TASK DESCRIPTION

This challenge is entity type prediction over linked data. In this challenge, 1897 URLs are provided and 1397 of them are provided with label information. The task is the classification of the 500 unlabeled URLs. The linked data about all the 1897 URLs are also provided.

1) TRAIN/TEST DATA DESCRIPTION

The labeled samples are provided in train.dat (http://www.jist2015.org/challenge/train.dat), and unlabeled samples are provided in test.dat (http://www.jist2015.org/challenge/test.dat).

Each line of the tain.dat is as follows. In detail, each line contains two columns, URL and the label, and the two columns are separated by tab.

URL	label
http://zhishi.me/baidubaike/resource/%FF%FE%CB%79%30%75%02%5E	city

Each line of test.dat is as follows. In detail, each line corresponds to an URL.

URL
http://zhishi.me/baidubaike/resource/%E6%88%91%E4%B8%BA%E6%AD%8C%E7%8B%82

2) LABEL DESCRIPTION

There are 10 labels. The distribution of these labels in the labeled samples is as follows: insect (124), university (157), game (143), politician (134), city (139), song (139), novel (150), scene (130), cartoon (134), actor (147)

3) LINKED DATA DESCRIPTION

There are four triple stores that can be used to predict the labels of the provided URLs. These triple stores were encoded from a portion of Zhishi.me, the linked data for Baidu Baike and Hudong Baike. All the triple stores are provided in *.nt files that can be read from many tools or API such as OWLAPI (version 3.0 or above).

Note: participants may also use other resources for this task besides the provided triple stores.

Triple store	Meaning
http://www.jist2015.org/challenge/zhishime_abstracts_zh_enc.nt	Abstracts
http://www.jist2015.org/challenge/zhishime_infobox_properties_zh_enc.nt	Infobox data

http://www.jist2015.org/challenge/zhishime_external_links_zh_enc.nt	External URLs
http://www.jist2015.org/challenge/zhishime_related_pages_zh_enc.nt	Related URLs

EVALUATION

The provided unlabeled samples are used as the evaluation set and the F-measure is used as the evaluation metric. This metric is the harmonic mean of precision and recall. All of them are defined as follows, where a_i is the number of URLs that are actually in label i and also predicted in label i, b_i is the number of URLs that are predicted in label i, c_i is the number of URLs that are actually in label i.

$$P = \sum_{i=1}^{10} \frac{a_i}{b_i} \quad R = \sum_{i=1}^{10} \frac{a_i}{c_i} \quad F = \frac{2PR}{P+R}$$

SUBMISSION GUIDELINE

Each submission include: 1) the result file, and 2) the associated documentation with environment setting and algorithm description. The above materials should be sent to both kliu@nlpr.ia.ac.cn and jfdu@gdufs.edu.cn before the deadline, and the email title should be 'teamname+datachallenge'.

The result file should be named "result.dat", and the format must be the same as that of train.dat, i.e. each line contains two columns, URL and the label, and the two columns are separated by tab.

The associated documentation should be named “datachallenge.pdf” and provided in the PDF format, using the style of the Springer Publications format for Lecture Notes in Computer Science (LNCS). The document must be no longer than *5* pages.

If you have any questions, please contact jfdu@gdufs.edu.cn.

IMPORTANT DATE

Submission Deadline: Nov 1, 2015
Notification Deadline: Nov 4, 2015

AWARDS

- **First Prize:** 2,000 RMB
- **Second Prize:** 1,000 RMB
- **Third Prize:** 500 RMB

The top 3 teams may be invited with free registration to present their solutions in the meeting.