

Trabajo 6

Análisis de Corpus mediante Integración de SQL y Procesamiento de Texto

Facultad De Ingeniería, Universidad De Cuenca

TEXT MINING

Freddy L. Abad L.

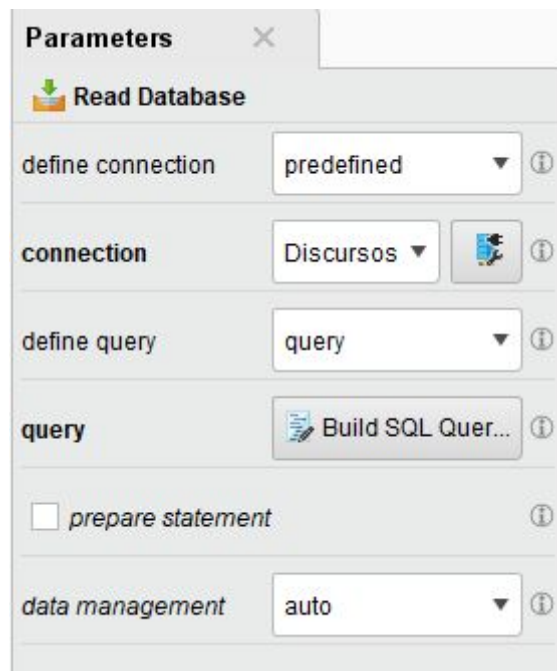
ffreddy.abadl@ucuenca.edu.ec

PASO A

Consultar y recuperar información relevante de la tabla tbldiscursos usando el operador Read Database



Configuración del proceso **Read Database**



Configuración del proceso **Read Database**, verificación de conexión con base de datos

Manage Database Connections

Within this dialog, you can create, edit and delete connections to databases.

Available Connections

- Discursos (PostgreSQL; localhost:5432)

Connection Details

Name: Discursos Advanced...

Database system: PostgreSQL

Host: localhost Port: 5432

Database scheme: discursosinaugurales

User: postgres

Password: *****

URL: jdbc:postgresql://localhost:5432/discursosinaugurales

☒ Connection ok Test

Save New Clone Delete OK Cancel

Ejecución de una sentencia SQL en el proceso Read Database

Sentencia :

```
SELECT "discurso"
FROM "public"."tblDiscursos"
WHERE "public"."tblDiscursos"."fecha" like '%2009%'
AND "public"."tblDiscursos"."presidente" like '%Obama%'
```

Build SQL Query: query

Build SQL Query: query
An SQL query.

Tables: public.tblDiscursos

Attributes:

Where Clause:

SQL Query

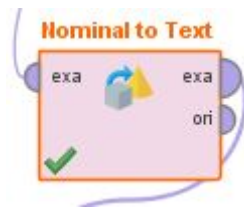
```
1 SELECT "discurso"
2 FROM "public"."tblDiscursos"
3 WHERE "public"."tblDiscursos"."fecha" like '%2009%'
4 AND "public"."tblDiscursos"."presidente" like '%Obama%'
```

Resultado

Row No.	discurso
1	<p>My fellow citizens, I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne b...

Paso B:

Procesamiento de texto de discurso de Nominal a Texto



Configuración del proceso **Nominal to Text**.

The screenshot shows the "Parameters" window for the "Nominal to Text" process. The window has a title bar with a close button. Below the title bar, there is a section for "Nominal to Text" with a green checkmark. The parameters are as follows:

- attribute filter type**: A dropdown menu set to "all".
- invert selection**: A checkbox that is unchecked.
- include special attributes**: A checkbox that is unchecked.
- 75% of users kept 'no':**: A progress bar showing 75% completion. The "yes" bar is grey and 25% full. The "no" bar is green and 75% full.

Extraccion y transformacion en tokens mediante el proceso **PROCESS DOCUMENT FROM DATA**

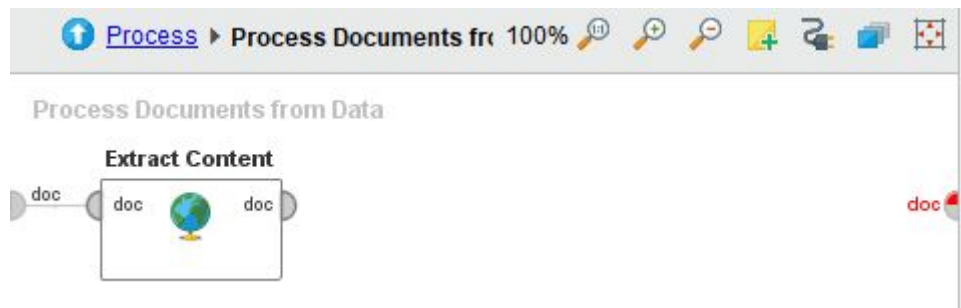


Configuración del proceso **PROCESS DOCUMENT FROM DATA**

The screenshot shows the "Parameters" window for the "Process Documents from Data" process. The window has a title bar with a close button. Below the title bar, there is a section for "Process Documents from Data" with a green checkmark. The parameters are as follows:

