SIGIR 2020 E-Commerce Workshop Data Challenge

Rakuten Institute of Technology

ACM Reference Format:
Rakuten Institute of Technology. 2020. SIGIR 2020 E-Commerce Workshop Data Challenge. In *SIGIR '20: ACM SIGIR, xxxx xx–xx, 20xx, xxxxx, xx*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Rakuten France Multi-modal Product Data Classification challenge is organized by Rakuten Institute of Technology, the research and innovation department of Rakuten group. This challenge focuses on the topic of large-scale multi-modal (text and image) classification, where the goal is to predict each product's `type code` as defined in the catalog of Rakuten France.

In the taxonomy of Rakuten France, products sharing the same product `type code` share the same exact array of attributes fields and possible values. Product `type codes` are numbers that match a generic product name, such as 1500 - `Watches`, 120 - `Laptops`, and so on. In that sense, the `type code` of a product is its category label.

The cataloging of product listings through some type of text or image categorization is a fundamental problem for any e-commerce marketplace, with applications ranging from personalized search and recommendations to query understanding. Manual and rule-based approaches to categorization are not scalable since commercial products are organized in many and sometimes thousands of classes. When actual users categorize product data, it has often been seen that not only the text of the title and description of the product is useful but also its associated images.

Advances in this area of research have been limited due to the lack of real product data from actual commercial catalogs. The challenge presents several interesting research aspects due to the intrinsic noisy nature of the product labels and images, the size of modern e-commerce catalogs, and a highly pronounced long tail phenomenon of the data distribution that follows a power law.

2 PROBLEM DESCRIPTION

The goal of this data challenge is to solve a fairly large-scale multi-modal (text and image) product data classification into `product type codes`.

For example, in the product catalog of Rakuten France, a product with a French title *Klarstein Présentoir 2 Montres Optique Fibre* is associated with an image and sometimes with an additional description. This product is categorized with a `product type code` of **1500**. There are other products with different titles, images and with possible descriptions, which are under the same `product type code`. Given these information on the products, like the example above, this challenge proposes participating teams build and submit systems that classify previously unseen products into their corresponding `product type codes`.

The main challenges for this task are as follows:

(1) **Multi-modal classification**. Given a training set of products and their `product type codes`, predict the corresponding `product type codes` for an unseen held out test set of products. The systems are free to use the available textual titles and/or descriptions whenever available and additionally the images to allow for true multi-modal learning.
(2) **Cross-modal retrieval**. Given an held-out test set of product items with their titles and (possibly empty) descriptions, predict the best image from among a set of test images that correspond to the products in the test set.

The difficulty in solving the tasks stems from the following observations:

- Highly imbalanced number of samples within the classes.
- Length of titles can vary – they can sometimes consist of one or two words.
- Descriptions, when present, may be a verbose representation of the product rather than a very specific one with precisely defined attributes for the product.
- Images may not be "clean". Some images could be of low quality, while some images may have text in them as often found in a banner.

3 DATA DESCRIPTION

For this challenge, **Rakuten France** will be releasing approximately 99K product listings in `tsv` format, including a training (84,916) and two test sets (9,372 samples for the classification and 4,440 samples for the retrieval task). The dataset consists of product titles, product descriptions, product images and their corresponding `product type code`.

The test set will be released towards the end of the data challenge. Furthermore, one can assume the test set has been generated from the same data distribution as the training set. This data challenge will be held in two phases which includes model building and model evaluation. Regarding each phase, there will be a separate test set for each task to be provided.

The data is divided using two criteria, forming six distinct sets: training or test, input or output for two tasks and two phases.

(1) `X_train.tsv`: training input file
(2) `Y_train.tsv`: training output file
(3) `x_test_task1_phase1.tsv`: test input file for task1 phase1
(4) `x_test_task1_phase2.tsv`: test input file for task1 phase2
(5) `x_test_task2_phase1.tsv`: test input file for task2 phase1
(6) `x_test_task2_phase2.tsv`: test input file for task2 phase2

It should be noted that each phase will be released relatively, and participants must produce:

SIGIR '20, xxxx xx–xx, 20xx, xxxxx, xx

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

y_test_task1_phase1_pred.tsv: phase1 task1 system output
y_test_task1_phase2_pred.tsv: phase2 task1 system output
y_test_task2_phase1_pred.tsv: phase1 task2 system output
y_test_task2_phase2_pred.tsv: phase2 task2 system output
for each task they may want to submit their results.

Additionally an image file is supplied containing all the images. It includes two sub-folders named image_training and image_test, containing training and test images respectively.

Also, catalog_english_taxonomy.tsv tab-separated file can demonstrate the correspondence between each product type code (abbreviated Prdtypecode) and its top level category in English. As an example:

Prdtypecode	Top level category
2280	Books
1280	Child

The first line of the input files contains the header, and the columns are tab-separated. The columns are:

(1) **Integer_id** - An integer Id for the product. This Id is used to associate the product with its corresponding product type code.
(2) **Title** - The product title, a short text summarizing the product.
(3) **Description** - A more detailed text describing the product. Not all the merchants use this field, so to retain originality of the data, *the description field can contain "NaN" values for many products.*
(4) **Image_id** - An unique Id for the image associated with the product.
(5) **Product_id** - An unique Id for the product.

The fields Image_id and Product_id are used to retrieve the images from the respective image folder. For a particular product, the image file name is identified as:

image_Image_id_product_Product_id.jpg

Table 1 displays three different lines of the training file, and Figure 1 shows the corresponding images for these three products. The examples are selected from the head, torso, and tail of the distribution, where two of which have descriptions and one without.

All the images corresponding to the products listed in training set, X_train.tsv can be found in image_training sub-folder, and all the images corresponding to the held-out test set can be found in image_test sub-folder.

