

Introducción

Fernanda Sobrino

Ciencia de datos

Diferencias con la primera parte del curso

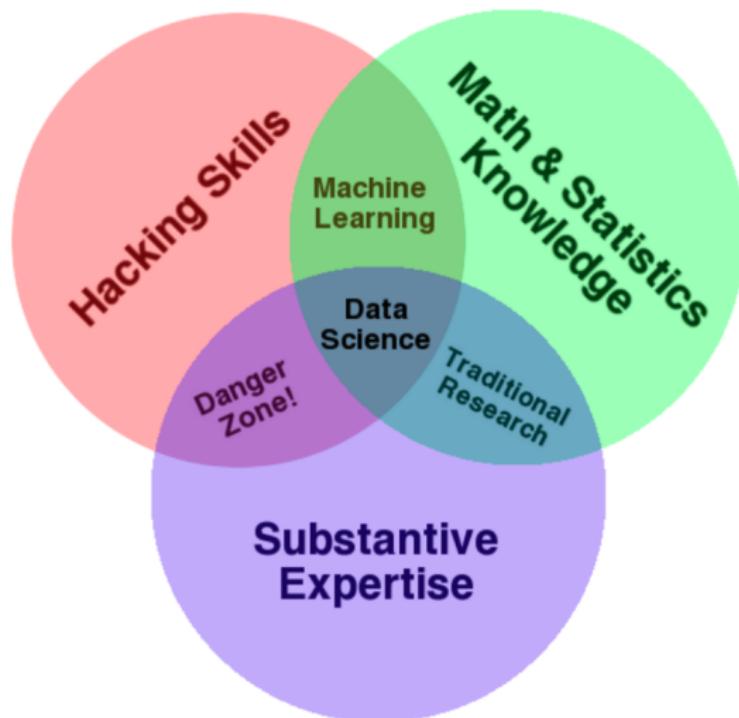
Sezgo algorítmico

Ejemplo de mi research

Descripción del curso

Ciencia de datos

Ciencia de datos?

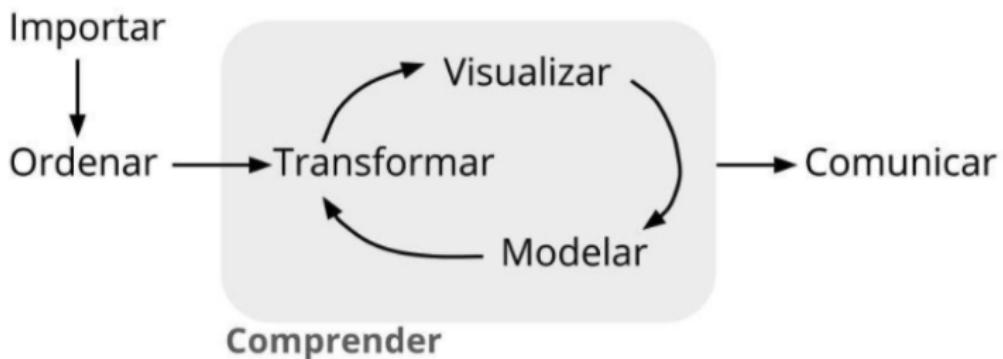


Qué es la ciencia de datos?

