CALVARIO STROIA

4 A)

=	≠	CAR	DOC
0	5	0	9
4	6	5	10
2	8	11	33
0	9	19	38
4	2	28	55
2	7	30	69
0	10	37	76

LAS POSIC EN LOS DOCS ESTÁN NUMERADAS DESDE POS 1, Y TODO CODIFICADO EN γ

LÉXICO CONCATENADO CON FRONT → APLICANDO CODING PARCIAL USANDO $n=3$ (CADA 3 TÉRMINOS ESCRIBE EL TÉRMINO COMPLETO)

MADRE IGUERA GISTRADO MALEFICO TANANTIAL MANIFIESTO

0 5 11 19 28 30 37

101 00100 11011010 100 100 1010101101 11111010 1111 010 101 101010 100 100 100 1000 1000 111 010

0 10 33 38 55 69 76

0101011010100100

10100 1001 1011 0101 00 100 101010 11111 0101 1111 0101 0110 1010 1001 0010 0100

0 10 33 38 55

01000 111 0100 0101 0110 1010 0100

69 76

↳ ESTOS SON LOS PUNTEROS A DOCS CODIFICADOS COMO GAMMA, PERO TAMBIÉN INCLUYE INFO SOBRE POSICIONES DE CADA TÉRMINO Y FRECUENCIA (DOC 1 FREC 1 P1... Pn DOC2...)

⇒ CON LOS DATOS DE LA TABLA ~~[illegible]~~ Y DEL LÉXICO CONCATENADO, PUEDO OBTENER LOS TÉRMINOS

MADRE → MADRIGUERA → MAGISTRADO → MALEFICO → MALETA → MANANTIAL → MANIFIESTO

↳ P/ OBTENER ESTOS TÉRMINOS, ~~[illegible]~~ SE REALIZA ~~[illegible]~~ UNA BÚSQUEDA BINARIA SOBRE EL ÍNDICE Y SE VA ACCEDIENDO A CADA POSICIÓN DEL LÉXICO Y LEYENDO TANTOS CARACTERES IGUALES O DISTINTOS DEL TÉRMINO ANTERIOR SEGÚN ME DIGA CADA POSICIÓN DEL ÍNDICE

AHORA TENGO QUE OBTENER LOS DOCUMENTOS, FRECUENCIAS Y POSICIONES DE CADA TÉRMINO SEGÚN LOS PUNTEROS CONCATENADOS → COMO ESTÁN CODIFICADOS EN GAMMA, SÉ QUE UNA FORMA DE DECODIFICAR GAMMA ES → LEER LA CANTIDAD DE CEROS CON LA QUE EMPIEZA CADA ~~[illegible]~~ NÚMERO

→ CONCATENACIÓN

$\gamma(x) = LEN(BINARIO(x)) - 1$ CEROS + BINARIO(x)

→ TODAS LAS DISTANCIAS (ENTRE DOCS Y ~~[illegible]~~) CODIFICADAS EN GAMMA

ENTRE LAS POSICIONES 0 y 10 → $\underbrace{1}_{1}\underbrace{010}_{2}\underbrace{010}_{2}\underbrace{011}_{3}$ → DOC 1, FREC 2, POSICIÓN 2 y 3

ENTRE LAS POSICIONES 10 y 33 → $\underbrace{011}_{3}\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{00100}_{4}\underbrace{1}_{1}\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{011}_{3}\underbrace{1}_{1}\underbrace{1}_{1}\underbrace{1}_{1}$

→ DOC 3, FREC 2, POS 1 y 4

→ DOC 4, FREC 2, POS 1 y 3

→ DOC 5, FREC 1, POS 1

ENTRE POS 33 y 38 → $\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{1}_{1}$ → DOC 2, FREC 1, POS 1

ENTRE POS 38 Y 55 → $\underbrace{1}_{1}\underbrace{1}_{1}\underbrace{1}_{1}\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{011}_{3}\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{010}_{2}$ → DOC 1, FREC 1, POS 1
→ DOC 3, FREC 1, POS 3
→ DOC 5, FREC 1, POS 2

→ ENTRE 55 Y 69 → $\underbrace{010}_{2}\underbrace{010}_{2}\underbrace{010}_{2}\underbrace{00100}_{4}$ → DOC 2, FREC 2, POS 2 Y 4

→ ENTRE 69 Y 76 → $\underbrace{011}_{3}\underbrace{1}_{1}\underbrace{010}_{2}$ → DOC 3, FREC 1, POS 2

→ POSICIONES MÁS ALLÁ DE 76 → $\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{011}_{3}\underbrace{010}_{2}\underbrace{1}_{1}\underbrace{00100}_{4}$ → DOC 2, FREC 1, POS 3
→ DOC 4, FREC 1, POS 4

ARMO LOS DOCUMENTOS → SON 5 DOCUMENTOS,

MADRE EN: DOC 1, FREC 2, POS 2 Y 3

MADRIGUERA EN: DOC 3, FREC 2, POS 1 Y 4; DOC 4, FREC 2, POS 1 Y 3; DOC 5, FREC 1, POS 1.

