Facultad de Ciencias

Máster en Data Science

Categorización de Textos: Asignación de temas

(Text categorization (assigning subject categories, topics))

Informe trabajo de Semántica

Autores: Julia Ruiz Salmón
José Ney Gandica Cárdenas

Abril 2019

- **Contenidos:**

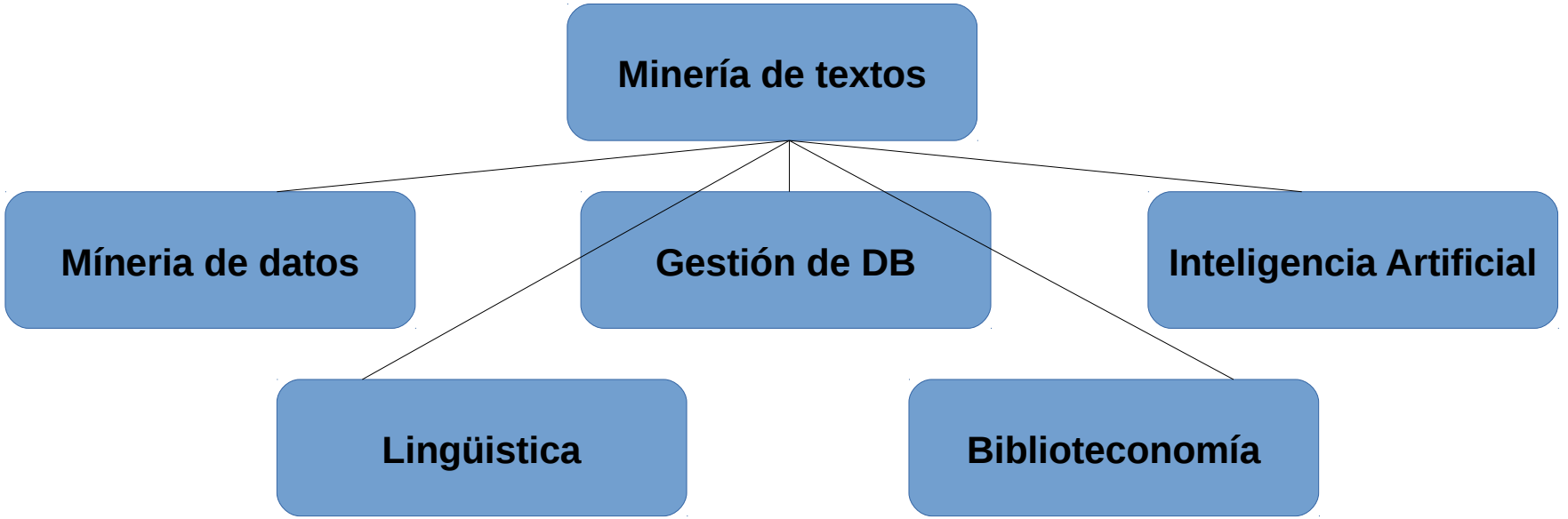
1) Introducción
2)
3) Algoritmo LDA (Latent Dirichlet Allocation)
4)
5) Categorización de Textos: Asignación de temas (Noticias – ejercicio práctico)

Introducción

¿Porqué es necesaria la minería de texto?

