EE412 Foundation of Big Data Analytics, Fall 2022
HW3

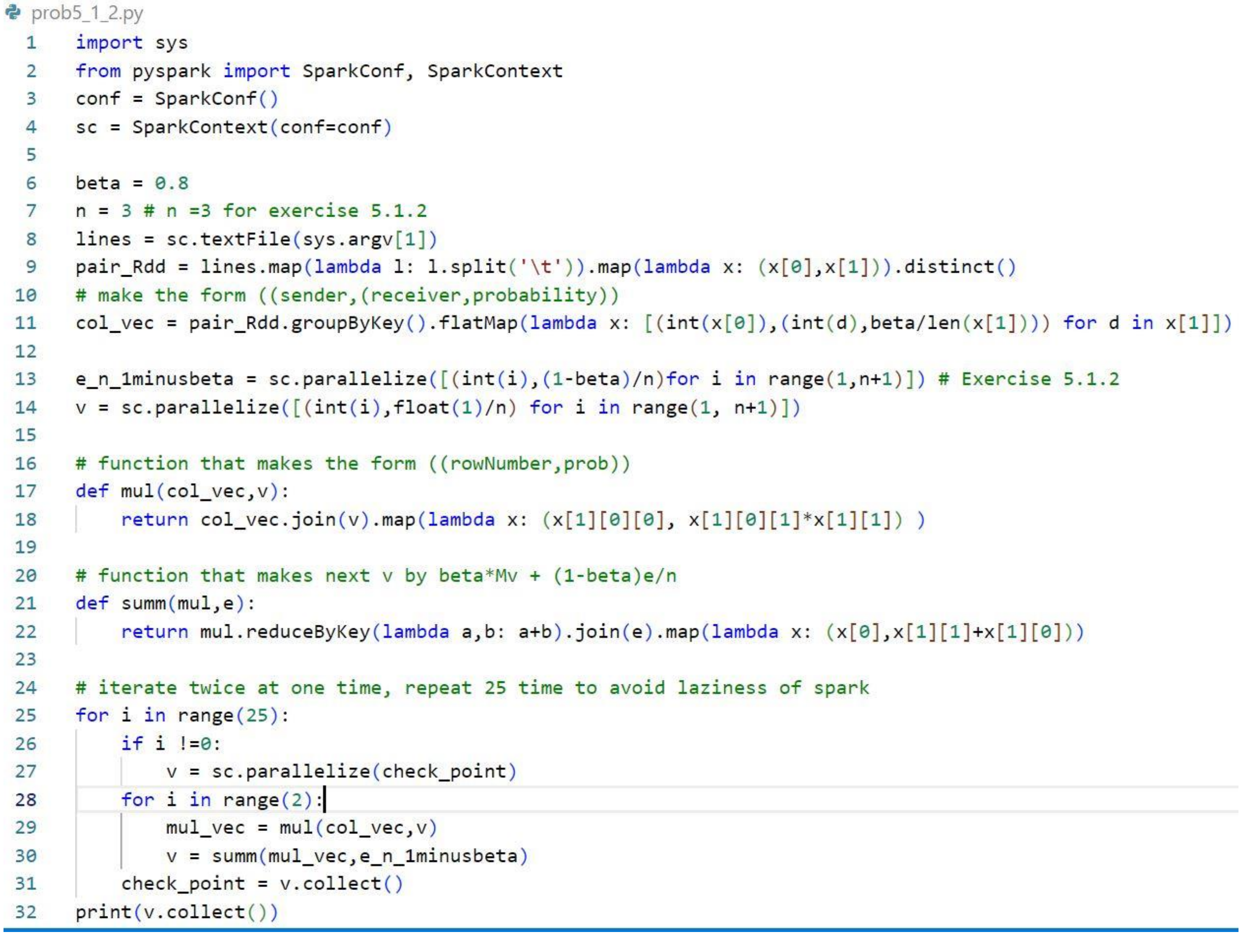
Name: 함태욱

Student ID: 20180716
Discussion Group (People with whom you discussed ideas used in your answers):
On-line or hardcopy documents used as part of your answers:

