CSc 110, Autumn 2017
Programming Assignment #12: Recommendation System
Due Tuesday, December 5, 2017, 7:00 PM

thanks to Susan Rodger of the Duke and Marty Stepp of Stanford

This program focuses on using data structures in complex ways. Turn in a file named `recommender.py` on the Projects section of the course web site. You will need several txt files containing ratings from the web site; place them in the same folder as your program.

Background:

If you've ever bought a book online, the bookseller's website has probably told you what other books you might like. This is handy for customers, but also very important for business. In 2010, online movie-rental company Netflix awarded one million dollars to the winners of the Netflix Prize. The competition simply asked for an algorithm that would perform 10% better than their own algorithm. Making good predictions about people's preferences was that important to this company. It is also an important current area of research in machine learning, which is part of the area of computer science called artificial intelligence.

What you will do:

So how might we write a program to make recommendations for books? Consider a user named Carlos. How is it that the program should predict books Carlos might like? The simplest approach would be to make almost the same prediction for every customer. In this case the program would simply calculate the average rating for all the books in the dataset and display the highest rated book. With this simple approach no information about Carlos is used.

We could make a better prediction about what Carlos might like by considering his actual ratings in the past and how these ratings compare to the ratings given by other customers. Consider how you decide on movie recommendations from friends. If a friend tells you about a number of movies that s(he) enjoyed and you also enjoyed them, then when your friend recommends another movie that you have never seen, you probably are willing to go see it. On the other hand, if you and a different friend always tend to disagree about movies, you are not likely to go to see a movie this friend recommends.

Your task for this project is to write a program that takes people's book ratings and makes book recommendations to them using both techniques described above. We will refer to the people who use the program as "users".

Implementation Details:

Your program will read user date from a file that contains sets of three lines. The first will contain a username, the second a book title and the third the rating that user gave the book. An example is displayed on the right.

```
Bob
Cats
1
Suelyn
Harry Potter
1
Carlos
Animal Farm
1
Kalid
Watership Down
3
Carlos
1984
-5
Bob
Harry Potter
-3
Suelyn
1984
1
Carlos
Harry Potter
3
Suelyn
Animal Farm
-3
Bob
The Hobbit
1
Kalid
Animal Farm
-3
Kalid
Harry Potter
1
```

When your program starts it should read through the file and create a list with one occurrence of each book in the file. For example, the file at right might produce the following list:

```
['1984', 'Cats', 'Harry Potter', 'Animal Farm', 'Watership Down', 'The Hobbit']
```

Note that the order of the books does not matter. We suggest that you create this list by putting all books in the file into a set and then casting that set to a list. If you have a variable called `data` that stores a `set`, you can turn it into a `list` by writing `data = list(data)`.

Once you have this list, create a dictionary to store the rating data and loop through the file again. This time add each person in the file as a key to the dictionary. The value that is associated with them should be a list the same length as the list of books you created in your first pass through the file. You should store the rating at the same index that book's name appears at in in the list of books. For example, when the first three lines of the file on the right are read, we would add a mapping from the key `Bob` to a value of `[1, 0, 0, 0, 0, 0]`. Books that the user has not rated (or whose ratings we have not read yet, as in this case) should be represented by `0`s.

The dictionary created with the file at right and the list of books written above would be as follows:

```
{ 'Kalid'  : [0, 0, 1, -3, 3, 0],  'Carlos' : [-5, 0, 3, 1, 0, 0],
  'Suelyn' : [1, 0, 1, -3, 0, 0],  'Bob'    : [0, 1, -3, 0, 0, 1]  }
```

Now your program is ready to make recommendations. It should output the following message:

```
Welcome to the CSC110 Book Recommender. Type the word in the
left column to do the action on the right.
recommend : recommend books for a particular user
averages  : output the average ratings of all books in the system
quit      : exit the program
next task?
```

The prompt should be repeated after every task is finished.