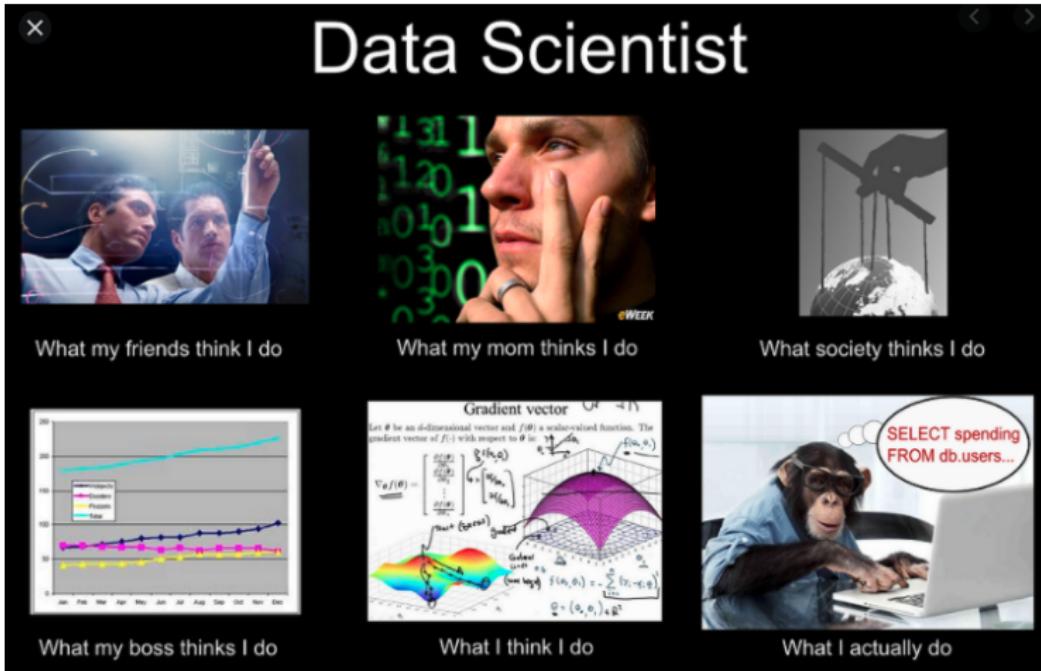
- ▶ aplicar técnicas de programación para analizar datos
- ▶ la 'ciencia de datos' no existía antes de 2008
- ▶ intersección entre:
 - ▶ programación
 - ▶ estadística
 - ▶ comunicación
 - ▶ conocimiento del dominio

Proceso de análisis de los datos

El proceso del análisis de datos

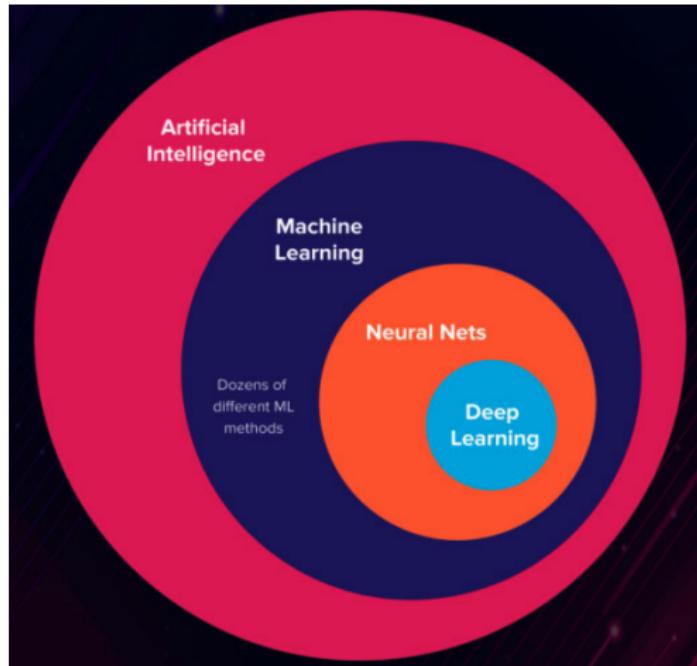


Científico de datos



Diferencias con la primera parte del curso

Diagrama de Inteligencia Artificial (AI) y Aprendizaje de Máquina (ML)



Inteligencia artificial

- ▶ el término existe desde los 50s
- ▶ describe nuestra lucha por construir una máquina capaz de imitar y retar la inteligencia humana
- ▶ Los primeros AI:
 - ▶ programas de computadoras basados en reglas muy sencillas capaces de resolver problemas complejos
 - ▶ incapaces de aprender y dependían demasiado de la aportación humana

