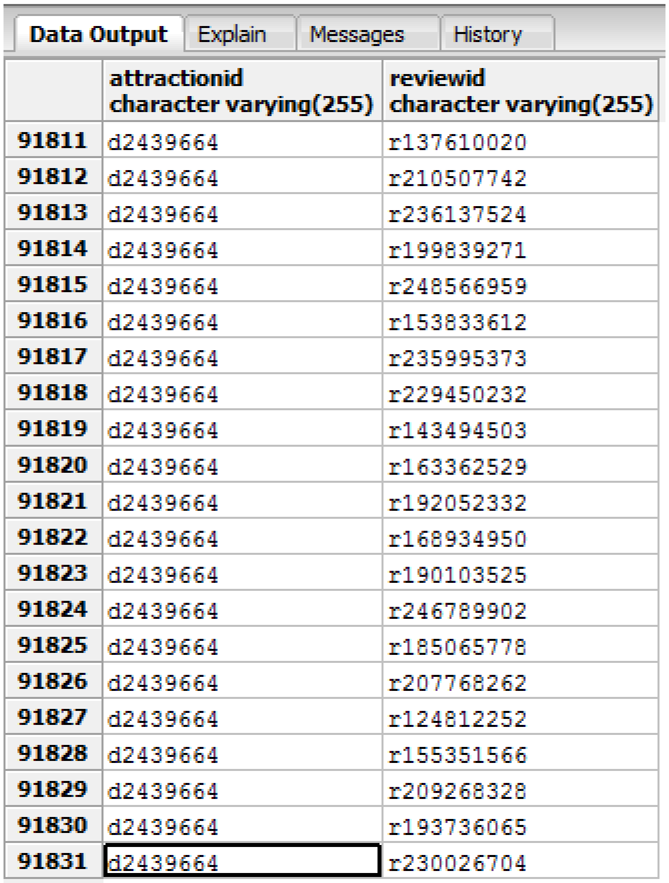
	attractioni	crawldate	description	imgsource	latitude	location	longtitude	attractionname	numofreviews	rating	url
	character	date	text	character v	characte	characte	character	character varyin	integer	real	character v
524	d4974874	2015-03-		www.tripa	1.29967	Singapo:	103.84513	LaserOPS	5	4.5	www.tripa
525	d4355449	2015-03-		www.tripa	1.30646	Singapo:	103.85231	Amrita Ayurved	1	4	www.tripa
526	d6706314	2015-03-	Discover I	www.tripa	1.28028	Singapo:	103.85645	Marina Bay Seg	22	4	www.tripa
527	d1837767	2015-03-		www.tripa	1.28539	Singapo:	103.86085	Marina Bay San	4147	4.5	www.tripa
528	d2348967	2015-03-	Located in	www.tripa	1.34551	Singapo:	103.79687	Green Fairways	3	3.5	www.tripa
529	d1523862	2015-03-	Bridging t	www.tripa	1.27966	Singapo:	103.80217	The Southern I	166	4.5	www.tripa
530	d2344350	2015-03-	The city t	www.tripa	1.43725	Singapo:	103.78716	Woodlands	13	3.5	www.tripa
531	d7189348	2015-03-				Singapo:		The Animal Res	1	3	www.tripa
532	d7054732	2015-03-		www.tripa	1.24940	Sentosa	103.83032	Song Of The Se	0	0	www.tripa
533	d4717342	2015-03-		www.tripa	1.33512	Singapo:	103.74619	IMM Building	27	3.5	www.tripa
534	d2344353	2015-03-	Experience	www.tripa	1.40319	Pulau Ub	103.9722	Chek Jawa	36	4.5	www.tripa
535	d2344356	2015-03-	Take in sw	www.tripa	1.27302	Singapo:	103.81878	Mount Faber	87	4	www.tripa
536	d2344355	2015-03-	Learning a	www.tripa	1.27917	Singapo:	103.79876	Hort Park	35	4.5	www.tripa
537	d4355450	2015-03-	Why is WOK	www.tripa	1.28116	Singapo:	103.84417	Wok n Stroll I	31	5	www.tripa
538	d7189343	2015-03-		www.tripa	1.30111	Singapo:	103.83549	Cuddles Cat Ca	4	3	www.tripa
539	d2338923	2015-03-	With five	www.tripa	1.29804	Singapo:	103.84457	Park Mall	7	2.5	www.tripa
540	d6199263	2015-03-	Phantom Jo	www.tripa	1.27814	Singapo:	103.84077	Phantom Joker	26	4.5	www.tripa
541	d2338921	2015-03-	Palais Ren	www.tripa	1.30635	Singapo:	103.82944	Palais Renais	2	5	www.tripa
542	d6974619	2015-03-			1.30110	Singapo:	103.86003	Khim s Collect	2	5	www.tripa
543	d1475614	2015-03-	Urban Fair	www.tripa	1.27738	Singapo:	103.84779	Urban Fairways	77	4.5	www.tripa
544	d6482566	2015-03-		www.tripa	1.24777	Singapo:	103.84197	Quayside Isle	8	4.5	www.tripa
545	d2439664	2015-03-	Singapore'	www.tripa	1.25544	Sentosa	103.8223	Universal Stud	6572	4.5	www.tripa

