SDM

$d_1 = (t_1\ t_2\ t_3\ t_4\ t_5\ t_6)$

$d_2 = (t_3\ t_4\ t_3\ t_1\ t_8\ t_2\ t_2\ t_2)$

$d_3 = (t_2\ t_4\ t_4\ t_1\ t_8\ t_2\ t_3\ t_1\ t_4)$

$q = (t_4\ t_3)$

$\mu = 6$

$\lambda_T = 0.85$

$\lambda_0 = 0.1$

$\lambda_u = 0.05$

$w = 4$

$D = \{d_1, d_2, d_3\}$

$|d|$ = length of a document

(a) what is the value of unigram feature function for "t_4" in document d_3?

$$f_T(q_i, d) = \log_2 P(q_i | \theta_d)$$

$$f_T(t_4, d_3) = \log_2 P(t_4 | \theta_{d_3})$$

$$= \log_2 \left(\frac{c(t_4, d_3) + \mu P(t_4 | \theta)}{|d_3| + \mu} \right)$$

$$= \log_2 \left(\frac{2 + 6\left(\frac{1+1+2}{6+8+9}\right)}{9+6} \right)$$

$$= \log_2 \left(\frac{2 + 6(4/23)}{15} \right)$$

$$= \boxed{-2.301}$$

(b) $$f_0(t_4, t_3, d_3) = \log_2 \left(\frac{c_0(t_4, t_3, d_3) + \mu P_0(t_4, t_3 | D)}{|d_3| + \mu} \right)$$

$$= \log_2 \left(\frac{1 + 6 \cdot \left(\frac{0+1+0}{6+8+9}\right)}{9+6} \right)$$

$$= \boxed{-5.345}$$

$\frac{1+2+1}{6+8+9} = 0.1789$

(c) $f_4('t_4, t_3'', d_3) = \log_2\left(\frac{0 + 6\left(\frac{1+2+1}{6+8+9}\right)}{9+6}\right) = -2.876$

(d) Total retrieval score for d_3 = score(d_3, q)

$= \lambda_7 \sum_{i=1}^{n} f_7(q_i, d_3) + \lambda_6 f_6(t_4, t_3, d_3) + \lambda_4 (t_4, t_3, d_3)$

$= 0.85(-2.301 - 2.876) + 0.1(-5.845) + 0.05(-2.876)$

$= \boxed{-5.129}$

(e) $f_7(t_4, d_2) = \log_2\left(\frac{1 + 6\left(\frac{1+1+2}{6+8+9}\right)}{8+6}\right) = -2.7769$

$f_7(t_3, d_2) = \log_2\left(\frac{2 + 6\left(\frac{1+2+1}{6+8+9}\right)}{8+6}\right) = -2.2022$

$f_6(t_4, t_3, d_2) = \log_2\left(\frac{1 + 6\left(\frac{0+1+0}{6+8+9}\right)}{8+6}\right) = -3.4771$

$f_4(t_4, t_3, d_2) = \log_2\left(\frac{2 + 6\left(\frac{1+2+1}{6+8+9}\right)}{8+6}\right) = -2.2015$

score$(d_2, q) = 0.85(-2.7769 - 2.2022) + 0.1(-3.4771) + 0.$

$= \boxed{-4.689}$

(f) $f_7(t_4, d_1) = \log_2\left(\frac{1 + 6\left(\frac{1+1+2}{6+8+9}\right)}{6+6}\right) = -2.5546$

$f_7(t_3, d_1) = \log_2\left(\frac{1 + 6(0.1789)}{12}\right) = -2.5546$

$f_6(t_4, t_3, d_1) = \log_2\left(\frac{0 + 6\left(\frac{0+1+0}{6+8+9}\right)}{12}\right) = -5.5395$

$f_4(t_4, t_3, d_1) = \log_2\left(\frac{1 + 6(0.174)}{12}\right) = -2.5538$

score$(d_1, q) = 0.85(-2.5546 - 2.5546) + 0.1(-5.5395) + 0.05$

$= -5.02367$

min-max Normalization

$$V' = \left(\frac{V - \min(A)}{\max(A) - \min(A)} \right) (\text{new\_max}(A) - \text{new\_min}(A)) + \text{new\_min}(A)$$

