NLPCC 2018 Shared Task Guideline: Automatic Tagging of Zhihu Questions

1 Task Definition

The task aims to tag questions in *Zhihu* with relevant tags from a collection of predefined ones. Accurate tags can benefit several downstream applications such as recommendation and search of *Zhihu* questions. In this task, you are challenged to build a multi-label model that assigning relevant tags to a given question. You will be using a dataset of questions collected from *Zhihu*'s online web site.

2 Data

In this task, we provide training, development, and testing data. Each question in the dataset contains a title, an unique id and an additional description. The labels are tagged collaboratively by users from the community question answering web site *Zhihu*. To improving the quality of the data, we removed infrequency tags, and relabeled manually to build development and testing dataset.

3 Evaluation

For each question in the test set, you can predict as much as five relevant tags, and the tags are sorted by their predicted probabilities. If the number of predicted tags for a specific question is less than 5, you should fill -1 on the corresponding slot.

The submission file should contain a header and have the following format:
question_id,tag@1,tag@2,tag@3,tag@4,tag@5
232324,341,296,450,62,9
7585873,297,6675,-1,-1,-1
etc.

Submissions are evaluated on the F_1 measure.
We compute the positional weighted precision. Let $correct\_num_{p_i}$ denote the correct count of predicted tags at position i, and $predict\_num_{p_i}$ denote the count of predicted tags at position i.
The precision, recall and F_1 measure are computed as following formulas:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{1}$$

$$P = \frac{\sum_{i=1}^{5} correct\_num_{p_i} / \log(i+2)}{\sum_{i=1}^{5} predict\_num_{p_i} / \log(i+2)} \tag{2}$$

$$R = \frac{\sum_{i=1}^{5} correct\_num_{p_i}}{ground\_truth\_num} \tag{3}$$

Here is an example.
Given a question: 雅思 7 分以上是什么概念?
The ground truth tag set: {雅思听力, 雅思阅读, 英语学习, 雅思, 出国}.
The predicted tag set of model A is {雅思, 托福, 雅思阅读, 英语学习 },
and set of model B is {雅思, 英语学习, 雅思阅读, 托福}.
We can calculate P, R, F_1 for model A and B: $R_A = R_B = 3/5$, $P_A = 0.7434$, $P_B = 0.8015$, $F_{1_A} = 0.6640$, $F_{1_B} = 0.6862$.