Review Table



	reviewid character var	author character	content text	date date	rating integer	title character varying(255)	wordcount integer
91820	r163362529	liketrave	A great	2013-06-	4	“Wonderful experience i	32
91821	r192052332	SingTrip	This wa	2014-01-	4	“Singapore Family Holid	42
91822	r168934950	ChandraJ	This is	2013-07-	3	“Transformer at USS”	50
91823	r190103525	AgniKutt	Plan fo	2014-01-	4	“Plan for a whole day t	65
91824	r246789902	xplorer9	My fami	2015-01-	4	“Fun for everyone”	178
91825	r185065778	Louiseia	Univers	2013-11-	4	“Universal studio overv	175
91826	r207768262	ashish19	It was	2014-05-	3	“Amazing trip to Univer	109
91827	r124812252	la revol	Here I	2012-02-	4	“One of the best theme	226
91828	r155351566	Nicifive	We had	2013-03-	4	“Great family day out”	33
91829	r209268328	maplesyr	We visi	2014-06-	2	“Extra super long lines	73
91830	r193736065	Review-S	My part	2014-02-	2	“Worth a look but bette	124
91831	r230026704	Abhi287	It is a	2014-09-	2	“Small and long lines”	67

AttractionReview Table



	attractionid character varying(255)	reviewid character varying(255)
91811	d2439664	r137610020
91812	d2439664	r210507742
91813	d2439664	r236137524
91814	d2439664	r199839271
91815	d2439664	r248566959
91816	d2439664	r153833612
91817	d2439664	r235995373
91818	d2439664	r229450232
91819	d2439664	r143494503
91820	d2439664	r163362529
91821	d2439664	r192052332
91822	d2439664	r168934950
91823	d2439664	r190103525
91824	d2439664	r246789902
91825	d2439664	r185065778
91826	d2439664	r207768262
91827	d2439664	r124812252
91828	d2439664	r155351566
91829	d2439664	r209268328
91830	d2439664	r193736065
91831	d2439664	r230026704

The numbers of records, words, and types (i.e., unique words) in the corpus

TripAdvisor Corpus

Number of Attractions: 545

Number of Reviews: 91,831 (in English) (Totally 114,795 reviews)

Number of words (for review): 7,012,188

Number of types (Unique words): 69517

Other statistics:

Crawling Elapsed time:	7h:51':18" (search 1 per second)
Links (Reviews) Not Retrieved (Non-English reviews):	22964
Total Reviews:	114795
Percentage of English reviews:	79.99564440959972%