Data Description

In this challenge, you are given a list of users along with their demographics, web session records, and some summary statistics. You are asked to predict which country a new user's first booking destination will be. All the users in this dataset are from the USA.

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. Please note that 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking.

The training and test sets are split by dates. In the test set, you will predict all the new users with first activities after 7/1/2014 (note: this is updated on 12/5/15 when the competition restarted). In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

File descriptions

* train\_users.csv - the training set of users
* test\_users.csv - the test set of users
  + id: user id
  + date\_account\_created: the date of account creation
  + timestamp\_first\_active: timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up
  + date\_first\_booking: date of first booking
  + gender
  + age
  + signup\_method
  + signup\_flow: the page a user came to signup up from
  + language: international language preference
  + affiliate\_channel: what kind of paid marketing
  + affiliate\_provider: where the marketing is e.g. google, craigslist, other
  + first\_affiliate\_tracked: whats the first marketing the user interacted with before the signing up
  + signup\_app
  + first\_device\_type
  + first\_browser
  + country\_destination: this is the target variable you are to predict
* sessions.csv - web sessions log for users
  + user\_id: to be joined with the column 'id' in users table
  + action
  + action\_type
  + action\_detail
  + device\_type
  + secs\_elapsed
* countries.csv - summary statistics of destination countries in this dataset and their locations
  + country\_destination
  + lat\_destination
  + lng\_destination
  + distance\_km
  + distance\_km2
  + destination\_language
  + language\_levenstein\_distance
* age\_gender\_bkts.csv - summary statistics of users' age group, gender, country of destination
* sample\_submission.csv - correct format for submitting your predictions