1. For example:

$V = (\underbrace{1000}_{a_1}, 2000, 3000, \underbrace{9000}_{a_4})$

$\min \qquad\qquad \max$

$V = 1000$

$\min(A) = 1000$

$\max(A) = 9000$

$\text{new\_min}(A) = 0$

$\text{new\_max}(A) = 1$

$\therefore a_1 = \frac{1000 - 1000}{9000 - 1000} \cdot (1-0) + 0 = 0$

$\therefore a_2 = \frac{2000 - 1000}{9000 - 1000} \cdot (1-0) + 0 = 0.125$

$\therefore a_3 = \frac{3000 - 1000}{9000 - 1000} \cdot (1-0) + 0 = 0.25$

$\therefore a_4 = \frac{9000 - 1000}{9000 - 1000} \cdot (1-0) + 0 = 1$

$V' = (0, 0.125, 0.25, 1)$

2. $V = (2, 12, 7)$

$V' = \left(\frac{2-2}{12-2} \cdot (1-0) + 0, \frac{12-2}{12-2}, \frac{7-2}{12-2} \right)$

$= (0, 1, 0.5)$

K-MEAN clustering.

1.

	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

K = 2

∴ Initial centroid

	X	Y
C1	185	72
C2	170	56

∴ Euclidean distance $[(x,y),(a,b)]$:

$= \sqrt{(x-a)^2 + (x-b)^2}$

$ED_{C1-C1} = \sqrt{(185-185)^2 + (72-72)^2} = 0$

$ED_{C2-C2} = \sqrt{(170-185)^2 + (56-72)^2}$

$= 21.93$

$ED_{C2-C2} = 0$

	centroid		
clustes	X	Y	Assignment
C1	0	21.93	1
C2	21.93	0	2

Let's take next dataset (168, 60) = N

$ED_{C1-N} = \sqrt{(168-185)^2 + (60-72)^2} = 20.808$

$ED_{C2-N} = \sqrt{(168-170)^2 + (60-56)^2} = 4.472$

	X	Y	Assi
N	20.808	4.472	2 (∵ Y<X)

recalculate C2.

(170 + 168)/2 (60 + 56)/2

169 58

C2 = (169, 58)

Now N = (179, 68)

$ED_{C1-N} = \sqrt{(179-185)^2 + (68-72)^2} = 7.211103$

$ED_{C2-N} = \sqrt{(179-169)^2 + (68-58)^2} = 14.14214$

∵ X < Y recalculate C1

∵ C1 = ((185+179)/2, (72+68)/2))

= (182, 70)

Now N = (182, 72)

$ED_{C1-N} = \sqrt{(182-182)^2 + (68-70)^2} = 2$

$ED_{C2-N} = \sqrt{(182-169)^2 + (72-58)^2} = 19.10$

new C1 = ((182+182)/2, (70+72)/2) = (182, 71)

Now N = (188, 77)

$ED_{C1-N} = \sqrt{(188-182)^2 + (77-71)^2} = 8.4852$

$ED_{C2-N} = \sqrt{(188-169)^2 + (77-58)^2} = 26.87$

new C1 = ((182+188)/2, (71+77)/2) = (185, 74)

Final assignment:

Dataset	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	77	1

2.

Dataset	w	x	y	z
P1	3	1	0	4
P2	2	0	1.5	1
P3	1	3	4.5	5
P4	2.5	2.5	5	3.5
P5	6	3	2	0

centroids	w	x	y	z
C1	3	5	4	4.5
C2	1.5	2.5	1	2
C3	2	3.5	1	0

For $P_1 = (3, 1, 0, 4)$

$ED_{P1-C1} = \sqrt{(3-3)^2 + (1-5)^2 + (0-4)^2 + (4-4.5)^2} = 32.25$

$ED_{P1-C2} = \sqrt{(3-1.5)^2 + (1-2.5)^2 + (0-1)^2 + (4-2)^2} = 9.5$

$ED_{P1-C3} = \sqrt{(3-2)^2 + (1-3.5)^2 + (0-1)^2 + (4-0)^2} = 24.25$

$\therefore$ assignment = C2 ($\because C2 < C1, C2 < C3$)

new C2 = $(\frac{1.5+3}{2}, \frac{2.5+1}{2}, \frac{0+1}{2}, \frac{2+4}{2}) = (2.25, 1.75, 0.5, 3)$

