寻找数据集中的相似用户

- import json
- import numpy as np
- from Chapter05.pearson_score import pearson_score
- '''定义一个函数，用于寻找与输入用户相似的用户。该函数有3个输入参数：数据库、输入用户和寻找的相似用户个数。第一步是查看该用户是否包含在数据库中。如果用户已经存在，则需要计算该用户与数据库中其他所有用户的皮尔逊相关系数'''def find_similar_users(dataset, user, num_users):
- if user not in dataset:
- raise TypeError('User ' + user + ' not present in the dataset')
- # 计算所有用户的皮尔逊相关度　scores = np.array([[x, pearson_score(dataset, user, x)] for x in dataset if user != x])
- # 评分按照第二列排序　scores_sorted = np.argsort(scores[:, 1])
- # 评分按照降序排列　scored_sorted_dec = scores_sorted[::-1]
- #提取出k个最高分　top_k = scored_sorted_dec[0:num_users]
- return scores[top_k]
- if __name__=='__main__':
- data_file = 'movie_ratings.json'　with open(data_file, 'r') as f:
- data = json.loads(f.read())
- user = 'John Carson'　print ("\nUsers similar to " + user + ":\n")
- similar_users = find_similar_users(data, user, 3)
- print ("User\t\t\tSimilarity score\n")
- for item in similar_users:
- print (item[0], '\t\t', round(float(item[1]), 2))