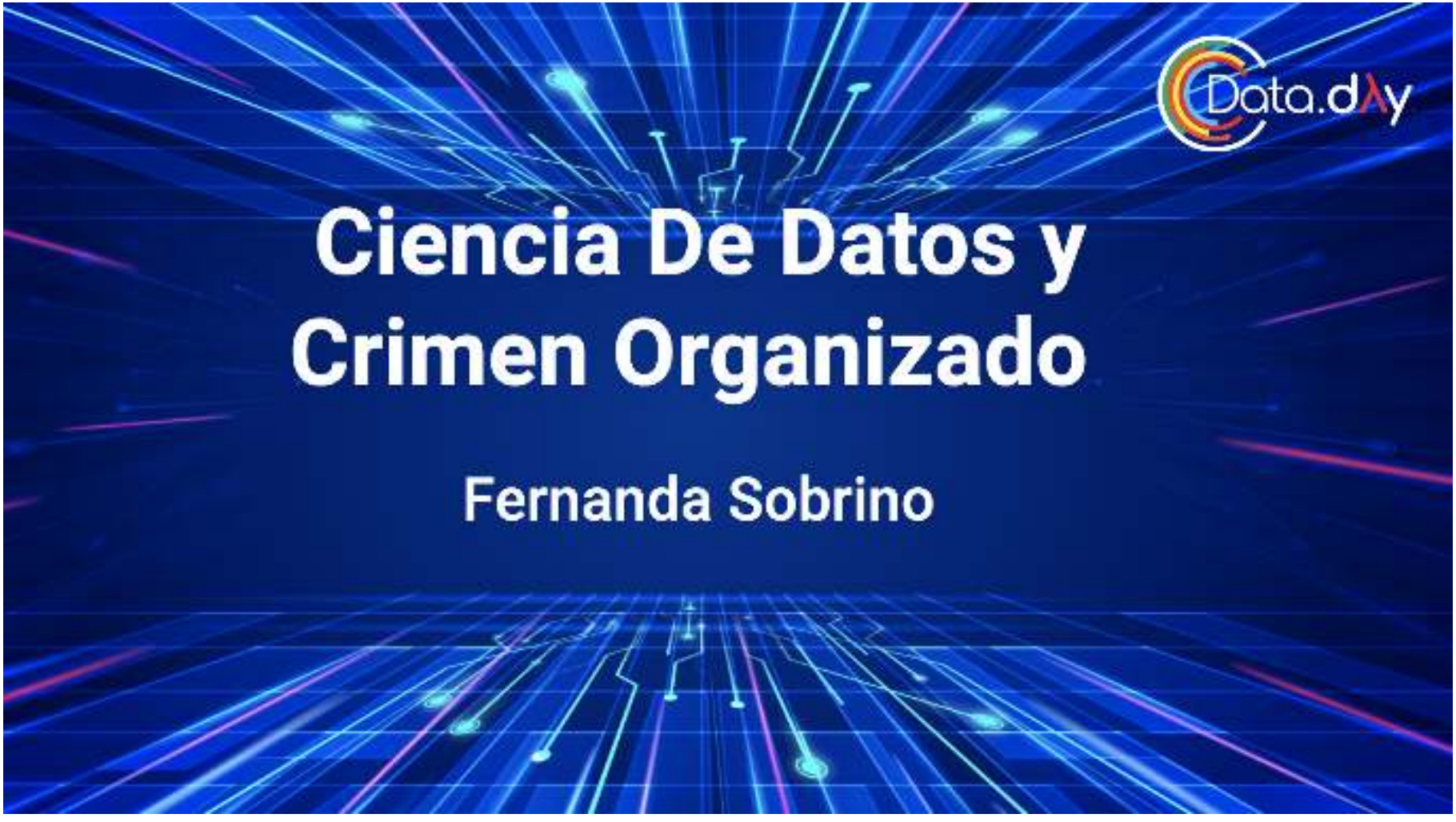
