Resumen extractivo de documentos. Un análisis comparativo de técnicas de puntuación.

Autora

Julieta Pilar Corvi

Directora

Dra. Laura Lanzarini

Asesor profesional

Dr. Augusto Villa Monte

UNIVERSIDAD NACIONAL DE LA PLATA

Agenda

1. Motivación y objetivo
2. Preparación de los datos
3. Modelo
4. Resultados

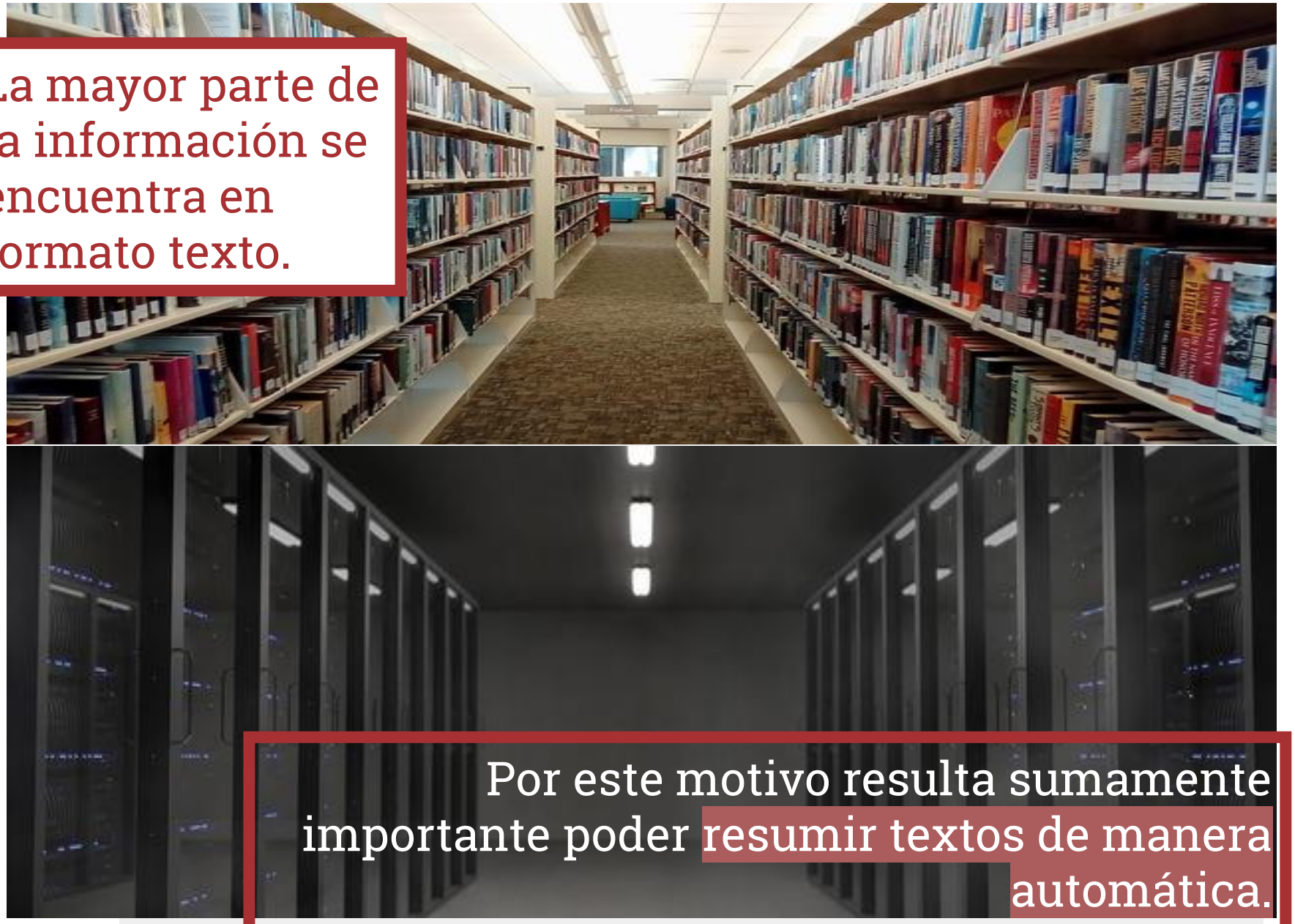
1. Motivación y objetivo

“Desde el inicio de la civilización hasta 2003 se generaron 5 exabytes de datos. La misma cantidad se genera cada 2 dias en la actualidad.”

Eric Schmidt, CEO de Google, Agosto 2010.

“El 90% de todos los datos existentes fueron creados en los últimos dos años y continúa creciendo. Para 2020 se espera que haya el doble.”

Michael Fork, Director IBM Cloud Platform, Agosto 2017.



La mayor parte de la información se encuentra en formato texto.

Por este motivo resulta sumamente importante poder resumir textos de manera automática.

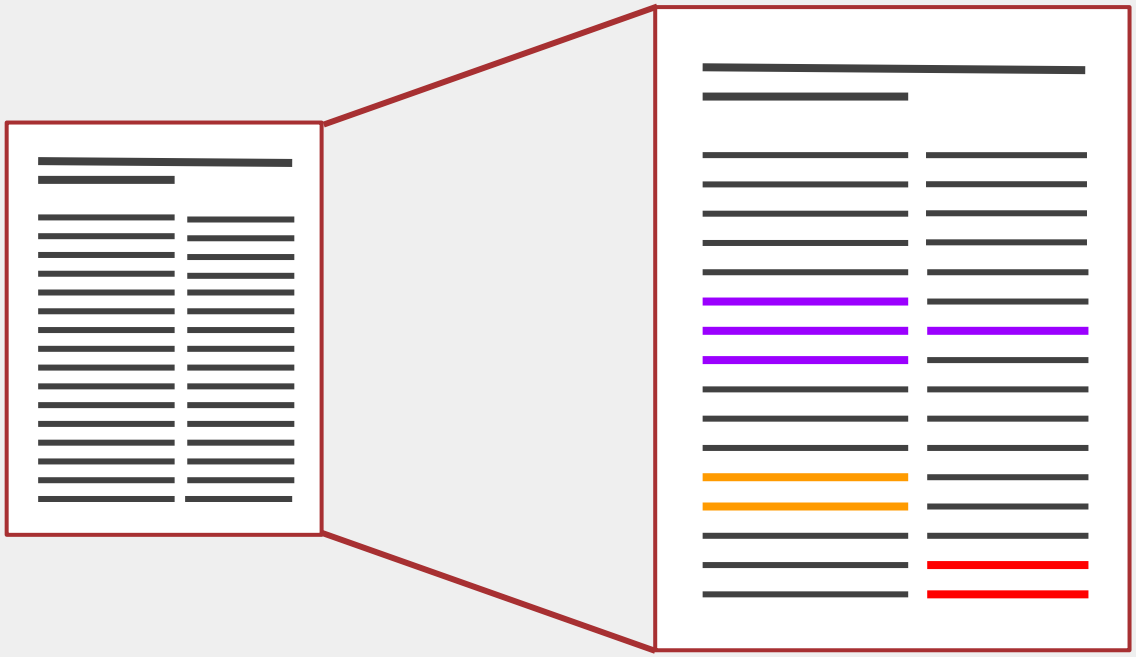
Resumen de un documento:

Punto intermedio entre el título y el cuerpo del documento. Descripción concisa del texto.

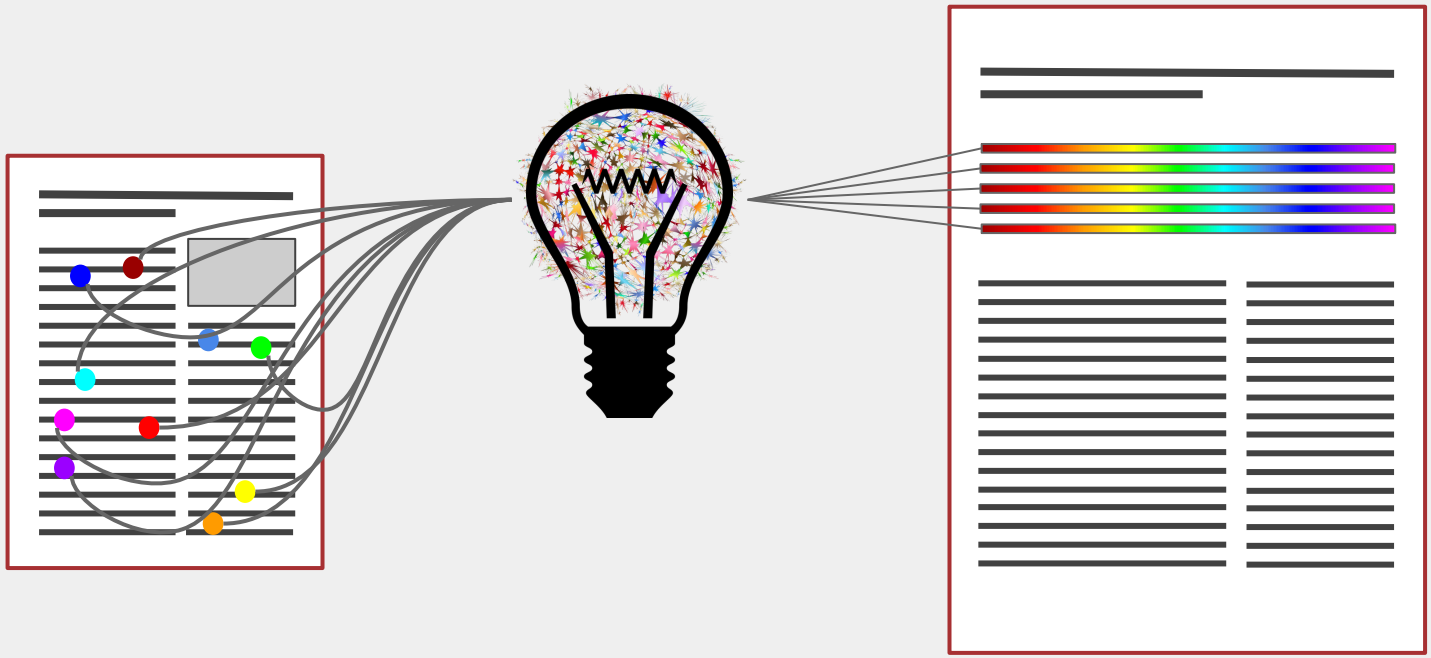
Resumen automático de un documento:

Proceso que, mediante una aplicación informática, reduce el documento obteniendo sus partes más importantes.

