Research Prediction Competition

Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

$25,000
Prize Money

Google · 112 teams · 2 months to go (2 months to go until merger deadline)

Training Data

key_1.csv.zip
sample_submission_1....
train_1.csv.zip

sample_submission_1.csv.zip 65.99 MB

Download

Data Introduction

The training dataset consists of approximately 145k time series. Each of these time series represent a number of daily views of a different Wikipedia article, starting from July, 1st, 2015 up until December 31st, 2016. The leaderboard during the training stage is based on traffic from January, 1st, 2017 up until March 1st, 2017.

The second stage will use training data up until September 1st, 2017. The final ranking of the competition will be based on predictions of daily views between September 10th, 2017 and November 10th, 2017 for each article in the dataset. You will submit your forecasts for these dates by September 10th.

For each time series, you are provided the name of the article as well as the type of traffic that this time series represent (all, mobile, desktop, spider). You may use this metadata and any other publicly available data to make predictions. Unfortunately, the data source for this dataset does not distinguish between traffic values of zero and missing values. A missing value may mean the traffic was zero or that the data is not available for that day.

To reduce the submission file size, each page and date combination has been given a shorter Id. The mapping between page names and the submission Id is given in the key files.

File descriptions

Files used for the first stage will end in '_1'. Files used for the second stage will end in '_2'. Both will have identical formats. The complete training data for the second stage will be made available prior to the second stage.

- **train_*.csv** - contains traffic data. This a csv file where each row corresponds to a particular article and each column correspond to a particular date. Some entries are missing data. The page names contain the Wikipedia project (e.g. en.wikipedia.org), type of access (e.g. desktop) and type of agent (e.g. spider). In other words, each article name has the following format: 'name_project_access_agent' (e.g. 'AKB48_zh.wikipedia.org_all-access_spider').
- **key_*.csv** - gives the mapping between the page names and the shortened Id column used for prediction
- **sample_submission_*.csv** - a submission file showing the correct format

---