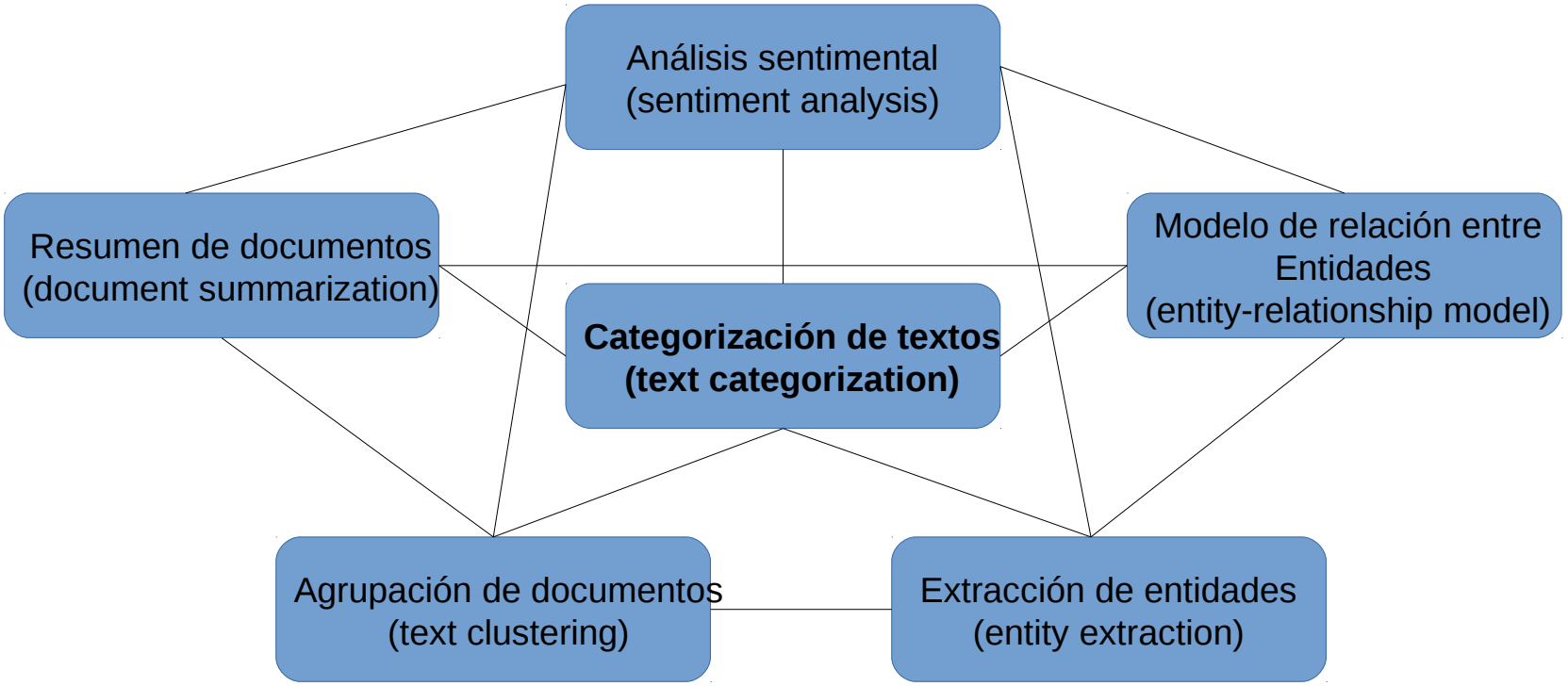
1) Elevado número de fuentes textuales
2) Datos no estructurados
3) Clasificación de documentos
4) Reconocimiento de patrones
5) Análisis predictivo

1)
Ramas del conocimiento que abarca:



Introducción

Ámbito de la minería de textos:



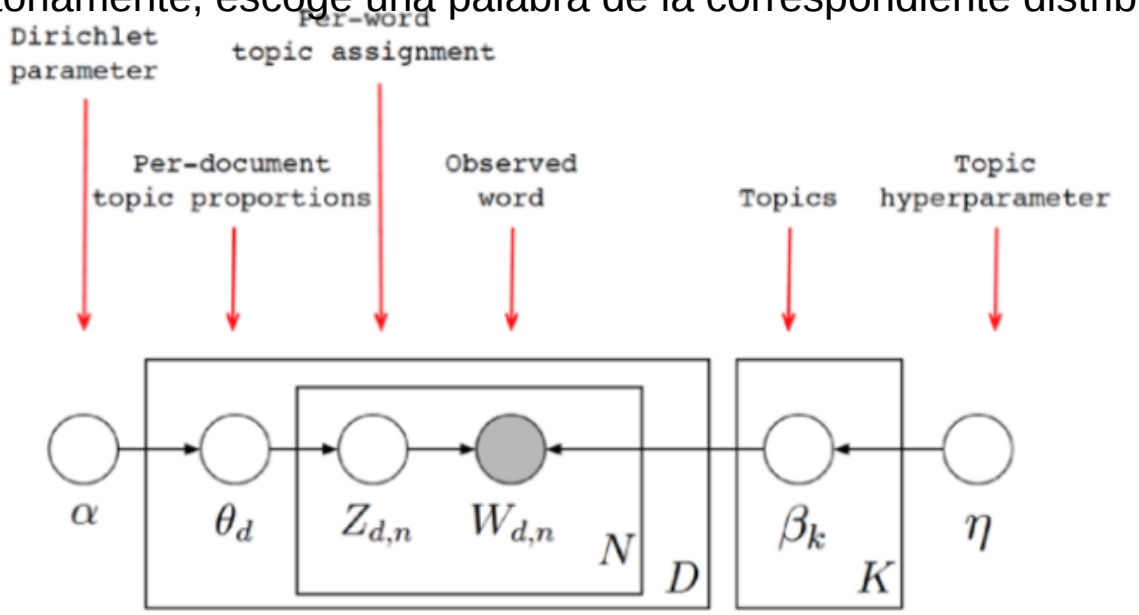
Algoritmo LDA (Latent Dirichlet Allocation)