Tipos de resúmenes
Enfoque extractivo



Tipos de resúmenes
Enfoque abstractivo



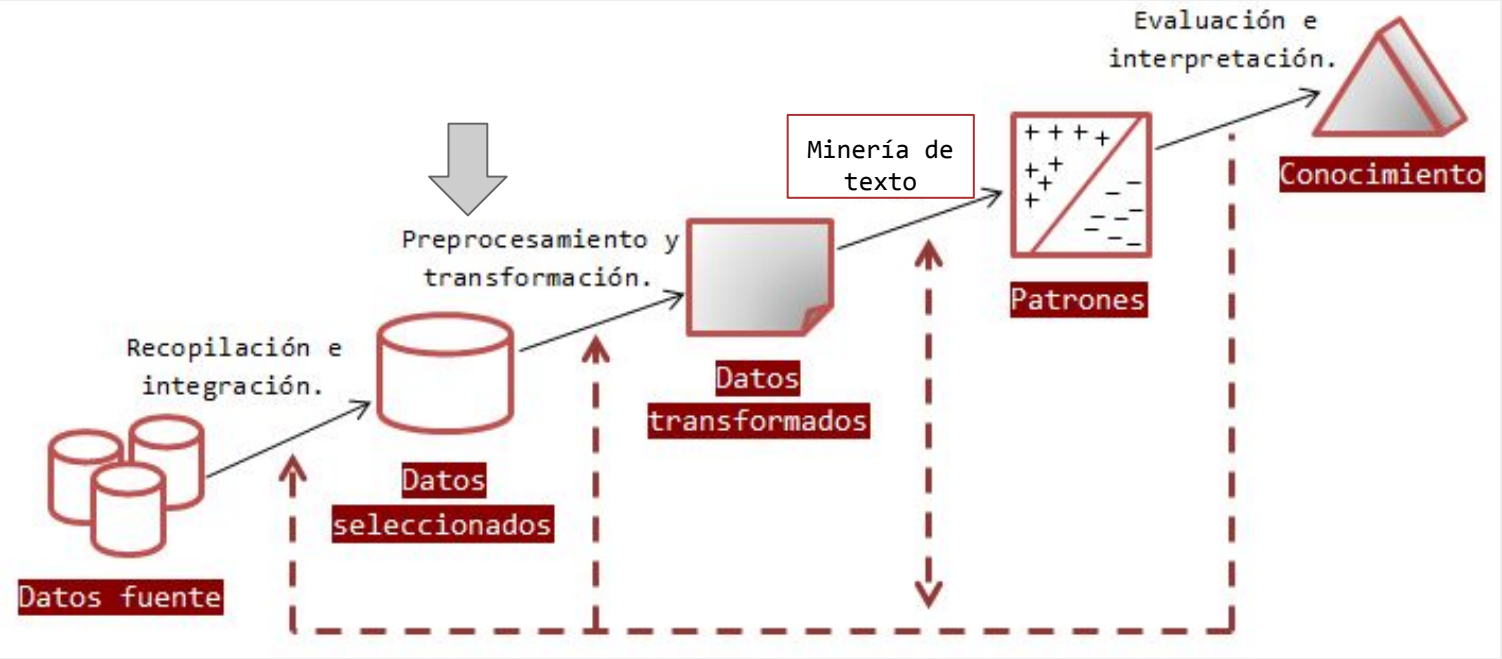
Tipos de resúmenes
Extractivo vs. abstractivo

Objetivo

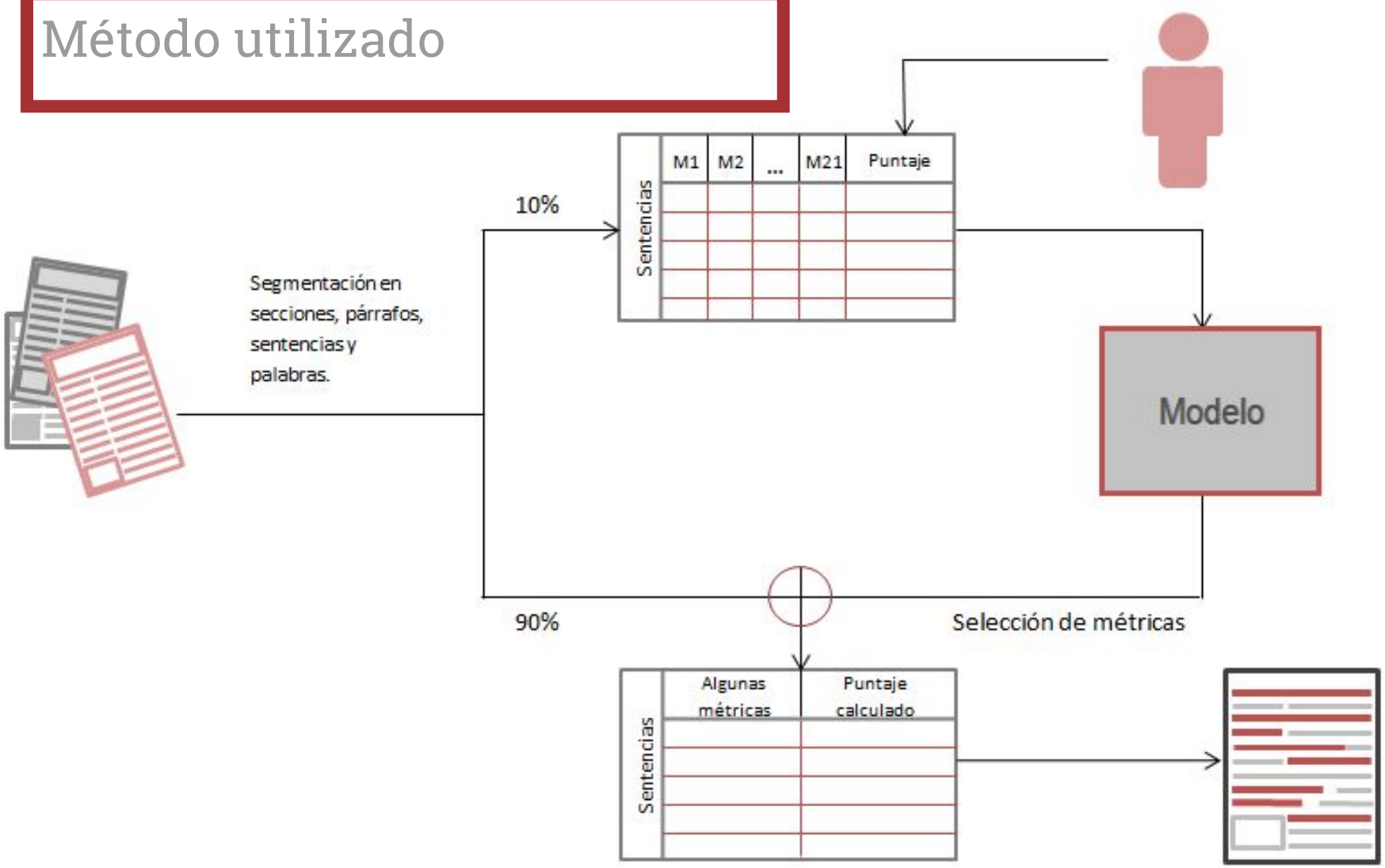
El objetivo de esta tesina es proponer una técnica de resumen automático extractivo basada en la opinión del usuario para seleccionar las partes importantes de un documento.

Proceso KDD

"KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos." (Fayyad, 1996)

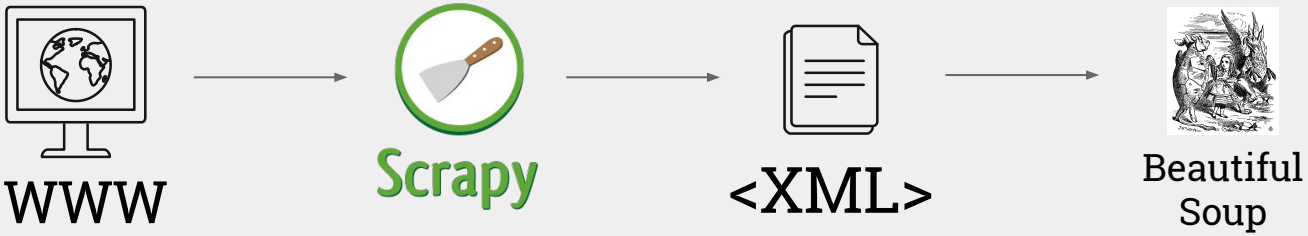


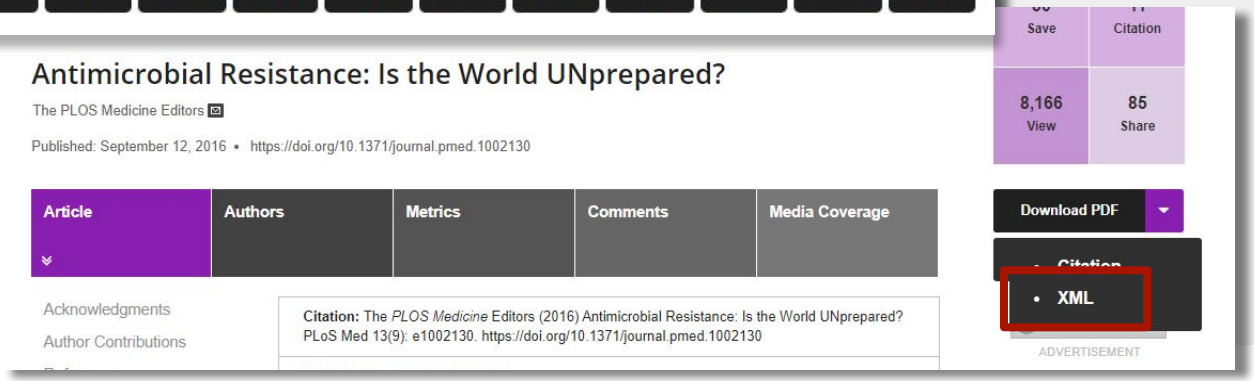
Método utilizado



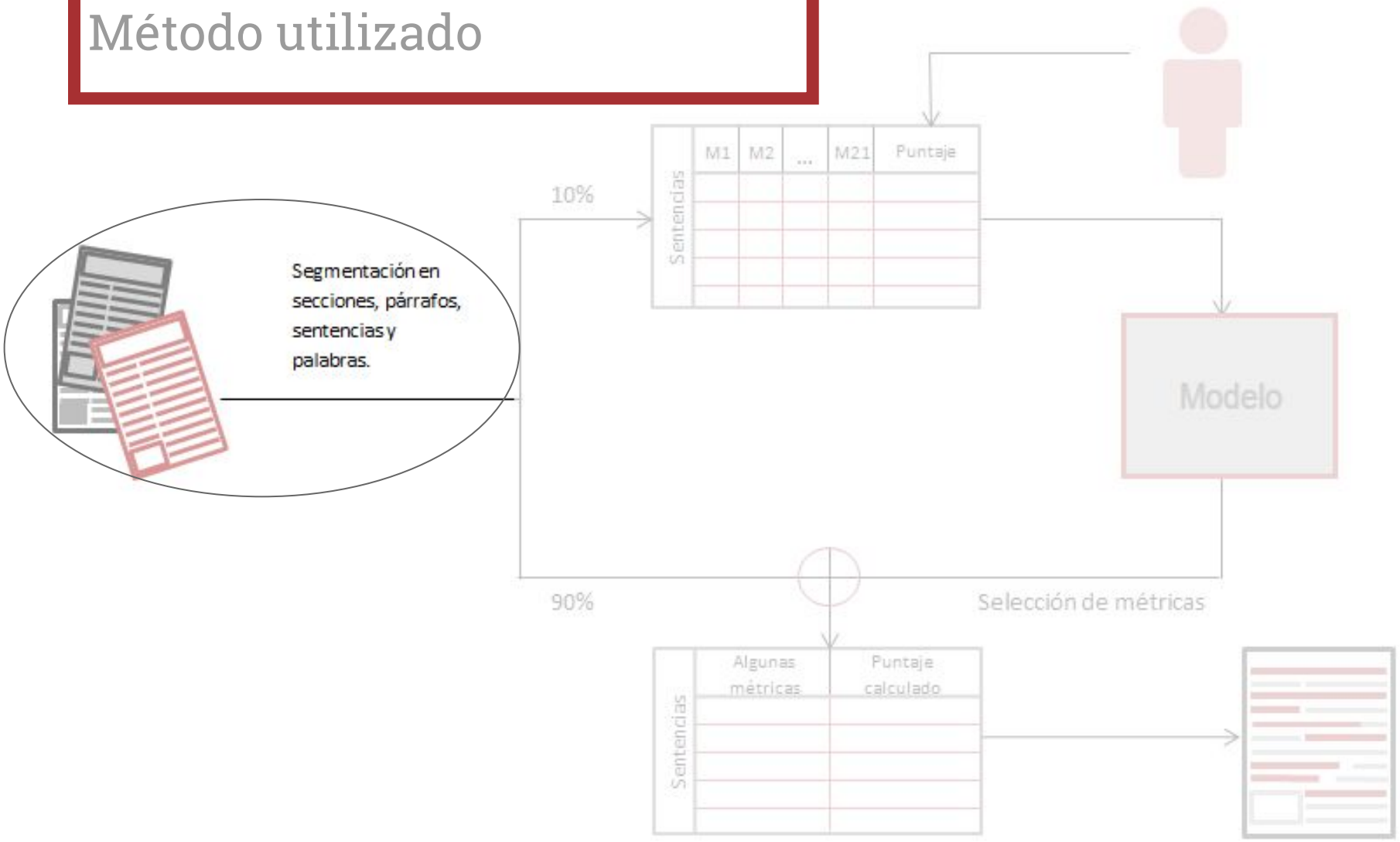
2. Preparación de datos

Recopilación e integración





Método utilizado



Preprocesamiento

La Minería de Texto necesita de un tratamiento particular de los datos previo a tener una *vista minable* adecuada.

