CS145 Howework 5

Important Note: HW5 is due on **11:59 PM PT, Jun 9 (Friday, Week 10)**. Please submit through GradeScope.

Print Out Your Name and UID

Name: Joshua Mares, UID: 005154394

Before You Start

You need to first create HW5 conda environment by the given `cs145hw5.yml` file.

```
conda env create -f cs145hw5.yml
conda activate hw4
conda deactivate
```

OR

```
conda env create --name NAMEOFYOURCHOICE -f cs145hw5.yml
conda activate NAMEOFYOURCHOICE
conda deactivate
```

To view the list of your environments, use the following command:

```
conda env list
```

In this notebook, you must not delete any code cells in this notebook. If you change any code outside the blocks (such as some important hyperparameters) that you are allowed to edit (between STRART/END YOUR CODE HERE), you need to highlight these changes. You may add some additional cells to help explain your results and observations.

In [1]:
```
import numpy as np
import pandas as pd
import sys
import random
import math
import matplotlib.pyplot as plt
from IPython.display import Image
from scipy.stats import multivariate_normal
%load_ext autoreload
%autoreload 2
```

If you can successfully run the code above, there will be no problem for environment setting.

1. Frequent Pattern Mining for Set Data - Aprior Algorithm (24 pts)

Table 1

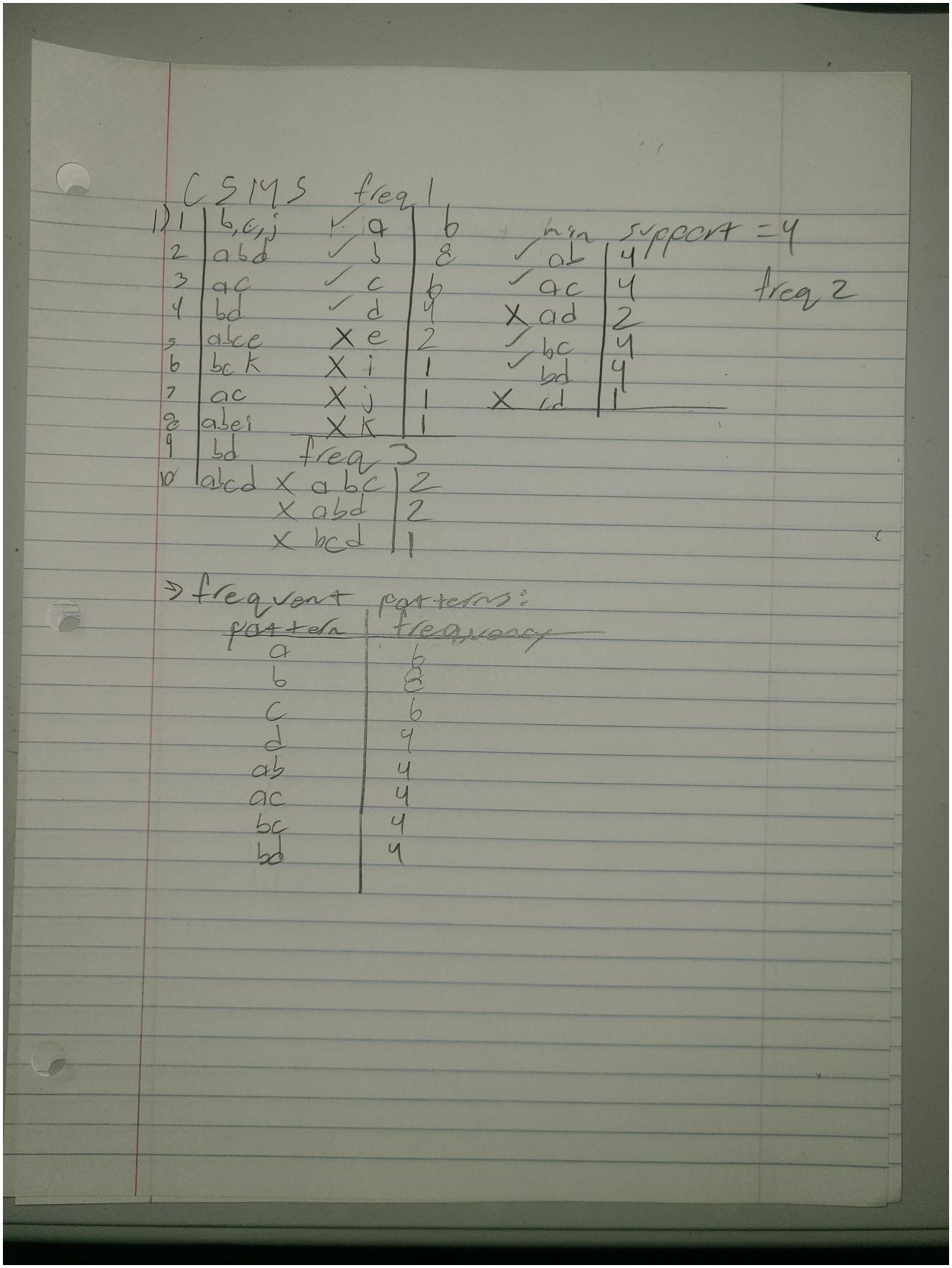
TID	Items
1	b,c,j
2	a,b,d
3	a,c
4	b,d
5	a,b,c,e
6	b,c,k
7	a,c
8	a,b,e,i
9	b,d
10	a,b,c,d