LDA consiste en:

1) Un modelo de temas no supervisado.
2) Extrae varios temas de un documento.
3) Cada tema tiene una correspondencia con una distribución sobre un vocabulario fijo. Esta distribución es la distribución de Dirichlet que se basa en el Teorema de Bayes.
4)

El proceso generativo del LDA consta de dos etapas:

5) Elige al azar una distribución sobre temas.
6) Para cada palabra en el documento realiza lo siguiente:
7) Aleatoriamente, elige un tema de la distribución sobre temas.
8) Aleatoriamente, escoge una palabra de la correspondiente distribución sobre vocabulario.



Fuente: *D.Blei (2012)*

Parámetros:

α prior per document topic distribution

η prior per topic word distribution

K número de topics

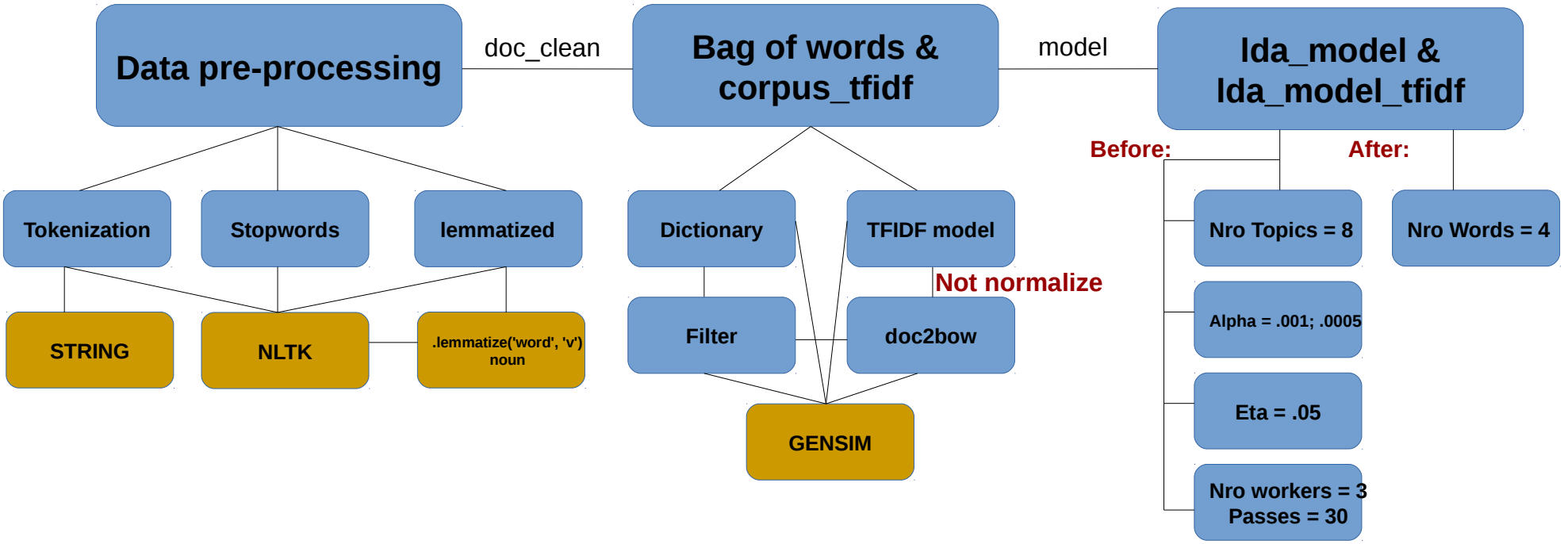
Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Fuente: BBC Mundo. Secciones: **economía, deporte, ciencia**. 13/04/2019
Corpus: 6 noticias recientes por cada tema, 18 en total.
Evaluación: 1 noticia de cada tema, 3 en total.