Answer to Problem 1

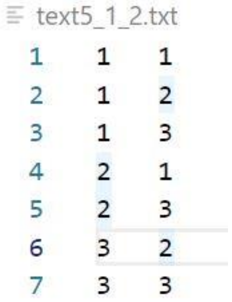
(a) **[Codes and its Page information text file]**

I reuse the source code and the text file format of problem 1-(b) that uses pyspark
In the text file each number means {1:'A', 2:'B', 3:'C', 4:'D'}
And for instance, line 1 2 means A(1) directs B(2) like graph.txt
In addition, in the problem 5_3_1, if the teleport set is [A,C], then the line 14 must be activated while line 13 must be not. The iteration is 50 times as same as 1-(b)



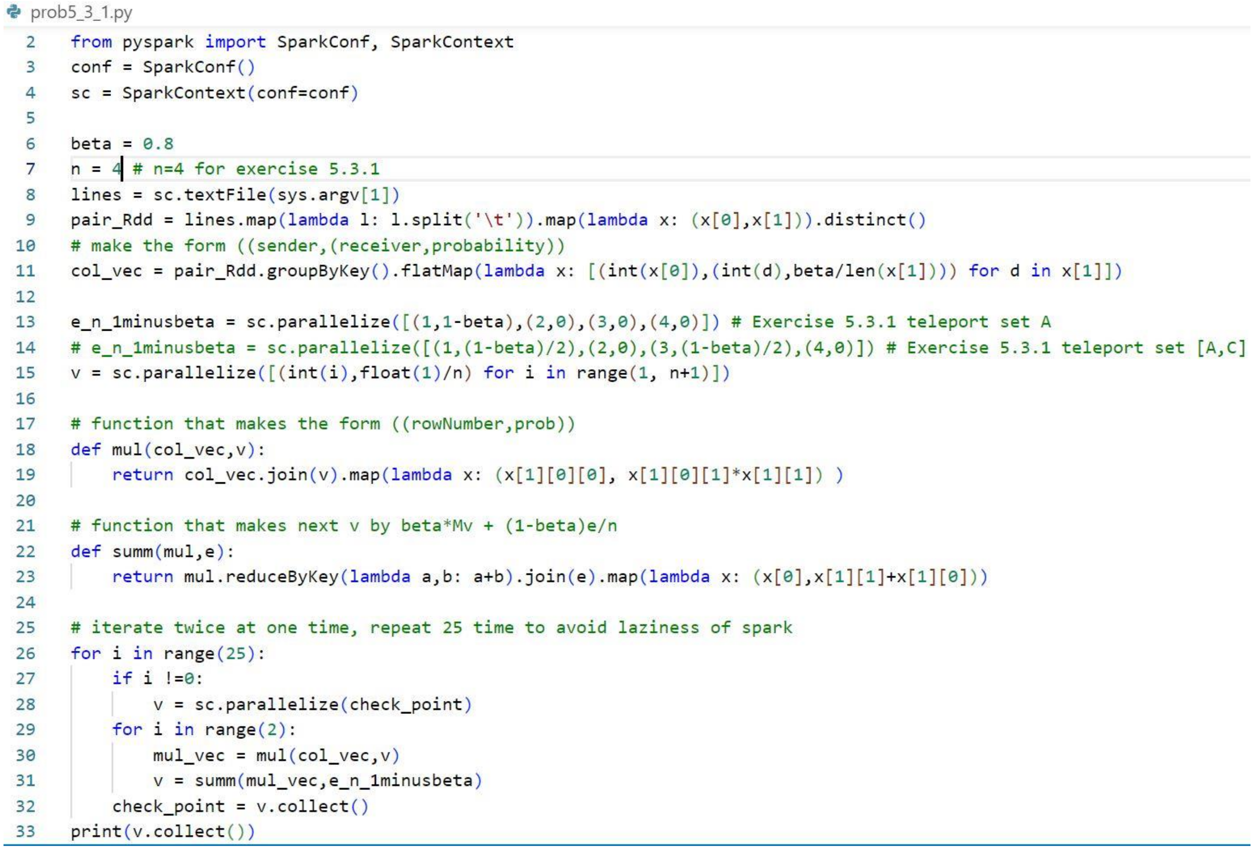
prob5_1_2.py

```
1   import sys
2   from pyspark import SparkConf, SparkContext
3   conf = SparkConf()
4   sc = SparkContext(conf=conf)
5
6   beta = 0.8
7   n = 3 # n =3 for exercise 5.1.2
8   lines = sc.textFile(sys.argv[1])
9   pair_Rdd = lines.map(lambda l: l.split('\t')).map(lambda x: (x[0],x[1])).distinct()
10  # make the form ((sender,(receiver,probability))
11  col_vec = pair_Rdd.groupByKey().flatMap(lambda x: [(int(x[0]),(int(d),beta/len(x[1]))) for d in x[1]])
12
13  e_n_1minusbeta = sc.parallelize([(int(i),(1-beta)/n)for i in range(1,n+1)]) # Exercise 5.1.2
14  v = sc.parallelize([(int(i),float(1)/n) for i in range(1, n+1)])
15
16  # function that makes the form ((rowNumber,prob))
17  def mul(col_vec,v):
18      return col_vec.join(v).map(lambda x: (x[1][0][0], x[1][0][1]*x[1][1]) )
19
20  # function that makes next v by beta*Mv + (1-beta)e/n
21  def summ(mul,e):
22      return mul.reduceByKey(lambda a,b: a+b).join(e).map(lambda x: (x[0],x[1][1]+x[1][0]))
23
24  # iterate twice at one time, repeat 25 time to avoid laziness of spark
25  for i in range(25):
26      if i !=0:
27          v = sc.parallelize(check_point)
28      for i in range(2):
29          mul_vec = mul(col_vec,v)
30          v = summ(mul_vec,e_n_1minusbeta)
31      check_point = v.collect()
32  print(v.collect())
```