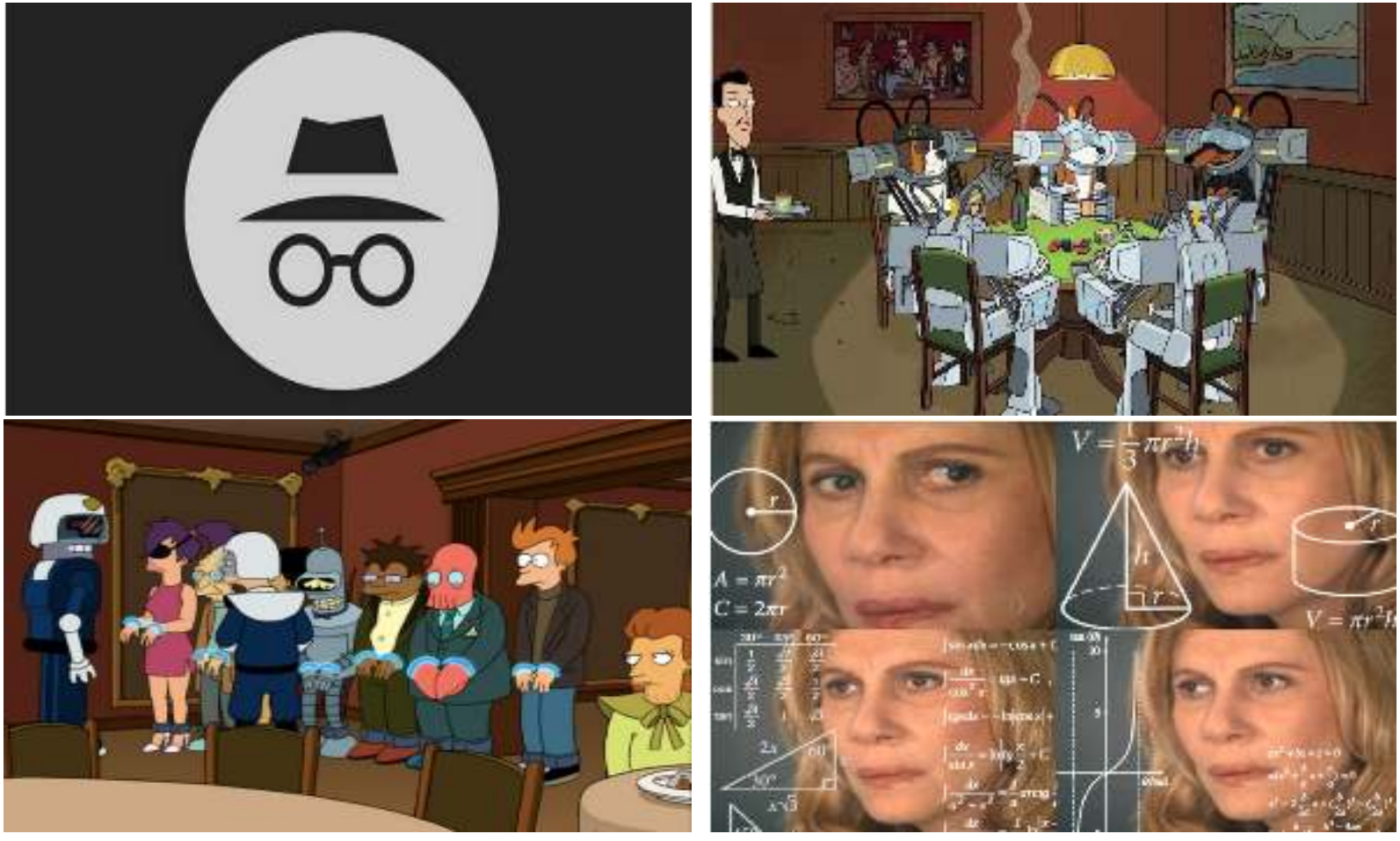