Librerias: gensim, #extract semantic topics from documents, models and transformations
Nltk, #programs for statistical natural language processing (NLP)
String, #paquete básico para trabajar con texto

Metodología:



Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Resultados en el corpus:

Bag of words

```
Palabras que más aparecen:
Palabra 5 ("américa") aparece 15 veces.
Palabra 151 ("argentina") aparece 12 veces.
Palabra 76 ("ingresos") aparece 11 veces.
Palabra 159 ("fútbol") aparece 10 veces.
Palabra 88 ("personas") aparece 10 veces.
Palabra 32 ("negros") aparece 9 veces.
Palabra 159 ("fútbol") aparece 9 veces.
Palabra 148 ("países") aparece 8 veces.
Palabra 85 ("negocio") aparece 8 veces.
Palabra 137 ("economía") aparece 8 veces.
Palabra 3 ("agujeros") aparece 7 veces.
Palabra 163 ("2020") aparece 7 veces.
Palabra 80 ("mercado") aparece 6 veces.
Palabra 12 ("contenido") aparece 6 veces.
Palabra 22 ("gravedad") aparece 6 veces.
```

lda_model

```
Algoritmo LDA usando un diccionario de palabras:
Topic: 1
Words: 0.058*"ingresos" + 0.027*"selección" + 0.027*"técnico" + 0.027*"argentina"
Topic: 2
Words: 0.066*"fútbol" + 0.047*"américa" + 0.041*"argentina" + 0.035*"presidente"
Topic: 3
Words: 0.076*"compañía" + 0.049*"mercado" + 0.049*"firma" + 0.042*"millones"
Topic: 4
Words: 0.078*"latina" + 0.062*"negocio" + 0.062*"ingresos" + 0.047*"unidos"
Topic: 5
Words: 0.068*"personas" + 0.056*"universidad" + 0.043*"investigación" + 0.031*"problemas"
Topic: 6
Words: 0.055*"número" + 0.055*"publicación" + 0.055*"hospital" + 0.028*"universidad"
Topic: 7
Words: 0.035*"negros" + 0.034*"millones" + 0.034*"científicos" + 0.029*"agujeros"
Topic: 8
Words: 0.056*"economía" + 0.056*"negocio" + 0.042*"mercado" + 0.035*"rival"
```

Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Resultados en el corpus:

Corpus_tfidf

```
[(12, 0.10812701334912835),
 (20, 0.0662977746971313),
 (23, 0.1546006072648786),
 (24, 0.1988933240913939),
 (33, 0.1325955493942626),
 (34, 0.10812701334912835),
 (35, 0.10812701334912835),
```

lda_model_tfidf

```
Algoritmo LDA usando un objeto tf-idf:
Topic: 1
Word: 0.004*"argentino" + 0.004*"director" + 0.004*"social" + 0.004*"comunicado"
Topic: 2
Word: 0.004*"argentino" + 0.004*"director" + 0.004*"social" + 0.004*"comunicado"
Topic: 3
Word: 0.028*"agujeros" + 0.028*"negros" + 0.026*"científicos" + 0.022*"grande"
Topic: 4
Word: 0.026*"ingresos" + 0.026*"publicación" + 0.026*"hospital" + 0.018*"medios"
Topic: 5
Word: 0.029*"fútbol" + 0.028*"selección" + 0.028*"argentina" + 0.022*"personas"
Topic: 6
Word: 0.027*"compañía" + 0.026*"negocio" + 0.022*"ingresos" + 0.021*"firma"
Topic: 7
Word: 0.030*"mercado" + 0.024*"presidente" + 0.019*"gravedad" + 0.018*"obtener"
Topic: 8
Word: 0.028*"argentino" + 0.028*"noche" + 0.024*"hora" + 0.023*"economía"
```

Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Resultados en el test:

