Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 1

Due: Jan 18, 2023(100 points)

Yu Wang
https://github.com/Titojojo/CS6220-Data-Mining

Coding Review

If it's outside the answer of a question (which is described below), this space can sometimes be useful as background information. For example, if you had a particular thought process that doesn't immediately answer the question, it could be written in a paragraph like this.

1. The cardinality of a set or collection of items is the number of unique items in that collection. Write a function called cardinality_items that takes a .csv text string file as input, where the format is as the above, and calculates the cardinality of the set of all the grocery items in any given dataset.

```
def cardinality_items(filename):
    """
    Takes a filename "*.csv" and returns an integer representing
    the number of unique items in the collection.
    """
    unique_items = set()

    with open(filename, 'r') as file:
        for line in file:
            items = line.strip().split(',')
            for item in items:
                unique_items.add(item.strip())

    return len(unique_items)
```

The cardinality in "basket data.csv" is: 21

2. Write a function called all_itemsets that takes a list of unique items and an integer k as input, and the output is a list of all possible unique itemsets with non-repeating k items. That is, the output is L = [S1, S2, · · · SN], a list of all possible sets, where each Si has k items.

```
def all_itemsets(item_list, k):
    """
    Takes a list of unique items and an integer k, return a
    list of all possible unique itemsets with non-repeating k
    items.
    """
    res = []
    n = len(item_list)

    def backtrack(comb, idx, remaining):
        if remaining == 0:
            res.append(comb[:])
        else:
            for i in range(idx, n):
                comb.append(item_list[i])
                backtrack(comb, i + 1, remaining - 1)
                comb.pop()

    backtrack([], 0, k)
    return res
```

Examining Our First Dataset

3. Let's review combined_data_*.txt.
   a) How many total records of movie ratings are there in the entire dataset (over all of combined_data_.txt)?

      100480507

   b) How many total unique users are there in the entire dataset (over all of combined_data_.txt)?

      480189

   c) What is the range of years that this data is valid over?

      1999 to 2005

4. Let's review movie_titles.csv.

a) How many movies with unique names are there? That is to say, count the distinct names of the movies.

17359

b) How many movie names refer to four different movies?

5

5. Let's review both.

a) How many users rated exactly 200 movies?

605

b) Of these users, take the lowest user ID and print out the names of the movies that this person liked the most (all 5 star ratings).

Spirit Lost
Butch and Sundance: The Early Days
A Slipping Down Life
Lawn Dogs
Training Day
Leprechaun
Rose Red
Poison Ivy 2
Jason Goes to Hell
Irreversible
Tank Girl
Swimming Pool
NYPD Blue: Season 2
The Main Event
Peculiarities of the National Fishing
In Living Color: Season 3
Dragon Ball: Red Ribbon Army Saga
Heidi
American Buffalo
Burn Up Scramble
Kill Me Again
Of Mice And Men
Barney's Super Singing Circus
Sweet Charity
The Best of Film Noir
Lust in the Dust
Miranda
Nature: Cloud: Wild Stallion of the Rockies
Eternal
Tying the Knot

She's Out of Control
Huey P. Newton Story
Inu-Yasha: The Movie 2: The Castle Beyond the Looking Glass
A Charlie Brown Thanksgiving / The Mayflower Voyagers
Home Improvement: Season 2
Now You See Him, Now You Don't
How to be a Woman and not Die in the Attempt
Billy Joel: Live at Yankee Stadium
Robin Hood: Prince of Thieves: Bonus Material
WMD: Weapons of Mass Deception
X-Men
Casanova's Big Night
Raging Bull
Lord of the Rings: The Return of the King
Monty Python and the Holy Grail
Raising Arizona
The Shawshank Redemption: Special Edition
Harold and Maude
Downfall
Lord of the Rings: The Return of the King: Extended Edition
Monster
Band of Brothers
Three Kings
Unforgiven
Maria Full of Grace
Days of Wine and Roses
Shakespeare in Love