- create word vector**: A checkbox that is checked.
- vector creation**: A dropdown menu set to "Term Occurrenc...".
- add meta information**: A checkbox that is checked.
- keep text**: A checkbox that is unchecked.
- prune method**: A dropdown menu set to "none".
- datamanagement**: A dropdown menu set to "double_sparse...".
- select attributes and weights**: A checkbox that is unchecked.

En el proceso **PROCESS DOCUMENT FROM DATA**, debe ingresarse subprocessos tal como el proceso



Configuración del subprocesso **Extract Content**

The screenshot shows a 'Parameters' window for the 'Extract Content' subprocess. It contains several settings:

- ☒ extract content
- minimum text block length: 5 (with a green checkmark icon)
- ☒ override content type information
- ☒ neglect span tags
- ☒ neglect p tags
- ☒ neglect b tags
- ☒ neglect i tags
- ☒ neglect br tags
- ☒ ignore non html tags

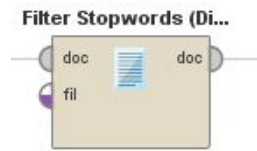
Proceso de dividir el discurso en palabras individuales (también llamadas tokens) usando el operador **Tokenize**



Proceso de eliminar palabras vacías del discurso en palabras individuales (también llamadas tokens) usando el operador **Filter Stopwords (English)**




Proceso de eliminar palabras del diccionario personalizado del discurso en palabras individuales (también llamadas tokens) usando el operador **Filter Stopwords (Dictionary)** tomando en cuenta las palabras a eliminar tomadas del archivo *discursostop.txt*





Configuración del subprocesso **Filter Stopwords (Dictionary)** teniendo en cuenta el archivo del diccionario personalizado

Parameters X

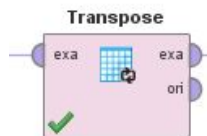
Filter Stopwords (Dictionary)

file  

☐ case sensitive 


encoding 

Proceso de transponer las filas en columnas



Resultado, se puede notar que se presenta el resultado de estos procesos en forma de columna, esto se debe al proceso **Transpose**, que únicamente presenta en formato de columna en lugar de fila única.

ExampleSet (Transpose) X

Open in  Turbo Prep  Auto Model

Row No.	id	att_1
1	Title	?
2	Language	?
3	Description	?
4	Keywords	?
5	Robots	?
6	America	5.0
7	Americans	2.0
8	Bush	1.0
9	Concord	1.0
10	Earth	2.0
11	Forty	1.0
12	Gettysburg	1.0
13	God	1.0
14	Homes	1.0
15	I	3.0

Proceso que permite guardar los resultados en un archivo de formato **CSV**.



Parameters [X]

Write CSV

csv file

column separator

☒ write attribute names

☒ quote nominal values

☒ format date attributes

☐ append to file

encoding

Resultado

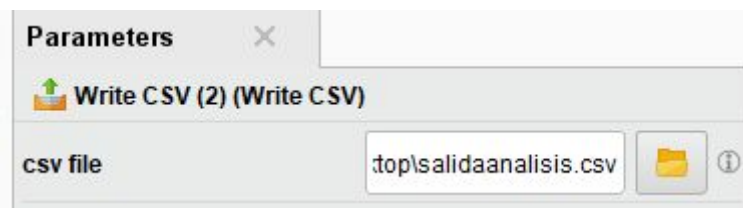
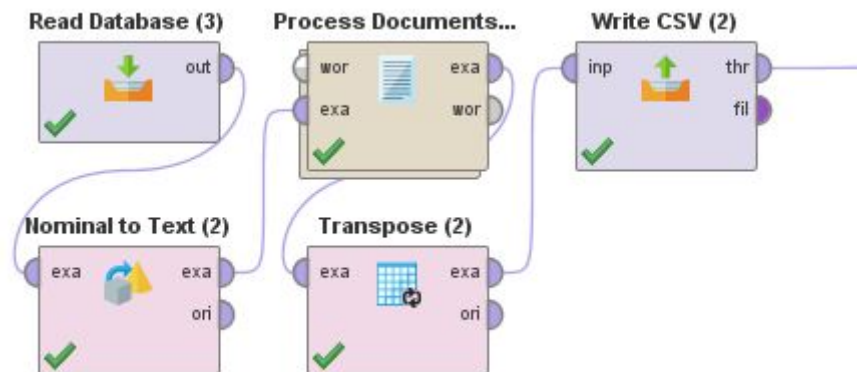
ExampleSet (Transpose) [X]

