UNIVERSIDAD IBEROAMERICANA

CONVOCATORIA ACADÉMICO DE TIEMPO COMPLETO

Departamento de Física y Matemáticas

EJERCICIO 2

DR. ANTONIO CEDILLO HERNÁNDEZ

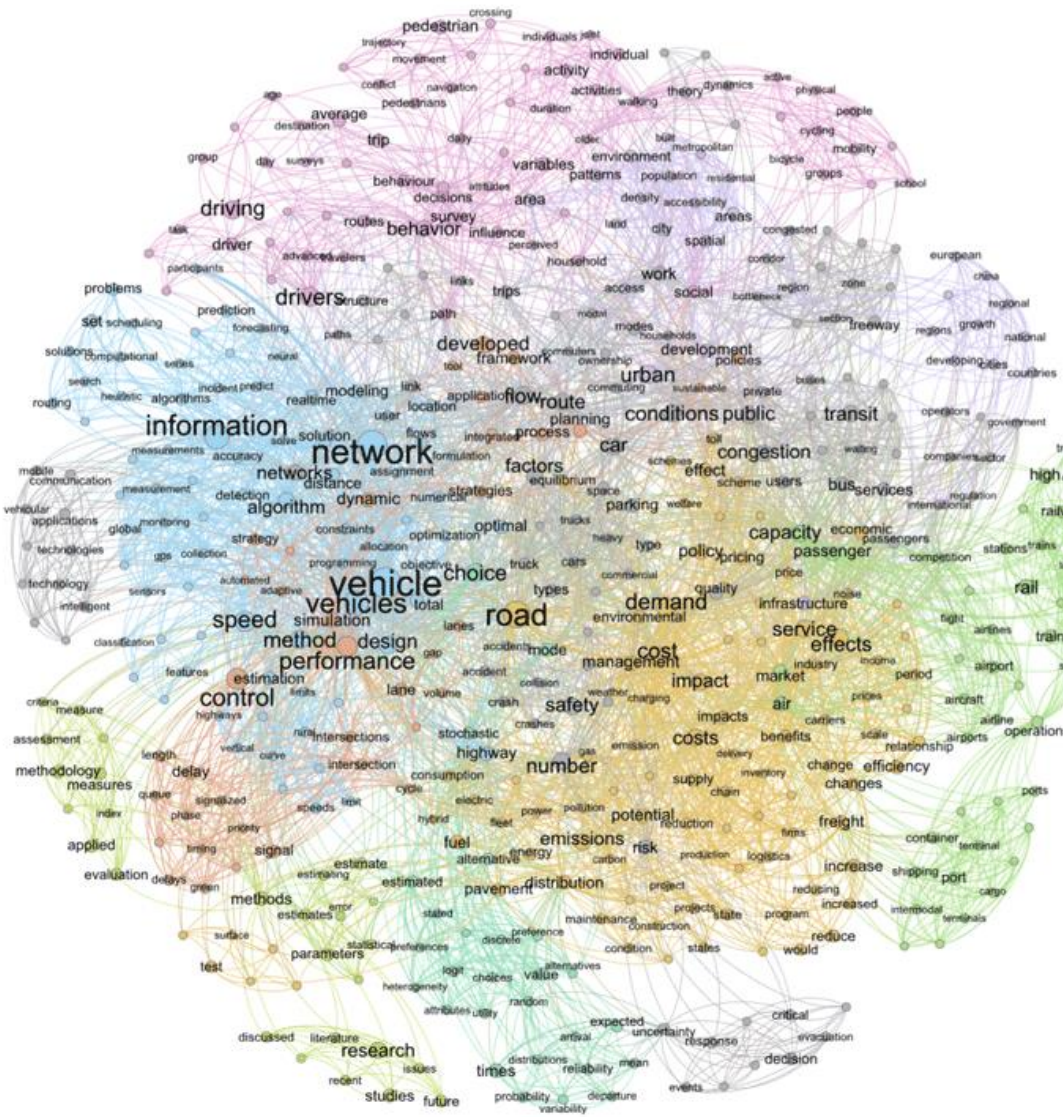
CONTENIDO

UNIVERSIDAD IBEROAMERICANA

- Topic Modelling
- Resultados (Topic Modelling)
- Análisis de sentimientos
- Resultados (Análisis de sentimientos)
- Código fuente