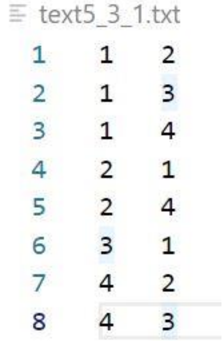
text5_1_2.txt

```
1	1
1	2
1	3
2	1
2	3
3	2
3	3
```



prob5_3_1.py

```python
from pyspark import SparkConf, SparkContext
conf = SparkConf()
sc = SparkContext(conf=conf)

beta = 0.8
n = 4 # n=4 for exercise 5.3.1
lines = sc.textFile(sys.argv[1])
pair_Rdd = lines.map(lambda l: l.split('\t')).map(lambda x: (x[0],x[1])).distinct()
# make the form ((sender,(receiver,probability))
col_vec = pair_Rdd.groupByKey().flatMap(lambda x: [(int(x[0]),(int(d),beta/len(x[1]))) for d in x[1]])

e_n_1minusbeta = sc.parallelize([(1,1-beta),(2,0),(3,0),(4,0)]) # Exercise 5.3.1 teleport set A
# e_n_1minusbeta = sc.parallelize([(1,(1-beta)/2),(2,0),(3,(1-beta)/2),(4,0)]) # Exercise 5.3.1 teleport set [A,C]
v = sc.parallelize([(int(i),float(1)/n) for i in range(1, n+1)])

# function that makes the form ((rowNumber,prob))
def mul(col_vec,v):
    return col_vec.join(v).map(lambda x: (x[1][0][0], x[1][0][1]*x[1][1]) )

# function that makes next v by beta*Mv + (1-beta)e/n
def summ(mul,e):
    return mul.reduceByKey(lambda a,b: a+b).join(e).map(lambda x: (x[0],x[1][1]+x[1][0]))

# iterate twice at one time, repeat 25 time to avoid laziness of spark
for i in range(25):
    if i !=0:
        v = sc.parallelize(check_point)
    for i in range(2):
        mul_vec = mul(col_vec,v)
        v = summ(mul_vec,e_n_1minusbeta)
    check_point = v.collect()
print(v.collect())
```



text5_3_1.txt

```
1	2
1	3
1	4
2	1
2	4
3	1
4	2
4	3
```

[Answer of the problems in 1-a]