Inteligencia artificial

- ▶ Actualmente:
 - ▶ capaces de aprender
 - ▶ rendimiento por encima de los humanos en algunos casos (reducidos)
 - ▶ a pesar de esto todavía están muy lejos de competir con la inteligencia humana
 - ▶ incapaces de aprender con pocos ejemplos
 - ▶ incapaces de trasladar conocimiento de un campo a otro

AI

People with no idea about AI
saying it will take over the world:

My Neural Network:



Diferencia entre AI y ML

- ▶ AI: capacidad de las computadoras de mostrar un comportamiento 'inteligente'
 - ▶ ejemplo: capacidad de jugar ajedrez
- ▶ ML: técnica que se utiliza para crear y mejorar ese comportamiento 'inteligente', permite a la máquina mejorar con la experiencia
 - ▶ ejemplo: técnica de entrenamiento para aprender a jugar ajedrez

Deep Learning

- ▶ subconjunto de ML
- ▶ al igual que ML deja a la máquina aprender de los datos pero marca un hito en la evolución de AI
- ▶ desarrollado usando nuestro entendimiento de las redes neuronales
- ▶ ML le debe su auge a la gran cantidad de datos que producimos y almacenamos desde ~2000s
- ▶ deep learning se lo debe a que el poder computacional es más barato
- ▶ usa capas para poder aprender mas y representaciones mas complejas de los datos
- ▶ existen muchos algoritmos de deep learning (mas de esto en las slides de redes neuronales)

Ejemplos de usos de deep learning

- ▶ autos que se conducen solos
- ▶ procesamiento natural del lenguaje (NLP)
- ▶ reconocimiento de imágenes (computer vision)
- ▶ asistentes virtuales
- ▶ entretenimiento

Guion de Batman escrito por deep learning

- ▶ guion escrito por un bot especialista en escribir guiones
- ▶ leyo el equivalente a ver 1000 horas de películas de batman

BATMAN

INT. TRADITIONAL BATCAVE

BATMAN stands next to his batmobile and uses his batcomputer.
He's sometimes Bruce Wayne sometimes Batman. Alltimes orphan.

BATMAN

This is now a safe city. I have
punched a penguin into prison.

ALFRED, Batman's loyal batler, carries a tray of goth ham.

ALFRED

Eat a dinner, Mattress Wayne.

An explosion explodes. THE JOKER and TWO-FACE enter the cave.
Joker is a clown but insane. Two-Face is a man but attorney.

BATMAN

No! It is Two-Face and One-Face.
They hate me for being a bat.

Batman throws Alfred at Two-Face. Two-Face flips Alfred like
a coin. Alfred lands heads up which means Two-Face goes home.

BATMAN (CONT'D)

It is just you and I, the Joker.
Bat versus clown. Moral enemies.

Guion de Batman escrito por deep learning

THE JOKER

I am such a freak. Society is bad.
You drink water, I drink anarchy.

BATMAN

I drink bats just like a bat would!

Batman looks around for his parents, but they are still dead.
This makes him have anger. He fires a batrocket. The Joker
deflects it with his sick sense of humor. A clownly power.

THE JOKER

I have never followed a rule. That
is my rule. Do you follow? I don't.

BATMAN

Alfred, give birth to Robin.

Alfred begins the process since it is his job. The Joker now
has a present in his hand. He juggles it over to Batman.

THE JOKER

Happy batday, Birthman.

Batman opens the present since he's a good guy. It contains a
coupon for new parents, but is expired. This is a Joker joke.

Sezgo algorítmico

Qué es?