For $P_2 = (2, 0, 1.5, 1)$

$ED_{P2-C1} = \sqrt{(2-3)^2 + (0-5)^2 + (1.5-4)^2 + (1-4.5)^2} = 6.670$

$ED_{P2-C2} = \sqrt{(2-2.25)^2 + (0-1.75)^2 + (1.5-0.5)^2 + (1-3)^2} = 2.850$

$ED_{P2-C3} = \sqrt{(2-2)^2 + (0-3.5)^2 + (1.5-1)^2 + (1-0)^2} = 3.674$

$\therefore$ assignment = C2 ($\because C2 < C1, C2 < C3$)

new C2 = $(\frac{2.25+2}{2}, \frac{1.75+0}{2}, \frac{1.5+0.5}{2}, \frac{1+3}{2})$

$= (2.125, 0.875, 1, 2)$

For $P_3 = (1, 3, 4.5, 5)$

$ED_{P3-C1} = \sqrt{(1-3)^2 + (3-5)^2 + (4.5-4)^2 + (5-4.5)^2} = 2.915$

$ED_{P3-C2} = \sqrt{(1-2.125)^2 + (3-0.875)^2 + (4.5-1)^2 + (5-2)^2} = 5.20$

$ED_{P3-C3} = \sqrt{(1-2)^2 + (3-3.5)^2 + (4.5-1)^2 + (5-0)^2} = 6.20$

assignment = C1 ($\because C_1 < C_2, C_1 < C_3$)

new C1 = $(\frac{3+1}{2}, \frac{5+3}{2}, \frac{4+4.5}{2}, \frac{4.5+5}{2}) = (2, 4, 4.25, 4.75)$

new $C1 = \left(\frac{2.5+2}{2}, \frac{2.5+4}{2}, \frac{5+4.25}{2}, \frac{3.5+4.75}{2}\right)$

$= (2.25, 3.25, 4.625, 4.125)$

Date
Page

For $P4 = (2.5, 2.5, 5, 3.5)$

$ED_{P4-C1} = \sqrt{(2.5-2)^2 + (2.5-4)^2 + (5-4.75)^2 + (3.5-4.75)^2}$

$= 2.150$

$ED_{P4-C2} = \sqrt{(2.5-2.125)^2 + (2.5-0.875)^2 + (5-1)^2 + (3.5-2)^2}$

$= 4.586$

$ED_{P4-C3} = \sqrt{(2.5-2)^2 + (2.5-3.5)^2 + (5-1)^2 + (3.5-0)^2}$

$= 5.431$

assignment = C1 ($\because C1 < C2, C1 < C3$)

For $P5 = (6, 3, 2, 0)$

$ED_{P5-C1} = \sqrt{(6-2.25)^2 + (3-3.25)^2 + (2-4.625)^2 + (0-4.125)^2}$

$= 6.167$

$ED_{P5-C2} = \sqrt{(6-2.125)^2 + (3-0.875)^2 + (2-1)^2 + (0-2)^2}$

$= 4.952$

$ED_{P5-C3} = \sqrt{(6-2)^2 + (3-3.5)^2 + (2-1)^2 + (0-0)^2}$

$= 4.153$

$\therefore$ assignment = C3 ($\because C3 < C1, C3 < C2$)

N = 1000
|k1| = 1.25
b = 0.8
avgdl = 50

BM25

	t1	t2	t3	t4	t5	t6	length
d1	3	0	8	1	10	5	21
d2	4	3	3	2	1	7	20
C	100	50	80	93	100	85	1000

$$BM25(q,d) = \sum_{t \in q} \frac{f_{t,d} \cdot (1+k_1)}{f_{t,d} + k_1 \left(1 - b + b \cdot \left[\frac{|d|}{avgdl}\right]\right)} \cdot \log_{10} \frac{N}{n_t}$$

q = "t2"

① $BM25(q, d_1) = 0$

② $BM25(q, d_2) = \frac{3 \cdot (1+1.25) \cdot \log_{10}\left(\frac{1000}{50}\right)}{3 + 1.25\left(1 - 0.8 + 0.8\left(\frac{20}{50}\right)\right)} = \frac{8.781952}{3.65}$

$= 2.40601$

q = "t2 t2 t5"