Data.dAy
Ciencia De Datos y
Crimen Organizado
Fernanda Sobrino

¿Qué es el crimen organizado?

Principales Actividades del Crimen Organizado:

Problemas con medir el crimen organizado:



Metodos alternativos para medir el crimen organizado

El caso del las organizaciones criminales en México

¿Qué haremos acá?

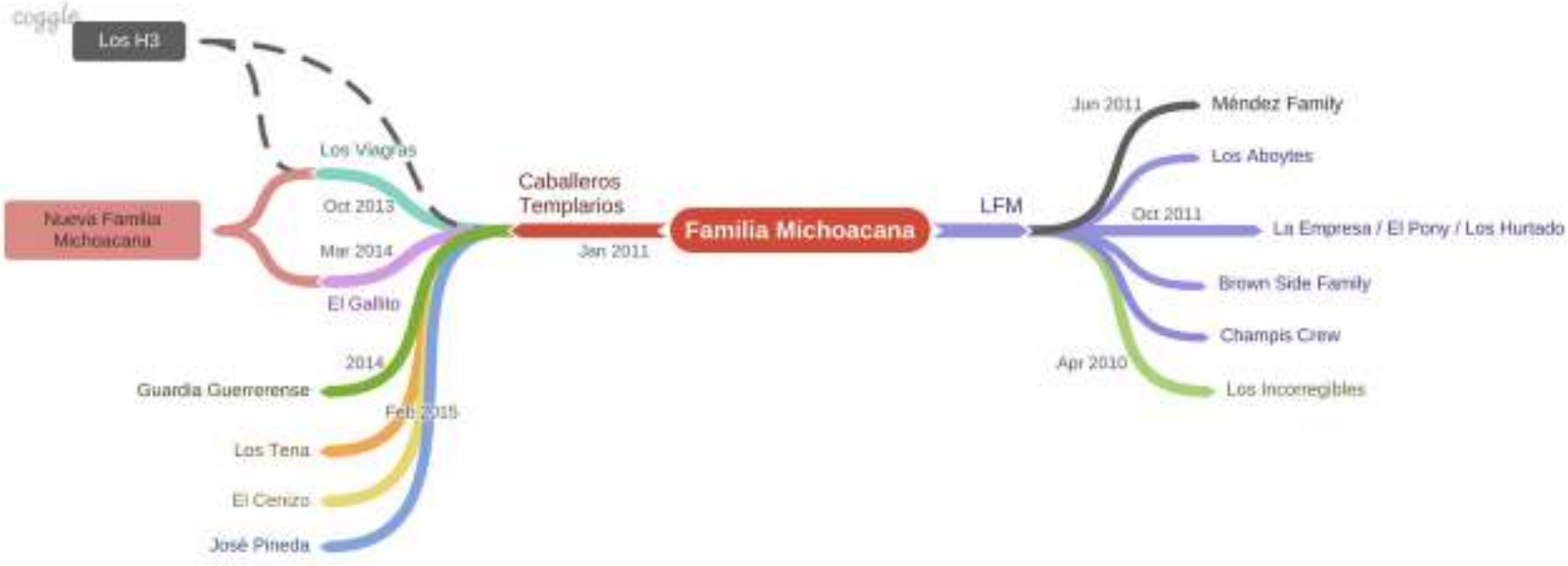
- Investigación cualitativa (nombres y evolución de los grupos)
- Web Scraping + NLP
  - Clasificación: noticias validas vs no válidas
  - Extracción de features: lugar, grupos

Nombres y evolución de los grupos

- 60 documentos oficiales de autoridades en USA y México
  - USA: DEA, Homeland Security, Treasury Office
  - México: PGR, Policia Federal, Ejercito y Marina
- 79 organizaciones criminales indepedientes hasta 2016

Nombres y evolución de los grupos: Beltrán Leyva

Nombres y evolución de los grupos: Familia Michoacana



Web Scraping

- Criterios de búsqueda:
  - 285 términos asociados a los 79 carteles
  - 3,198 asociados a los 2,456 municipios
  - 4 periodos: 1990-2004, 2005-2009, 2010-2015 y ¿2015
- 3'645,720 terminos de busqueda únicos

Google Bot


"Cartel de Sinaloa" "Badiraguato"
All
News
Videos
Images
Maps
More
Search tools
Tools
All news
Jan 1, 2006 – Dec 31, 2010
Sorted by relevance
Clear
El gran enemigo de "El Chapo"
Dec 17, 2008
BBC Mundo
BBC Mundo | América Latina | El millonario más buscado
Mar 12, 2009
Detienen en Calexico a supuesto Beltrán Leyva
Mar 28, 2009

