

Trabajo 8

Similaridad de Documentos - RapidMiner

Facultad De Ingeniería, Universidad De Cuenca

TEXT MINING

Freddy L. Abad L.

freddy.abadl@ucuenca.edu.ec

El objetivo de esta práctica es determinar la similaridad de varios documentos usando una medida de similitud

Datos: A modo de ejemplo use los siguientes documentos

Documento 1: noticias1.txt

A civil society organization on Tuesday said there was an urgent need to improve government schools across the country so that a maximum number of children can get quality education instead of joining the army of unskilled labor force as they grow up.

Documento 2: noticias2.txt

Sometimes you just don't want to think. You only want mindless entertainment where nothing is logical and it doesn't matter.

Documento 3: noticias3.txt

Google announced its Street View trekker has plunged under the waters of the Galapagos and tracked across its islands, bringing to the internet 360-degree images of the isolated landscape and world's largest living tortoises that inspired Darwin to develop his theory of evolution.

Documento4: noticias4.txt

Pakistan skipper Misbah-ul-Haq believes his team's prospects in the Champions Trophy will depend on his players' ability to adapt to English conditions.

Documento 5: noticias5.txt (texto repetido del Documento4-noticias4.txt)

Pakistan skipper Misbah-ul-Haq believes his team's prospects in the Champions Trophy will depend on his players' ability to adapt to English conditions.

Con este objetivo se usará los siguientes operadores.

Lectura de Archivos

Read Document



Read Document (4)



Read Document (3)



Read Document (2)



Read Document (5)

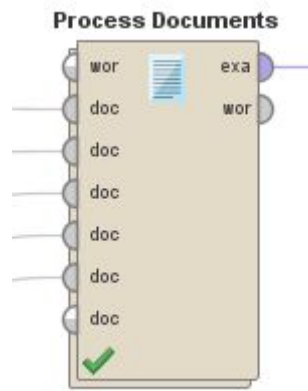


Process Document from Data: Use este operador para leer información de varios documentos.

Entonces configure este operador para determinar cada etiqueta de clase y mencionar el directorio

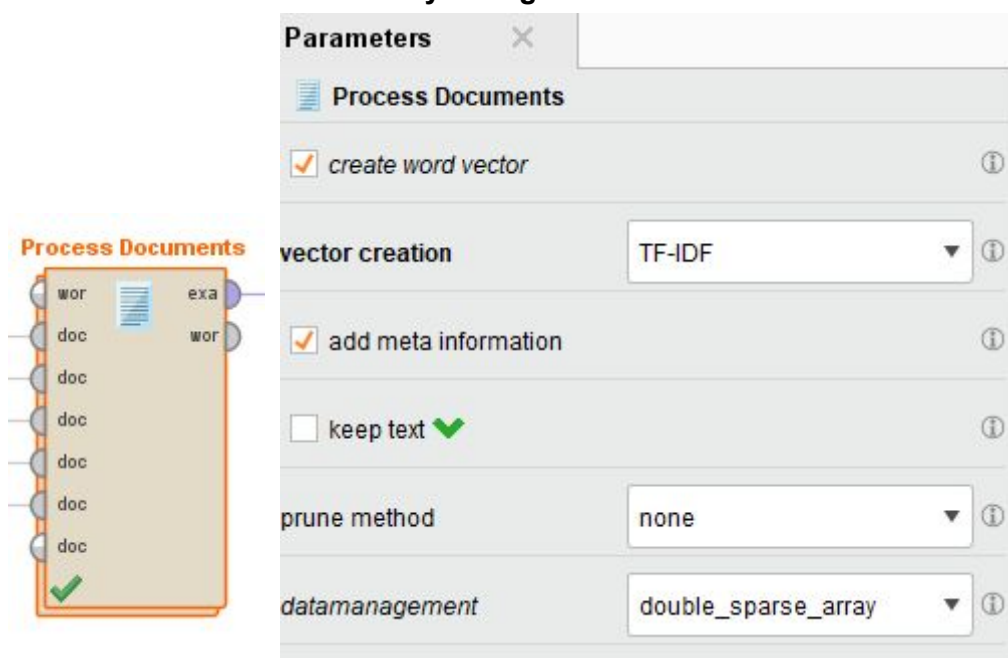
de origen. Dentro de este operador crear un subprocesso que permita ejecutar las siguientes acciones

1. Dividir las frases en tokens
2. Filtras los tokens por tamaño
3. Transformar a mayúsculas
4. Eliminar Palabras Vacías
5. Reducir palabras similares a sus bases (stem).