UNIVERSIDAD IBEROAMERICANA

TOPIC MODELLING



UNIVERSIDAD IBEROAMERICANA

Definición

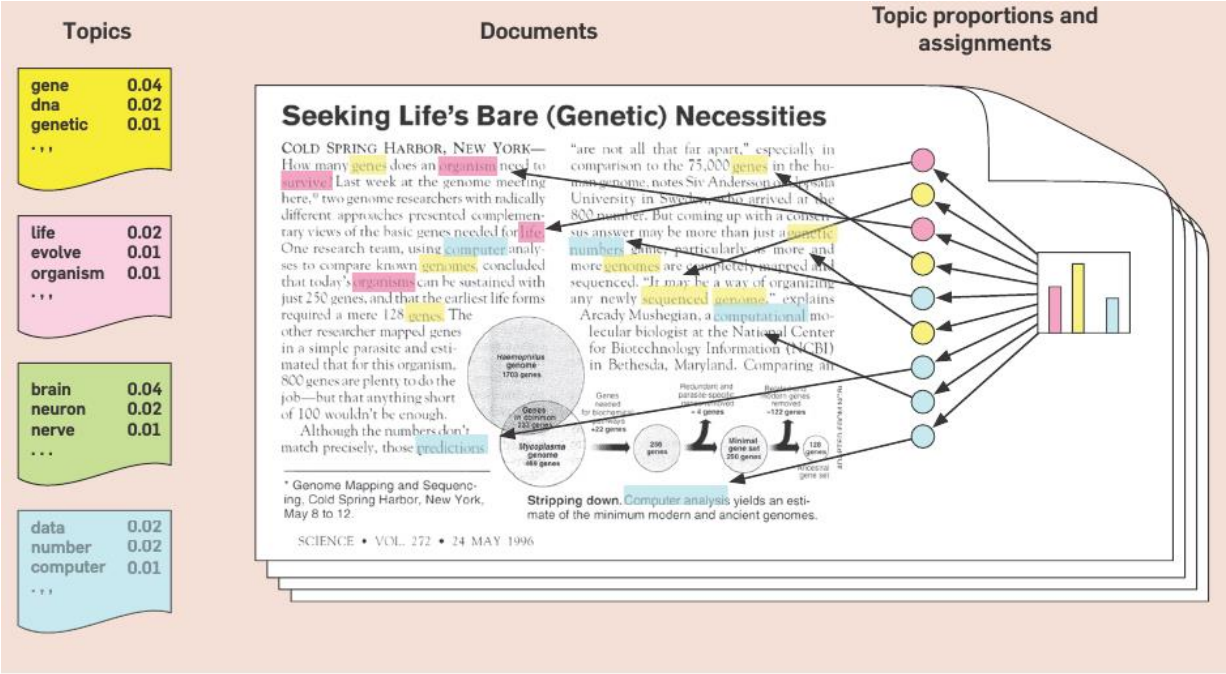
Topic modeling es una herramienta de análisis textual proveniente de la rama informática de Minería de datos.

- A partir solo de texto (sin utilizar diccionarios o representaciones semánticas) ***Topic modeling*** usa complejos modelos estadísticos para reconstruir temas.

UNIVERSIDAD IBEROAMERICANA

Latent Dirichlet Allocation (LDA)

El autor más influyente sobre ***Topic modeling*** es ***David Blei***, quien propuso (2011) su modelo llamado ***Latent Dirichlet Allocation (LDA)***



UNIVERSIDAD IBEROAMERICANA

Latent Dirichlet Allocation (LDA)

- Las palabras en un texto siguen una hipótesis de *bolsa de palabras*, es decir, que el orden no importa
- El uso de una palabra es ser parte de un *tema (**topic**)*
- Una palabra comunica la misma información sin importar donde se encuentre en el texto

UNIVERSIDAD IBEROAMERICANA

Pasos a seguir:

- Extracción de datos desde archivo
- Preparar datos para el análisis (Pre-procesamiento)
  - Borrar signos de puntuación
  - Dividir el texto en TOKENS
  - Quitar palabras de parada (*stop words*)
  - Eliminar palabra de menos de 2 caracteres (*shortWords*)
  - Eliminar palabras de más de 15 caracteres (*longWords*)
- Crear bolsa de palabras
- Remover palabras que no aparecen más de 2 veces en el texto y vacías
- Análisis estadístico LDA

RESULTADOS (TOPIC MODELLING)