Google Bot: Resultados

- 5'620,109 enlaces susceptibles de ser scrapeados; descartamos 674,413 por estar detrás paywalls
- NewsPlease procesó el 80% de estos enlaces.
- Usamos una heurística propia que utiliza el enlace para predecir la estructura del HTML para el restante 20%
- NewsPlease + heuristica lograron extraer el 95% de los textos
- Se extrajeron 4'698,411 textos; de estos, solo 2'137,341 están en español y contienen al menos una referencia a una organización criminal en el texto principal.

Textos válidos

Clasificación: motivación

- El narcotráfico es un tema común en los medios mexicanos
- Muchos artículos periodísticos mecionan grupos criminales y municipios pero no implica presecia
- Asignar presecia a cualquier artículo sobre-estimaria la presencia

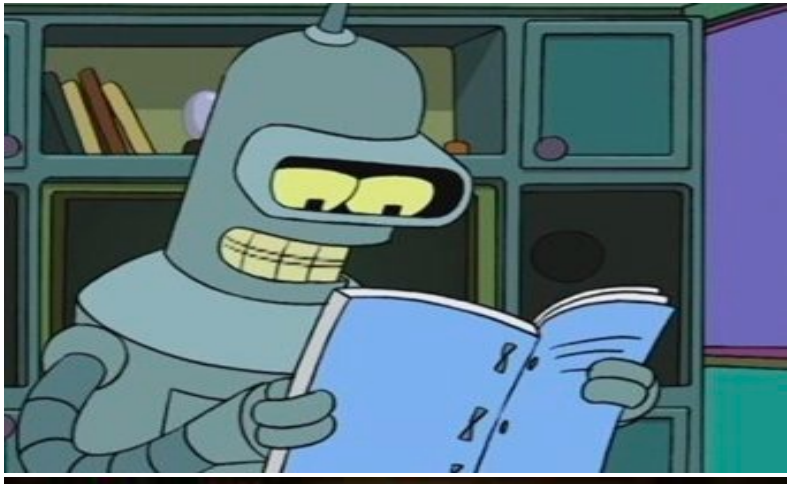
Ejemplos: no presencia

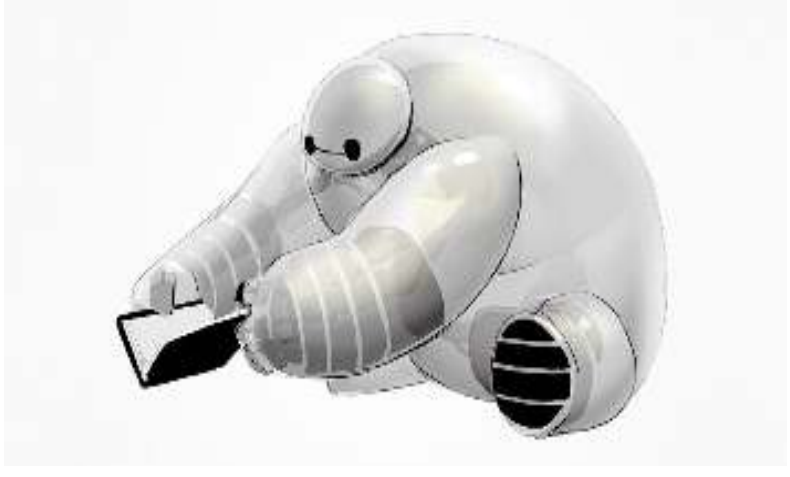
- "Una iglesia que venera a Jesus Malverde, tratado como santo patrono por los miembros del cartel de Sinaloa, fue cerrada en Pachuca Hidalgo" (La Silla Rota 2012)
- "Edgar Jimenez, un mexico-americano de 16 años de edad miembro del Cartel de Sinaloa, fue extraditado esta noche del centro de detencion juvenil en Cuernavaca a uno en El Paso, Texas" (Animal Politico 2013)
- "El pirata de Culiacan (una celebridad de Youtube) se burlo de Nemesio Oseguera Cervantes, lider del Cartel de Jalisco Nueva Generacion, en Baridaguato"(Cambio de Michoacan 2016)
- "Melissa Plancarte, cantante e hija de uno de los lideres de los Caballeros Templarios, grabo un video musical en Colon, Queretaro" (Diario Cambio 2014)