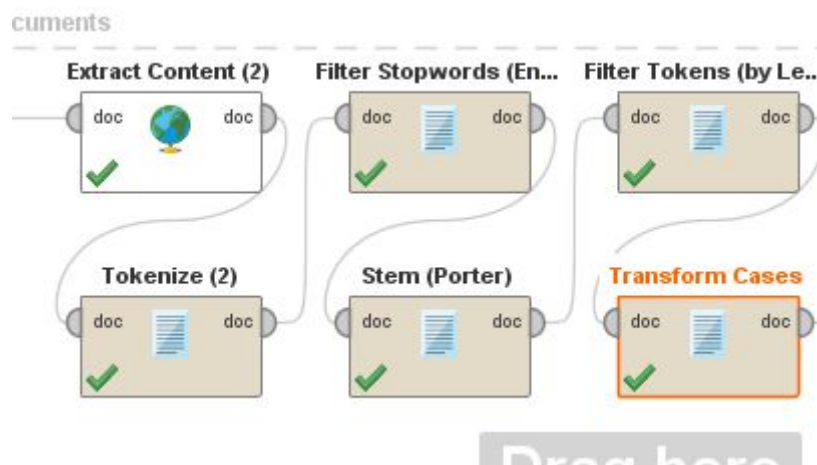


En este proceso, configure el método de creación de vectores a ser utilizado. El operador a ser usado es TF-IDF, el cual permite asignar pesos a cada una de las palabras que conforman el documento.



Proceso Process Documents y configuracion de la creación de vectores



Procesos dentro del proceso Process Documents

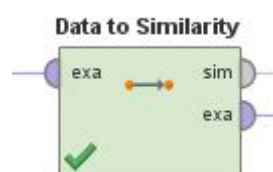


Resultado


ExampleSet (Transpose) X						
Open in  Turbo Prep  Auto Model						
Row No.	id	att_1	att_2	att_3	att_4	att_5
1	ABIL	0.0	0.0	0.0	0.277350098...	0.277350098...
2	ADAPT	0.0	0.0	0.0	0.277350098...	0.277350098...
3	ANNOUNC	0.0	0.0	0.208514414...	0.0	0.0
4	ARMI	0.213200716...	0.0	0.0	0.0	0.0
5	BELIEV	0.0	0.0	0.0	0.277350098...	0.277350098...
6	BRING	0.0	0.0	0.208514414...	0.0	0.0
7	CHAMPION	0.0	0.0	0.0	0.277350098...	0.277350098...
8	CHILDREN	0.213200716...	0.0	0.0	0.0	0.0
9	CIVIL	0.213200716...	0.0	0.0	0.0	0.0
10	CONDIT	0.0	0.0	0.0	0.277350098...	0.277350098...
11	COUNTRI	0.213200716...	0.0	0.0	0.0	0.0
12	DARWIN	0.0	0.0	0.208514414...	0.0	0.0
13	DEGRE	0.0	0.0	0.208514414...	0.0	0.0
14	DEPEND	0.0	0.0	0.0	0.277350098...	0.277350098...
15	DEVELOP	0.0	0.0	0.208514414...	0.0	0.0


Tareas:



Identifique el operador necesario para calcular la similaridad entre documentos usando una función de distancia. Pruebe de usar la distancia euclidiana numérica y la medida del coseno.






Medida de distancia de Coseno

 **Data to Similarity**

measure types 

NumericalMeasures



numerical measure 

CosineSimilarity



Resultado

First	Second	Similarity
1.0	2.0	0
1.0	3.0	0
1.0	4.0	0
1.0	5.0	0
2.0	3.0	0
2.0	4.0	0
2.0	5.0	0
3.0	4.0	0
3.0	5.0	0
4.0	5.0	1

Medida de distancia de Euclidiana

Data to Similarity

measure types

NumericalMeasures

numerical measure

EuclideanDistance

Resultado

First	Second	Distance
1.0	2.0	1.414
1.0	3.0	1.414
1.0	4.0	1.414
1.0	5.0	1.414
2.0	3.0	1.414
2.0	4.0	1.414
2.0	5.0	1.414
3.0	4.0	1.414
3.0	5.0	1.414
4.0	5.0	0

Medida de distancia de Euclidiana

Data to Similarity

measure types

NumericalMeasures

numerical measure

ManhattanDistance

Resultado

ExampleSet (Process Documents)		SimilarityMeasureObject (Data to Similarity)
First	Second	Distance
1.0	2.0	7.024
1.0	3.0	9.486
1.0	4.0	8.296
1.0	5.0	8.296
2.0	3.0	7.129
2.0	4.0	5.939
2.0	5.0	5.939
3.0	4.0	8.401
3.0	5.0	8.401
4.0	5.0	0

En cada caso identifique los valores que permiten determinar cuándo dos documentos son similares

Dados estos resultados, se puede identificar que los documentos 4 y 5 son los mismos, ya que la distancia euclidiana nos da 0, la distancia de coseno nos da 1 ($\cos 0 = 1$) y así mismo la distancia manhattan nos da 0.