En la Minería de Texto se busca extraer conocimiento y analizar a partir de texto.

Eliminación de ruido

```
<sec id="sec001" sec-type="kdd-proceso">
    <title>KDD: Knowledge Discovery in Database</title>
    <p>El KDD es el proceso de descubrir conocimiento a partir de información almacenada. Este proceso comienza con la recolección de la información, ya que no siempre se encuentra ubicada en un mismo punto y finaliza con el conocimiento extraído a partir de ella. Se trata de un proceso en fases que requiere de revisiones continuas, especialmente relacionada con los resultados intermedios obtenidos y la mirada puesta en el tipo de respuesta esperada.</p>
    <p>Según (Fayyad, Piatetsky-Shapiro y Smyth, 1996), "KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos". El resultado de este proceso es lo que le da significado al almacenamiento de la información y ayudará a poder tomar decisiones en base a esa información.</p>
</sec>
```

Eliminación de ruido

KDD: Knowledge Discovery in Database

El KDD es el proceso de descubrir conocimiento a partir de información almacenada. Este proceso comienza con la recolección de la información, ya que no siempre se encuentra ubicada en un mismo punto y finaliza con el conocimiento extraído a partir de ella. Se trata de un proceso en fases que requiere de revisiones continuas, especialmente relacionada con los resultados intermedios obtenidos y la mirada puesta en el tipo de respuesta esperada.

Según (Fayyad, Piatetsky-Shapiro y Smyth, 1996), "KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos". El resultado de este proceso es lo que le da significado al almacenamiento de la información y ayudará a poder tomar decisiones en base a esa información.

Segmentación y tokenización

El KDD es el proceso de descubrir conocimiento a partir de información almacenada. Este proceso comienza con la recolección de la información, ya que no siempre se encuentra ubicada en un mismo punto y finaliza con el conocimiento extraído a partir de ella. Se trata de un proceso en fases que requiere de revisiones continuas, especialmente relacionada con los resultados intermedios obtenidos y la mirada puesta en el tipo de respuesta esperada.

Según (Fayyad, Piatetsky-Shapiro y Smyth, 1996), "KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos". El resultado de este proceso es lo que le da significado al almacenamiento de la información y ayudará a poder tomar decisiones en base a esa información.

Filtrado palabras vacías

[Según] [Fayyad] [Piatetsky-Shapiro] [y] [Smyth] [1996] [KDD] [es] [el] [proceso] [no] [trivial] [de] [identificar] [patrones] [válidos] [novedosos] [potencialmente] [útiles] [y] [en] [última] [instancia] [comprensibles] [a] [partir] [de] [los] [datos] [El] [resultado] [de] [este] [proceso] [es] [lo] [que] [le] [da] [significado] [al] [almacenamiento] [de] [la] [información] [y] [ayudará] [a] [poder] [tomar] [decisiones] [en] [base] [a] [esa] [información]

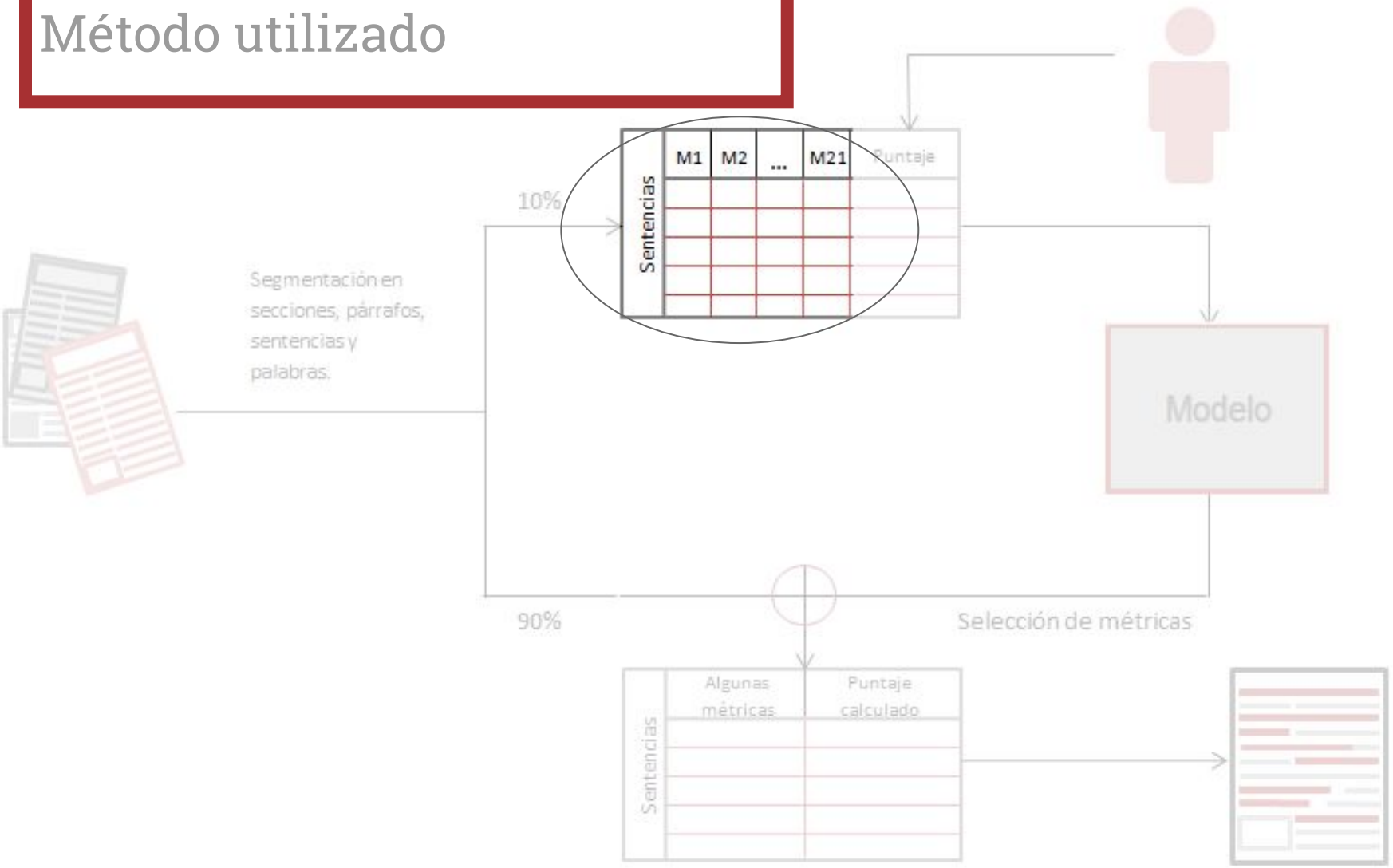
Normalización: Stemming

"proceso" "procesos"
"procesamiento" "procesador"

"casa" "caso" "casar"

"investigar" "investigación"
"investigadora" "investigaron"

Método utilizado



Métricas

Ponderar la importancia de las unidades textuales

Métricas de posición

2.4. Métricas

Como se mencionó anteriormente para poder realizar un resumen automático, es necesario ponderar numéricamente la importancia de las unidades textuales seleccionadas respecto del documento. Es necesario aclarar en este punto que en el presente trabajo se seleccionaron las oraciones de los documentos de texto como unidades textuales para formar parte del resumen.

En esta sección se desarrollarán veintiún técnicas de medición que van a servir como medida para almacenar en la base de datos. A partir de allí se podrán utilizar herramientas de la minería de datos para buscar patrones dentro de estos resultados.

1 POS_F

2 POS_L

3 POS_B

Métricas de posición

2.4. Métricas

Como se mencionó anteriormente para poder realizar un resumen automático, es necesario ponderar numéricamente la importancia de las unidades textuales seleccionadas respecto del documento. Es necesario aclarar en este punto que en el presente trabajo se seleccionaron las oraciones de los documentos de texto como unidades textuales para formar parte del resumen.

En esta sección se desarrollarán veintiún técnicas de medición que van a servir como medida para almacenar en la base de datos. A partir de allí se podrán utilizar herramientas de la minería de datos para buscar patrones dentro de estos resultados.

1 POS_F

2 POS_L

3 POS_B

Métricas de posición

2.4. Métricas

Como se mencionó anteriormente para poder realizar un resumen automático, es necesario ponderar numéricamente la importancia de las unidades textuales seleccionadas respecto del documento. Es necesario aclarar en este punto que en el presente trabajo se seleccionaron las oraciones de los documentos de texto como unidades textuales para formar parte del resumen.

En esta sección se desarrollarán veintiún técnicas de medición que van a servir como medida para almacenar en la base de datos. A partir de allí se podrán utilizar herramientas de la minería de datos para buscar patrones dentro de estos resultados.

1 POS_F

2 POS_L

3 POS_B

Métricas de longitud

15 palabras - 139 caracteres

2.4. Métricas

Como se mencionó anteriormente para poder realizar un resumen automático, es necesario ponderar numéricamente la importancia de las unidades textuales seleccionadas respecto del documento. Es necesario aclarar en este punto que en el presente trabajo se seleccionaron las oraciones de los documentos de texto como unidades textuales para formar parte del resumen.

En esta sección se desarrollarán veintiún técnicas de medición que van a servir como medida para almacenar en la base de datos. A partir de allí se podrán utilizar herramientas de la minería de datos para buscar patrones dentro de estos resultados.

1 LEN_W

2 LEN_CH

Métricas de título

Métricas de título (TITLE)

La idea principal de aplicar el método que mide en base a los títulos, es que el autor genera el titulo como el texto representativo del sujeto del documento. Esto vale también para título de secciones y párrafos. El método de Edmundson (Edmundson, 1969) asigna un valor positivo a una sentencia basándose en la cantidad de palabras en común que tiene con el título del documento, sección o párrafo, según se prefiera.

Las siguientes métricas están basadas en las medidas de similitud mencionadas en la sección anterior 2.4.4. Se considera a la similitud entre palabras como palabras iguales.

Métricas de frecuencia

1 TF

Calcula el promedio de las frecuencias de los términos de una oración.

Problema: La palabra “auto” dentro de un documento de la industria automotriz

2 TF-ISF

Métricas de frecuencia

1 TF

2 TF-ISF

Se basa en que si un término aparece varias veces en una oración, pero no con tanta frecuencia en el resto del documento, la oración es representativa en el documento.

Métricas de frecuencia

1 TF

2 TF-ISF

Muchas más...

