



Summary of Assignment 1:

Memory-based CF for Rating Prediction

Note

- Do data analysis **first**
- Implement the simple baselines
 - In rating prediction, they could be global average rating scores, average user rating scores, and average item rating scores.
 - Advantages: help to check your methods.
- Similarity computation (for users and items)
 - Value vectors V.S. Set
- Local test sets
 - After determining hyper-parameters, it could be merged with training data for training models



Assignment 2:

Content-based Recommendation

Project Report

- Main content
 - Detailed implementation introduction
 - Key code explanation
 - Result analysis
 - Hyper-parameter influence
 - Table or plot
 - How to run your code
 - Environments (E.g., python 2.X or python 3.x, c++?)
 - Dependency libraries

Source Code

- Programming language
 - Python is preferred.
- Terminal command
 - E.g., `python2.7 xxx.py test.file output.file`
- Source code files
 - How the method is implemented

Data Format

- Train/train.tsv
 - Date : 时间
 - History: 用户点击历史
 - Impression : 当前展示页面所包含的新闻, 以及用户的点击情况。在train和test中会有所区别
- Train/train_news.tsv
 - Nid: 新闻id
 - Category : 新闻类别
 - SubCategory: 新闻子类别
 - Title : 新闻标题 (可为空)
 - Abstract : 新闻摘要 (可为空)

Data Format

- Click and prediction

下图为train的Impression示例，表示页面包含3条新闻，用户点击了N49685

N49685-1 N28047-0 N29128-0

下图为test的Impression示例，表示页面包含3条新闻。任务是预测用户会点击其中哪几条（一条及以上）。

N49685 N28047 N29128

Submission

- Submission file
 - Zip/Rar file, named with 'ID'
 - 注意：只能有一个test.tsv。如有多种方法请选择你认为最好，其余的在报告中说明即可。



预测结果在所给的test.tsv的最后插入新的列，列名为'Predict'，数据格式为字符串。如下图所示，假设预测结果为用户点击了第一个新闻，后两个都没点击，则数据格式为'100'。

- Submission URL
 - <http://xzc.cn/l4Wsc6H7eH>

How to evaluate?

- Code
 - Executable
 - Clear, easy-to-understood
- Prediction results
 - Effectiveness
 - How is the predicted score compared with ground-truth?
 - $(\text{Precision@1} + \text{Precision@2} + \text{Precision@3})/3$
- Report
 - Integrity
 - Readability
 - Highlights

Data and Deadline

- **Data**

- <https://pan.baidu.com/s/1rc8U9q1pKVBmD1KA3tAYOw>
- 提取码: japn

- **Deadline**

- 24:00, 2020-5-6
- Submission after that time might be penalized