NLPCC 2018 Shared Task Guideline:

User Profiling and Recommendation

1. Introduction

User profiling and personalized recommendation is essential for business decisions, such as targeting advertisement and personalized news feed. Since user behavioral data is heterogeneous, it's still challenging to effectively leverage the heterogeneous information for user profiling and recommendation.

2. Description of the Task

We provide a social media dataset including the following heterogenous information: users' profiles (gender, province, city, tags), social ties (following relationship), users' published tweets, and users' location visits. This shared task includes the following two subtasks:

1) **User Tags Prediction (UTP)**, given users' other information except tags, predict which tags are related to a user.

2) **User Following Recommendation (UFR)**, given users' following relationship and other provided information, predict the users a user would like to follow in the future.

3. Data

The data, collected from a social media platform, contains the following five aspects:

1) profile.txt describes users' profiles. Currently gender, province and city are provided.

user	gender	province	city	tags

2) tags.txt describes users' tags. Each line contains a user and related tag.

user	tag

3) social.txt describes users' following relationship, where user1 follows user2 on this social media platform.

user 1	user2

4) tweets.txt describes what user posted. Each line contains a user and the posted tweet.

user	tweet

5) checkins.txt describes users' location visits. The format is as follows, where POI is the location user visits, cate1, cate2, cate3 is the category of the POI in a hierarchical level. lat and lng is the latitude and longitude information and Name is the location name.

user	POI	cate1	cate2	cate3	lat	lng	name

All the files are UTF-8 encodes and tab separated.

4. Evaluation Metric

The quality of **User Tags Prediction (UTP)** and **User Following Recommendation (UFR)** subtasks will both be evaluated by F1@K,

$$P_i@K = \frac{|H_i|}{K} \ , \quad R_i@K = \frac{|H_i|}{|V_i|} \ , \quad F1_i@K = \frac{P_i@K * R_i@K}{P_i@K + R_i@K}$$
$$\text{F1}@K = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i@K$$

where $|H_i|$ is the correctly predicted item set (item refers to tag in **UTP** and user in **UFR**) for user i 's top K prediction, $|V_i|$ is the ground truth item set for user i. $P_i@K, R_i@K$ and $F1_i@K$ is the precision, recall and F1 for a user i.

In **UTP**, we set $K = 3$.
In **UFR**, we set $K = 10$.

5. Contact Information

For any questions about this shared task, please contact Fuzheng Zhang from Microsoft Research. Email: fuzzhang@microsoft.com