Palabras clave

Cobertura

Grafos

Representación

	POS_F	POS_L	POS_B	LEN_W	LEN_CH	TF	...	LUHN
S_1	1	158	1	24	167	10,875		13
S_2	2	157	0,5	27	187	9,6296		12
...								
S_M	158	1	1	8	49	5		4

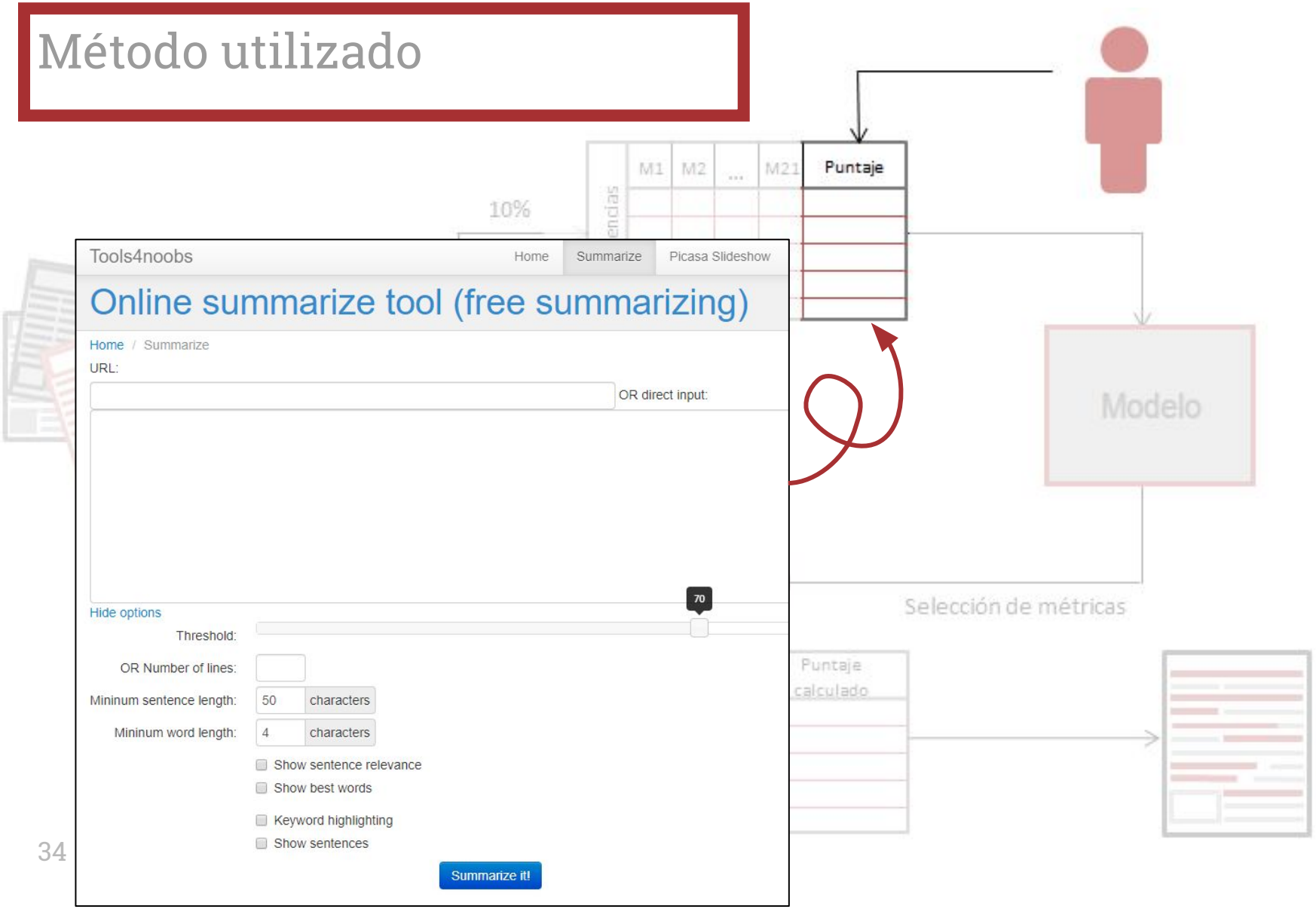
21 métricas

Representación

Normalización Z

	POS_F	POS_L	POS_B	LEN_W	LEN_CH	TF	...	LUHN
S_1	1.452	-1.456	1.431	-1.282	-1.325	-0.627		-1.263
S_2	1.432	-1.437	1.048	-0,824	-0,686	-0,997		-0.794
...								
S_M	-1.531	1.469	1.431	-1.435	-1.576	0.719		-0.560