1-a

5.1.2 $v' = \beta M v + (1-\beta) e/n$

$$M = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \qquad v' = 0.8 \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} v + 0.2 \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

After 50 iterations, $\begin{bmatrix} 0.25926 \\ 0.30864 \\ 0.43210 \end{bmatrix}$

5.3.1

teleport set = A only.

$$\begin{bmatrix} 0.42857 \\ 0.19048. \\ 0.19048, \\ 0.19048 \end{bmatrix}$$

teleport set A C only.

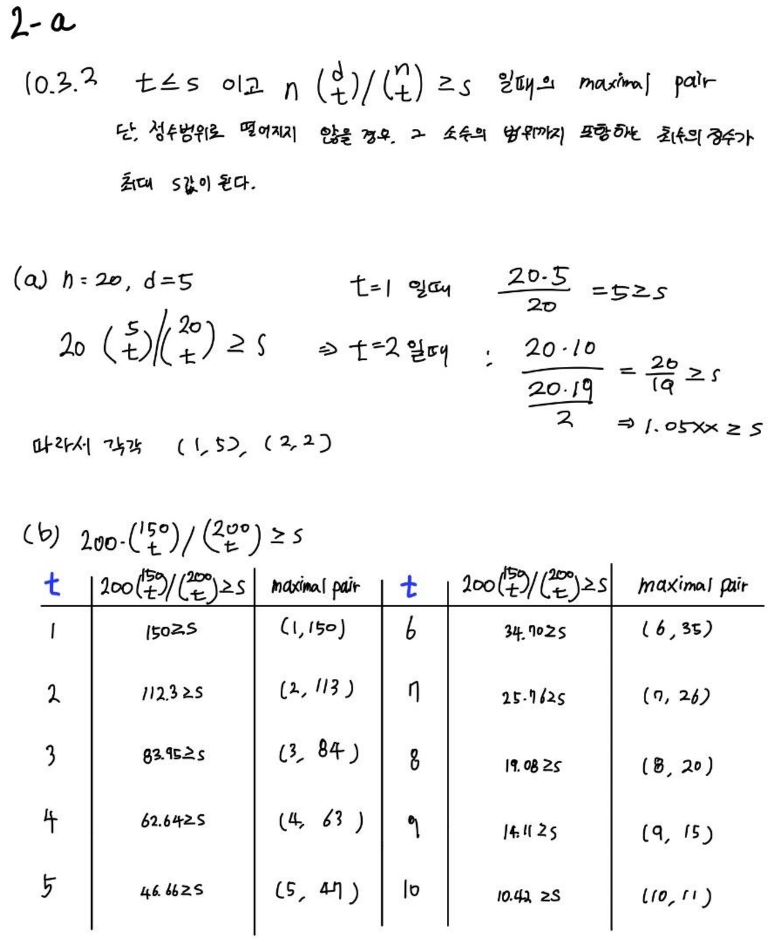
$$\begin{bmatrix} 0.38571 \\ 0.17143 \\ 0.27143 \\ 0.17143. \end{bmatrix}$$

(b)

263	0.00216
537	0.00212
965	0.00206
243	0.00197
255	0.00194
285	0.00193
16	0.00191
126	0.00190
747	0.00190
736	0.00189

Answer to Problem 2

(a)



2-a

10.3.2 $t \leq s$ 이고 $n \binom{d}{t} / \binom{n}{t} \geq s$ 일때의 maximal pair

단, 정수범위로 떨어지지 않을 경우, 그 소수의 범위까지 포함하는 최소의 정수가 최대 s값이 된다.