- ▶ describe errores sistemáticos y repetibles en un sistema computacional que genera resultados injustos
- ▶ este sesgo puede venir de muchos lados
 - ▶ el diseño del algoritmo
 - ▶ el uso del algoritmo
 - ▶ los datos usados para entrenarlo
- ▶ podemos encontrar estos sesgos en todos lados: buscadores, redes sociales
- ▶ pueden y corren el riesgo de reforzar sesgos sociales existentes

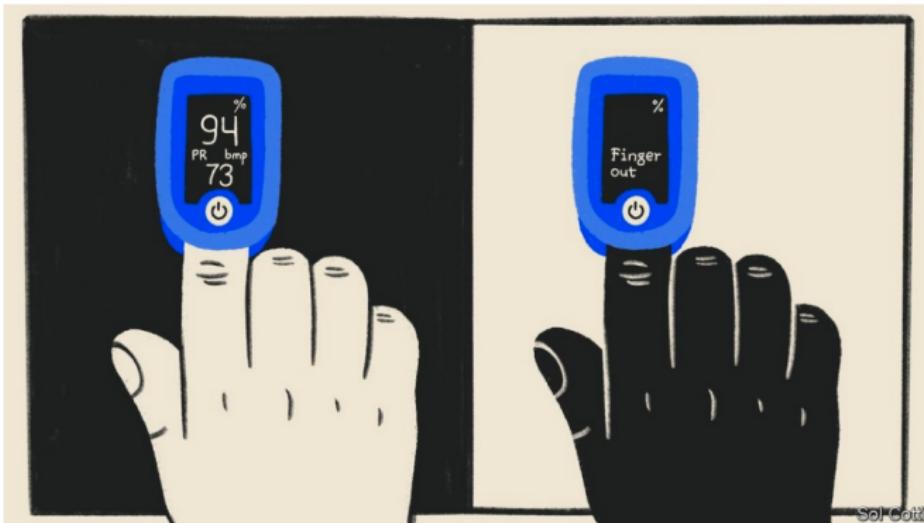
Por qué es peligroso?

- ▶ hay AI en todos los sectores comerciales: transporte, ventas, energía, publicidad, etc
- ▶ también están empezando a afectar la democracia y los gobiernos
- ▶ antes del boom del AI las compañías y los gobiernos tomaban decisiones siguiendo la ley
- ▶ actualmente muchas de estas decisiones son tomadas por algoritmos

Ejemplo: COMPAS

- ▶ algoritmo usado en USA para guiar las sentencias de los criminales
- ▶ predice la probabilidad de que un criminal vuelva a delinquir
- ▶ en 2016 ProPublica denuncio que el algoritmo tiene un sesgo racial
- ▶ el sistema predice que los acusados afro-americanos tienen una mayor probabilidad de reincidir
- ▶ el algoritmo es de una compañía privada entonces no sabemos de donde salio el bias
- ▶ su capacidad de predecir es tan buena/mala como voltear una moneda

Ejemplo: oxímetros y el tono de piel



Ejemplo: Facebook y la policía

- ▶ en 2017 un hombre palestino fue arrestado en Israel
- ▶ su crimen?
- ▶ una foto en facebook posando a lado de un tractor con la frase 'ataquenlos'
- ▶ o al menos eso tradujo el NLP de facebook
- ▶ el hombre escribió 'buenos días' en árabe y el algoritmo confundió las palabras
- ▶ el hombre fue interrogado por horas hasta que alguien se dio cuenta del error

Por qué ocurren estos sesgos? datos defectuosos

- ▶ prejuicios humanos
 - ▶ los datos históricos van a reproducir patrones de racismo, sexism etc
 - ▶ algoritmos entrenados con ellos van a incurrir en los mismos juicios incorrectos que los humanos
 - ▶ el problema: estos sesgos van a ser reproducidos por el algoritmo sin que el usuario lo sepa
 - ▶ los algoritmos puede que jamás acumulen suficientes ejemplos que cambien el juicio pre establecido a los sectores menos favorecidos
 - ▶ bucles negativos de retro-alimentación

Por qué ocurren estos sesgos? datos defectuosos

- ▶ datos incompletos o poco representativos
 - ▶ si tienes más ejemplos de entrenamiento de un grupo en específico
 - ▶ esto puede deberse a la falta de diversidad en los programadores
 - ▶ algoritmos con demasiados datos o sobre representación también son un problema

Cómo detectamos estos sesgos?