Given a transaction database shown in Table 1, answer the following questions. Let the parameter `min_support` be 4.

Question

Note: This is a "question-answer" style problem. You do not need to code anything and you are required to calculate by hand (with a scientific calculator).
Find all the frequent patterns with min_support 4 using Apriori Algorithm (**sort them in ascending order w.r.t. the length.**)

Answer


CSMS
freq 1
1) 1 b,c,j
2 abd
3 ac
4 bd
5 abce
6 bck
7 ac
8 abei
9 bd
10 abcd
a 6
b 8
c 6
d 4
e 2
i 1
j 1
k 1
min support = 4
ab 4
ac 4
ad 2
bc 4
bd 4
cd 1
freq 2
freq 3
abc 2
abd 2
bcd 1
⇒ frequent patterns:
pattern
frequency
a 6
b 8
c 6
d 4
ab 4
ac 4
bc 4
bd 4

2. Apriori for Yelp (50 pts)

In `apriori.py` , fill the missing lines with the parameters (already set in the code): `min_suppport=50` and `min_conf` = 0.25, and `ignore_one_iter_set=True` . Use the Yelp data `yelp.csv` and `id_nams.csv` , run the following cell and report the frequent patterns and rules associated with it. The code takes around 1m15s to finish on a M1 chip.

In [2]:
```
#No need
from hw5code.apriori import *
input_file = read_data('./data/yelp.csv')
min_support = 50
min_conf = 0.25
items, rules = run_apriori(input_file, min_support, min_conf)
name_map = read_name_map('./data/id_name.csv')
print_items_rules(items, rules, ignore_one_item_set=True, name_map=name_map)
```

```
item:
"Holsteins Shakes & Buns","Wicked Spoon" 51
item:
"Secret Pizza","Wicked Spoon" 52
item:
"Earl of Sandwich","Wicked Spoon" 52
item:
"The Cosmopolitan of Las Vegas","Wicked Spoon" 54
item:
"Mon Ami Gabi","Wicked Spoon" 57
item:
"Bacchanal Buffet","Wicked Spoon" 63

------------------------ RULES:
Rule:
"Secret Pizza" "Wicked Spoon" 0.2561576354679803
Rule:
"The Cosmopolitan of Las Vegas" "Wicked Spoon" 0.27692307692307694
Rule:
"Holsteins Shakes & Buns" "Wicked Spoon" 0.3148148148148148
```

3. Correlation Analysis (10 pts)

Note: This is a "question-answer" style problem. You do not need to code anything and you are required to calculate by hand (with a scientific calculator).