③ $BM25(q, d_1) = 0 + 0 + \frac{10 \cdot (1+1.25) \cdot \log_{10}\left(\frac{1000}{100}\right)}{10 + 1.25\left(1 - 0.8 + 0.8\left(\frac{21}{50}\right)\right)} = \frac{22.5}{10.67}$

$= 2.10871$

④ $BM25(q, d_2) = 2.40601 + 2.40601 + \frac{1 \cdot (1+1.25) \cdot \log_{10}\left(\frac{1000}{100}\right)}{1 + 1.25\left(1 - 0.8 + 0.8\left(\frac{20}{50}\right)\right)}$

$= 6.175$

	in	out	
A	1	1	$B/2$
B	2	2	$C + \frac{F}{2}$
C	0	1	0
D	1	[illegible]	$\frac{D}{[illegible]}$
E	3	1	$\frac{B}{2} + \frac{F}{2} + \frac{D}{[illegible]}$
F	1	2	$\frac{A}{1}$

0 **1**

A 0.167 $(\frac{0.2}{6}) + 0.8(\frac{0.164}{2}) = 0.1$

B 0.167 $(\frac{0.2}{6}) + 0.8(0.167 + \frac{0.167}{2}) = 0.233$

C 0.167 $(\frac{0.2}{6}) + 0.8(0) = 0.033$

D 0.167 $(\frac{0.2}{6}) + 0.8(\frac{0.167}{[illegible]}) = 0.169$

E 0.167 $(\frac{0.2}{6}) + 0.8(\frac{0.167}{2} + \frac{0.167}{2} + 0.167) = 0.3$

F 0.167 $(\frac{0.2}{6}) + 0.8(0.167) = 0.167$

2

A $(\frac{0.2}{6}) + 0.8(\frac{0.233}{2}) = 0.127$

B $(\frac{0.2}{6}) + 0.8(0.033 + \frac{0.167}{2}) = 0.127$

C $(\frac{0.2}{6}) + 0.8(0) = 0.033$

D $(\frac{0.2}{6}) + 0.8(0.167) = 0.167$

E $(\frac{0.2}{6}) + 0.8(\frac{0.233}{2} + \frac{0.167}{2} + 0.167) = 0.327$? 0.433

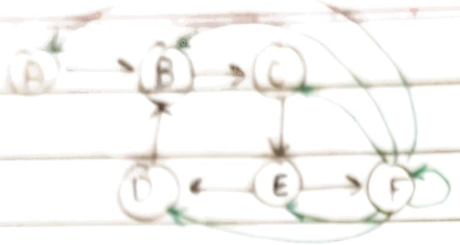
F $(\frac{0.2}{6}) + 0.8(0.1) = 0.113$

PageRank

$q = 0.4$

$T = 6$

$1-q = 0.6$



F does not have outgoing link so sink needs to be fixed.

Out		P(in|om)	i=0	i=1
1	A	F/6	$1/6 = 0.167$	$(\frac{0.4}{6}) + 0.6(\frac{0.167}{6}) = 0.083$
1	B	$A + \frac{F}{6} + D$	0.167	$(\frac{0.4}{6}) + 0.6(0.167 + \frac{0.167}{6} + 0.167) = 0.283$
1	C	$B + \frac{F}{6}$	0.167	$(\frac{0.4}{6}) + 0.6(0.167 + \frac{0.167}{6}) = 0.183$
1	D	$E/2 + F/6$	0.167	$(\frac{0.4}{6}) + 0.6(0.167/2 + \frac{0.167}{6}) = 0.133$
2	E	$C + \frac{F}{6}$	0.167	$(\frac{0.4}{6}) + 0.6(0.167 + \frac{0.167}{6}) = 0.183$
6	F	$F/6 + E/2$	0.167	$(\frac{0.4}{6}) + 0.6(0.167/6 + \frac{0.167}{2}) = 0.133$

i=2

A $(\frac{0.4}{6}) + 0.6(\frac{0.133}{6}) = 0.79$

B $(\frac{0.4}{6}) + 0.6(0.083 + \frac{0.133}{6} + 0.133) = 0.209$

C $(\frac{0.4}{6}) + 0.6(0.283 + \frac{0.133}{6}) = 0.249$

D $(\frac{0.4}{6}) + 0.6(\frac{0.183}{2} + \frac{0.133}{6}) = 0.135$

E $(\frac{0.4}{6}) + 0.6(0.183 + \frac{0.133}{6}) = 0.189$

F $(\frac{0.4}{6}) + 0.6(\frac{0.133}{6} + \frac{0.183}{2}) = 0.135$

...ification Evaluation

Instance	True Class	A	B
1	N	Y	N
2	N	N	Y
3	Y	Y	Y
4	N	Y	N
5	Y	N	Y
6	Y	Y	Y
7	N	Y	N
8	Y	Y	N

