Práctica 7

TF*IDF Cálculo Manual

Facultad De Ingeniería, Universidad De Cuenca
TEXT MINING
Freddy L. Abad L.
{freddy.abadl}@ucuenca.edu.ec

Términos

t1 = Pekin **t2** =plato **t3** =pato **t4** =conejo **t5** =receta **t6** =asado

Documentos

D1 = "Si camina como un pato y patea como un pato, debe ser un pato".
D2 = "El pato de Pekín es muy apreciado por la piel fina y crujiente. Existen versiones auténticas del plato que sirve principalmente la piel de pato".
D3 = "La ascensión de Bugs al estrellato también hizo que los animadores de Warner volvieran a mostrar al Pato Daffy como el rival del conejo, intensamente celoso y decidido a recuperar el foco de atención mientras Bugs permanecía indiferente ante los celos del pato, o lo usó para su ventaja. Esto resultó ser la receta para el éxito del dúo".
D4 = "6:25 PM 1/7/2007 entrada de blog: Encontré esta gran receta para Conejo Estofado en Vino en cookingforengineers.com."
D5 = "La semana pasada, se mostró cómo hacer pato Sechuan. Hoy vamos a hacer albóndigas chinas (Jiaozi), un plato popular que tuve la oportunidad de probar el verano pasado en Pekin. Hay muchas recetas para Jiaozi. "

Frecuencia de Términos en cada documento

Documentos	t1	t2	t3	t4	t5	t6
D1	0	0	3	0	0	0
D2	1	1	2	0	0	0
D3	0	0	2	1	1	0
D4	0	0	0	1	1	0
D5	1	1	1	0	1	0
Suma()	2	2	4	2	3	0

Documento 1

Términos	TF	TF Normalizado	IDF	W
Pekín	0	0/3 = 0	Log(5/2) = 0.39	0
plato	0	0/3=0	Log(5/2) = 0.39	0
pato	3	3/3=1	Log(5/4) = 0.09	0.09
conejo	0	0/3=0	Log(5/2) = 0.39	0

receta	0	0/3=0	Log(5/3) = 0.22	0
asado	0	0/3=0	Log(5/0) = indefinido	0

Documento 2				
Términos	**TF**	**TF Normalizado**	**IDF**	**W**
Pekín	1	1/2 = 0.5	Log(5/2) = 0.39	0.19
plato	1	1/2=0.5	Log(5/2) = 0.39	0.19
pato	2	2/2=1	Log(5/4) = 0.09	0.09
conejo	0	0/2=0	Log(5/2) = 0.39	0
receta	0	0/2=0	Log(5/3) = 0.22	0
asado	0	0/2=0	Log(5/0) = indefinido	0

Documento 3				
Términos	**TF**	**TF Normalizado**	**IDF**	**W**
Pekín	0	0/2 = 0	Log(5/2) = 0.39	0
plato	0	0/2=0	Log(5/2) = 0.39	0
pato	2	2/2=1	Log(5/4) = 0.09	0.09
conejo	1	1/2=0.5	Log(5/2) = 0.39	0.19
receta	1	1/2=0.5	Log(5/3) = 0.22	0.11
asado	0	0/2=0	Log(5/0) = indefinido	0

Documento 4				
Términos	**TF**	**TF Normalizado**	**IDF**	**W**
Pekín	0	0/1 = 0	Log(5/2) = 0.39	0
plato	0	0/1=0	Log(5/2) = 0.39	0
pato	0	0/1=0	Log(5/4) = 0.09	0
conejo	1	1/1=1	Log(5/2) = 0.39	0.39
receta	1	1/1=1	Log(5/3) = 0.22	0.22
asado	0	0/1=0	Log(5/0) = indefinido	0