(a) $n = 20$, $d = 5$

$20 \binom{5}{t} / \binom{20}{t} \geq s$

$t=1$ 일때 $\frac{20 \cdot 5}{20} = 5 \geq s$

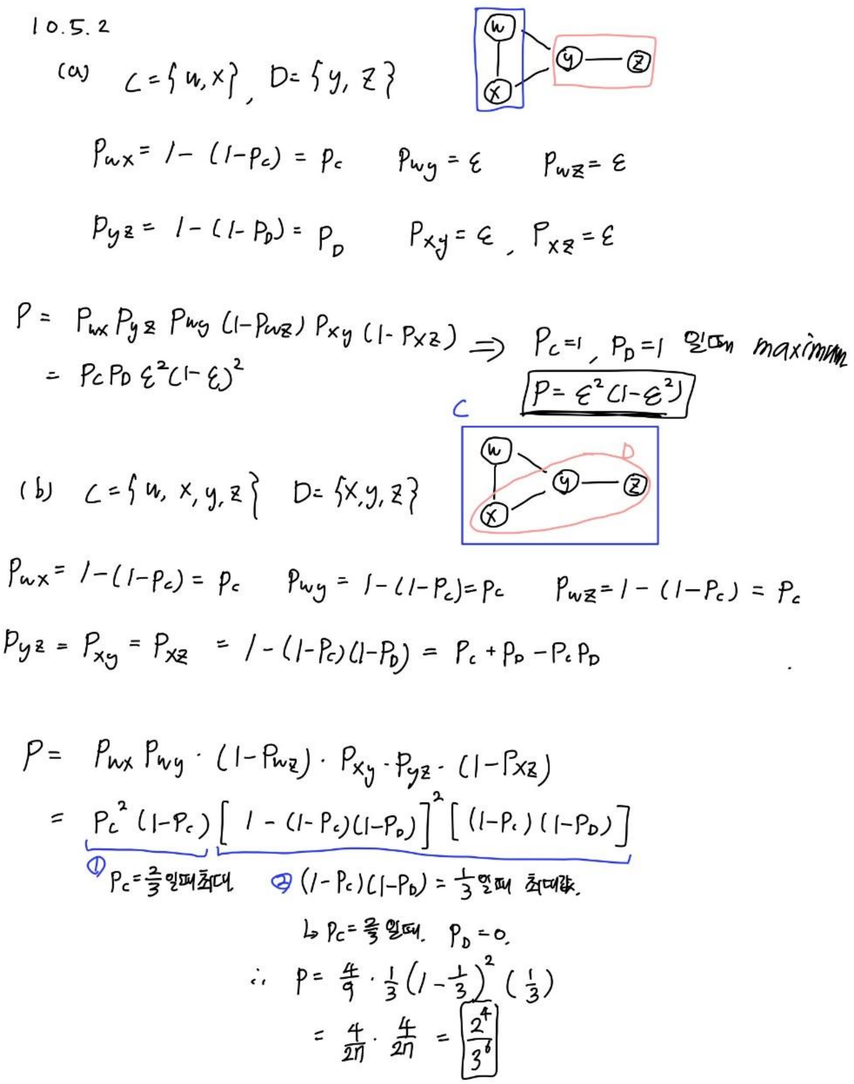
$\Rightarrow t=2$ 일때 : $\frac{20 \cdot 10}{\frac{20 \cdot 19}{2}} = \frac{20}{19} \geq s$

$\Rightarrow 1.05xx \geq s$

따라서 각각 (1, 5), (2, 2)

(b) $200 \cdot \binom{150}{t} / \binom{200}{t} \geq s$

t	$200 \binom{150}{t} / \binom{200}{t} \geq s$	maximal pair	t	$200 \binom{150}{t} / \binom{200}{t} \geq s$	maximal pair
1	$150 \geq s$	(1, 150)	6	$34.70 \geq s$	(6, 35)
2	$112.3 \geq s$	(2, 113)	7	$25.762 \geq s$	(7, 26)
3	$83.95 \geq s$	(3, 84)	8	$19.08 \geq s$	(8, 20)
4	$62.642 \geq s$	(4, 63)	9	$14.11 \geq s$	(9, 15)
5	$46.66 \geq s$	(5, 47)	10	$10.42 \geq s$	(10, 11)