A

	−	+
−	TN = 1	FP = 3
+	FN = 1	TP = 3

B

	−	+
−	3	1
+	1	3

A $P = \frac{TP}{TP+FP} = \frac{3}{3+3} = \frac{3}{6} = 0.5$

$R = \frac{TP}{FN+TP} = \frac{3}{1+3} = \frac{0.75}{\text{same}}$

$F1 = \frac{2PR}{P+R} = \frac{2 \cdot (0.5)(0.75)}{0.5+0.75} = 0.6$

B $P = \frac{3}{1+3} = \boxed{0.75}$

$R = \frac{3}{4} = 0.75$ same.

$F1 = \frac{2(0.75)(0.75)}{0.75+0.75} = \boxed{0.75}$

actual	predicted
N	Y
Y	Y
N	Y
N	Y
N	N
Y	Y
Y	Y
Y	N
N	N
Y	N

actual \ Predicted	+	−
+	3	2
−	3	2

Classification Evaluation

Date
Page

predicted

actual	N	Y
N	Y	
Y	Y	
Y	Y	
N	Y	
N	N	
N	N	
Y	Y	
Y	N	
N	N	
N	Y	

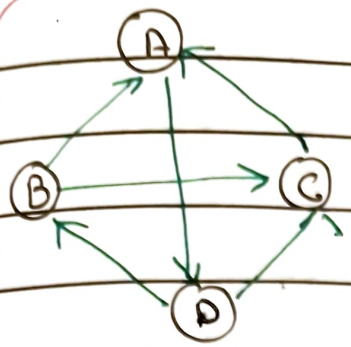
	−	+
−	TN = 3	FP = 3
+	FN = 1	TP = 3

$$P = \frac{TP}{TP + FP} = \frac{3}{3+3} = 0.5$$

$$R = \frac{TP}{FN + TP} = \frac{3}{1+3} = 0.75$$

$$F1 = \frac{2 \cdot PR}{P+R} = \frac{2(0.5)(0.75)}{0.5+0.75} = 0.6$$

PageRank



$d = 0.2$

$T = 4$

$1 - d = 0.8$

Out		P(in)	0	1	2
1	A	B/2 + C/1	0.25	$(\frac{0.2}{4}) + 0.8(\frac{0.25}{2} + 0.25) = 0.35$	0.31
2	B	D/2	0.25	$(\frac{0.2}{4}) + 0.8(\frac{0.25}{2}) = 0.15$	0.15
1	C	B/2 + D/2	0.25	$(\frac{0.2}{4}) + 0.8(\frac{0.25}{2} + \frac{0.25}{2}) = 0.25$	0.21
2	D	A	0.25	$(\frac{0.2}{4}) + 0.8(0.35) = 0.25$	0.33

Language Modeling Retrieval
Dirichlet smoothing

	d_1	d_2	d_3	d_4	$\|c\|$	$P(t_i, C)$
t_1	1	1	2	1	5	$\frac{5}{22} = 0.227$
t_2	0	2	0	1	3	$\frac{3}{22} = 0.136$
t_3	2	0	1	0	3	$\frac{3}{22} = 0.136$
t_4	4	0	1	2	7	$\frac{7}{22} = 0.318$
t_5	1	2	1	0	4	$\frac{4}{22} = 0.182$
	8	5	5	4	22	

① $\text{Empirical\_P}(t_5|d_1) = \frac{C_{t_5,d_1}}{|d_1|} = \frac{1}{8} = 0.125$

② $\text{background\_P}(t_4|C) = \frac{\sum_{d'} C_{t,d'}}{\sum_{d'} |d'|} = \frac{7}{22} = 0.318$