Economía

lda_model

```
Score: 0.5112519860267639       Topic: 0.076*"compañía" + 0.049*"mercado" + 0.049*"firma" + 0.042*"millones"
Score: 0.2332012802362442       Topic: 0.035*"negros" + 0.034*"millones" + 0.034*"científicos" + 0.029*"agujer
os"
Score: 0.1384214162826538       Topic: 0.066*"fútbol" + 0.047*"américa" + 0.041*"argentina" + 0.035*"president
e"
Score: 0.06454111635684967      Topic: 0.078*"latina" + 0.062*"negocio" + 0.062*"ingresos" + 0.047*"unidos"
Score: 0.052524253726005554     Topic: 0.058*"ingresos" + 0.027*"selección" + 0.027*"técnico" + 0.027*"argenti
na"
```

lda_model_tfidf

```
Score: 0.6342630386352539
Topic: 0.027*"compañía" + 0.026*"negocio" + 0.022*"ingresos" + 0.021*"firma"

Score: 0.18710722029209137
Topic: 0.028*"agujeros" + 0.028*"negros" + 0.026*"científicos" + 0.022*"grande"

Score: 0.1136845201253891
Topic: 0.029*"fútbol" + 0.028*"selección" + 0.028*"argentina" + 0.022*"personas"

Score: 0.06490522623062134
Topic: 0.030*"mercado" + 0.024*"presidente" + 0.019*"gravedad" + 0.018*"obtener"
```

Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Resultados en el test:

Deporte

lda_model

```
Score: 0.4928743243217468       Topic: 0.066*"fútbol" + 0.047*"américa" + 0.041*"argentina" + 0.035*"president
e"
Score: 0.20012427866458893      Topic: 0.058*"ingresos" + 0.027*"selección" + 0.027*"técnico" + 0.027*"argenti
na"
Score: 0.1964830905199051       Topic: 0.035*"negros" + 0.034*"millones" + 0.034*"científicos" + 0.029*"agujer
os"
Score: 0.11039335280656815      Topic: 0.055*"número" + 0.055*"publicación" + 0.055*"hospital" + 0.028*"univer
sidad"
```

lda_model_tfidf

```
Score: 0.4107641279697418
Topic: 0.029*"fútbol" + 0.028*"selección" + 0.028*"argentina" + 0.022*"personas"

Score: 0.24102327227592468
Topic: 0.027*"compañía" + 0.026*"negocio" + 0.022*"ingresos" + 0.021*"firma"

Score: 0.18949885666370392
Topic: 0.026*"ingresos" + 0.026*"publicación" + 0.026*"hospital" + 0.018*"medios"

Score: 0.1072831004858017
Topic: 0.028*"agujeros" + 0.028*"negros" + 0.026*"científicos" + 0.022*"grande"

Score: 0.05138380452990532
Topic: 0.030*"mercado" + 0.024*"presidente" + 0.019*"gravedad" + 0.018*"obtener"
```

Categorización de Textos: Asignación de temas

(Noticias – ejercicio práctico)

Resultados en el test:

Ciencia

lda_model

```
Score: 0.7739781141281128       Topic: 0.035*"negros" + 0.034*"millones" + 0.034*"científicos" + 0.029*"agujer
os"
Score: 0.1496550738811493       Topic: 0.068*"personas" + 0.056*"universidad" + 0.043*"investigación" + 0.031*
"problemas"
Score: 0.06052404269576073      Topic: 0.056*"economía" + 0.056*"negocio" + 0.042*"mercado" + 0.035*"rival"
Score: 0.015815572813153267     Topic: 0.066*"fútbol" + 0.047*"américa" + 0.041*"argentina" + 0.035*"president
e"
```

lda_model_tfidf

```
Score: 0.5811505913734436
Topic: 0.028*"agujeros" + 0.028*"negros" + 0.026*"científicos" + 0.022*"grande"

Score: 0.17511488497257233
Topic: 0.029*"fútbol" + 0.028*"selección" + 0.028*"argentina" + 0.022*"personas"

Score: 0.13132624328136444
Topic: 0.030*"mercado" + 0.024*"presidente" + 0.019*"gravedad" + 0.018*"obtener"

Score: 0.05026784539222717
Topic: 0.028*"argentino" + 0.028*"noche" + 0.024*"hora" + 0.023*"economía"

Score: 0.03475542739033699
Topic: 0.027*"compañía" + 0.026*"negocio" + 0.022*"ingresos" + 0.021*"firma"

Score: 0.027378197759389877
Topic: 0.026*"ingresos" + 0.026*"publicación" + 0.026*"hospital" + 0.018*"medios"
```

A todos muchas gracias por vuestra atención…

A quien le guste el tema le recomendamos:

Dynamic topic modelling for cryptocurrency community forums

¿PREGUNTAS?