

Northeastern University, Khoury College of Computer Science



CS 6220 Data Mining | Assignment 1

Due: January 18, 2024(100 points)



Dharun Suryaa Nagarajan

<https://github.com/dharun4772/CS6220/tree/main/Assignment%201>

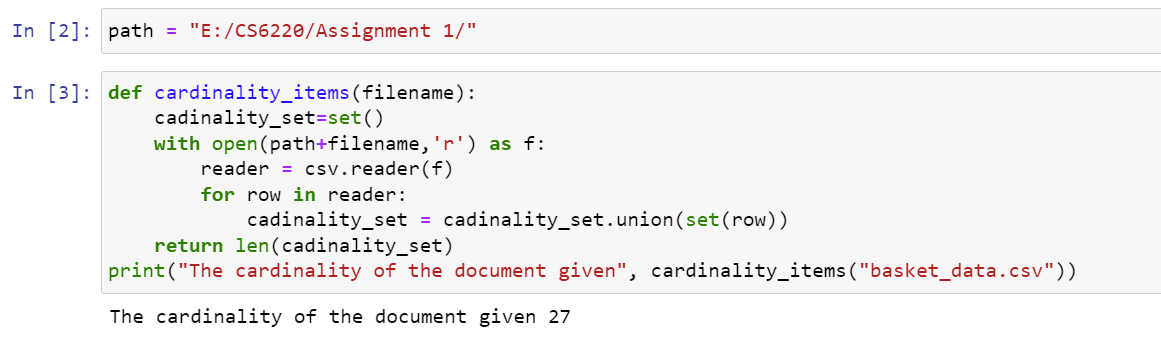
Coding Review

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.

Solution:

The provided code aims to determine the cardinality of items within a document, specifically a CSV file named "basket\_data.csv." Cardinality, in this context, refers to the number of unique items present in the document.

1. The function takes a filename as a parameter, representing the CSV file to be analyzed.
2. A set named cadinality\_set is initialized to store unique items found in the document.
3. Read the CSV file using opencv and iterate through the rows
4. For each row, the unique items are added to the cadinality\_set using the union operation.
5. The function returns the length of the final cadinality\_set, representing the cardinality of unique items in the document.



1. 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.

Solution:

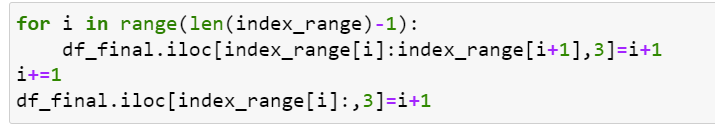
The given Python code aims to generate all possible k-length itemsets from a list of items. The function all\_itemsets takes a list of items and an integer k as input parameters and returns a list of all possible k-length itemsets.

1. The function takes a list of items (items) and an integer (k) as parameters.
2. The base cases are defined. If k is 0, an empty list (representing an empty itemset) is returned. If k is greater than the length of the items list, an empty list is returned, as it's not possible to form k-length itemsets.
3. If k is greater than 0 and less than or equal to the length of items, the function proceeds to generate k-length itemsets recursively. For each item in the list, it considers the current item, removes it from the remaining items, and recursively generates (k-1)-length itemsets. The current item is then added to each subset, forming k-length itemsets.
4. The function returns the list of all k-length itemsets generated.

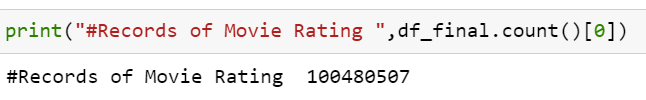


1. Let’s review combined\_data\_\*.txt:

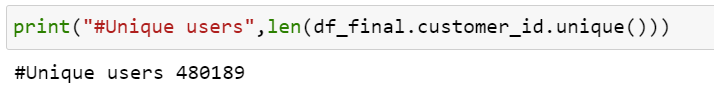
*Data Cleaning:* Since the data was not in a tabular format with movie as a separate row followed by the all customers rating for the movie. I used the pandas to first read the csv file using the three columns (Customer\_id, rating, date) and then use transformation logic to remove the movie rows and add it as a column along each customer rating row under the name movie in the dataset using the indexing.