③ $\text{Smoothed\_P}(t_2|d_3) = \frac{C_{t_2,d_3} + \mu P(t_2|C)}{|d_3| + \mu} = \frac{0 + 6(\frac{3}{22})}{5+6} = 0.074$

④ $P(t_1|d_2) = \frac{1+6(\frac{5}{22})}{5+6} = 0.214$ $\quad P(t_4|d_2) = \frac{0+6(\frac{7}{22})}{5+6} = 0.173$

$P(t_2|d_2) = \frac{2+6(\frac{3}{22})}{5+6} = 0.256$ $\quad P(t_5|d_2) = \frac{2+6(\frac{4}{22})}{5+6} = 0.280$

$P(t_3|d_2) = \frac{0+6(3/22)}{5+6} = \boxed{0.074}$ lowest

⑤ $Q = "t_1\ t_3\ t_5"$

$\text{score}(d_1, Q) = \log\left(\frac{1+6(\frac{5}{22})}{8+6}\right)\cdot 1 + \log\left(\frac{2+6(\frac{3}{22})}{8+6}\right)\cdot 1 + \log\left(\frac{1+6(\frac{4}{22})}{8+6}\right)\cdot 1$

$= -2.2944$

$\text{score}(d_2, Q) = \log\left(\frac{1+6(5/22)}{5+6}\right)\cdot 1 + \log\left(\frac{0+6(\frac{3}{22})}{5+6}\right)\cdot 1 + \log\left(\frac{2+6(\frac{4}{22})}{5+6}\right)\cdot 1$

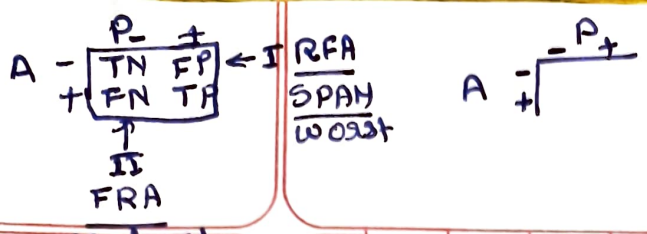
$= -2.3476$

$\text{score}(d_3, Q) = \log\left(\frac{2+6(\frac{5}{22})}{5+6}\right)\cdot 1 + \log\left(\frac{1+6(\frac{3}{22})}{5+6}\right)\cdot 1 + \log\left(\frac{1+6(\frac{4}{22})}{5+6}\right)\cdot 1$

$= -2.0173 \leftarrow$ highest

$\text{score}(d_4, Q) = \log\left(\frac{1+6(\frac{5}{22})}{4+6}\right)\cdot 1 + \log\left(\frac{0+6(\frac{3}{22})}{4+6}\right)\cdot 1 + \log\left(\frac{0+6(\frac{4}{22})}{4+6}\right)\cdot 1$

$= -2.6757$



Date
Page

$$ACC = \frac{TP+TN}{TN+FN+FP+TP} = \text{Fof CC items}$$

$$ER = \frac{FN+FP}{TN+FN+TP+FP} = \text{Fof ICC items.}$$

$$P = \frac{TP}{TP+FP} = \text{Fof CC items - a - pos - from - total Pos predicted.}$$

$$R = \frac{TP}{TP+FN} = \text{Fof CC items - a - pos - from actual pos.}$$

sensitivity
True Pos Rate

$$\text{Specificity} = \frac{TN}{FP+TN}$$

$$F1 = \frac{2 \cdot PR}{P+R.} = \text{mean of P \& R.}$$

$$FPR = E\text{-}I = \frac{FP}{FP+TN} = \text{Fof WC items - pos - out of - actual - neg}$$

$$FNR = E\text{-}II = \frac{FN}{FN+TP} = \text{Fof WC items - neg out of actual pos}$$

$$P(t|C) = \frac{n(t,C)+1}{\sum_{t'} n(t',C)+|C|}$$

NB

	t1	t2	t3	t4	t5	t6	t7	t8	class
d1	1	9	7	5	2	1	3	0	C1
d2	0	2	1	3	0	1	0	2	C2
d3	1	1	0	0	4	5	2	1	C1
d4	2	1	4	0	5	2	7	0	C3
d5	0	1	5	0	1	4	0	2	C2
d6	6	3	0	3	2	1	3	0	C3
d7	2	1	3	1	8	1	0	9	C1
d8	1	1	0	1	0	9	0	0	C3