10.5.2

(a) $C = \{w, x\}$, $D = \{y, z\}$

$P_{wx} = 1 - (1 - P_c) = P_c$ $\quad P_{wy} = \varepsilon$ $\quad P_{wz} = \varepsilon$

$P_{yz} = 1 - (1 - P_D) = P_D$ $\quad P_{xy} = \varepsilon$, $P_{xz} = \varepsilon$

$$P = P_{wx} P_{yz} P_{wy} (1 - P_{wz}) P_{xy} (1 - P_{xz}) \Rightarrow P_C = 1,\ P_D = 1 \text{ 일때 maximum}$$

$$= P_C P_D \varepsilon^2 (1 - \varepsilon)^2 \qquad \boxed{P = \varepsilon^2 (1 - \varepsilon^2)}$$

(b) $C = \{w, x, y, z\}$ $\quad D = \{x, y, z\}$

$P_{wx} = 1 - (1 - P_c) = P_c$ $\quad P_{wy} = 1 - (1 - P_c) = P_c$ $\quad P_{wz} = 1 - (1 - P_c) = P_c$

$P_{yz} = P_{xy} = P_{xz} = 1 - (1 - P_c)(1 - P_D) = P_c + P_D - P_c P_D$

$$P = P_{wx} P_{wy} \cdot (1 - P_{wz}) \cdot P_{xy} \cdot P_{yz} \cdot (1 - P_{xz})$$

$$= \underbrace{P_c^2 (1 - P_c)}_{①} \underbrace{\left[1 - (1 - P_c)(1 - P_D) \right]^2 \left[(1 - P_c)(1 - P_D) \right]}_{②}$$

① $P_c = \frac{2}{3}$ 일때 최대 $\quad$ ② $(1 - P_c)(1 - P_D) = \frac{1}{3}$ 일때 최대값.

↳ $P_c = \frac{2}{3}$ 일때. $P_D = 0$.

$$\therefore\ P = \frac{4}{9} \cdot \frac{1}{3} \left(1 - \frac{1}{3}\right)^2 \left(\frac{1}{3}\right)$$

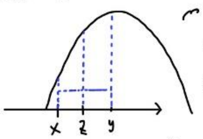
$$= \frac{4}{27} \cdot \frac{4}{27} = \boxed{\frac{2^4}{3^6}}$$

(b)
3501542

Answer to Problem 3

3-a

12.5.3



concave

Gini $= 1-\sum_{i=1}^{n} p_i^2$

Entropy $= \sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i})$

two class $\begin{cases} A: x \\ B \quad 1-x \end{cases}$

「To proof the concavity, we can show $f'' < 0$」

Gini $f(x) = 1 - x^2 - (1-x)^2$

$= 1 - x^2 - (x^2 - 2x + 1)$

$= 2x - 2x^2$

$= 2x(1-x)$

"Concave graph"

$f'(x) = -4x+2$

$f''(x) = -4 < 0.$

Entropy $f(x) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x} \qquad 0 < x < 1$

$= x\frac{-\ln x}{\ln 2} + (1-x)\frac{-\ln(1-x)}{\ln 2}$

$= \frac{1}{\ln 2}\left(-x\ln x - (1-x)\ln(1-x)\right)$

$= \frac{1}{\ln 2}\left(-x\ln x - \ln(1-x) + x\ln(1-x)\right)$

$\frac{1}{\ln 2}\left[x\ln\frac{1-x}{x} - \ln(1-x)\right]$

$f'(x) = \frac{1}{\ln 2}\left(-\ln x - x\frac{1}{x} \underbrace{- \frac{-1}{1-x} + \ln(1-x) + x\frac{-1}{1-x}}_{+1}\right)$

$= \frac{1}{\ln 2}\left(-\ln x + \ln(1-x)\right)$

$f''(x) = \frac{1}{\ln 2}\left(-\frac{1}{x} - \frac{1}{1-x}\right) < 0.$ always negative.

If a function $\hat{f}$ is concave then its double prim $f'' < 0$, vice versa.

So entropy $f(x)$ is concave func.

(b)

0.832833333333

0.7

0.001