Open in Turbo Prep Auto Model

Row No.	id	att_1
1	Title	?
2	Language	?
3	Description	?
4	Keywords	?
5	Robots	?
6	America	5.0
7	Americans	2.0
8	Bush	1.0
9	Concord	1.0
10	Earth	2.0
11	Forty	1.0
12	Gettysburg	1.0
13	God	1.0
14	Homes	1.0
15	I	3.0

PROCESO 2

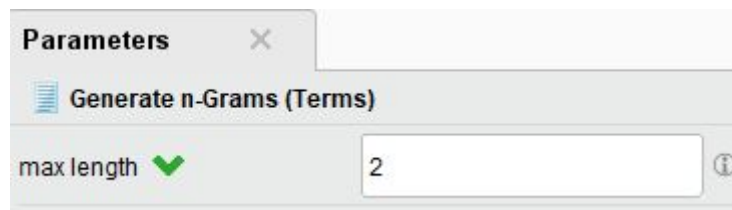
El requerimiento del proceso 2, es similar al del proceso 1, con las peculiaridades de tener un texto más filtrado, esto se logra añadiendo subproceso al **Process Document**



El proceso **Generate n-Grams (Terms)** permite filtrar las palabras mayores a un n dado.



Configuración del proceso **Generate n-Grams (Terms)**



El proceso **Filter Tokens by Content** filtra las palabras por expresiones regulares



Configuración del proceso **Filter Tokens by Content**

Parameters

Filter Tokens (by Content)