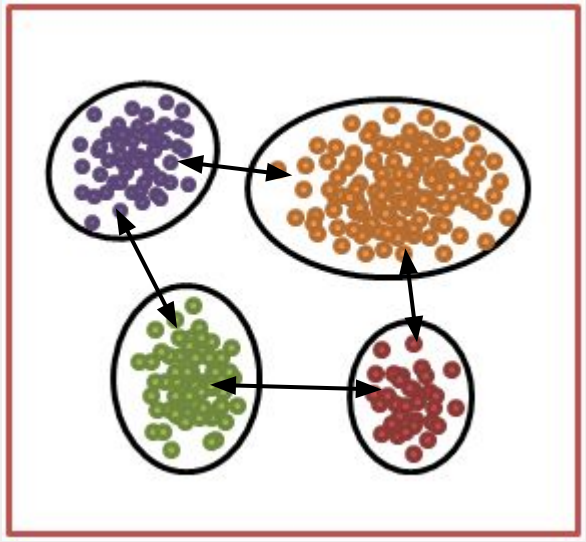
Método utilizado



3. Modelo

Modelos descriptivos

Modelo descriptivo



Agrupamiento partitivo

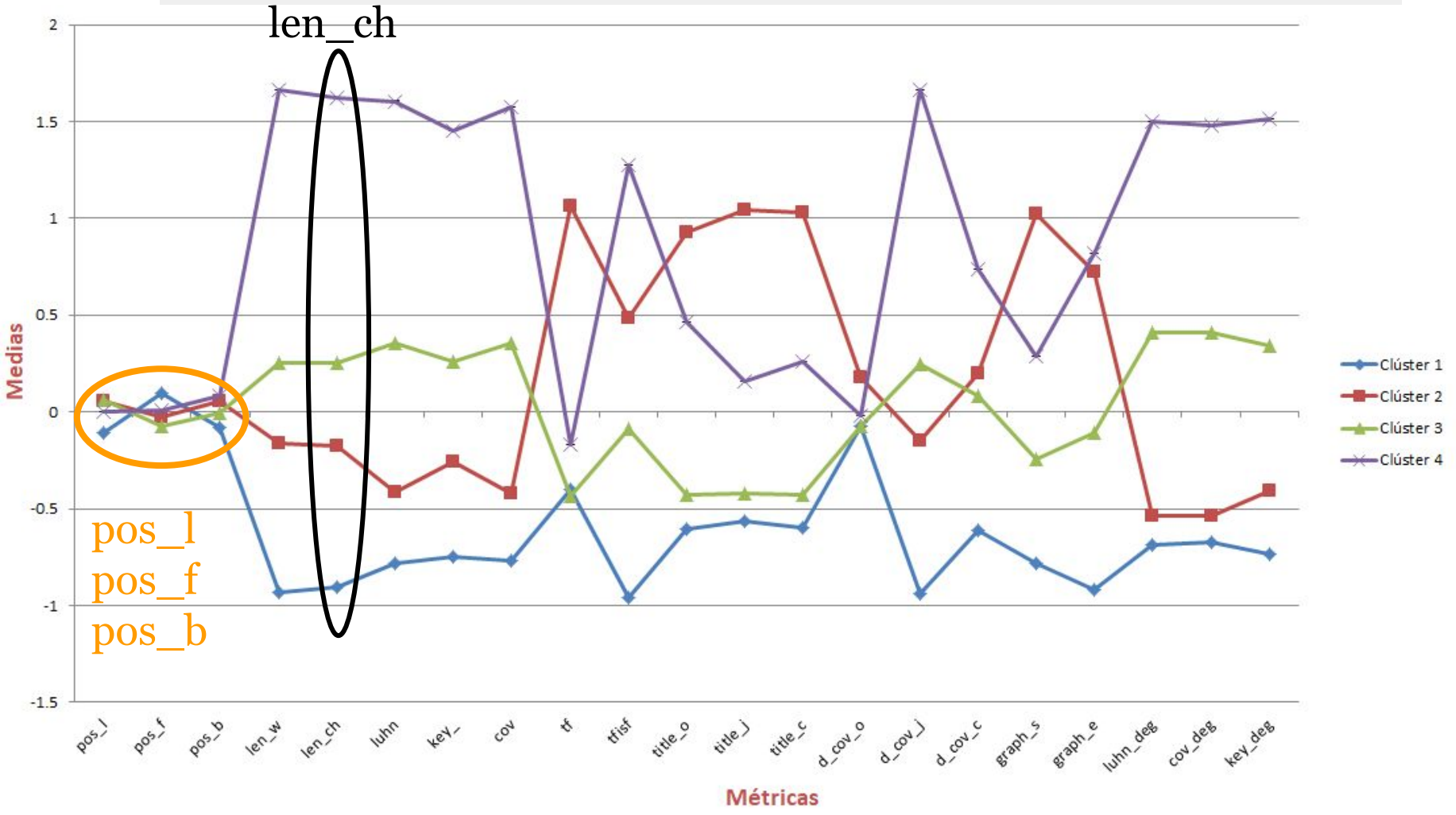
K-Medias

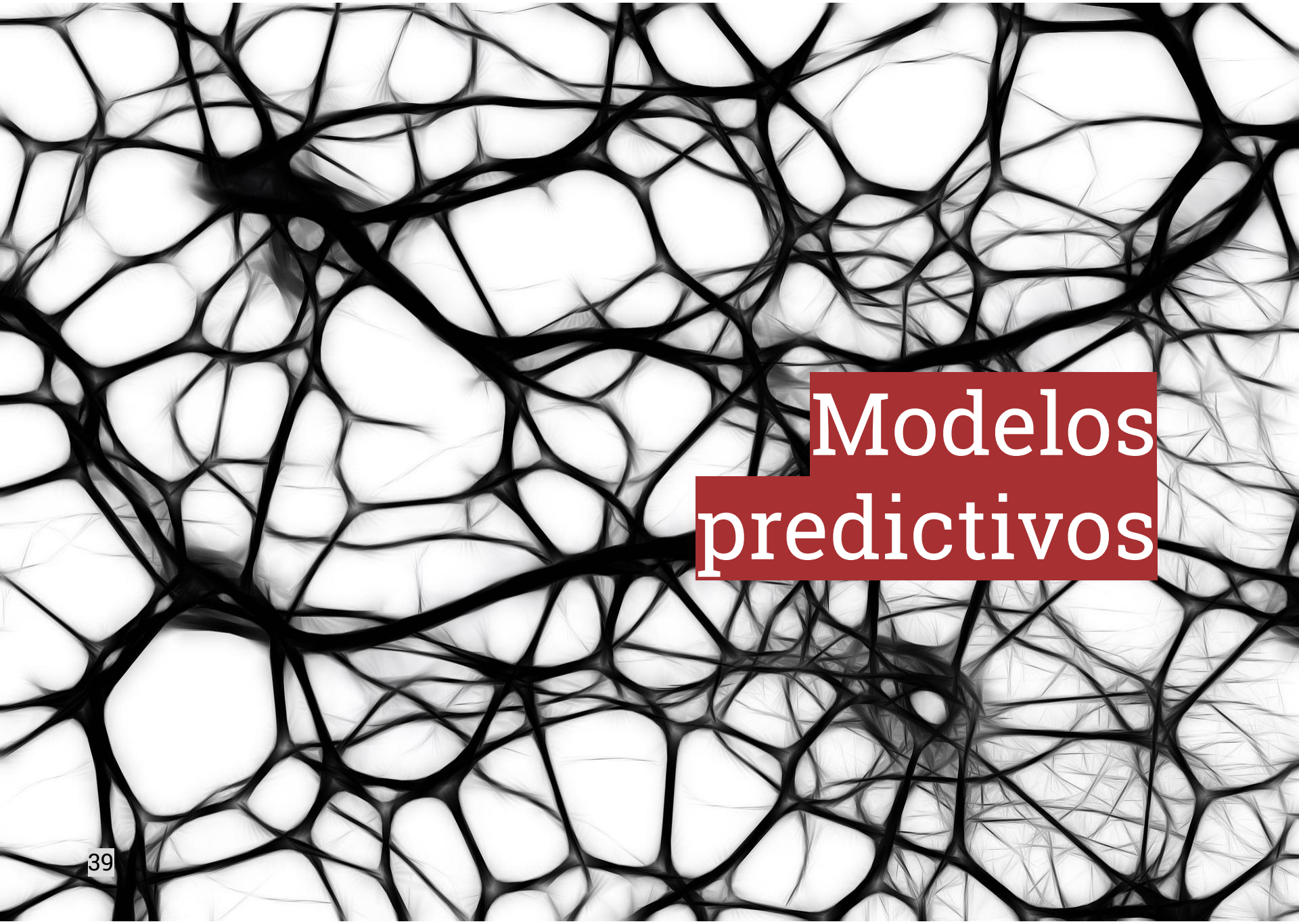
↓

Davies Bouldin

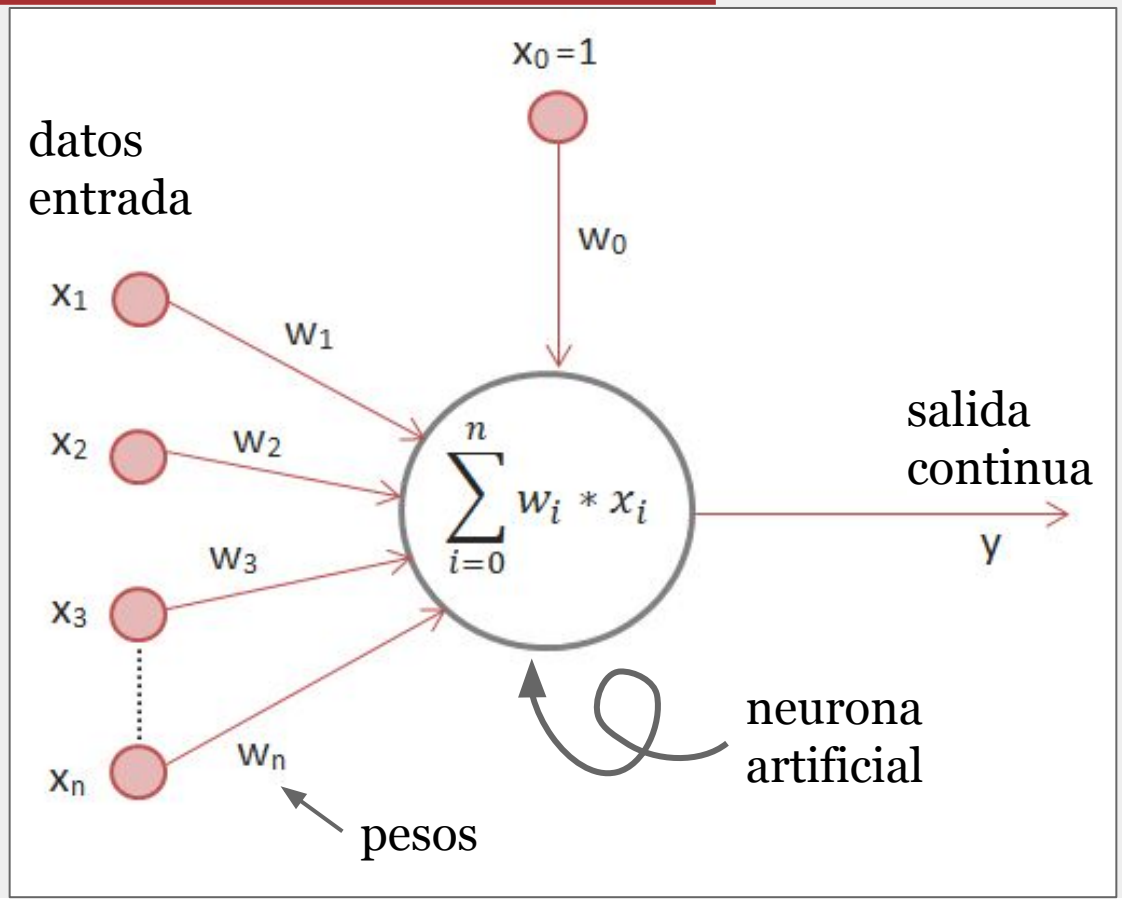
↓

K = 4

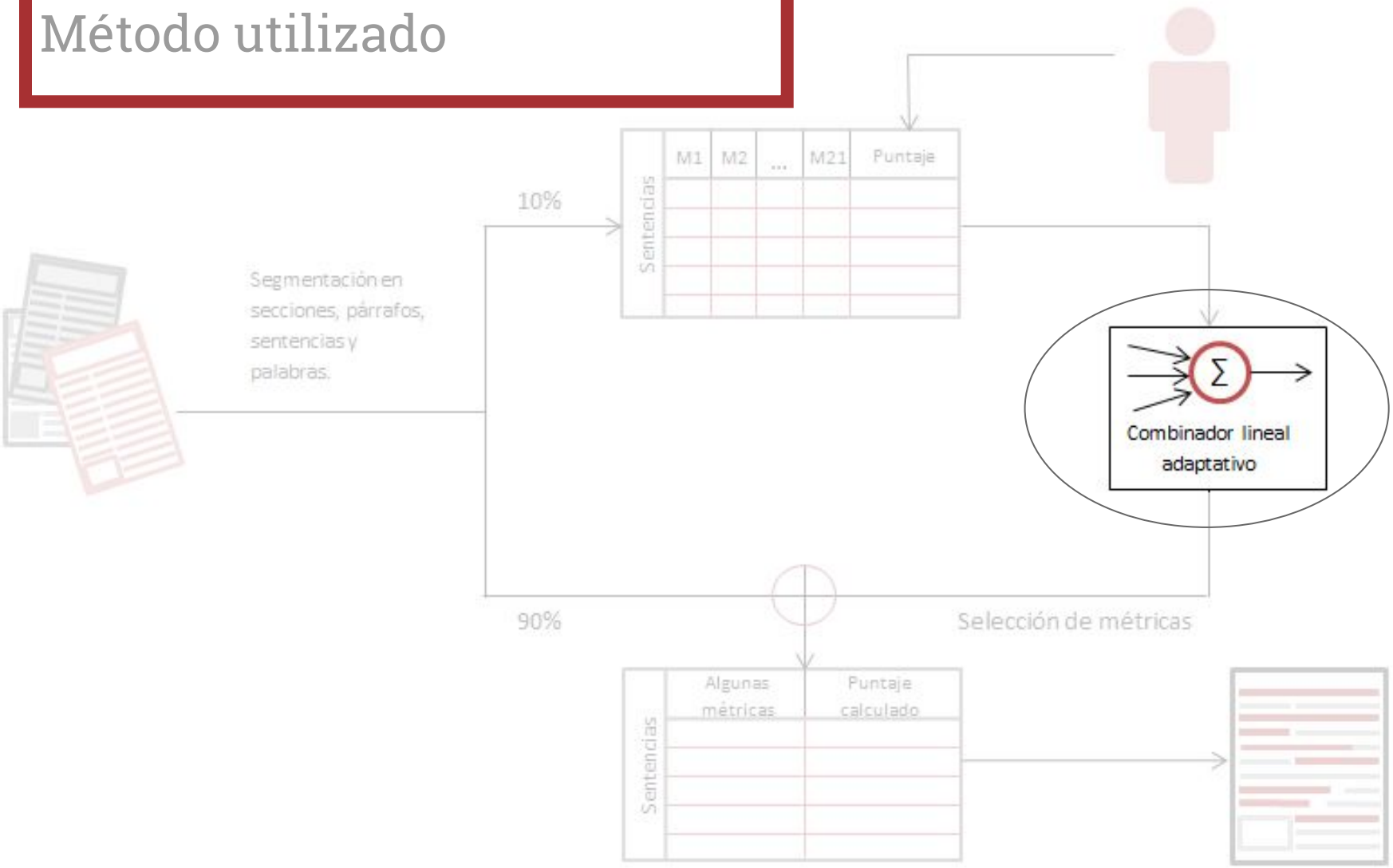

len_ch
pos_l
pos_f
pos_b
Medias
Métricas
2
1.5
1
0.5
0
-0.5
-1
-1.5
pos_l
pos_f
pos_b
len_w
len_ch
luhn
key_
cov
tf
tfisf
title_o
title_j
title_c
d_cov_o
d_cov_j
d_cov_c
graph_s
graph_e
luhn_deg
cov_deg
key_deg
Clúster 1
Clúster 2
Clúster 3
Clúster 4


Modelos
predictivos

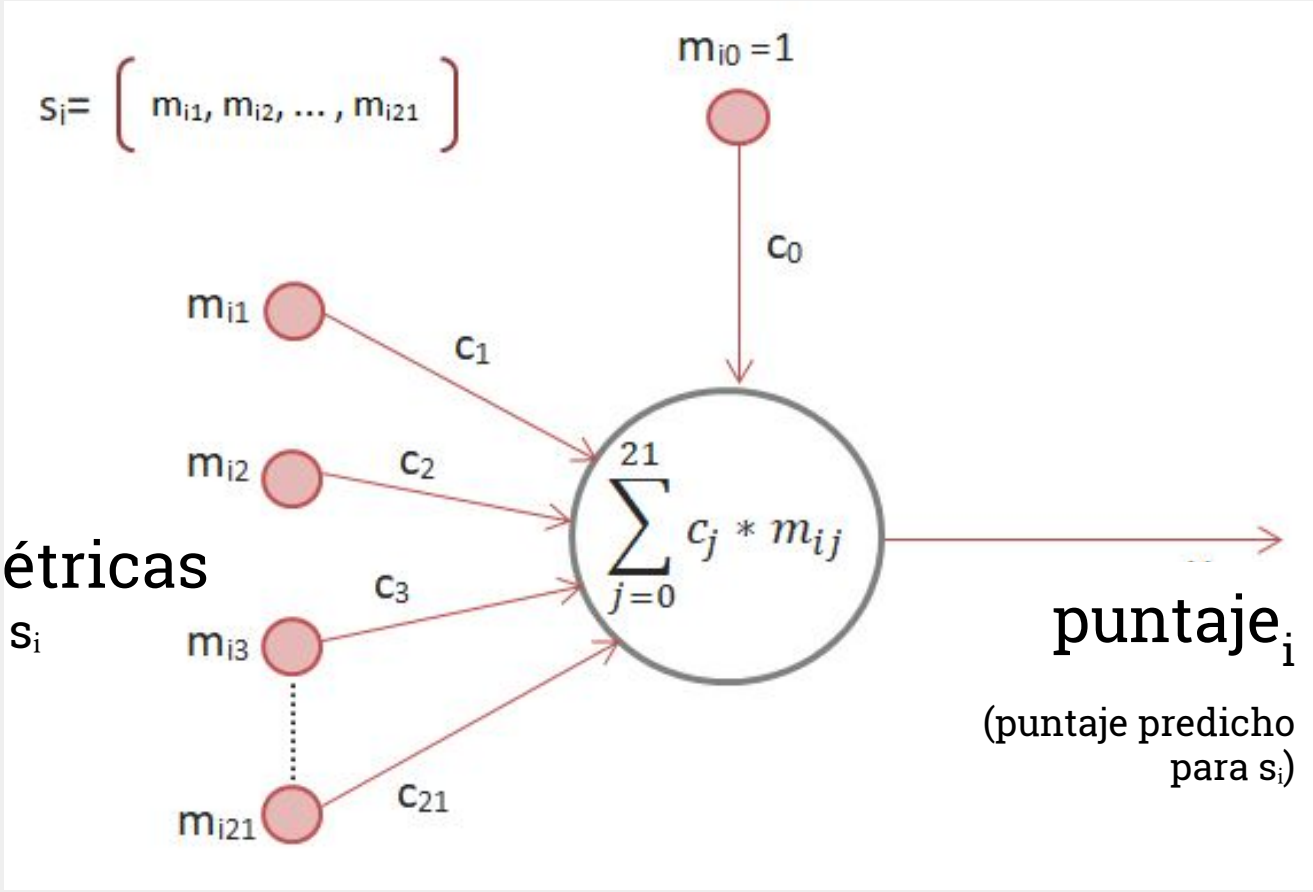
Redes neuronales
Combinador lineal adaptativo