MAGISTRADO EN: DOC 2, FREC 1, POS 1

MALEFICO EN: DOC 1, FREC 1, POS 1; DOC 3, FREC 1, POS 3; DOC 5, FREC 1, POS 2

MACETA EN: DOC 2, FREC 2, POS 2 Y 4

MANANTIAL EN: DOC 3, FREC 1, POS 2

MANIFIESTO EN: DOC 2, FREC 1, POS 3; DOC 4, FREC 1, POS 4

D1: MALEFICO MADRE MADRE
D2: MAGISTRADO, MACETA, MANIFIESTO, MACETA
D3: MADRIGUERA, MANANTIAL, MALEFICO, MADRIGUERA
D4: MADRIGUERA, , MADRIGUERA, MANIFIESTO → HABRÁ ALGÚN ERROR YA QUE NO APARECE LA POS 2 DEL DOC 4.
D5: MADRIGUERA, MALEFICO,

TF-IDF → TF: TERM FREQUENCY (CANT DE VECES QUE APARECE CADA TÉRMINO DE Q EN CADA DOC

IDF → BUSCA DARLE UN PESO INVERSAMENTE PROPORCIONAL A SU FREC (LOS TÉRMINOS QUE APARECEN EN MUCHOS DOCS VAN A SER MENOS IMPORTANTES QUE LOS Q' APARECEN EN POCOS DOCS)

Q: "MADRIGUERO MALEFICO"

$IDF = \log\left(\frac{N+1}{F_{Ti}}\right)$ → N: CANT. TOTAL DE DOCS
→ CANT DE DOCS QUE POSEEN EL TÉRMINO

	TF					IDF
	D1	D2	D3	D4	D5	
MADRIGUERA	0	0	2	2	1	log (6/3) = 0,301
MALEFICO	1	0	1	0	1	log (6/3) = 0,301

PARA RANKEAR, CALCULAMOS → $R(Q, D_J) = \Sigma F_{Ti} \cdot IDF$ (PARA CADA TÉRMINO DE Q Y EN CADA DOC)

D1 → $0 * 0{,}301 + 1 * 0{,}301 = \overset{4°}{0{,}301}$; D2 → 0; D3 → $2 * 0{,}301 + 0{,}301 = \overset{1°}{0{,}903}$; D4 → $2 * 0{,}301 = \overset{2°}{0{,}602}$

D5 → $1 * 0{,}301 + 1 * 0{,}301 = \overset{3°}{0{,}602}$ RANKING → 1°: D3, 2°: D4, 3°: D5, 4°: D1, ① Ⓐ

```
MUERTES = SC.TEXTFILE('MUERTES.CSV')
PORCENTAJE_RAZA = SC.TEXTFILE('RAZAS.CSV')

# FILTRO LAS MUERTES POR RAZA NEGRA
MUERTES_NEGROS = MUERTES.FILTER(LAMBDA X: X[3]=="BLACK")\
.MAP(LAMBDA X: (X[4], 1))    → (ESTADO, 1)

# SUMO LAS MUERTES
MUERTES_NEGROS = MUERTES_NEGROS.REDUCEBYKEY(LAMBDA X,Y: X+Y)

# ME QUEDO CON EL ESTADO DE MAYOR CANTIDAD DE MUERTES YA QUE ESE VA A TENER EL
MAYOR PORCENTAJE DE MUERTES

MUERTES_NEGROS.REDUCE(LAMBDA X,Y: X IF X[1] > Y[1] ELSE Y)
```

Ⓑ

```
                SHARE.BLACK
# (STATE, (~~RAZA NEGRA~~, 1))
                                                            SUMO LOS PORCENTAJES
                                                            DE LAS CLASES Y EL
RDD1 = PORCENTAJE_RAZA.MAP(LAMBDA X: (X[0], (X[3], 1)))      CONTADOR
RDD1 = RDD1.REDUCEBYKEY(LAMBDA X,Y: (X[0]+Y[0], X[1]+Y[1]))

# CALCULO EL PROMEDIO POR ESTADO
                                           PROMEDIO
RDD1 = RDD1.MAP(LAMBDA X: (X[0], X[1][0] / X[1][1]))

# ~~[illegible]~~ CALCULO EL PORCENTAJE DE LAS MUERTES
MUERTES_NEGROS = MUERTES_NEGROS.MAP(LAMBDA X: (X[0], X[1]/MUERTES_NEGROS.COUNT()))
# JOINEO RDD1 CON MUERTES_NEGROS

RDD2 = RDD1.JOIN(MUERTES_NEGROS) → # (ESTADO, (PORCENTAJE_MUERTES, PROMEDIO))
```