RESULTADOS

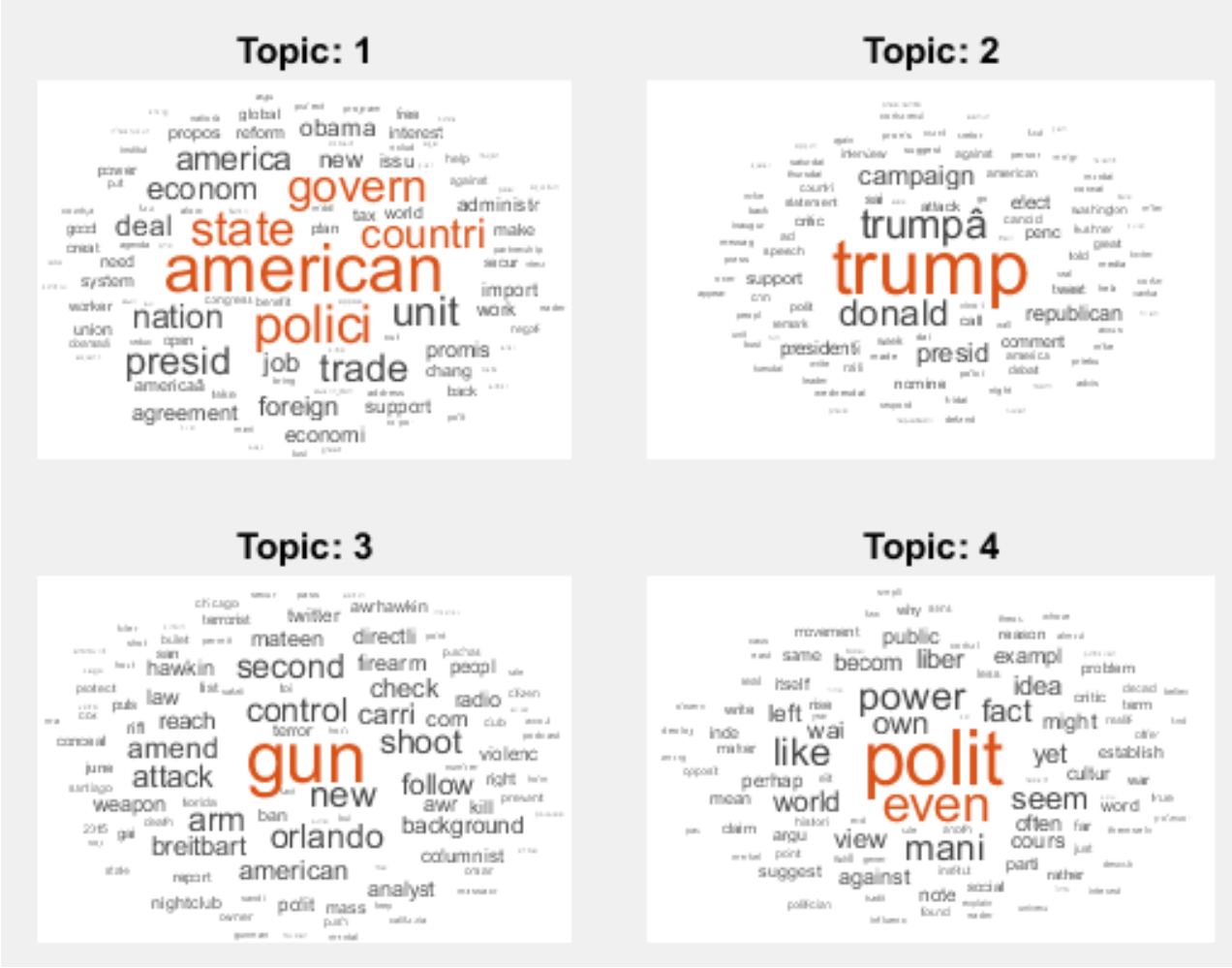
UNIVERSIDAD IBEROAMERICANA

Articles1.csv

Iteration	Time per iter., s	Relative Delta log(L)	Training perplexity	Topic concentr.	Concentr. iterations
0	162.40	Inf	4.725e+03	15.000	0
1	392.22	5.9061e-02	2.948e+03	15.000	0
2	393.08	5.0375e-03	2.832e+03	15.000	0
3	400.85	1.1688e-03	2.806e+03	15.000	0
4	383.37	5.1683e-04	2.795e+03	15.000	0
5	373.12	3.1379e-04	2.788e+03	15.000	0
6	375.35	2.3803e-04	2.782e+03	15.000	0
7	374.38	2.0697e-04	2.778e+03	15.000	0
8	379.45	1.8618e-04	2.774e+03	15.000	0
9	387.46	1.5602e-04	2.770e+03	15.000	0
10	376.71	1.3634e-04	2.767e+03	15.000	0
11	379.58	1.4564e-04	2.764e+03	14.013	7
12	388.35	4.1665e-04	2.755e+03	13.780	5
13	384.79	2.3641e-04	2.750e+03	13.724	4
14	379.59	1.6867e-04	2.746e+03	13.718	4
15	388.62	1.4912e-04	2.743e+03	13.731	3
16	385.62	1.4581e-04	2.740e+03	13.740	3
17	376.74	1.2549e-04	2.737e+03	13.739	3
18	397.51	1.1271e-04	2.735e+03	13.717	3
19	413.82	1.2796e-04	2.732e+03	13.710	3
20	400.59	1.0925e-04	2.730e+03	13.699	3
21	391.76	1.2777e-04	2.727e+03	13.667	3
22	402.83	8.1057e-05	2.725e+03	13.638	3