Método utilizado



Combinador lineal utilizado



Sea ***i*** el número de sentencia y ***k*** la cantidad de métricas

$$puntaje_i = \sum_{j=0}^{k} (c_j * m_{ij})$$

1 $$acumulado_i = \sum_{j=1}^{k} abs(c_j * m_{ij})$$

2 $$participa_{ij} = \frac{abs(c_j * m_{ij})}{acumulado_i}$$

3 $$p_j = \frac{\sum_{i=1}^{M} participa_{ij}}{M}$$

Siendo ***M*** la cantidad de sentencias

	m1	m2	m3
S1	0.1	0.5	0.4
S2	0.8	0.1	0.1
S3	0.6	0.2	0.2

1.5 / 3

$p_1 = 0.5$

Sea M la cantidad de sentencias y k la cantidad de métricas

$$p_j = \frac{\sum_{i=1}^{M} participa_{ij}}{M}$$

$$promedio \text{ de participación general} = \frac{\sum_{j=1}^{k} p_j}{k}$$

	m1	m2	m3
S1	0.1	0.5	0.4
S2	0.8	0.1	0.1
S3	0.6	0.2	0.2
	1.5/3	0.8/3	0.7/3
	$p_1 = 0.5$	$p_2 = 0.3$	$p_3 = 0.2$

Si $p_j > 0.\bar{3}$

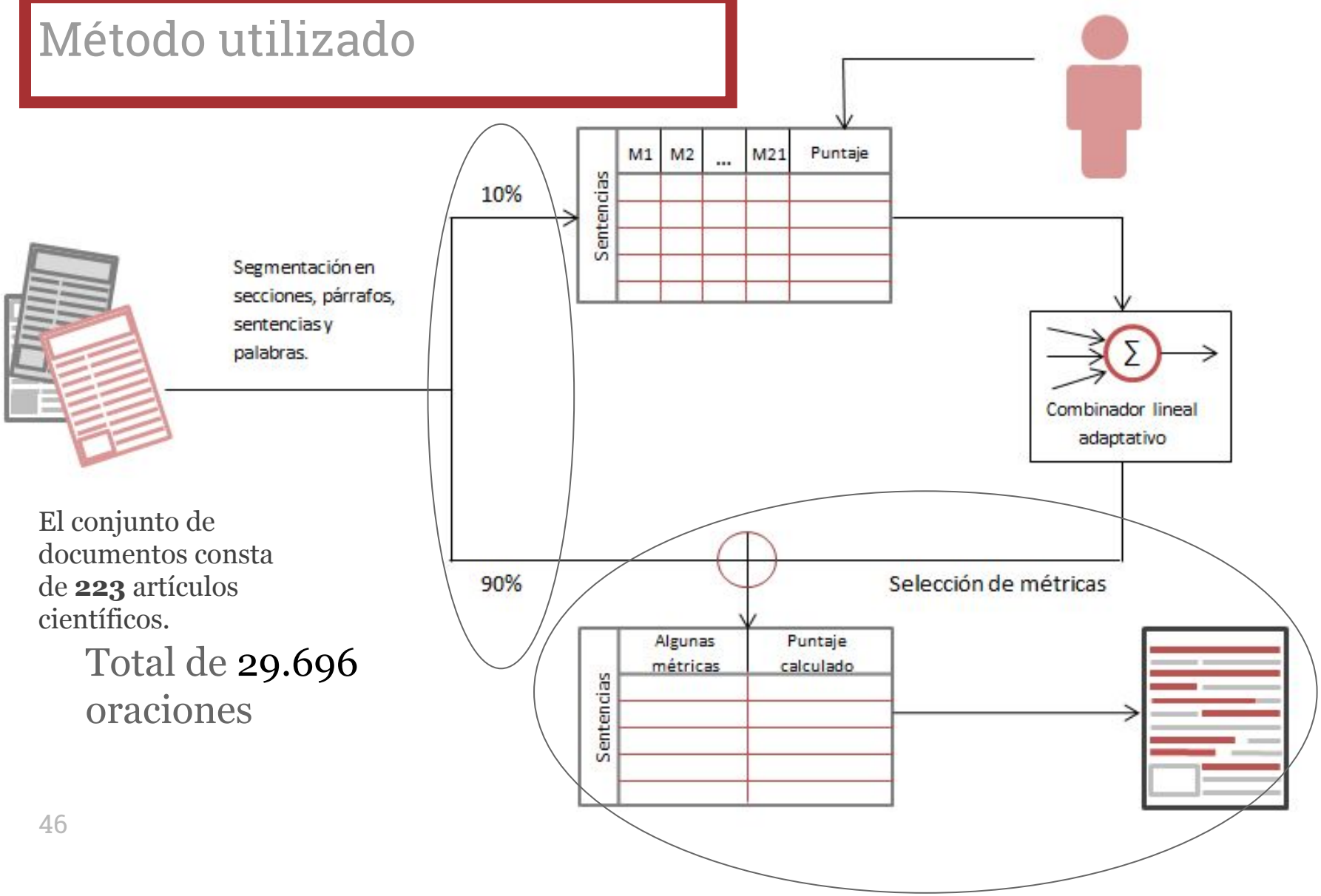
⇨ La métrica j es relevante

⇨ En el ejemplo ***m1*** es la única seleccionada

0.5 + 0.3 + 0.2 = 1/3 = $0.\bar{3}$

4. Resultados

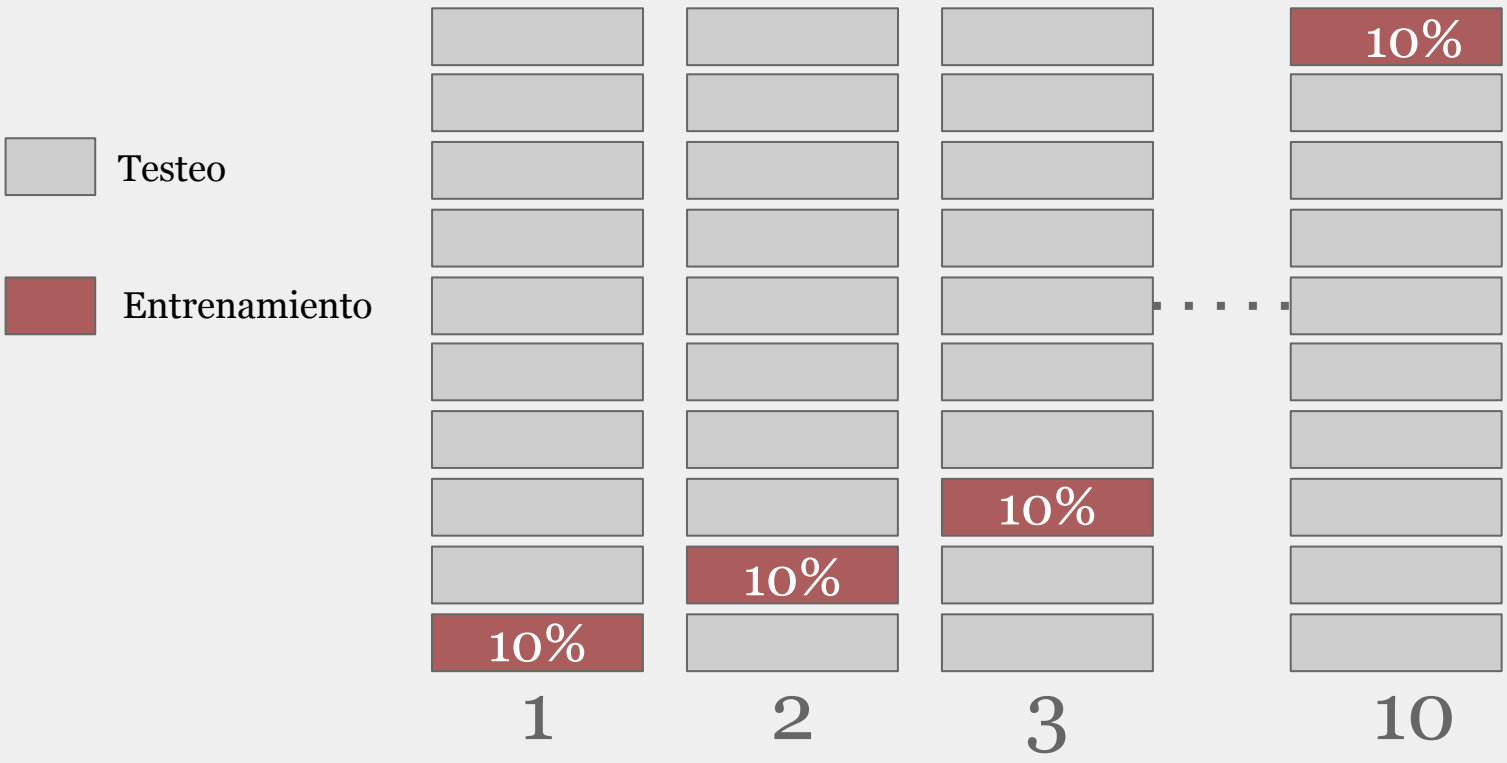
Método utilizado



El conjunto de documentos consta de **223** artículos científicos.

Total de **29.696** oraciones

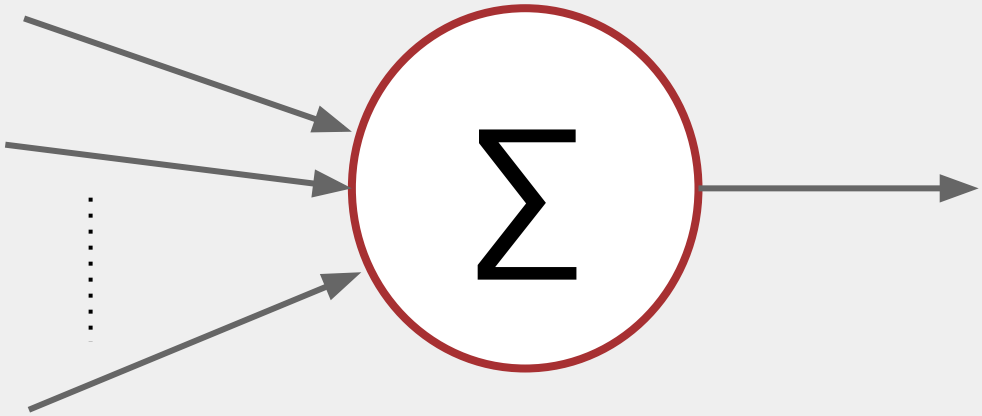
Método utilizado
Validación cruzada



Se realizaron 40 ejecuciones de este proceso.

