kaggle Host Competitions Datasets Scripts Jobs Community ▾ Rupak Chakraborty Logout

Completed • $5,000 • 625 teams

StumbleUpon Evergreen Classification Challenge

Fri 16 Aug 2013 – Thu 31 Oct 2013 (2 years ago)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline
- Winners

Forum

Leaderboard
- Public
- Private

My Submissions

Leaderboard

1. fchollet
2. Maarten Bosma
3. michaelp
4. Marco Lui
5. Guru
6. Abhishek
7. Yevgeniy
8. JedF
9. Issam Laradji
10. tks

Forum (52 topics)

usage of data for acedemic purpose
3 months ago

Beating the Benchmark (Leaderboard AUC ~0.878)
6 months ago

Beating the benchmark and getting AUC = 0.75109
12 months ago

Improving on R
16 months ago

Data Ranges: compression_ratio and embed_ratio

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
raw_content	.zip (157.13 mb)
sampleSubmission	.csv (21.53 kb)
test	.tsv (8.99 mb)
train	.tsv (20.96 mb)
Data Access Agreement	.docx (130.30 kb)

Note: researchers who wish to use this data outside the competition should download and read the data access agreement.

There are two components to the data provided for this challenge:

The first component is two files: **train.tsv** and **test.tsv**. Each is a tab-delimited text file containing the fields outlined below for 10,566 urls total. Fields for which no data is available are indicated with a question mark.

- **train.tsv** is the training set and contains 7,395 urls. Binary evergreen labels (either evergreen (1) or non-evergreen (0)) are provided for this set.
- **test.tsv** is the test/evaluation set and contains 3,171 urls.

The second component is **raw_content.zip**, a zip file containing the raw content for each url, as seen by StumbleUpon's crawler. Each url's raw content is stored in a tab-delimited text file, named with the urlid as indicated in train.tsv and test.tsv.

The following table includes field descriptions for train.tsv and test.tsv:

FieldName	Type	Description
url	string	Url of the webpage to be classified
urlid	integer	StumbleUpon's unique identifier for each url
boilerplate	json	Boilerplate text
alchemy_category	string	Alchemy category (per the publicly available Alchemy API found at www.alchemyapi.com)
alchemy_category_score	double	Alchemy category score (per the publicly available Alchemy API

20 months ago

Is it possible to get the true labels of test data?
22 months ago

teams

players

entries

		found at www.alchemyapi.com)
avglinksize	double	Average number of words in each link
commonLinkRatio_1	double	# of links sharing at least 1 word with 1 other links / # of links
commonLinkRatio_2	double	# of links sharing at least 1 word with 2 other links / # of links
commonLinkRatio_3	double	# of links sharing at least 1 word with 3 other links / # of links
commonLinkRatio_4	double	# of links sharing at least 1 word with 4 other links / # of links
compression_ratio	double	Compression achieved on this page via gzip (measure of redundancy)
embed_ratio	double	Count of number of <embed> usage
frameBased	integer (0 or 1)	A page is frame-based (1) if it has no body markup but have a frameset markup
frameTagRatio	double	Ratio of iframe markups over total number of markups
hasDomainLink	integer (0 or 1)	True (1) if it contains an <a> with an url with domain
html_ratio	double	Ratio of tags vs text in the page
image_ratio	double	Ratio of <img> tags vs text in the page
is_news	integer (0 or 1)	True (1) if StumbleUpon's news classifier determines that this webpage is news
lengthyLinkDomain	integer (0 or 1)	True (1) if at least 3 <a> 's text contains more than 30 alphanumeric characters
linkwordscore	double	Percentage of words on the page that are in hyperlink's text
news_front_page	integer (0 or 1)	True (1) if StumbleUpon's news classifier determines that this webpage is front-page news
non_markup_alphanum_characters	integer	Page's text's number of alphanumeric characters
numberOfLinks	integer	Number of <a> markups
numwords_in_url	double	Number of words in url
parametrizedLinkRatio	double	A link is parametrized if it's url contains parameters or has an attached onClick event
spelling_errors_ratio	double	Ratio of words not found in wiki (considered to be a spelling mistake)
label	integer (0 or 1)	User-determined label. Either evergreen (1) or non-evergreen (0); available for train.tsv only

About Our Team Careers Terms Privacy Contact/Support