Rakuten Data Challenge

Taxonomy Classification for eCommerce-scale Product Catalogs

SIGIR eCom workshop, 2018

Call For Participation

The SIGIR eCom workshop is organizing a Data Challenge as part of the workshop. The data is provided by Rakuten Institute of Technology, Boston (RIT-Boston), a dedicated R&D organization for the Rakuten group.

The dataset has **1 million** titles and **~3400 labels**, unbalanced class sizes.

Challenge website: https://sigir-ecom.github.io/data-task.html

Important Dates:

- Data Challenge Registration Deadline - May 15, 2018
- System Description Paper Submission - June 1, 2018
- Paper Acceptance Notification - June 15, 2018
- Final Leaderboard - June 24, 2018
- SIGIR eCom Full day Workshop - July 12, 2018

Task Description:

This challenge focuses on the topic of large-scale taxonomy classification where the goal is to predict each product's category as defined in the taxonomy tree given product's title. The cataloging of product listings through taxonomy categorization is a fundamental problem for any e-commerce marketplace, with applications ranging from personalized search recommendations to query understanding.

For example, in the Rakuten.com catalog, "Dr. Martens Air Wair 1460 Mens Leather Ankle Boots" is categorized under the "Clothing, Shoes & Accessories -> Shoes -> Men -> Boots" leaf. However, manual and rule based approaches to categorization are not scalable since commercial product taxonomies are organized in tree structures with three to ten levels of depth and thousands of leaf nodes.

Advances in this area of research have been limited due to the lack of real data from actual commercial catalogs. The challenge presents several interesting research aspects due to the intrinsic noisy nature of the product labels, the size of modern eCommerce catalogs, and the typical unbalanced data distribution.

Participation and Data

The data challenge is open to everyone.

As part of this challenge, Rakuten will be releasing 1M product listings in tsv format, including the train (0.8M) and test set (0.2M), consisting of product titles and their corresponding category ID paths. Details about evaluation metrics and other aspects of the task can be found at the website: https://sigir-ecom.github.io/data-task.html