condition

matches

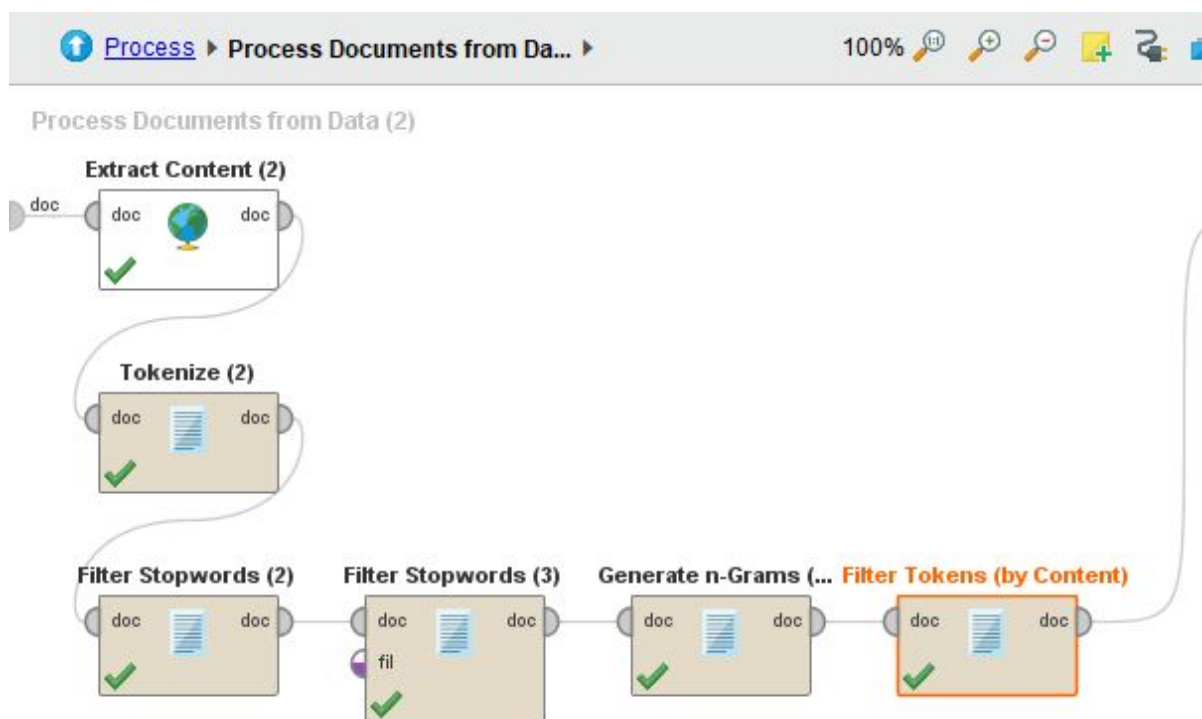
regular expression

.*_.*

☒ case sensitive

☐ invert condition

Vista general de los subprocessos y sus relaciones



Resultado del proceso

ExampleSet (Transpose (2))		
<div>Open in</div> <div>Turbo Prep</div> <div>Auto Model</div>		
Row No.	id	att_1
1	Title	?
2	Language	?
3	Description	?
4	Keywords	?
5	Robots	?
6	America_big...	1.0
7	America_carr...	1.0
8	America_ever...	1.0
9	America_met	1.0
10	America_s	1.0
11	Americans_c...	1.0
12	Americans_t...	1.0
13	Bush_service	1.0
14	Concord_Get...	1.0
15	Earth_fought	1.0

TRABAJO FINAL

Variar los resultados obtenidos aplicando diferentes operadores dentro del operador **Process Documents From Data**, incluyendo por ejemplo el operador de filtrado por longitud (**Filter Tokens (by length)**), verificando partes del habla tales como nombres o verbos (operador **Filter Tokens (by POS Tags)**), o palabras derivadas mediante el operador (Stem (Porter)). Sin embargo, el punto clave no es mostrar cuán sofisticado es su modelo de minería de textos o cuántos operadores han incluido; es únicamente identificar los beneficios de la inclusión de estos operadores para su análisis general.



Parameters X

Filter Tokens (by Length)

min chars ✓ 4 ⓘ

max chars 25 ⓘ

Expresión Regular para verbos:

JJ.*|NN.*



Configuración del proceso **Filter Tokens by POS Tags**

Parameters X

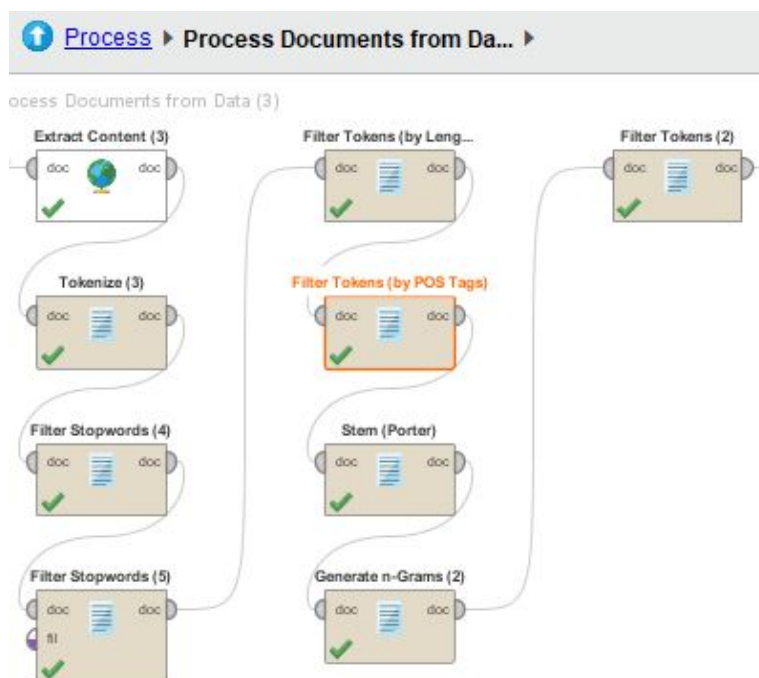
Filter Tokens (by POS Tags)

language ✓ English ⓘ

expression JJ.*|NN.* ⓘ



Vista general de los subprocessos y sus relaciones



Resultado del filtrado.

ExampleSet (Transpose (3))		
Open in  Turbo Prep  Auto Model		
Row No.	id	att_1
1	Title	?
2	Language	?
3	Description	?
4	Keywords	?
5	Robots	?
6	action_bold	1.0
7	adversari_thr...	1.0
8	ambit_greater	1.0
9	ambit_suggest	1.0
10	america_big...	1.0
11	america_carri	1.0
12	america_decl...	1.0
13	america_ever...	1.0
14	america_gat...	1.0
15	american_crisi	1.0