Pruebas realizadas

CL-All



21 métricas

un solo entrenamiento

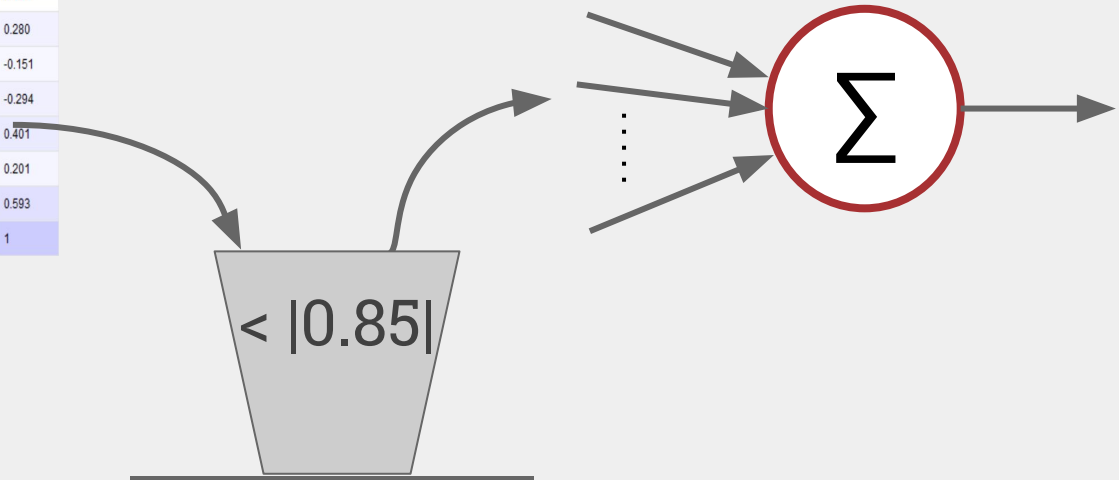
Pruebas realizadas

CL-Corr

un solo entrenamiento

Métricas	pos_f	pos_b	luhn	tf	tfisf	title_c	d_cov_o	d_cov_c	graph_e
pos_f	1	-0.037	-0.049	-0.043	-0.056	0.127	-0.007	0.066	0.198
pos_b	-0.037	1	0.040	0.034	-0.028	0.166	0.019	0.100	0.008
luhn	-0.049	0.040	1	0.160	-0.160	0.003	0.038	0.311	0.280
tf	-0.043	0.034	0.160	1	0.097	-0.121	-0.052	-0.078	-0.151
tfisf	-0.056	-0.028	-0.160	0.097	1	-0.165	-0.037	-0.235	-0.294
title_c	0.127	0.166	0.003	-0.121	-0.165	1	0.153	0.428	0.401
d_cov_o	-0.007	0.019	0.038	-0.052	-0.037	0.153	1	0.242	0.201
d_cov_c	0.066	0.100	0.311	-0.078	-0.235	0.428	0.242	1	0.593
graph_e	0.198	0.008	0.280	-0.151	-0.294	0.401	0.201	0.593	1