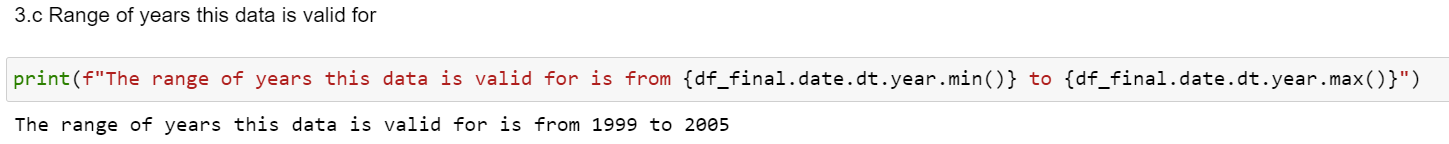
1. How many total records of movie ratings are there in the entire dataset (over all of combined\_data\_\*.txt)?



1. How many total unique users are there in the entire dataset (over all of combined\_data\_\*.txt)?



1. What is the range of years that this data is valid over?

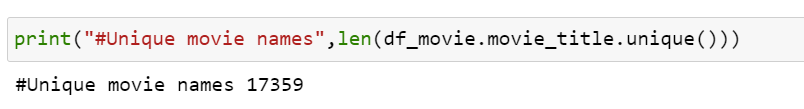


1. Let’s review movie\_titles.csv:

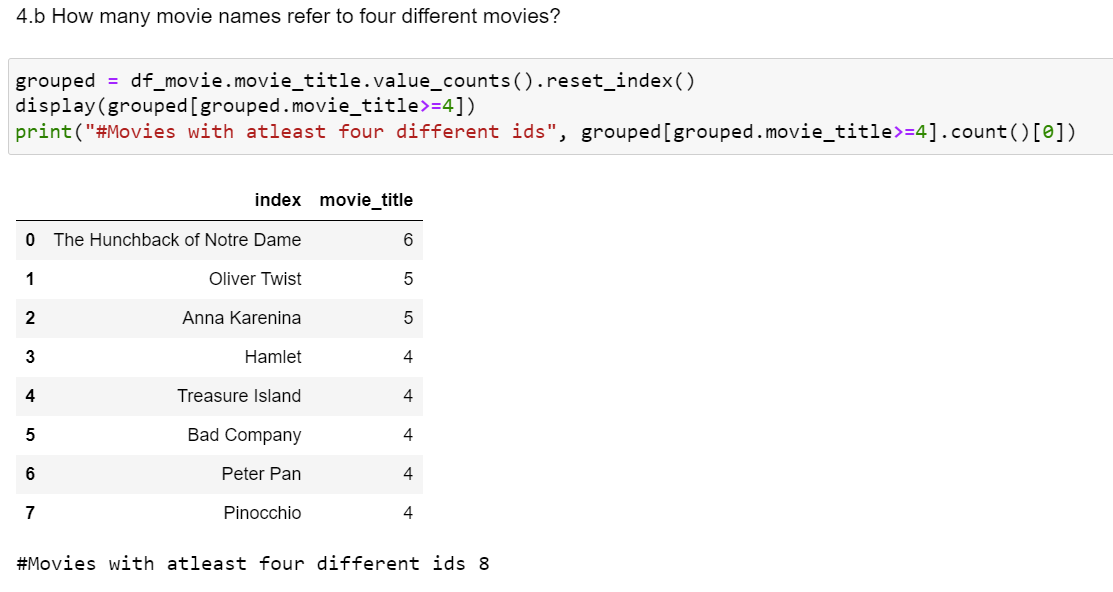
*Data Cleaning:* Like hinted in the question the movie\_titles file has commas in the movie title column that cannot be read in the typical way so no csv reader will work. Instead I used the inbuilt open function of python to read the file and then pull out information from each row separately using just two commas as separators and append the three variables into their respective columns. Finally used the columns to create a DataFrame structure in Python.



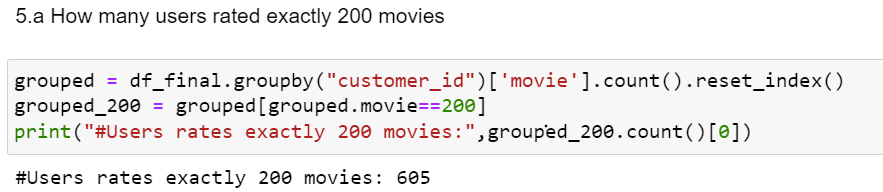
1. How many movies with unique names are there? That is to say, count the distinct names of the movies.



1. How many movie names refer to four different movies? 5 movies



1. Let’s review both:
2. How many users rated exactly 200 movies?



1. 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).