Table 2

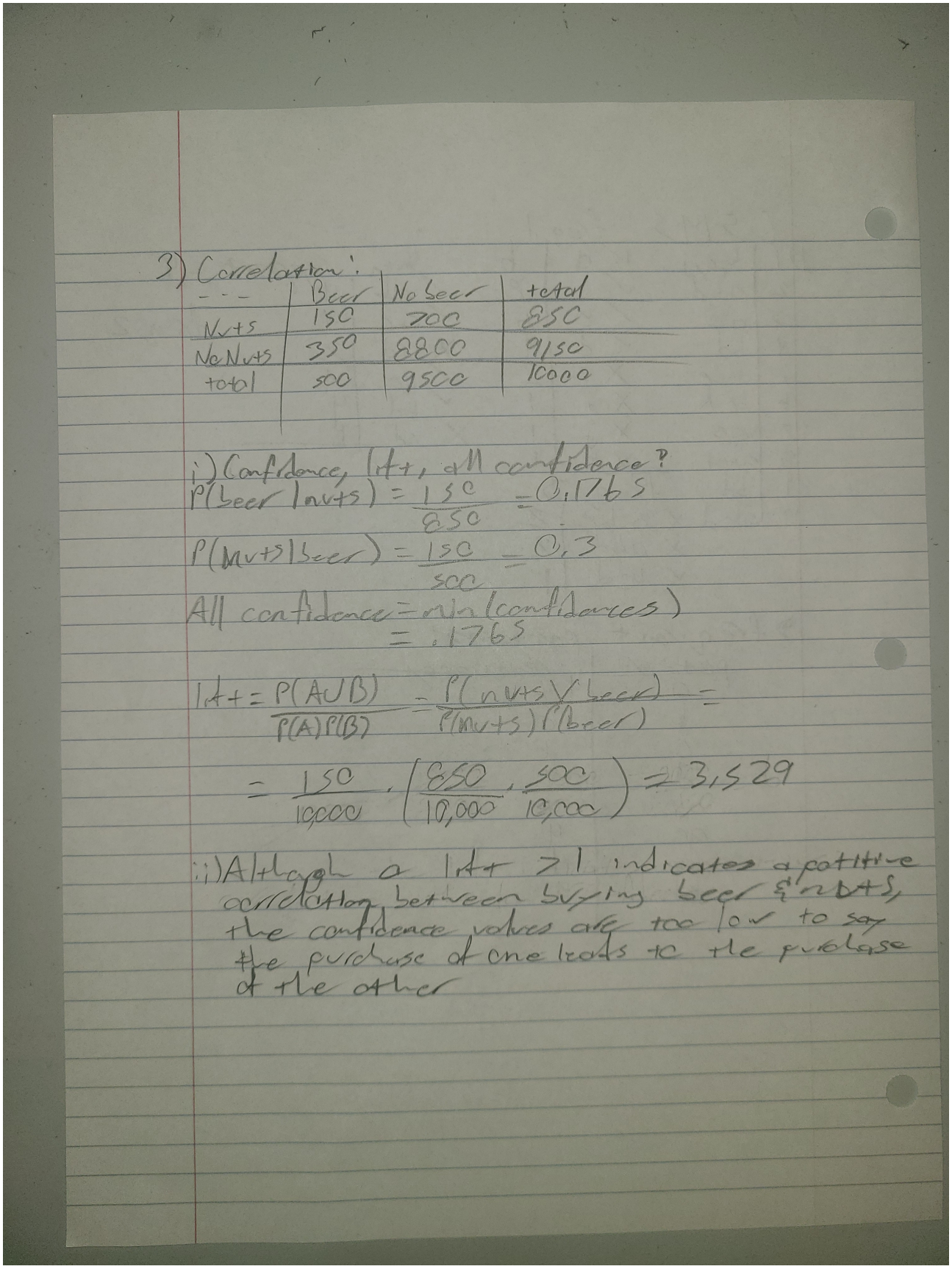
---	Beer	No Beer	Total
Nuts	150	700	850
No Nuts	350	8800	9150
Total	500	9500	10000

Table 2 shows how many transactions containing beer and/or nuts among 10000 transactions.

Answer the following questions:

3.1 Calculate `confidence` , `lift` and `all_confidence` between buying beer and buying nuts.
3.2 What are you conclusions of the relationship between buying beer and buying nuts? Justify your conclusion with the previous measurements you calculated ealier.

Answer



3) Correlation:

	Beer	No beer	total
Nuts	150	700	850
No Nuts	350	8800	9150
total	500	9500	10000

i) Confidence, lift, all confidence?

$P(\text{beer} \mid \text{nuts}) = \frac{150}{850} = 0.176$

$P(\text{Nuts} \mid \text{beer}) = \frac{150}{500} = 0.3$

All confidence = min(confidences)
= .176

$\text{lift} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(\text{nuts} \cap \text{beer})}{P(\text{nuts})P(\text{beer})} =$

$= \frac{150}{10000} \cdot \left(\frac{850}{10,000} \cdot \frac{500}{10,000}\right) = 3,529$

ii) Although a lift > 1 indicates a positive correlation between buying beer & nuts, the confidence values are too low to say the purchase of one leads to the purchase of the other

