Task Definition for Large Scale Text Categorization at NLPCC 2014

1. Overview

The main goal for the Large Scale Hierarchical Text Categorization Evaluation at NLPCC 2014 is to prompt research in and to evaluate the current development of techniques in automatically categorizing text documents into a predefined taxonomy.

2. The Task

In this Text Categorization task, given a news document and a predefined hierarchy of categories with a depth of 2, a system is required to provide the IDs of the categories which this document belongs to. Note that a document may have more than one category ID, and this year, we will assume that each document can be labeled with up to 2 category IDs and participant systems should sort multiple IDs in descending order with respect to their confidence scores.

3. Data

In this evaluation, we use ***the Classification and Code of News in Chinese*** **(CCNC)** as the predefined hierarchy of categories, which will be available along with the training data from the NLPCC 2014 website in late April. This hierarchy of categories consists of at most 5 levels of subdivisions, specifically, which includes 24 main entries and 367 entries in the first and the second levels, respectively. This year, we will focus on the top two levels only.

The text document corpus used in this task includes about 30,000 news articles in Chinese with careful annotations and is kindly provided by courtesy of the Xinhua News Agency. An example document with category annotation in XML format is:

```
<doc id="1">
        <title>博尔特、纳达尔等体坛名将获劳伦斯奖提名</title>
        <content >新华网吉隆坡 2 月 26 日体育专电（记者赵博超）经全球媒体提名投票，博尔特、纳达尔、小威廉姆斯、老虎 伍兹等体坛名将获 2014 年劳伦斯世界体育奖提名。其中，博尔特和小威廉姆斯已经赢得过 3 次劳伦斯奖，F1 冠军维特尔是第五次获得该奖的提名，而老虎 伍兹则在 2000 年就获得过首届劳伦斯奖。另外，此次纳达尔和伊辛巴耶娃则在劳伦斯奖下的两个分奖项均获得了提名。..... </content>
        <ccnc_cat id ="1">39.14</ccnc_cat>
        <ccnc_label id ="1">体育|体育奖</ccnc_label>
</doc>
```

In this example, the news article talks about the nominations of the Laureus Sports Word Awards 2014. The associated CCNC category annotation **39.14** means that this news article should be classified into 体育 (Sports, category code: **39**) in the first level and 体育奖 (Sports Award, category code: **39.14**) in the second level.

4. Scoring Metric

In this evaluation, we will use the macro mean of precision, recall, F1 score, in both the first and second level, to evaluate a text categorization system.

Precision of the category *j* is defined as:

$$P_j = \frac{correct_j}{predict_j} \times 100\%$$

where $correct_j$ is the number of documents correctly classified as category *j*, $predict_j$ is the total number of documents which are classified as category *j* by the system.

Recall of the category *j* is defined as:

$$R_j = \frac{correct_j}{true_j} \times 100\%$$

where $correct_j$ is the number of documents correctly classified as category *j*, $true_j$ is the total number of documents whose true category is *j*.

F1 score of category *j* is the harmonic mean of precision and recall of category *j* :

$$F1_j = \frac{P_j \times R_j \times 2}{P_j + R_j}$$

The macro precision and recall are the arithmetic means of the precisions and recalls of all categories:

$$MacroP = \frac{1}{n}\sum_{j=1}^{n} P_j \quad \text{and} \quad MacroR = \frac{1}{n}\sum_{j=1}^{n} R_j$$

where *n* is the number of documents in the dataset.

The macro F1 score is the harmonic mean of the macro precision and the macro recall:

$$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$$

Note that for items with multiple categories, we will only use the one with the highest confidence to compute the precision, recall and F1 score.

5. Submission

Each submission should be formulated as the following form:

id	team-tag	run-tag	doc-id	cat-id	ccnc-cat
1	TeamXYZ	XYZ-1	xhn-1	1	01.17
2	TeamXYZ	XYZ-1	xhn-2	1	11.21
3	TeamXYZ	XYZ-1	xhn-4	1	21.16
4	TeamXYZ	XYZ-1	xhn-4	2	35.01

…

Each row will correspond to a news article, with fields corresponding to the result id, team id, system id, document id, category id and the CCNC category, separated by the \tab symbol.