# make-ipinyou-data

This project is to formalise the iPinYou RTB data into a standard format for further researches.

## 百度云数据地址

http://pan.baidu.com/s/1kTwX2mF

## Step 0

Go to data.computational-advertising.org to download `ipinyou.contest.dataset.zip`. Unzip it and get the folder `ipinyou.contest.dataset`.

## Step 1

Update the soft link for the folder `ipinyou.contest.dataset` in `original-data`. `weinan@ZHANG:~/Project/make-ipinyou-data/original-data$ ln -sfn ~/Data/ipinyou.contest.dataset ipinyou.contest.dataset` Under `make-ipinyou-data/original-data/ipinyou.contest.dataset` there should be the original dataset files like this: `weinan@ZHANG:~/Project/make-ipinyou-data/original-data/ipinyou.contest.dataset$ ls algo.submission.demo.tar.bz2 README testing2nd training3rd city.cn.txt region.cn.txt testing3rd user.profile.tags.cn.txt city.en.txt region.en.txt training1st user.profile.tags.en.txt files.md5 testing1st training2nd` You do not need to further unzip the packages in the subfolders.

## Step 2

Under `make-ipinyou-data` folder, just run `make all`.

Note: The user agent parser used in `formalizeua.py` uses the python package user-agents.

After the program finished, the total size of the folder will be 14G. The files under `make-ipinyou-data` should be like this: `weinan@ZHANG:~/Project/make-ipinyou-data$ ls 1458 2261 2997 3386 3476 LICENSE mkyzxdata.sh python schema.txt 2259 2821 3358 3427 all Makefile original-data README.md` Normally, we only do experiment for each campaign (e.g. `1458`). `all` is just the merge of all the campaigns. You can delete `all` if you think it is unuseful in your experiment.

## Use of the data

We use campaign 1458 as example here. `weinan@ZHANG:~/Project/make-ipinyou-data/1458$ ls featindex.txt test.log.txt test.yzx.txt train.log.txt train.yzx.txt` * `train.log.txt` and `test.log.txt` are the formalised string data for each row (record) in train and test. The first column is whether the user click the ad or not.

The 14th column is the winning price for this auction. * `featindex.txt` maps the features to their indexes. For example, `8:115.45.195.* 29` means that the 8th column in `train.log.txt` with the string `115.45.195.*` maps to feature index `29`. * `train.yzx.txt` and `test.yzx.txt` are the mapped vector data for `train.log.txt` and `test.log.txt`. The format is y:click, z:wining_price, and x:features. Such data is in the standard form as introduced in iPinYou Benchmarking.

For any questions, please report the issues or contact Weinan Zhang.