$n(t,C)$	t1	t2	t3	t4	t5	t6	t7	t8	$n(t,C)$ Total	P(C)
C1	4	11	10	6	14	7	5	10	67	3/8 = 0.375
C2	0	3	6	3	1	5	0	4	22	2/8 = 0.25
C3	9	5	4	4	7	12	10	0	51	3/8 = 0.375

smoothed $P(t,C)$	t1	t2	t3	t4	t5	t6	t7	t8
C1	$\frac{4+1}{67+3}$ = 0.071	0.171	0.157	0.1	0.214	0.114	0.086	0.157
C2	$\frac{1}{25}$ = 0.04	0.16	0.28	0.16	0.08	0.24	0.04	0.2
C3	$\frac{9+1}{51+3}$ = 0.186	0.11	0.093	0.093	0.148	0.24	0.203	0.018

$P(C2 \mid "t2\ t3\ t6")$

$= P(C2) \cdot P("t2" \mid C2) \cdot P("t3" \mid C2) \cdot P("t6" \mid C2)$

$= 0.25 \times 0.16 \times 0.28 \times 0.24$

$= 0.0027$

$P(C \mid "t2\ t7\ t8")$ is the highest? C1

$P(C1 \mid "t2\ t7\ t8") = 0.375 \times 0.171 \times 0.086 \times 0.157 = 0.00086$

$P(C2 \mid "t2\ t7\ t8") = 0.25 \times 0.16 \times 0.04 \times 0.2 = 0.00032$

$P(C3 \mid "t2\ t7\ t8") = 0.375 \times 0.11 \times 0.203 \times 0.018 = 0.00015$

$$P(t|C) = \frac{n(t,C)+1}{\sum_{t'} n(t',C)+|C|}$$ NB

Date
Page

	t1	t2	t3	t4	t5	t6	t7	class
d1	2	0	1	2	0	2	4	C1
d2	0	0	0	0	3	2	2	C3
d3	3	4	0	2	0	0	2	C2
d4	4	0	3	1	1	1	0	C3
d5	1	0	0	3	1	2	0	C2
d6	0	1	1	0	3	4	1	C1

n(t,C)

								n(t',C)	P(C)
C1	2	1	2	2	3	6	5	21	$\frac{3}{6}=0.35$
|C| C2	4	4	0	5	1	2	2	18	$\frac{2}{6}=0.33$
=3 C3	4	0	3	1	4	3	2	17	$\frac{2}{6}=0.33$
	10	5	5	8	8	11	9	56	

smoothed P(t|C)

C1	$\frac{3}{24}=0.125$	0.833	0.125	0.125	0.167	0.291	0.25
C2	$\frac{5}{21}=0.238$	0.238	0	0.286	0.095	0.143	0.143
C3	$\frac{5}{20}=0.25$	0	0.2	0.1	0.25	0.2	0.15

① $P(C2) = 0.333$

② $P(\text{"t4"}|C2) = 0.285$

③ $P(C1|\text{"t1"}) = P(C1)\cdot P(\text{"t1"}|C1) = 0.33 \times 0.125 = 0.041$

④ $P(C3|\text{"t1 t4 t5"}) = P(C3)\cdot P(\text{"t1"}|C3)\cdot P(\text{"t4"}|C3)\cdot P(\text{"t5"}|C3)$

$= 0.33 \times 0.25 \times 0.1 \times 0.25$

$= 0.0020625$

⑤ $P(C1|\text{"t4 t4 t5"}) = 0.33 \times 0.125 \times 0.125 \times 0.167 = 0.000861$

$P(C2|\text{"t4 t4 t5"}) = 0.33 \times 0.285 \times 0.285 \times 0.095 = 0.002546$