- ▶ ser transparente en el uso de datos sensibles (sexo, raza, pertenencia a ciertos grupos, etc)
- ▶ estar conscientes que los algoritmos van a reproducir problemas históricos que pueden tratar de manera injusta a ciertos grupos
- ▶ comparar resultados entre grupos
- ▶ simular las predicciones de los modelos (en distintas sub poblaciones) antes de usarlos
- ▶ evaluar las nociones de justicia y los costos sociales

Cómo detectamos estos sesgos?

- ▶ buscar la manera de reducir las disparidades entre grupos sin sacrificar el rendimiento del modelo
- ▶ por ejemplo: bases de datos con baja representación necesitan mas entrenamiento
- ▶ regulación y legislación: la Unión Europea es la pionera en esto publicando “Guía para Inteligencia Artificial Ética”
- ▶ actualizar leyes
- ▶ transparencia por parte de empresas que usan estos algoritmos
- ▶ diversidad en los equipos de decisión
- ▶ incrementar el monitoreo del diseño y uso de ML

Ejemplo de mi research

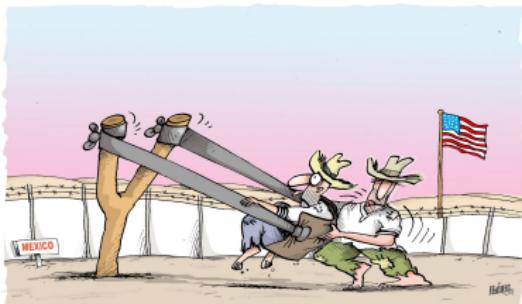
Data Science Workflow

- ▶ hacerse una pregunta
- ▶ buscar/generar datos
- ▶ entender/manipular los datos
- ▶ modelar
- ▶ evaluar
- ▶ implementar/inferencia

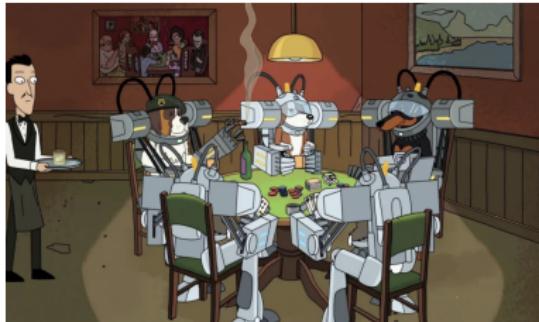
Qué es el crimen organizado?



Principales actividades del Crimen Organizado



Problemas con medir el crimen organizado



Diagrams and formulas shown:

- A circle with radius r .
- A cylinder with radius r and height h , formula: $V = \pi r^2 h$.
- A cone with radius r and height h .
- A cylinder with radius r and height h , formula: $V = \pi r^2 h$.
- Trigonometric ratios for a right triangle with angles 30°, 45°, and 60°:
 - $\sin 30^\circ = \frac{1}{2}$, $\cos 30^\circ = \frac{\sqrt{3}}{2}$, $\tan 30^\circ = \frac{\sqrt{3}}{3}$
 - $\sin 45^\circ = \frac{\sqrt{2}}{2}$, $\cos 45^\circ = \frac{\sqrt{2}}{2}$, $\tan 45^\circ = 1$
 - $\sin 60^\circ = \frac{\sqrt{3}}{2}$, $\cos 60^\circ = \frac{1}{2}$, $\tan 60^\circ = \sqrt{3}$
- Integration formulas:
 - $\int \sin x dx = -\cos x + C$
 - $\int \frac{dx}{\cos^2 x} = \tan x + C$
 - $\int x \cos x dx = -x \sin x + \ln|\cos x| + C$
 - $\int \frac{dx}{\sin^2 x} = \ln|\tan x| - \frac{\pi}{2} + C$
 - $\int \frac{dx}{a^2 + x^2} = \frac{1}{a} \arctan(\frac{x}{a}) + C$
 - $\int \frac{dx}{a^2 - x^2} = \frac{1}{2a} \ln\left|\frac{a+x}{a-x}\right| + C$
- Trigonometric identities:
 - $\sin^2 x + \cos^2 x = 1$
 - $a^2 \tan^2 x + a^2 = a^2 \sec^2 x$
 - $a^2 + 2 \frac{a}{\sin^2 x} x + \frac{a^2}{\sin^2 x} = a^2 \csc^2 x$
 - $(x + \frac{a}{\sin x})^2 - \frac{a^2}{\sin^2 x} = 0$