Ejemplos: presencia

- "El fiscal general de Nayarit ordeno una operacion que resulto en la captura de dos operadores de los Beltran-Leyva en Tepic"(Debate, 2016)
- "Mientras investigaban un caso de allanamiento la policia federal termino enfrentandose con miembros del Cartel de Jalisco Nueva Generacion" (SDP Noticias 2015)
- "De acuerdo a informes policiacos, la comunidad de Tancitaro se encuentra bajo el control de La Familia Michoacana desde 2004" (Proceso 2006)
- "El 16 de diciembre del 2009, la mrina detecto la presencia de Arturo Beltrán Leyva en la ciudad de Cuernavaca" (El Universal 2010)

Procesamiento del Lenguage Natural





Enseñarle a la maquina a leer

1. Clasificación manual
2. Pre-procesamiento del texto
3. Transformar el texto en algo que la máquina pueda leer
4. Escoger algoritmos para la tarea especifica: clasificación binaria
5. Entrenar los algorítimos
6. Probar, evaluar y escoger el mejor

Clasificación manual

1. Aleatoriamente se sacaron 11,000 parrafos
2. 3 personas clasificaron estos 11,000 (para poder romper empates)
3. 1 el parrofo implica presencia
4. 0 no implica presencia
5. La muestra esta balanceada 5500 son 1's y 5500 son 0's

Pre-procesamiento del Texto

- Dividimos los 2'137,341 textos en párrafos.
- Un párrafo es definido como un fragmento de texto de máximo 200 palabras debido a que la mayoría de los LLM se limitan a 512 tokens.
- Utilizamos una ventana de 50 tokens para mantener el contexto del texto original.
- Se procuró preservar la estructura general del texto, dividiendo los párrafos en puntos convenientes y evitando dejar fragmentos de oraciones.
- Resultando en 7'292,204 párrafos que mencionan al menos un cártel y un municipio.

