kaggle Host Competitions Datasets Scripts Jobs Community ▾ Rupak Chakraborty Logout

Completed • $20,000

Predict Closed Questions on Stack Overflow

Tue 21 Aug 2012 – Sat 3 Nov 2012 (3 years ago)

Dashboard

Home
Data

Information
Description
Evaluation
Rules
Prizes
Jobs
Submission Instructions
Timeline
Visualization Prospect
Winners

Forum

Leaderboard
Public
Private

Visualization

Leaderboard

1.	Malacka
2.	nbu
3.	James
4.	F# with Composite Networks
5.	MaBu
6.	vikas
7.	Daniel Velkov
8.	Glen
9.	SquaredLoss
10.	Marco Lui

Forum (76 topics)

multiclass logloss in R
8 months ago

Access to the solutions now that the competition is closed?
2 years ago

Sharing my solution (Ranked #10)
2 years ago

Aftermath & What now?
2 years ago

Data Files

File Name	Available Formats
train	.7z (846.21 mb)
	.csv (3.47 gb)
	.zip (1.20 gb)
public_leaderboard	.7z (20.78 mb)
	.csv (88.11 mb)
	.gz (29.36 mb)
	.zip (30.65 mb)
train-sample	.csv (133.20 mb)
train-sample	.7z (33.30 mb)
	.gz (46.39 mb)
	.zip (48.49 mb)
basic_benchmark	.csv (7.58 mb)
uniform_benchmark	.csv (7.06 mb)
prior_benchmark	.csv (7.48 mb)
2012-07 Stack Overflow	.7z (6.20 gb)
test	.csv (797.72 kb)
private_leaderboard	.7z (23.38 mb)
	.csv (99.67 mb)
	.gz (32.92 mb)
	.zip (34.40 mb)
train_October_9_2012	.7z (923.43 mb)
	.csv (3.83 gb)
	.gz (1.30 gb)
	.zip (1.31 gb)
train-sample_October_9_2012_v2	.7z (42.93 mb)

Visualization Contest - any announcement date?
3 years ago

i am an idiot
3 years ago

teams

players

entries

	.csv (171.84 mb)
	.gz (59.76 mb)
	.zip (62.47 mb)

You only need to download one format of each file.
Each has the same contents but use different packaging methods.

The code for the benchmarks is on Github.

The training data contains data through July 31st UTC, and the public leaderboard data goes from August 1 UTC to August 14 UTC.

The train.csv file contains post text and associated metadata at the time of post creation which will serve as inputs to your solution. The state of the post as of July 31st is also included. It contains the following fields (not in this order):

- Input
  - PostCreationDate
  - OwnerUserId
  - OwnerCreationDate
  - ReputationAtPostCreation
  - OwnerUndeletedAnswerCountAtPostTime
  - Title
  - BodyMarkdown
  - Tag1
  - Tag2
  - Tag3
  - Tag4
  - Tag5
- Output
  - OpenStatus
- Additional Data
  - PostId
  - PostClosedDate

The public leaderboard data contains all of the above fields, except for the target field OpenStatus and PostClosedDate.

The file train-sample.csv is a stratified sample of the training data: it contains every closed question and an equally-sized random sample of the open questions in the training data.

All questions will have a value in Tag1, but Tags 2 through 5 are optional.

To convert the user submitted Markdown found in BodyMarkdown to HTML if desired, our open source implementations in C# and Javascript may be useful.

Additional data can be found in "2012-07 Stack Overflow.7z", which contains an entire public data dump of Stack Overflow. Descriptions of the values can be found in the archive itself as well as on Meta Stack Overflow. This data will not be available as inputs, but may be useful in building your solution. As it is rather large (6GB) you may find it easier to download as a torrent, more details can be found in this forum post.