Each line in the tab separated training output file, Y_train.tsv, contains the Prdtypecode, the category for the classification task corresponding to the Integer_id, Image_id, and Product_id tuple in the training input file (X_train.tsv). Here also the first line of the file is the header and columns are tab-separated.

An example readout of the training output file corresponding to the examples in Table 1 is shown below:

Integer_id	Image_id	Product_id	Prdtypecode
2	938777978	201115110	50
40001	1273112704	3992402448	1920
84915	684671297	57203227	2522

For the tab separated test input file, x_test_task1_phase1.tsv or x_test_task1_phase2.tsv , participants need to provide a test output file ,y_test_task1_phase1_pred.tsv or y_test_task1_phase2_pred.tsv, in the same format as the training output file (associating each Integer_id, Image_id, and Product_id tuple with the predicted Prdtypecode). The first line of this test output file should contain the header:

Integer_id Image_id Product_id Prdtypecode

It should be noted that the product titles and descriptions are for the vast majority written in French (99%), although, one can find some outlying samples related to other languages like English, German, and Spanish. The images are all squares of dimensions 500 x 500 px^2, which can have white or black borders included.

The second task is **cross-modal retrieval**, where the systems are asked to match the textual part of the product to the corresponding image. To accomplish this, we will supply a separate test input file and a separate test images sub-folder within the images folder:

(1) x_test_task2_phase1.tsv: phase1 task2 test input file
(2) x_test_task2_phase2.tsv: phase2 task2 test input file
(3) image_test.cross-modal sub-folder of images for cross-modal retrieval

This tab separated test input file will have the following fields:

Integer_id Title Description Product_id

The image filenames in the sub-folder of images for cross-modal retrieval will be identified as: image_Image_id.jpg. Note that the Product_id suffixes will be missing from these image file names.

For each phase, a tab separated test output file named y_test_task2_phase1_pred.tsv or y_test_task2_phase2_pred.tsv needs to be produced by each participating system. This file will have the following header as the first line:

Integer_id Product_id Image_id

where the Integer_ids and Product_ids will be those found in the test input file and the corresponding Image_ids will be the ids from the filenames of the most relevant images predicted.

3.1 Evaluation Phases

Stage 1 - Model Building (April 20 - July 15) Participants build and test models on the training data. The leader board only shows the model performance on a SUBSET of the test set according to your LATEST submission. Each team can submit at most 4 times per day (UTC time) in this stage.

Stage 2 - Model Evaluation (July 15 - July 23) The final leaderboard will freeze on July 23, and show the model performance on the remaining held out test set according to your LATEST submission. In this stage each team can submit at most 7 times during the time period that the evaluation is open, and there must be a period of 24 hours between two submissions.

3.2 Submission instructions

The submission is team-based, so only team leader can submit the prediction files. There is no limit on maximum team size. After the registration, participants will be provided with a unique ID to be used later for the submission purpose. The prediction file for each task has to be the same tsv format with the same name that mentioned before otherwise system will reject the submission.

Table 1: Three samples in the X_train.tsv file

Integer_id	Title	Description	Image_id	Product_id
2	Grand Stylet Ergonomique Bleu Gamepad ...	PILOT STYLE Touch Pen ...	938777978	201115110
40001	Drapeau Américain Vintage Oreiller ...	Vintage American Flag Pillow Cases ...	1273112704	3992402448
84915	Gomme De Collection 2 Gommes Pinguin ...	NaN	684671297	57203227

Figure 1: Images of the three example products shown in Table 1.

(a) `image_938777978_product_201115110.jpg`; Category: Entertainment

(b) `image_1273112704_product_3992402448.jpg`; Category: Household

(c) `image_684671297_product_57203227.jpg`; Category: Books

Each team can submit the results for one or both tasks concerning 4 times limitation per day.

3.3 Evaluation Metric

Since in this challenge, we are dealing with many classes with highly asymmetric number of samples, an item weighted metric used to rank the participants will not reveal the deficiencies of the classification algorithms.

Task 1: We will use the **macro-F1 score** to evaluate `product type code` classification on held out test samples. The score is understood as the arithmetic average of per-`product type code` F1 score.

Task 2: For the cross-modal retrieval task, the systems will be evaluated on **recall at 1 (R@1)** on held out test samples. The score is understood to be the average of the per-sample scoring of 1 if the image returned matches the title and 0 otherwise.

We will further make the evaluation scripts available with the Data Challenge.

4 LEGAL NOTICE

By express derogation from any preexisting or future contractual documents and/or terms and conditions pertaining to the Rakuten Data Challenge occurring on the occasion of the SIGIR 2020 Workshop on eCommerce ("Rakuten Data Challenge"), the participant ("Participant") acknowledges that the study data ("Study Data") uploaded by Rakuten France (the "Provider") on the occasion of the Rakuten Data Challenge is strictly considered as confidential information. The Participant unreservedly undertakes to (i) hold in strict confidence and not disclose to any third party all or part of the Study Data, (ii) use the Study Data for the sole purpose of the good performance of the Rakuten Data Challenge (the "Purpose"), (iii) not use, apply, reveal, report, publish, extract all or part of the Study Data or otherwise disclose all or part of the Study Data in any circumstances for a purpose other than the Purpose, excluding notably commercial or technical use of any kind. As of the termination of the Rakuten Data Challenge, the Participant shall immediately cease any use of the Study Data unless otherwise agreed by the Provider. The present specific terms shall remain in full force and effect until the termination of the Purpose and for a period of two (2) years following the termination date of the Purpose.

5 CONTACT

If you have any question, please contact:
Hesam Amoualian (hesam.amoualian@rakuten.com) or
Parantapa Goswami (parantapa.goswami@rakuten.com)