Data Challenges – Documentation

All Data Challenges: Start Saturday Apr 13 1pm BST (12 noon GMT+1) and End: Sunday Apr 14 1pm BST (12noon GMT + 1)

Please read the "*Data Challenges Rules*" document first.

1. Data Science Challenge

This challenge will be run as a Kaggle competition.

Log into your Kaggle account or sign up in Kaggle if you don't have an account.

Please refer to the documentation in Kaggle for all matters concerning this Data Science Challenge

The challenge is to Predict human judgments about who is more influential on social media. Please see section: "Prizes, Data Science Challenge"

Use the training and test datasets available in Kaggle. The dataset provided by **Peerindex** as a standard pairwise preference learning task. Each datapoint describes two individuals. Pre-computed, standardised features based on twitter activity (such as volume of interactions, number of followers, etc) will be provided for each individual. The discrete label represents a human judgement about which one of the two individuals is more influential. The goal of the challenge is to train a machine learning model which, for a pair of individuals, predicts the human judgement on who is more influential with high accuracy. Labels for the dataset have been collected by **PeerIndex** using an application similar to the one described in this post. This competition challenge will be hosted and run as a Kaggle competition.

The winners will be the Top 3 teams that deliver the most accurate models based on the measure benchmark specified in Kaggle.

2. Data Visualization Challenge

The challenge is to develop a cool, fun or interesting data visualization based on data provided by Peerindex with rich social influence data about the 300 most influential people in UK.

Please use the data visualization dataset loaded in your Windows Azure account.

This data represent the ~300 most influential twitter accounts in the UK. The data is in json format (1 json object per line, representing an influencer)

There are four files provided
twitter_lookup.json is the returned data from the lookup call to twitter's API. THis contains the number of followers, friends, lists, names, image urls, etc.
extended.json contains data returned by the extended call to the PeerIndex API. This includes PeerIndex influence scores, as well as Activity, Authority and Audience measures. All of these numbers are between 0 and 100.
topics.json Contains topic data for 8 topics, like Technology, News or Sports. The topic_score is also between 0 and 100
graph.json Contains a snapshot of the influence graph, with top 20 inbound and outbound influencers for each person.

Note that in general files are not ordered in the same order, but they can be joined by peerindex_id or twitter_id.

The Top 2 winners will be selected by a panel of judges appointed by Data Science London and UKWAG.

3. **2. Free Style Data Challenge**

The challenge is to develop a cool, fun, interesting, or innovative data hack, app or whatever you like with the free style datasets.

Please use the free style datasets loaded in your Windows Azure account.

The winners will be the Top 3 teams as selected solely on the criteria of a panel of judges appointed by Data Science London and UKAWUG.

Please use the data visualization dataset loaded in your Windows Azure account. These datasets are:

- The Bitcoin Dataset
  http://compbio.cs.uic.edu/data/bitcoin/
- The Twitter Census DataSets
  http://www.infochimps.com/datasets/twitter-census-developer-tools-mapping-from-twitter-user-search-
- The Friendster Dataset
  http://snap.stanford.edu/data/com-Friendster.html
- Gowalla Location based dataset
  http://snap.stanford.edu/data/loc-gowalla.html
- Social Memes dataset
  http://snap.stanford.edu/data/memetracker9.html
- Last.fm dataset
- http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html
- Beer Reviews dataset
  http://snap.stanford.edu/data/web-BeerAdvocate.html
- Wine Reviews dataset
  http://snap.stanford.edu/data/web-CellarTracker.html
- Movie reviews dataset
  http://snap.stanford.edu/data/web-Movies.html
- The Malicious URLs dataset
  http://sysnet.ucsd.edu/projects/url/