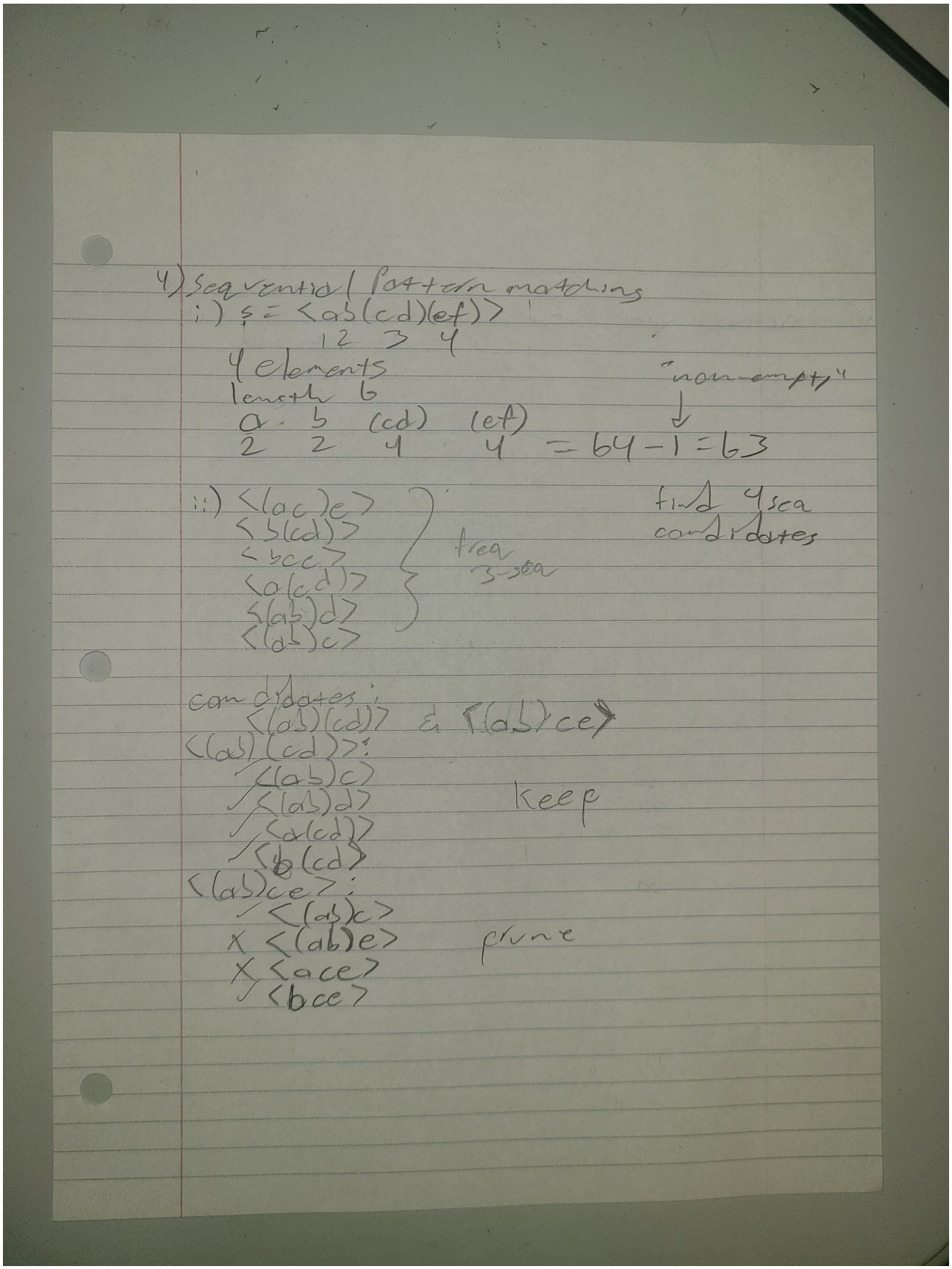
4. Sequential Pattern Mining (GSP Algorithm) (16 pts)

Note: This is a "question-answer" style problem. You do not need to code anything and you are required to calculate by hand (with a scientific calculator).

4.1 For a sequence $s =< ab(cd)(ef) >$, how many events or elements does it contain? What is the length of s? How many non-empty subsequences does s contain?

4.2 Suppose we have $L_3 = \{< (ac)e >, < b(cd) >, < bce >, < a(cd) >, < (ab)d >, < (ab)c >\}$, as the frequent 3-sequences, write down all the candidate 4-sequences C_4 with the details of the join and pruning steps.

Answers



End of Homework 5