

# 实验3 推荐系统和社会网络

Web信息处理与应用 2021 / 011179.01

## 知识库

### Attention

1. 在实验截止时间以前，知识库可能会继续更新。
2. 本知识库的正确打开方式：**当作字典，按需取用。**

## GPU计算资源

如果你需要GPU计算资源，以下线上的免费算力可能有帮助，但几乎都会有时间限制。

### Google Colaboratory (推荐)

Colab可以视作存储在Google Drive上的Jupyter Notebook，无需配置、轻松共享。Colab默认使用TensorFlow (可参考<https://www.tensorflow.org/>)，也支持PyTorch (可参考<https://pytorch.org/>)，免费用户一般使用Tesla K80显卡。你可以参考<https://colab.research.google.com/> 获取更多信息。

### Kaggle

Kaggle平台是当今最流行的数据科学社区，隶属于Google。Kaggle一般提供每周30小时的GPU算力，显卡是Tesla P100，支持PyTorch和TensorFlow。你可以参考<https://www.kaggle.com/docs/efficient-gpu-usage>获取更多信息。

### AI Studio

AI Studio是百度提供的开源深度学习平台。运行环境包含CPU基础版 (直接使用) 和GPU高级版 (需要算力卡)，算力卡可以通过多种方式获得 (如运行项目自动获取每天12小时算力卡、新手任务可获得100小时算力卡)。AI Studio提供Tesla V100显卡，但仅支持PaddlePaddle框架 (可参考<https://www.paddlepaddle.org.cn/>)，TAs不建议没有相关基础的同学使用。你可以参考<https://aistudio.baidu.com/> 获取更多信息。

## 教程

根据实验部分调研的反馈，TAs为大家提供了一些公开、常见的教程。特别说明，网络上各式各样、或简或繁的相关教程非常之多，希望大家选择真实适合自己的。

### 编程语言

关于Python快速入门可参考<https://www.runoob.com/python3/python3-tutorial.html>。

### 深度学习框架

- PyTorch
  - a. 官方提供60min快速入门视频，可参考[https://pytorch.org/tutorials/beginner/deep\\_learning\\_60min\\_blitz.html](https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html)。
  - b. PyTorch中文文档，可参考<https://pytorch-cn.readthedocs.io/zh/latest/>。
  - c. PyTorch Handbook是一本中文开源的入门书籍，可参考<https://handbook.pytorch.wiki/>。
  - d. 一个大多数模型代码少于30行的PyTorch教程，可参考<https://github.com/yunjey/pytorch-tutorial>。
- TensorFlow

- a. 一个文档、视频、代码相结合的教程，可参考<https://github.com/instillai/TensorFlow-Course>。
- b. TensorFlow文档，可参考[https://www.tensorflow.org/api\\_docs](https://www.tensorflow.org/api_docs)。
- c. 适用于初学者的经典模型库，可参考<https://github.com/aymericdamien/TensorFlow-Examples>。

## 机器学习知识

- 无数学子的机器学习入门课——Andrew Ng (Stanford University) CS229课程。可参考[https://www.youtube.com/watch?v=jGwO\\_UgTS7I](https://www.youtube.com/watch?v=jGwO_UgTS7I) 或<https://www.bilibili.com/video/BV1JE411w7Ub>。
- 非常经典的DeepLearning.AI深度学习课程。可参考[https://mooc.study.163.com/university/deeplearning\\_ai#/c](https://mooc.study.163.com/university/deeplearning_ai#/c) 或<https://www.deeplearning.ai/program/deep-learning-specialization/>。

## 开源库

根据USTC评课社区 (<https://icourse.club/course/18319/>) 的反馈，正式实验前的调研非常重要，TAs深表赞同。为了帮助大家减轻负担，TAs这次也非常善意地分享一些开源代码资源。

如果你要设计**非深度学习的推荐模型**，以下开源库可能对你有帮助。

- **Surprise**是一个用于处理包含显式评分数据的推荐系统的Python工具包。可参考<http://surpriselib.com/>。
- **Crab**基于 `numpy`、`scipy`、`matplotlib` 实现了许多经典的协同过滤算法。可参考<https://github.com/muricoca/crab>。
- **Case Recommender**是一个基于Python的推荐系统库，主要包含了一些相对早期的推荐系统模型，如课程中涉及的NMF、SVD等经典模型。可参考<https://github.com/caserec/CaseRecommender>。
- **SVDFeature**专门用于解决基于特征的协同过滤，可以用于评分预测和排序。可参考<https://apex.sjtu.edu.cn/projects/33>。

如果你要设计**基于深度学习的推荐模型**，以下开源库可能对你有帮助。

- **DeepRec**是一个基于Python和TensorFlow的深度学习推荐系统库，包含了评分预测、Top-N推荐和序列推荐等3类。可参考<https://github.com/cheungdaven/DeepRec>。
- **NeuRec**使用Python和TensorFlow实现了33个模型，涵盖一般、社交、序列推荐三大类。可参考<https://github.com/wubinzzu/NeuRec>。
- **RecBole**是一个基于Python和PyTorch 实现的推荐系统代码库。实现了72个模型，涵盖一般、序列、基于上下文、基于知识的推荐四大类。可参考<https://github.com/RUCAIBox/RecBole>。

## 数据统计

数据统计分析是数据挖掘类问题的重中之重，你可以试着这样对数据进行一些分析，TAs闲着没事儿又帮了大家。其中音乐数据不包括在线评测系统所用的测试数据。

	音乐	电影	书籍
用户数	23599	23570	16583
物品数	21602	68408	162346
交互数	1425957	5911479	1101122
平均每个用户交互数	60.42	250.81	66.40
平均每个物品交互数	66.10	86.42	6.78
拥有打分的交互数	990652	4588066	779278
打分比例	69.4728%	77.6128%	70.7713%
稀疏度	0.2797%	0.3666%	0.0409%

(Last updated on Jan 1, 2022 by TA Team)