$P(C3|\text{"t4 t4 t5"}) = 0.33 \times 0.1 \times 0.1 \times 0.25 = 0.000825$

Weighted TF-IDF

N=4	t_1	t_2	t_3	t_4	t_5	t_d
d_1	0	5	6	0	1	12
d_2	1	6	3	3	5	18
d_3	7	0	3	8	0	18
d_4	4	4	9	0	2	19
N_t	12	15	21	11	8	
n_t	3	3	4	2	3	

$$w_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log_2 \frac{N+1}{n_t}$$

w_{d_1}	$\frac{0}{12}\cdot\log_2(\frac{4+1}{3})$	$\frac{5}{12}\cdot\log_2(\frac{5}{3})$	$\frac{6}{12}\cdot\log_2(\frac{5}{4})$	0	$\frac{1}{12}\cdot\log_2(\frac{5}{3})$
w_{d_2}	$\frac{1}{18}\cdot\log_2(\frac{5}{3})$	$\frac{6}{18}\cdot\log_2(\frac{5}{3})$	$\frac{3}{18}\cdot\log_2(\frac{5}{4})$	$\frac{3}{18}\log_2(\frac{5}{2})$	$\frac{5}{18}\cdot\log_2(\frac{5}{3})$
w_{d_3}	$\frac{7}{18}\cdot\log_2(\frac{5}{3})$	0	$\frac{3}{18}\cdot\log_2(\frac{5}{4})$	$\frac{8}{18}\cdot\log_2(\frac{5}{2})$	0
w_{d_4}	$\frac{4}{19}\cdot\log_2(\frac{5}{3})$	$\frac{4}{19}\cdot\log_2(\frac{5}{3})$	$\frac{9}{19}\cdot\log_2(\frac{5}{4})$	0	$\frac{2}{19}\cdot\log_2(\frac{5}{3})$

w_{d_1}	0	0.307	0.161	0	0.061
w_{d_2}	0.041	0.246	0.054	0.220	0.205
w_{d_3}	0.287	0	0.054	0.587	0
w_{d_4}	0.155	0.155	0.152	0	0.078

$$\text{Cosine Similarity}_{(d,d')} = \frac{\sum_t w_{t,d} \cdot w_{t,d'}}{\sqrt{\sum_t (w_{t,d})^2} \cdot \sqrt{\sum_t (w_{t,d'})^2}} = \frac{d \cdot d'}{\|d\| \cdot \|d'\|}$$

$$\text{Sim}(d_2, d_4) = \frac{[1,6,3,3,5] \cdot [4,4,9,0,2]}{\sqrt{1^2+6^2+3^2+3^2+5^2} \cdot \sqrt{4^2+4^2+9^2+0^2+2^2}} = 0.672$$

$$\text{Sim}(w_{d_2}, w_{d_4}) = \frac{[0.041, 0.246, 0.054, 0.220, 0.205] \cdot [0.155, 0.155, 0.152, 0, 0.078]}{\sqrt{0.041^2+0.246^2+0.054^2+0.220^2+0.205^2} \cdot \sqrt{0.155^2+0.155^2+0.152^2+0+0.078^2}}$$

$$= 0.627$$

BM25

	t1	t2	t3	t4	t5	t6	Length
d_1	3	0	5	2	10	5	25
d_2	4	3	5	2	1	5	20
C	100	50	80	93	100	25	1000

$avgdl = 50$

$k_1 = 1.2$

$b = 0.75$

$$BM25_{(q,d)} = \sum_{t \in q} \frac{f_{t,d} \cdot (1+k_1)}{f_{t,d} + k_1 (1 - b + b \cdot \frac{|d|}{avgdl})} \cdot \log \frac{N}{n_t}$$

$q =$ "terme"

$BM25_{(q,d_1)} = \boxed{0}$

$$BM25_{(q,d_2)} = \frac{3 \cdot (1+1.2)}{3 + 1.2(1 - 0.75 + 0.75(\frac{20}{50})} \cdot \log_{10}\left(\frac{1000}{50}\right)$$

$$= \boxed{2.346}$$

$q =$ "terme terme terme"

$$BM25_{(q,d_1)} = 0 + 0 + \frac{10 \cdot (1+1.2) \cdot \log_{10}\left(\frac{1000}{100}\right)}{10 + 1.2(1 - 0.75 + 0.75(\frac{25}{50})} = \boxed{2.046}$$

$$BM25_{(q,d_2)} = 2.346 + 2.346 + \frac{1 \cdot (1+1.2) \cdot \log_{10}\left(\frac{1000}{100}\right)}{1 + 1.2(1 - 0.75 + 0.75(\frac{20}{50})}$$

$$= \boxed{6.017501}$$