Expedia Dataset

Welcome! This dataset was provided by the Expedia, an online travel agency (OTA) website: with logs of customer behavior. These include what customers searched for, how they interacted with search results (click/book), whether or not the search result was a travel package. **The data in this dataset is a random selection from Expedia and is not representative of the overall statistics.**

Context

Expedia Group is the world's travel platform, with an extensive brand portfolio that includes some of the world's most trusted online travel brands since 1996.

Expedia has in-house algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center, etc) are grouped together. These hotel clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data.

The datasets are split based on time: one from 2013 and 2014, while the other are from 2015. 2013-2014 data includes all the users in the logs, including both click events and booking events. 2015 data only includes booking events.

destinations.csv data consists of features extracted from hotel reviews text.

Note that some srch\_destination\_id's in the dataset files don't exist in the destinations.csv file. This is because some hotels are new and don't have enough features in the latent space. You should be able to handle this missing information whenever necessary.

## File descriptions

* 2013-2014.csv - the dataset about users in the logs between 2013 and 2014
* 2015.csv - the dataset including booking events in 2015
* destinations.csv - hotel search latent attributes

## Data fields

**2013-2014/2015.csv**

| Column name | Description | Data type |
| --- | --- | --- |
| date\_time | Timestamp | string |
| site\_name | ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...) | int |
| posa\_continent | ID of continent associated with site\_name | int |
| user\_location\_country | The ID of the country the customer is located | int |
| user\_location\_region | The ID of the region the customer is located | int |
| user\_location\_city | The ID of the city the customer is located | int |
| orig\_destination\_distance | Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated | double |
| user\_id | ID of user | int |
| is\_mobile | 1 when a user connected from a mobile device, 0 otherwise | tinyint |
| is\_package | 1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise | int |
| channel | ID of a marketing channel | int |
| srch\_ci | Checkin date | string |
| srch\_co | Checkout date | string |
| srch\_adults\_cnt | The number of adults specified in the hotel room | int |
| srch\_children\_cnt | The number of (extra occupancy) children specified in the hotel room | int |
| srch\_rm\_cnt | The number of hotel rooms specified in the search | int |
| srch\_destination\_id | ID of the destination where the hotel search was performed | int |
| srch\_destination\_type\_id | Type of destination | int |
| hotel\_continent | Hotel continent | int |
| hotel\_country | Hotel country | int |
| hotel\_market | Hotel market | int |
| is\_booking | 1 if a booking, 0 if a click | tinyint |
| cnt | Number of similar events in the context of the same user session | bigint |
| hotel\_cluster | ID of a hotel cluster | int |

**destinations.csv**

| Column name | Description | Data type |
| --- | --- | --- |
| srch\_destination\_id | ID of the destination where the hotel search was performed | int |
| d1-d149 | latent description of search regions | double |