**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.

1. **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.

1. **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 |

1. tags.txt describes users’ tags. Each line contains a user and related tag.

|  |  |
| --- | --- |
| user | tag |

1. social.txt describes users’ following relationship, where user1 follows user2 on this social media platform.

|  |  |
| --- | --- |
| user 1 | user2 |

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

|  |  |
| --- | --- |
| user | tweet |

1. 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.

1. **Evaluation Metric**

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

where is the correctly predicted item set (item refers to tag in **UTP** and user in **UFR**) for user ‘s top prediction, is the ground truth item set for user . *,* andis the precision, recall and F1 for a user *.*

In **UTP**, we set .

In **UFR**, we set .

For **UTP** subtask**,** the participants need to predict the tags for all the users in **predict\_tags\_uids.txt,** where the submitted file format will be like this:

User1, tag1, tag2, …

User2, tag3, tag4, …

Where all the candidate tags are in **tags\_candidates.txt**.

For **UFR** subtask**,** the participants need to predict the following users for all the users in **predict\_socials\_uids.txt,** where the submitted file format will be like this:

User1, following user1, following user2, …

User2, following user3, following user4, …

Where all the users in **predict\_socials\_uids.txt** appeared in the training data, and for each user, the predicted following user candidates are in **socials\_following\_candidates.txt**, and the predicted following relationship should not appear in the training data.

1. **Contact Information**

For any questions about this shared task, please contact [Fuzheng Zhang](https://www.microsoft.com/en-us/research/people/fuzzhang/) from Microsoft Research. Email: [fuzzhang@microsoft.com](mailto:fuzzhang@microsoft.com)