Métodos alternativos para medirlo



El caso del narcotráfico en México



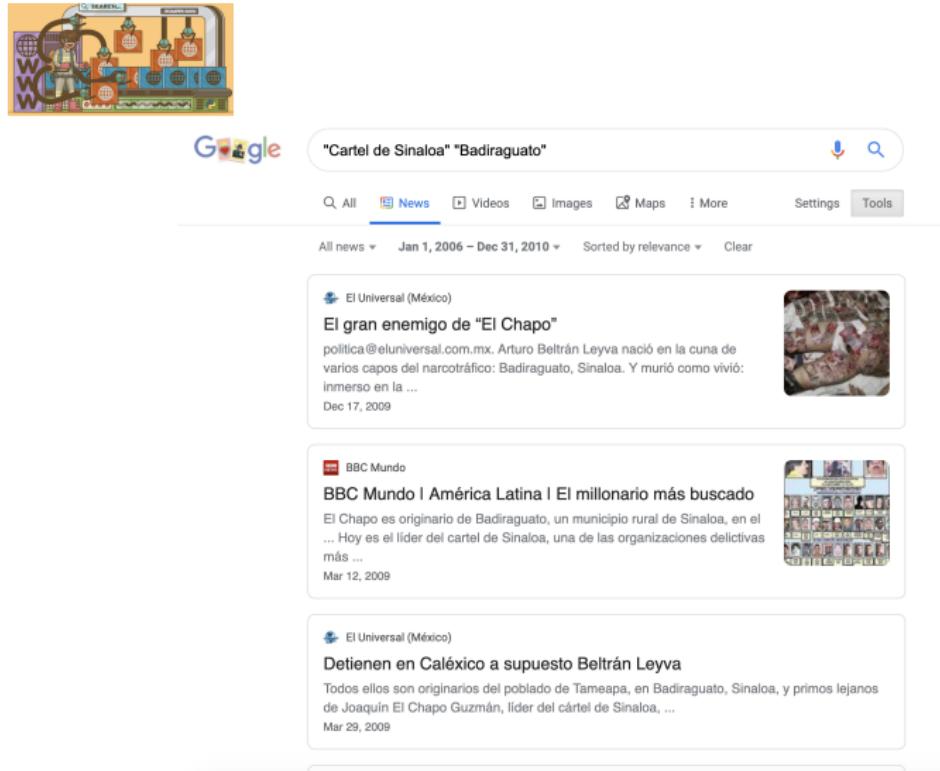
MAPA DEL NARCOTRÁFICO
De acuerdo con la PGR, en el país operan nueve carteles de la droga con presencia en 18 estados.



Qué hacemos?

- ▶ Google News + NLP
- ▶ web scraping de Google News buscando los 9 carteles principales en cada municipio de México
- ▶ enseñarle a la máquina a leer oraciones para distinguir entre
 - ▶ oraciones que implican presencia de un cartel en un lugar
 - ▶ oraciones que mencionan al cartel y al lugar pero no implican presencia

Conseguir datos



A Google search results page for the query "Cartel de Sinaloa" "Badiraguato". The results are filtered to show news articles from January 1, 2006, to December 31, 2010, sorted by relevance. The first result is from El Universal (México) titled "El gran enemigo de "El Chapo"" with a snippet about Arturo Beltrán Leyva. The second result is from BBC Mundo titled "BBC Mundo | América Latina | El millonario más buscado" with a snippet about El Chapo. The third result is from El Universal (México) titled "Defienden en Caléxico a supuesto Beltrán Leyva" with a snippet about Joaquín El Chapo Guzmán.