Documento 5				
Términos	**TF**	**TF Normalizado**	**IDF**	**W**
Pekín	1	1/1 = 1	Log(5/2) = 0.39	0.39
plato	1	1/1=1	Log(5/2) = 0.39	0.39
pato	1	1/1=1	Log(5/4) = 0.09	0.09
conejo	0	0/1=0	Log(5/2) = 0.39	0
receta	1	1/1=1	Log(5/3) = 0.22	0.22
asado	0	0/1=0	Log(5/0) = indefinido	0

Vector de la consulta						
Docs	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**
Consulta	1	0	1	0	1	0

Consulta				
Términos	**TF**	**TF Normalizado**	**IDF**	**W**
Pekín	1	1/1 = 1	Log(5/2) = 0.39	0.39
plato	0	0/1= 0	Log(5/2) = 0.39	0
pato	1	1/1=1	Log(5/4) = 0.09	0.09
conejo	0	0/1=0	Log(5/2) = 0.39	0
receta	1	1/1=1	Log(5/3) = 0.22	0.22
asado	0	0/1=0	Log(5/0) = indefinido	0

Comparación							
Consulta	0.39	0	0.09	0	0.22	0	-------
Doc1	0	0	0.09	0	0	0	--------
doc1*q	0	0	81	0	0	0	**81**
||doc1||	0	0	81	0	0	0	**0.0081 sqr = 0.09**
||q||	1.521	0	81	0	48	0	**0.2082 sqr = 0.456**

Cos(doc1,q) = (doc1*q) / ||doc1||*||q||
Cos(doc1,q) = 0.0081 / (0.09*0.456)
Cos(doc1,q) = 0.1973

Consulta	0.39	0	0.09	0	0.22	0	-------
Doc2	0.19	0.19	0.09	0	0	0	--------
doc2*q	74	0	81	0	0	0	**821**
||doc2||	36	36	81	0	0	0	**0.0802** sqr = **0.283**
||q||	1.521	0	81	0	48	0	**0.2082** sqr = **0.456**

Cos(doc2,q) = (doc2*q) / ||doc2||*||q||
Cos(doc2,q) = 0.0821 / (0.283*0.456)
Cos(doc2,q) = 0.0821/ 0.129 = 0.63

Consulta	0.39	0	0.09	0	0.22	0	-------
Doc3	0	0	0.09	0.19	0.11	0	--------
doc3*q	0	0	81	0	24	0	**32**
||doc3||	0	0	81	36	12	0	**0.056** sqr = **0.236**
||q||	1.521	0	81	0	48	0	**0.2082** sqr = **0.4562**

Cos(doc3,q) = (doc3*q) / ||doc3||*||q||
Cos(doc3,q) = 0.0032 / (0.236*0.456)
Cos(doc3,q) = 0.0032/ 0.129 = 0.30

Consulta	0,39	0	0,09	0	0,22	0	-------
Doc4	0	0	0	0,39	0,22	0	--------
doc4*q	0	0	0	0	0,0484	0	**32**
||doc4||	0	0	32	36	12	0	**0.056** sqr = **0.236**
||q||	1.521	0	81	0	48	0	**0.2082** sqr = **0.4562**
Cos(doc4,q) = (doc4*q) / ||doc4||*||q||							
Cos(doc4,q) = 0.0032 / (0.236*0.456)							
Cos(doc4,q) = 0.0032/ 0.129 = 0.23							

Consulta	0,39	0	0,09	0	0,22	0	-------
Doc5	0	0	0	0,39	0,22	0	--------
doc5*q	0	0	0	0	0,0484	0	**32**
||doc5||	0	0	32	36	12	0	**0.056** sqr = **0.236**
||q||	1.521	0	81	0	48	0	**0.2082** sqr = **0.4562**
Cos(doc5,q) = (doc5*q) / ||doc5||*||q||							
Cos(doc5,q) = 0.0032 / (0.236*0.456)							
Cos(doc5,q) = 0.0032/ 0.129 = 0.76							

Dado esta comparación, se llega a la conclusion que el Documento 5 es el que mayor similitud con la consulta tiene.