

Search kaggle

Q

Competitions

Datasets

Kernels

Discussion

Jobs



Crowdflower Search Results Relevance

Predict the relevance of search results from eCommerce sites

\$20,000 · 1,326 teams · 2 years ago

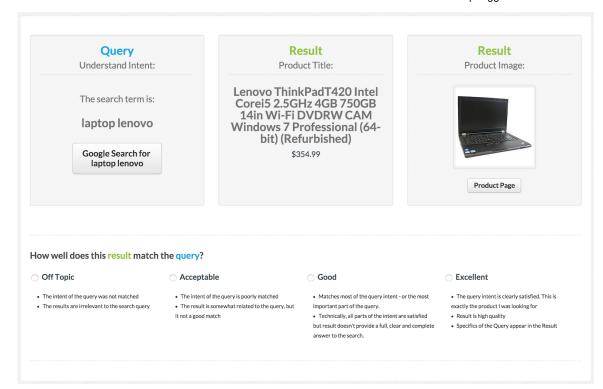
Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Competition Data		Edit
sampleSubmission.csv test.csv.zip train.csv.zip	train.csv.zip 1.87 MB	å Download

Data Description

See this script for a quick exploration of the data

To evaluate search relevancy, CrowdFlower has had their crowd evaluate searches from a handful of eCommerce websites. A total of 261 search terms were generated, and CrowdFlower put together a list of products and their corresponding search terms. Each rater in the crowd was asked to give a product search term a score of 1, 2, 3, 4, with 4 indicating the item completely satisfies the search query, and 1 indicating the item doesn't match the search term.



The challenge in this competition is to predict the relevance score given the product description and product title. To ensure that your algorithm is robust enough to handle any noisy HTML snippets in the wild real world, the data provided in the product description field is raw and contains information that is irrelevant to the product.

To discourage hand-labeling the data, CrowdFlower has also provided extra data that was not labeled by the crowd in the test set. This data is ignored when calculating your score.

Ready to explore the data? Scripts is the most frictionless way to get familiar with the competition dataset! See the data at a glance here. No download needed to start publishing and forking code in R and Python. It's already preloaded with our favorite packages and ready for you to start competing!

File and data descriptions

- train.csv the training data set includes:
 - o id: Product id
 - o query: Search term used
 - product_description: The full product description along with HTML formatting tags
 - o median_relevance: Median relevance score by 3 raters. This value is an integer between 1 and 4.
 - relevance_variance: Variance of the relevance scores given by raters.
- test.csv the test set
 - o id: Product id
 - o query: Search term used
 - product description: The full product description along with HTML formatting tags
- sampleSubmission.csv a sample submission file in the correct format

External data, such as dictionaries, thesaurus, language corpuses, are allowed. However, they must not be directly related to this specific dataset. The source of your external data must be posted to the forum to ensure fairness for all the participants in the community.

© 2017 Kaggle Inc

Our Team Terms Privacy Contact/Support