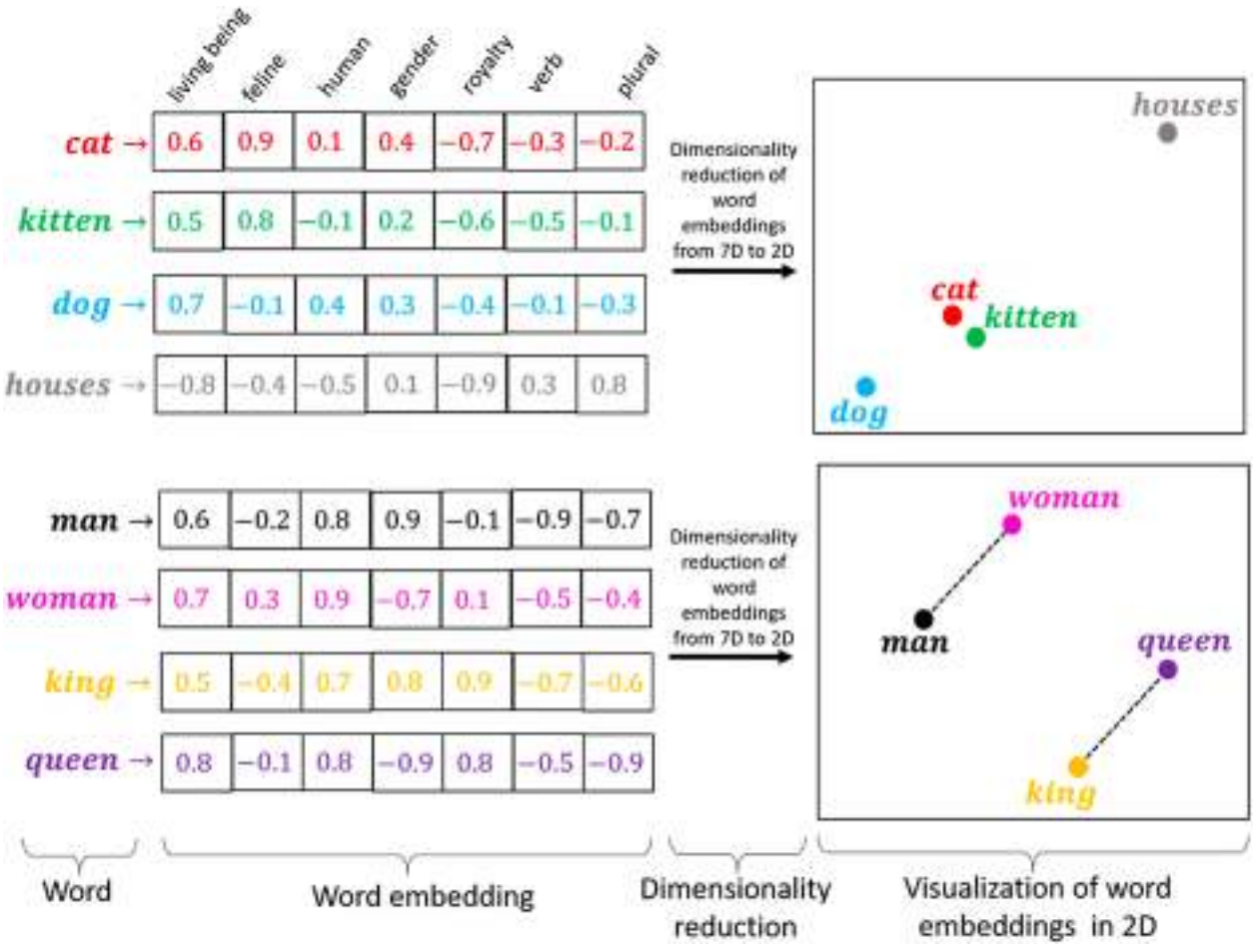
Transformar el parrafo en algo que la compu pueda leer

- binario : 1 si la palabra está
- frecuencia : cuantas veces aparece cada palabra
- hashing: convierte cada palabra en una representación numérica
- tf-idf: contiene información en las palabras mas y menos importantes de todos los textos
- Embeddings: el contexto es tomado en cuenta

Embeddings

- Representación vectorial
- Similitud semántica
- Reducción de dimensionalidad
- Ejemplos de algoritmos que calculan embeddings: Word2Vec, GloVe y FastText
- Transfer Learning
- Contextual Embeddings