Scanned with CamScanner

```
#AHORA CALCULO LA DIFERENCIAS ENTRE PORCENTAJES Y PROTESTA

RDD2 = RDD2.MAP(LAMBDA X: (X[0], ABS(X[1][0] - X[1][1]))
                          ESTADO  DIFERENCIA

#ME QUEDO CON EL TOP10

RDD2.TAKEORDERED(10, LAMBDA X: -X[1])
```

Scanned with CamScanner

```
(2) A) IMPORT PANDAS AS PD

METRICAS = PD.READ_CSV('METRICS.CSV')
CLIENTES = PD.READ_CSV(' ~~METRICA~~ 'CLIENTS.CSV')

# ME QUEDO CON CLIENTES CON EL NUMERO DE CUENTA MAYOR AL DADO

CLIENTES_FILTER = CLIENTES[CLIENTES['ACCOUNT_NUMBER'] > 25679247]

# ME QUEDO CON LAS METRICAS PEDIDAS

METRICAS_PEDIDAS = METRICAS[(METRICAS['METRIC'] == 'AWS.VPC.NETWORK_IN') \
                            OR (METRICAS['METRIC'] == 'AWS.VPC.NETWORK_RATE') \
                            OR (METRICAS['METRIC'] == 'AWS.VPC.NETWORK-OUT')]

# HAGO MERGE 'INNER' POR CLIENT_ID

MERGED = CLIENTES_FILTER.MERGE(METRICAS_PEDIDAS, ON='CLIENT_ID')

# AGRUPO POR METRICA Y ACCOUNT_NUMBER

MERGED = MERGED.GROUPBY([['METRIC', 'ACCOUNT_NUMBER']).AGG({'VALUE':'MEAN'}) ~~RESET_INDEX()~~
         ~~[illegible]~~

MERGED.UNSTACK().T.RENAME(COLUMNS={'AWS.VPC.NETWORK_IN': 'AWS.VPC.NETWORK_IN_MEAN',
                                   'AWS.VPC.NETWORK_OUT': 'AWS.VPC.NETWORK_OUT_MEAN',
                                   'AWS.VPC.NETWORK_RATE': 'AWS.VPC.NETWORK_RATE_MEAN'})
```

Scanned with CamScanner

CALTABIANO SERGIO

HOJA N°

FECHA

③ B)

$V_1 = [1,5,3,2]$, $V_2 = [0,0,1,2]$, $V_3 = [4,4,5,0]$, $V_4 = [5,1,0,1]$

CALCULAR REGLA DE HIPERPLANOS CON $b=1$, $r=4$. HALLAR HIPERPLANOS CORRESP PARA QUE EL + SIMILAR ~~a~~ A V1 SEA V3

CANTIDAD DE MH → $r \times b = 4$

PARA QUE LOS 2 VEC'S SEAN SIMILARES, TIENE QUE DAR IGUAL EL MISMO VALOR DE PRODUCTO INTERNO ENTRE LOS VEC'S Y LOS HIPERPLANOS

$u_1 = [1,-1,-1,1]$ → $V_1 \cdot u_1 < 0$ → $MH_1 = 1$
$V_2 \cdot u_1 > 0$ → $MH_2 = -1$
$V_3 \cdot u_1 < 0$ → $MH_3 = 1$
$V_4 \cdot u_1 > 0$ → $MH_3 = -1$
} V1 SIMILAR A V3

$u_2 = [-1,1,1,-1]$ → $V_1 \cdot u_2 > 0$, $V_2 \cdot u_2 < 0$, $V_3 \cdot u_2 > 0$, $V_4 \cdot u_2 < 0$ } $MH_1 = 1$, $MH_2 = -1$, $MH_3 = 1$, $MH_4 = -1$ } V1 SIMILAR A V3

$u_4 = [1,1,1,-1]$ → $V_1 \cdot u_4 = > 0$
$V_2 u_4 < 0$
$V_3 u_4 > 0$
$V_4 u_4 > 0$

$u_3 = [-1,-1,-1,-1]$ → $V_1 \cdot u_3 < 0$, $V_2 \cdot u_3 < 0$, $V_3 \cdot u_3 < 0$, $V_4 \cdot u_3 < 0$ } → COMO YA ENCONTRE 2 HIPERPLANOS ACERTADOS QUE HACEN QUE V1 Y V3 SEAN SIMILARES, NO CAMBIA EL RESULTADO

	MH1	MH2	MH3	MH4
V1	1			
V2				
V3				
V4				

	V1	V2	V3	V4
u1	1	-1	1	-1
u2	1	-1	1	-1
u3	-1	-1	-1	-1
u4	1	-1	1	1

V1 Y V3 SIMILARES EN TODOS SUS MH USADOS

$u_1 = [1,-1,-1,1]$
$u_2 = [-1,1,1,-1]$
$u_3 = [-1,-1,-1,-1]$
$u_4 = [1,1,1,-1]$