Matriz de correlación

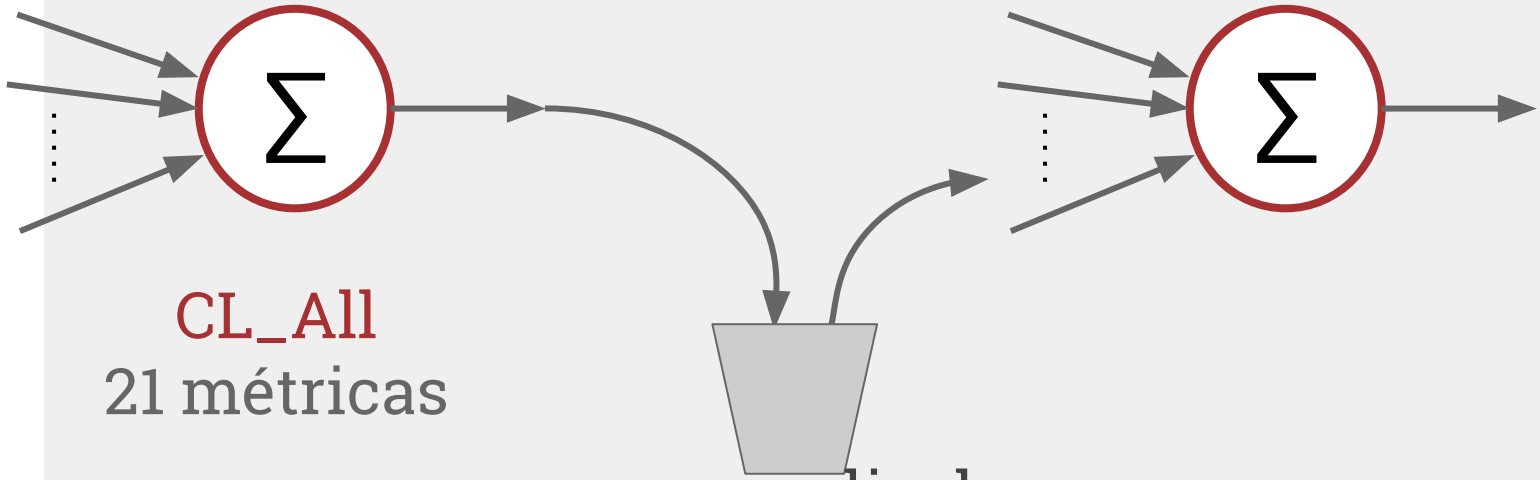


14 métricas

Pruebas realizadas

CL-Pruned

un solo entrenamiento



CL_All
21 métricas

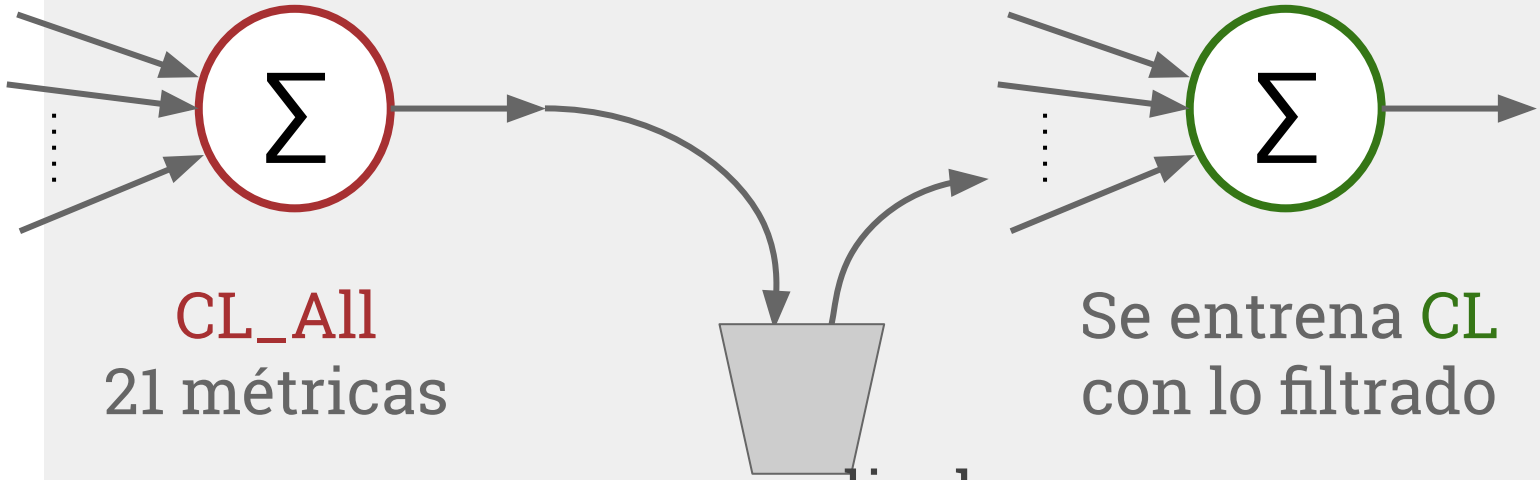
> promedio de participación

9 métricas

Pruebas realizadas

CL-Sel-Train

dos entrenamientos



CL_All
21 métricas

> promedio de participación

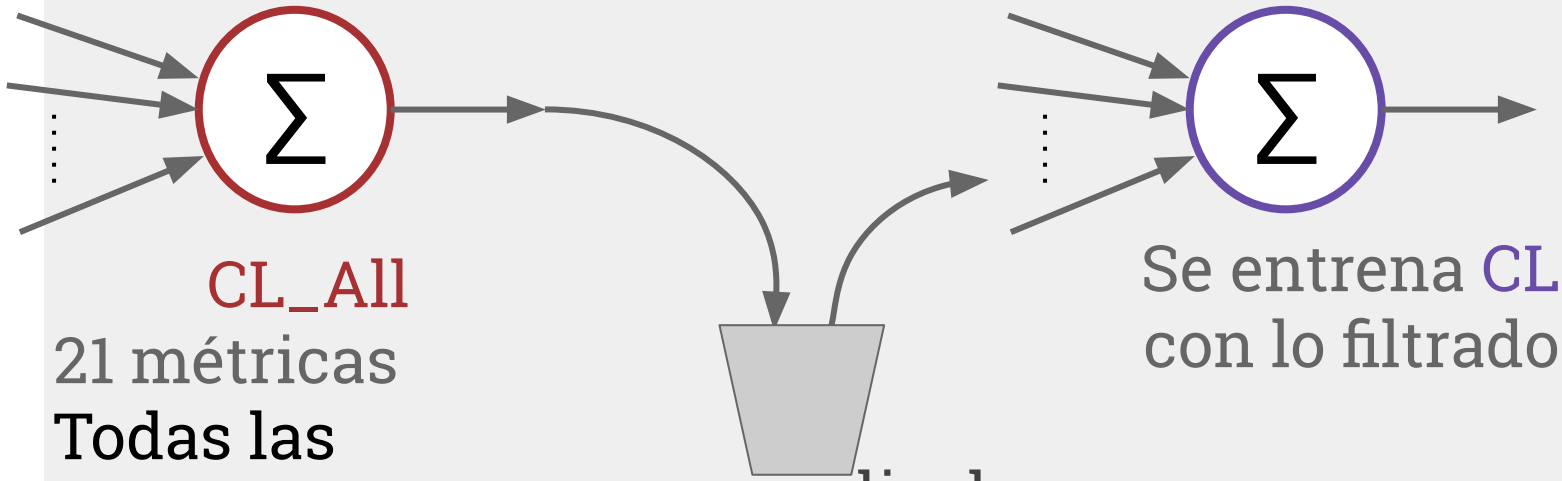
9 métricas

Se entrena CL con lo filtrado

Pruebas realizadas

CL-Selecc

dos entrenamientos



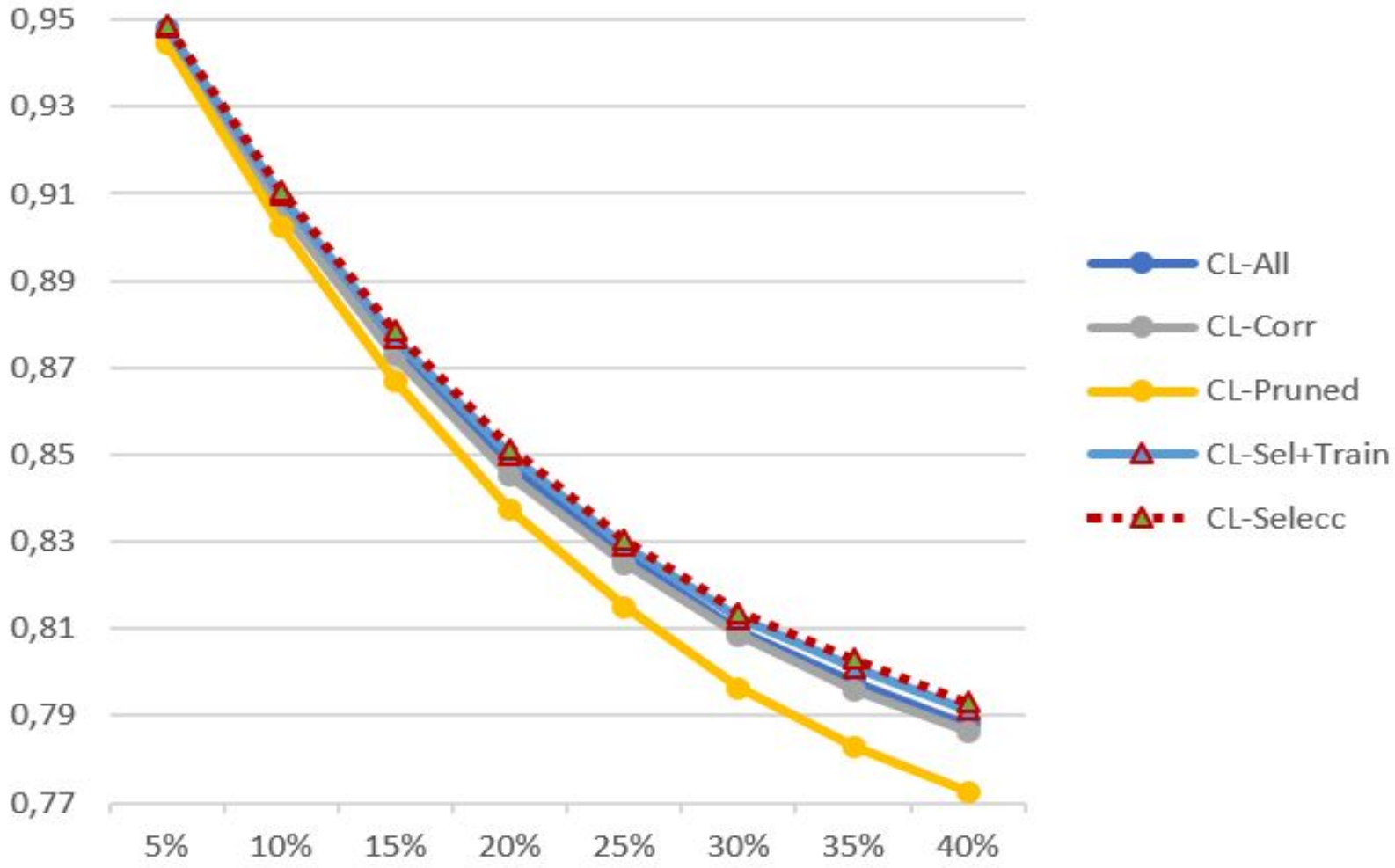
CL_All
21 métricas
Todas las sentencias

> promedio de participación

9 métricas

Se entrena CL con lo filtrado

Tasa de acierto

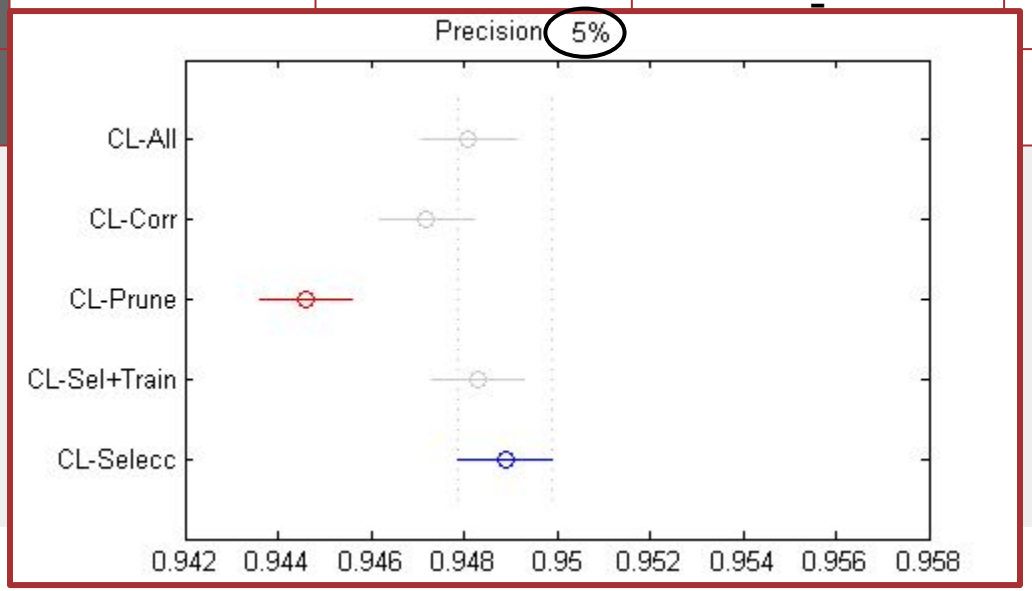


Intervalos de confianza
Test ANOVA

	CL-Pruned	CL-Corr	CL-Sel-Train	CL-Selecc
CL-All	△	-	-	-
CL-Pruned		▽	▽	▽
CL-Corr			-	-
CL-Sel-Train				-

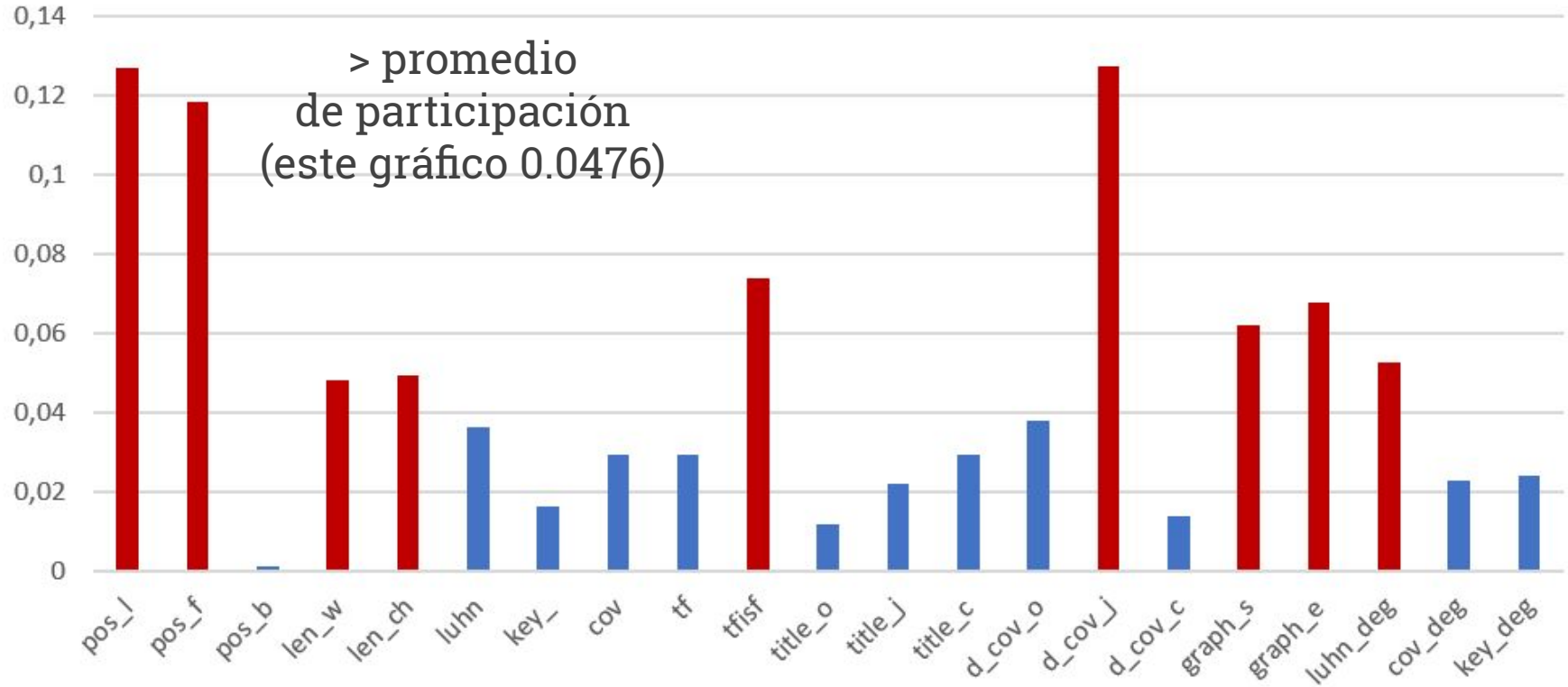
Intervalos de confianza
Test ANOVA

	CL-Pruned	CL-Corr	CL-Sel-Train	CL-Selecc
CL-All	△	-	-	-
CL-Pruned		▽	▽	▽
CL-Corr			-	-
CL-Sel-Train				-



Todos los cortes dieron semejantes resultados.

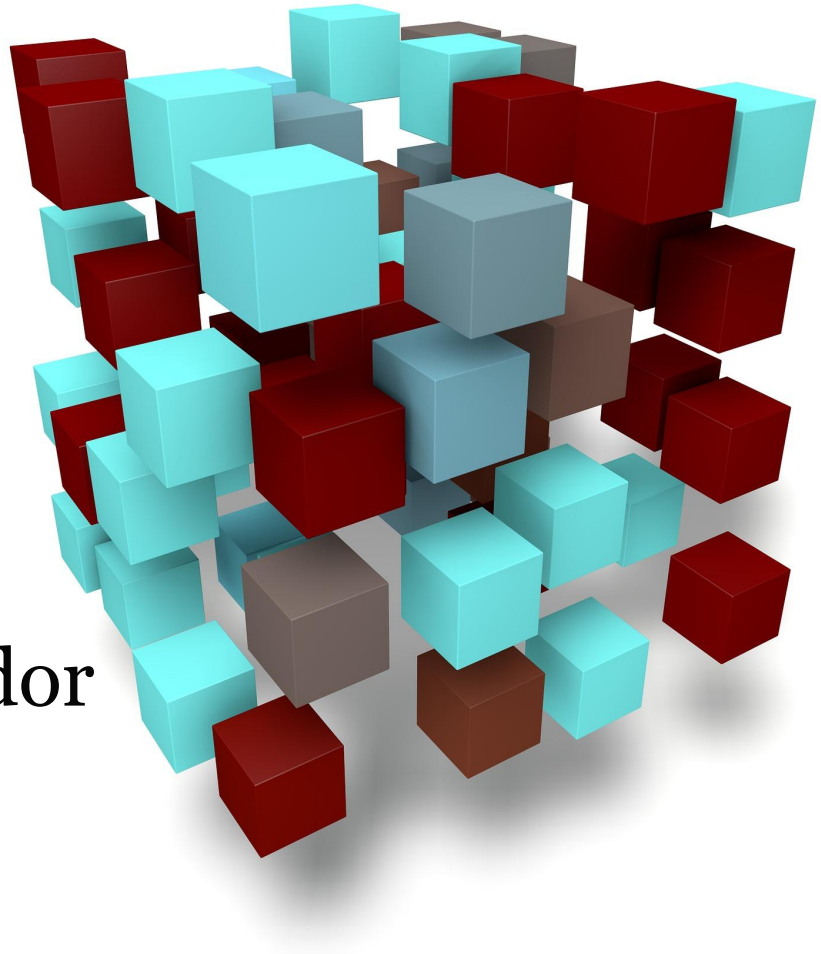
Métricas seleccionadas



POS_L
POS_F
LEN_CH
LEN_W
TFISF
GRAPH_S
GRAPH_E
D_COV_J
LUHN_DEG

Conclusiones generales

- Documentos
- Modelo relacional
- Cálculo de métricas
- Alternativas combinador



Trabajos futuros

1. Relación métricas y documentos
2. Ampliar y mejorar métricas
3. Desarrollar otros modelos y comparar

¡Muchas gracias por su atención!

Julieta Pilar Corvi
julieta.corvi@gmail.com