Embeddings


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Algorítmos de clasificación

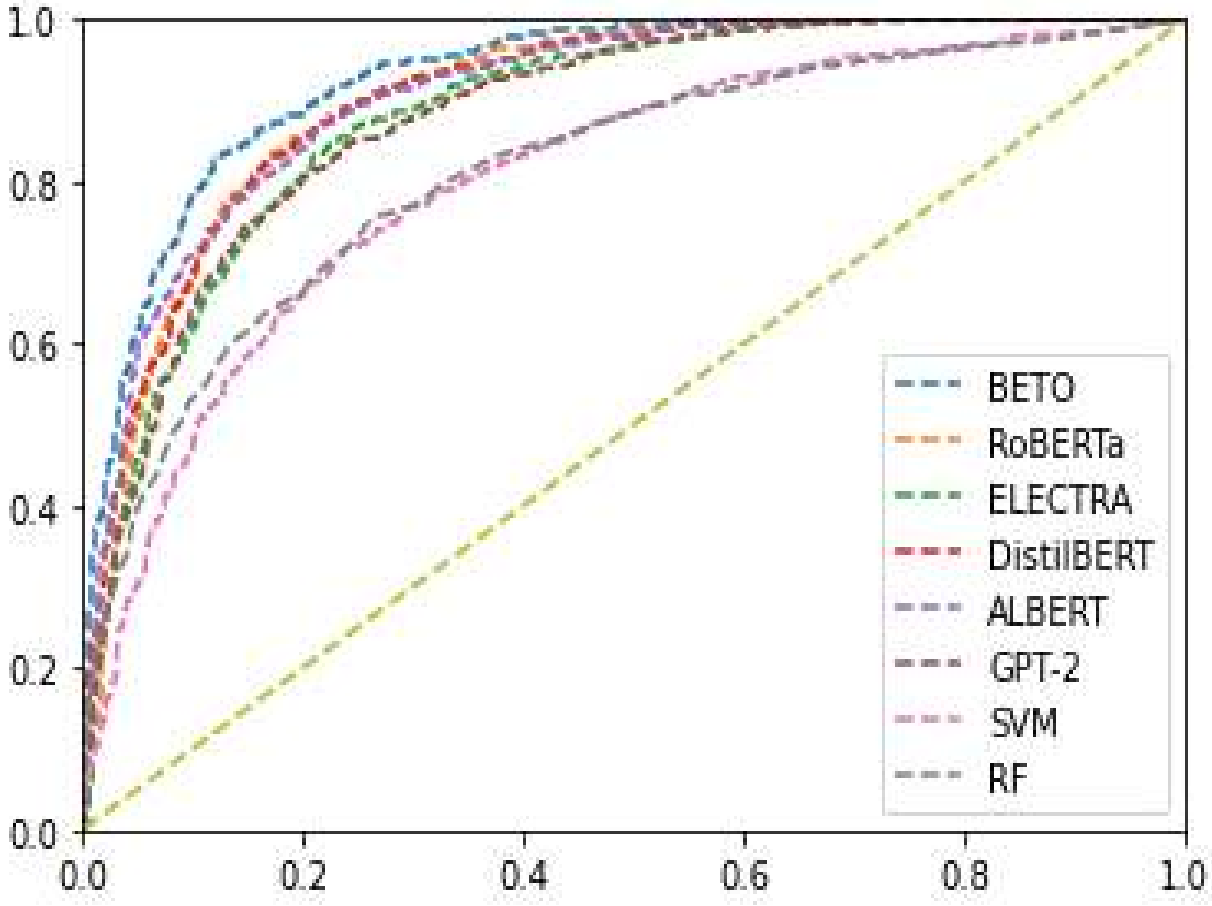
Modelos entrenados:

- Usando GloVe Embeddings en Español: SVM, RF, LR
- Transfer Learning: BETO, RoBERTa, DistilBERT, ALBERT, ELECTRA, y GPT-2 + Classification head
- Se optimizaron todos los modelos buscando en grids de hiper parámetros 50 por cada uno
- Metricas a comparar: accuracy, F1 y AUC

Metricas

Model	Accuracy	F1	AUC
SVM	0.72	0.75	0.80
RF	0.75	0.78	0.81
BETO	0.91	0.91	0.92
RoBERTa	0.84	0.84	0.90
DistilBERT	0.82	0.83	0.90
ELECTRA	0.79	0.81	0.88
ALBERT	0.83	0.81	0.90
GPT-2	0.79	0.79	0.88

Roc curves


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
BETO
RoBERTa
ELECTRA
DistilBERT
ALBERT
GPT-2
SVM
RF

Clasificación usando BETO

- El 68% de los parrafos son asignados como presencia 5 millones
- Reproduce la distribución original de los datos
- Ya terminamos? NO :(, falta asignar cada uno de los parrafos a municipios, carteles y un momento en el tiempo

