Python 数据分析小项目
AI 程序设计@NJU

1. 基于 MovieLens 100k 数据集中男性女性对电影的评分来判断男性还是女性电影评分的差异性更大。

(1) 数据来源

数据集下载： http://files.grouplens.org/datasets/movielens/ml-100k.zip

数据含义（具体可参见数据集中的说明文件）:

u.data 表示 100k 条评分记录，每一列的数值含义是:

user id | item id | rating | timestamp

u.user 表示用户的信息，每一列的数值含义是:

user id | age | gender | occupation | zip code

u.item 文件表示电影的相关信息，每一列的数值含义是:

movie/item id | movie title | release date | video release date |IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy |Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |Thriller | War | Western |

(2)可能用到的函数/方法包括如： pd.read_table()， pd.merge()， pd.pivot_table()，df.query()

具体功能请参考 Python 系统帮助信息或 API 文档

http://pandas.pydata.org/pandas-docs/stable/

2. 美国农业部(USDA)制作了一份有关食物营养信息的数据库，Ashley Williams 制作了该 数据的 JSON 版本。ndb.zip 中抓取了其中的 3378 条数据（01001.json-12108.json)，每个文件是一种食品的营养信息，文件名为该食品的 id。请基于这一份食品数据进行挖掘，例如可找出所有食品中某一种成分含量最高的若干种食品并可视化。

PS：数据转换成 DataFrame 提示

```
# 需要的数据转换成 DataFrame
 >>> import pandas as pd
>>> df01001 = pd.DataFrame(n01001['report']['food']['nutrients'])
# 删除不需要的列
>>> del df01001['measures']
```