Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 0

Due: February 15, 2023(100 points)

Taiwei Cui
https://github.com/ccttww117/datamining

Question 1

Answer:

Cardinality is 21

Method:

Use csv reader to read the csv file, and for every element, insert it into a set, calculate the length of the set.

Question 2

Answer:

all_itemsets(["a", "b", "c", "d", "e"], 3): [['c', 'b', 'a'], ['d', 'b', 'a'], ['e', 'b', 'a'], ['d', 'c', 'a'], ['e', 'c', 'a'], ['e', 'd', 'a'], ['d', 'c', 'b'], ['e', 'c', 'b'], ['e', 'd', 'b'], ['e', 'd', 'c']]

Method:

Use recursion, write a aux_all_itemsets(items,idx,k) for recursion, which will return all possible permutations which start from idx and have a length of k. If k is one, directly return items[idx], else invoke aux_all_itemsets(items, idx+i, tmp), where i is from 0 to n-k.

Question 3

a:

Answer:

Number of total records of movies is100480507

Method:

Traversing all records in combined_data_*.txt, then determine whether current line is a record of rating by checking whether it ends with ‘:’.

b:

Answer:

Number of unique users is 480189

Method:

Put all id of user into a set, then return the size of set.

c:

Answer:

Range of years is 1999-2005

Question 4

a and b:

Answer:

Movies with unique names is 17359 and movie names refer to four different movies is 5

Method:

Open the .csv file and then devide every row into 3 string with its first two commas, then the third string is the name of the movie. Use a dictionary to restore the name of movie and how many times it has occurred. The size of dictionary is number of unique name, and number of values equal to 4 is number of movie names that refer to 4 different movies.

Question 5

a:

Answer:

Number of users that rated 200 movies is 605

Method:

Use a dictionary, which takes user id as key, and an array of number of rating record and names of 5 star rating movies. Then counting number of users rated exactly 200 movies.

b:

Answer:

The movies that the user with lowest id of there users are: ['High Fidelity\n', "Monty Python's The Meaning of Life: Special Edition\n", 'American Beauty\n', 'Roger & Me\n', 'Eternal Sunshine of the Spotless Mind\n', 'Being John Malkovich\n', 'Vietnam: A Television History\n', 'Super Size Me\n', 'Lord of the Rings: The Fellowship of the Ring\n', 'This Is Spinal Tap\n', 'The Pianist\n', 'The Silence of the Lambs\n', 'Sideways\n', 'Whale Rider\n', 'Garden State\n', 'Bowling for Columbine\n', 'Gandhi\n', 'Apocalypse Now Redux\n', 'To Die For\n', "Monty Python's Life of Brian\n", 'The Manchurian Candidate\n', 'Memento\n', 'Amelie\n', 'Apocalypse Now\n', 'The Usual Suspects\n', 'Lord of the Rings: The Two Towers: Extended Edition\n', 'The Lord of the Rings: The Fellowship of the Ring: Extended Edition\n', 'Touching the Void\n', 'Minority Report\n', 'The Royal Tenenbaums\n', 'Election\n', 'Good Will Hunting\n', 'L.A. Confidential\n', 'Taxi Driver\n', 'Lord of the Rings: The Two Towers\n', 'Cabaret\n', 'Adaptation\n', 'The Accused\n', 'Lost in Translation\n', "Boys Don't Cry\n", 'To Be and To Have\n', "Schindler's List\n", 'Raging Bull\n', 'Lord of the Rings: The Return of the King\n', 'Monty Python and the Holy Grail\n', 'Raising Arizona\n', 'The Shawshank Redemption: Special Edition\n', 'Harold and Maude\n', 'Downfall\n', 'Lord of the Rings: The Return of the King: Extended Edition\n', 'Monster\n', 'Band of Brothers\n', 'Three Kings\n', 'Unforgiven\n', 'Maria Full of Grace\n', 'Days of Wine and Roses\n', 'Shakespeare in Love\n']

Method:

Use same dictionary, find the user with 200 rated 200 movies and give out this user’s 5 star ratings.