Asignación a lugares y grupos del crimen organizado

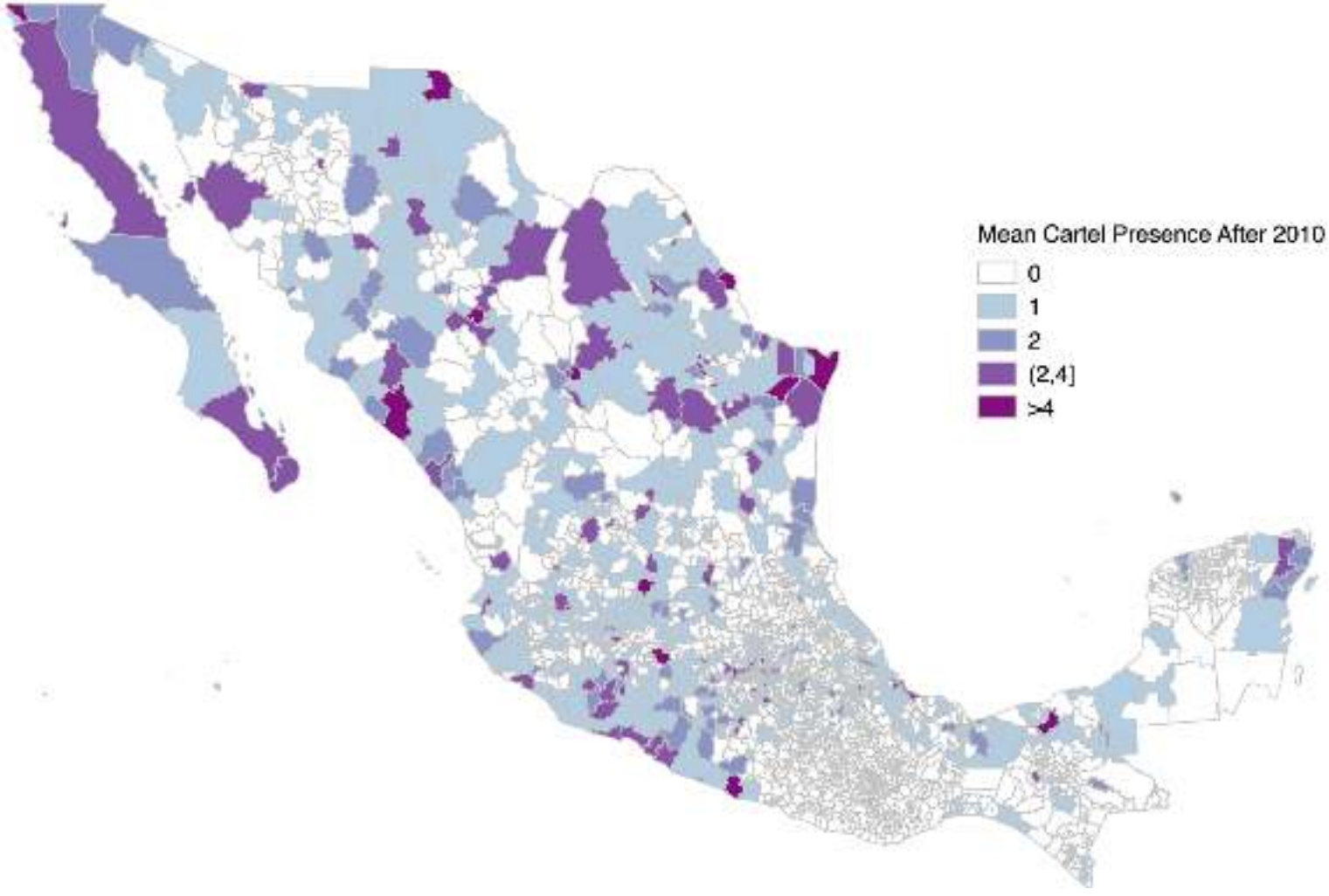
Una tarde de Difuntos MISC conoció al piloto a sueldo del cártel de Juárez ORG , en Culiacán LOC .

- Exact matches usando listas de municipios y carteles
- Flair NER sin fine-tuning
- Mil parrafos tageados a mano para medir cual lo hace mejor

Métricas Municipios y Cárteles

Model	Elemento	Precision	Recall	F1
Exact Matches	Municipios	0.93	0.71	0.81
Flair NER	Municipios	0.90	0.95	0.92
Exact Matches	Carteles	0.85	0.9	0.87
Flair NER	Carteles	0.92	0.57	0.70

Datos Finales



¿Cómo se compara con bases similares?

Externamente; comparando con datos oficiales agregados y otras bases de datos generadas con noticas

- DEA a nivel estatal 0.70
- Datos del Prof. Víctor Manuel Sanchez 0.78
- Coscia-Rios 0.38

Escuela de Gobierno y
Transformación Pública
Tecnológico de Monterrey

¡Gracias!

Fernanda Sobrino
Profesora-investigadora
@fersobrino

@egobiernoytp