Averages

If the user selects the averages option the program should output all of the books in the file sorted by average rating from highest to lowest. For example, for the file on the previous page you should output the listing on the right.

```
Watership Down 3.0
The Hobbit 1.0
Cats 1.0
Harry Potter 0.5
Animal Farm -1.6666666666666667
1984 -2.0
```

We suggest that you figure this out by building up a list of tuples containing the average rating for a book first and the title of that book second. You can build up this list by going through the list of books one at a time and for each person in the dictionary adding up their rating of that book and counting how many people in the dictionary rated it something other than 0. The average score for that book is the sum scores divided by the count of non-zero ratings. Once you have created this list you can sort it by using the list's `sort` function.

Since the averages will stay the same throughout the run of the program you may want to calculate them once at the start of the program.

Recommendations

When the user selects the recommend option the program should first prompt the user for the name of the user that the program wants recommendations for as follows:

```
user?
```

If the name that the user types in isn't in the dictionary of ratings, the program should output the same list of books as when the user selects the averages option.

If the name the user inputs is in the dictionary of ratings you should use the data in the dictionary to find the other users that are most similar to the user you are looking for recommendations for.

The first step to do this is to calculate the similarities between your user and the other users. We will use the dot product between the users' lists of ratings to calculate their similarity. This means that we will multiply each element in your user's list with the element at the same index in the other user's list and sum the result. For example, if we were looking for a recommendation for Kalid we would do the following to calculate his similarity to Carlos:

```
(0 * -5) + (0 * 0) + (1 * 3) + (-3 * 1) + (3 * 0) + (0 * 0) = 0 + 0 + 3 + -3 + 0 + 0 = 0
```

Compute this similarity for each user in the dictionary. Store tuples containing first the similarity number and second the name of the other user in a list. You can use the list `sort` function to sort this list.

Note that the user that you are looking for will always be in the dictionary. We are not interesting in how similar the user is to themselves. You may find it helpful not to add the user to the list or to remove them after sorting. Note that the user will always be most similar to themselves.

Now that we have a list of the most similar users, we can use this to figure out which books to recommend. To generate recommendations take an average of the ratings of the three users with the highest similarity to the user you are looking for.

To average the ratings create a new list the same length as the list of books and filled with `0`s. Loop through this list. For every index of this list loop through the first three users, add up their ratings and then divide by the number of non-zero ratings. If there are no non-zero ratings for a book you should not divide as you will get a divide by 0 error.

Once you have calculated these averages, create a list of tuples that contain first the average rating and then the book title for all books that have non-zero ratings in the averages list. Then, sort this list. Now you have a list of books to recommend.

Note that it must be possible for the user to choose options multiple times in any order and get the correct results each time.

Your program should match the expected output files provided on the course web page exactly.

Development Strategy and Hints:

We suggest that you complete the assignment in the order described above.

You will find the list `sort` function very helpful for this assignment. The `sort` function sorts elements in a list from smallest to largest. If the list contains tuples it sorts by the first element in the tuple. Therefore, it is very important to order your tuple contents as described in the instructions above. If you have a variable called `my_list`, the following code will sort it: `my_list.sort()`. Note that `sort` alters the list, it doesn't return a new list.

Use `print` statements to view the state of your structures. Create small input files for yourself to help you debug.

You should use our Output Comparison Tool to see that your outputs match what is expected.

Style Guidelines and Grading:

Part of your grade will come from appropriately using data structures and following the algorithms described in this document.

Redundancy is always a major grading focus; avoid redundancy and repeated logic as much as possible in your code. Divide your code into a set of functions that captures the structure of the program.

Follow good general style guidelines such as: appropriately using control structures like loops and `if`/`else` statements; avoiding redundancy using techniques such as functions, loops, and `if`/`else` factoring; good variable names, and naming conventions; and not having any lines of code longer than 80 characters. You may have no global variables (except constants) and you may not nest functions inside other functions

Comment descriptively at the top of your program, each function, and on complex sections of your code. Comments should explain each function's behavior, parameters and returns.