NLPCC 2017 Shared Task Guideline:

Social Media User Modeling

1. Introduction

User modeling on social media is essential for business decisions, such as user segmentation and targeting advertisement. Since user behavioral data on social media is heterogeneous, it's still challenging to effectively leverage the heterogeneous information for user modeling.

2. Description of the Task

We provide a social media dataset including the following heterogenous information: users' profiles (gender), social ties (following relationship), users' tags, users' published tweets, and users' location visits. The user modeling task include the following two subtasks:

1) **Interested Location Prediction**, given users' some historical location visits and other provided information, predict what locations a user is interested to visit in the future.

2) **User Profiling**, given users' other information expect profiles, predict each user's profile information.

3. Data

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

1) checkins.txt describes users' location visits. The format is as follows, where POI is the location id 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

2) profile.txt describes users' profiles. Currently only gender is provided.

user	gender

3) social.txt describes users' social tie, where User1 follows User2 on this social media platform.

user 1	user2

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

user	tag

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

tweet	user

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

4. Evaluation Metric

1) The submit file format of **Interested Location Prediction** subtask will be like this:
   User1,POI1,POI2,….
   User2,POI3,POI4…

   Where all the users in the test data appeared in training data, and for each user, the predicted POI should be contained in the training data, but not appear in this user's visit history.

   The quality of this subtask will be evaluated by F1@K (K=10), where $|H_i|$ is the correctly predicted locations for user i 's top K prediction, $P_i@K$, $R_i@K$ and $\mathrm{F1}_i@K$ is the precision, recall and F1 for a user i.

$$P_i@K = \frac{|H_i|}{K} \quad , \quad R_i@K = \frac{|H_i|}{|V_i|} \quad , \quad \mathrm{F1}_i@K = \frac{P_i@K * R_i@K}{P_i@K + R_i@K}$$

$$\mathrm{F1}@K = \frac{1}{N}\sum_{i=1}^{N} \mathrm{F1}_i@K$$

2) **User Profiling** subtask focus on gender prediction, the submit file format will be like this:
   User1,m
   User2,f
   The quality of this subtask will be evaluated by accuracy, where $Label_i$ is the ground truth gender and $Predict_i$ is the predicted gender, δ is the indicator function where $Label_i$ and $Predict_i$ is the same.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \delta(Label_i, Predict_i)$$

5. Contact Information

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