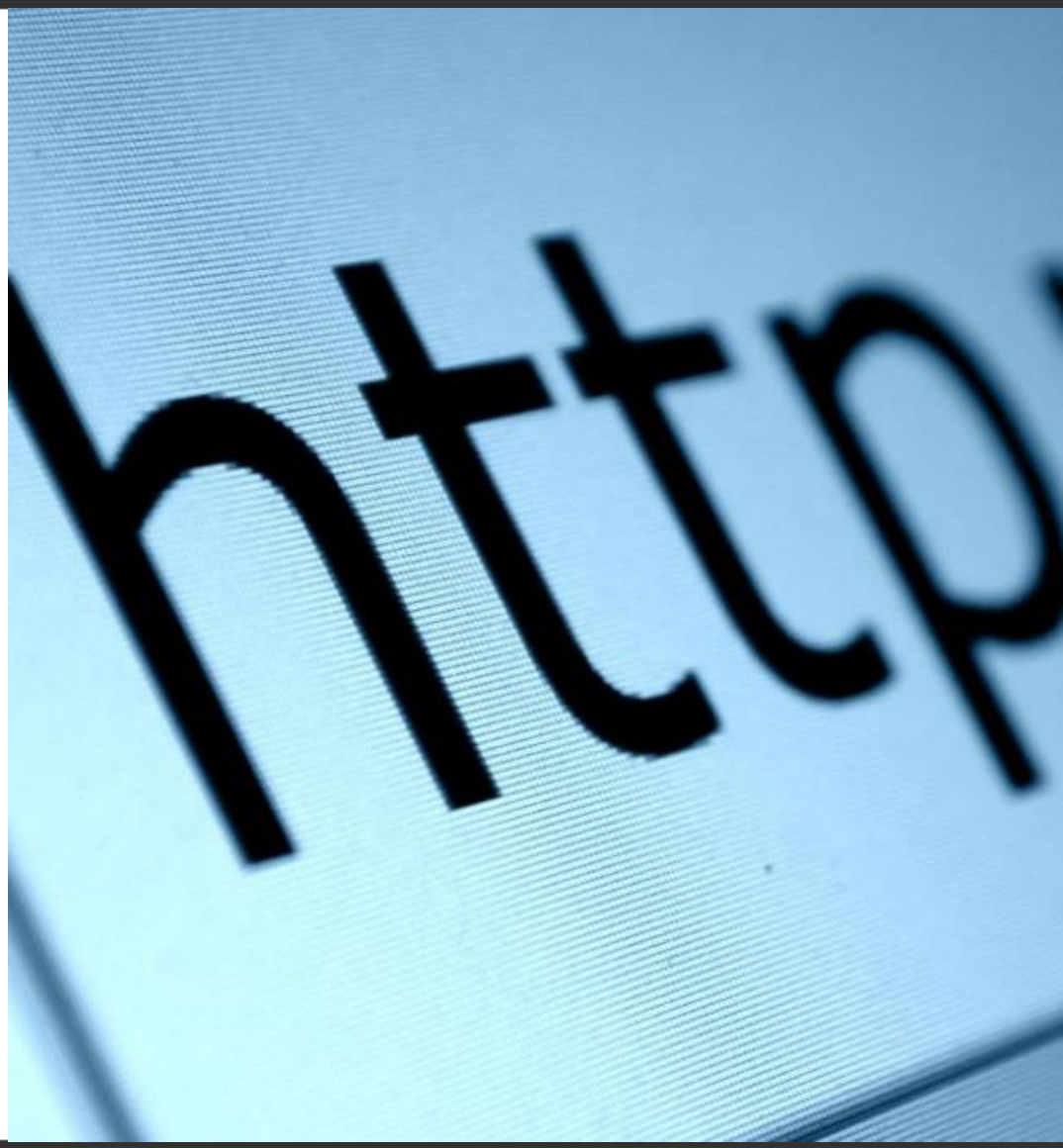
UNIVERSIDAD IBEROAMERICANA

Articles1.csv



UNIVERSIDAD IBEROAMERICANA

CÓDIGO FUENTE



UNIVERSIDAD IBEROAMERICANA

URL GITHUB

https://github.com/antoniochz/ibero