kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs Sign In

Photo Quality Prediction

Given anonymized information on thousands of photo albums, predict whether a human evaluator would mark them as 'good'.

$5,000 · 200 teams · 6 years ago

Overview Data Discussion Leaderboard Rules

Competition Data

Edit

- example_entry.csv
- test.csv
- training.csv

training.csv 5.45 MB

Download

Data Description

Data format

For anonymity reasons, the caption, title and description texts have been broken down into word tokens, and the most common (excluding stop words) have been encoded as numbers. The fields are:

- *latitude*: The integrally-rounded latitude of the location of the album
- *longitude*: The integrally-rounded longitude of the location of the album
- *width*: The width of the images in the album, in pixels
- *height*: The height of the images in the album, in pixels
- *size*: The number of photos in the album
- *name*: The common tokens in the name of the album
- *description*: The common tokens in the description of the album
- *caption*: The common tokens in all the captions of the photos within the album
- *good*: 1 means the human reviewer liked the album, 0 means they didn't. This is the variable we want to predict.

Notes

From our own experiments, we know there's obvious correlations that make sense, such as areas in Africa rich in wildlife having a high proportion of good photos, and certain words that are associated with poor albums. Our goal is to turn some of those informal correlations into a more rigorous model we can reuse on much larger data sets.

To encode the text I'm creating a whitelist of words, consisting of non-stop-words that occur more than twenty times in different albums. For each word in that list, I assign a numerical value, and write out the numbers of all the words that appear in each album's title, description and photo captions.

The main goal of this exercise is to ensure that there's not enough information to reconstruct the text, whilst still having enough content for a prediction algorithm to sink its teeth into. I've done some preliminary work to make sure there are promising signals in the data that's remaining, and there do seem to be strong correlations, so I'm hopeful it's a good compromise.

Why don't we make the image data available? Our application needs to make quality decisions on very large volumes of externally-hosted images in a short amount of time, so image processing solutions aren't feasible for us. We're hopeful that the meta-data will provide a good enough approximation to be useful

© 2017 Kaggle Inc

Our Team Terms Privacy Contact/Support