Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 1
Due: Jan 18, 2024(100 points)

Sibo Wu
LINK TO REPOSITORY

Coding Review

1.

Answer:

The cardinality of the set of all grocery items in the provided "basket_data.csv" file is **21**

Code:

```python
def cardinality_items(filename):
    """
    Takes a filename "*.csv" and returns an integer representing the
cardinality of the set of all grocery items.
    """
    cardinality = set()
    with open(filename, 'r') as file:
        for line in file:
            items = line.strip().split(',')
            cardinality.update([item.strip() for item in items])
    return len(cardinality)
```

2.

Answer:

```python
def all_itemsets(items, k):
    """
    Generate all unique itemsets of size k from the given list of
items.

    :param items: List of unique items.
    :param k: Size of each itemset.
    :return: List of all possible unique itemsets of size k.
    """
    if k == 0:
        return [[]]
    if len(items) < k:
        return []

    result = []
    for i in range(len(items)):
        # Generate all combinations that include items[i] and have size
k
        for subset in all_itemsets(items[i+1:], k-1):
            result.append([items[i]] + subset)

    return result
```

3. Netflix Challenge

3A: Total movie rating records: 100480507
3B: Total unique users: 480189
3C: Year range: 1999 to 2005
4A: Number of Unique movie names: 17359
4B: Movies with four different names: 5
5A: Number of users rated exactly 200 movies: 605
5B: Lowest user ID who rated exactly 200 movies: 1001192