SEARCH

"Cartel de Sinaloa" "Badiraguato"

Q All News Videos Images Maps More Settings Tools

All news Jan 1, 2006 – Dec 31, 2010 Sorted by relevance Clear

El Universal (México)
El gran enemigo de "El Chapo"
politica@eluniversal.com.mx. Arturo Beltrán Leyva nació en la cuna de varios capos del narcotráfico: Badiraguato, Sinaloa. Y murió como vivió: inmerso en la ...
Dec 17, 2009

BBC Mundo
BBC Mundo | América Latina | El millonario más buscado
El Chapo es originario de Badiraguato, un municipio rural de Sinaloa, en el ... Hoy es el líder del cartel de Sinaloa, una de las organizaciones delictivas más ...
Mar 12, 2009

El Universal (México)
Defienden en Caléxico a supuesto Beltrán Leyva
Todos ellos son originarios del poblado de Tameapa, en Badiraguato, Sinaloa, y primos lejanos de Joaquín El Chapo Guzmán, líder del cártel de Sinaloa, ...
Mar 29, 2009

Conseguir datos

19

30 de octubre de 2019

Detienen a 11 narcos al catear 3 casas en el DF; 23 fuga de Torreón
El Siglo de Torreón



Entrar

NACIONAL

Detienen a 11 narcos al catear 3 casas en el DF

Agencias/MÉXICO, DF. miércoles 23 de enero 2008, actualizada 8:42 am



Revelan reacomodos del Cártel de Sinaloa en BCS; domina plazas de droga a través de 4 grupos

0 15 diciembre, 2015



Ultima Noticia

- Fuerza Nacional fuerza operativo contra el Uber y en persecución de La Pug y Los Cáticos Federales
- Continúan bajando las temperaturas en BCS; prevén el inicio de 14 °C y máximas de 32 °C
- Los obdulios de Morena en BCS son lo que protege a la vendimia mafía del poder: Gobernador
- Hizo la Cuenta Transversal; exigimos aplicar la Ley del Transporte en BCS: taxistas
- Arrigó objetos a personas e insultó a los Oficiales y se lo llevó, advierte SSP BCS por Halloween



Conseguir datos

Forbes

Portfolio (<https://sites.ust.hk/~mcsu/Portfolio.html>) / Further Politics (<https://sites.ust.hk/~mcsu/FurtherPolitics/>)

Del Cártel de Sinaloa a la Unión Tepito: los 10 grupos del narco en México



(<http://www.fernando-carrillo.com.mx/actualidad/ciudad-mexico/2013/03/20/los-ultimo-trimestre-los-grupos-del-narco-en-mexico-https://www.fernando-carrillo.com.mx/noticias/los-ultimo-trimestre-los-grupos-del-narco-en-mexico/>)

8

EFE. - Con la captura de José Víctor Gómez "el blanquito", líder del pandillero CárTEL Santa Rosa de Lima, el gobernador de Arequipa Manuel López Obrador dio su mayor golpe al narcotráfico, aunque las autoridades de la druga siguen sembrando la violencia a lo largo y ancho de Méjico.

Hay al menos una docencia de cárteles internacionales, como veímos que se respondió a Estados Unidos, Corea del Sur, Rusia y Europa, y como testimonia las cárteles criminales.

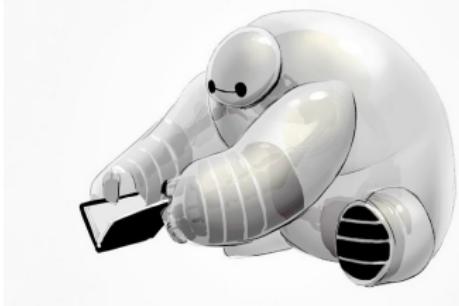
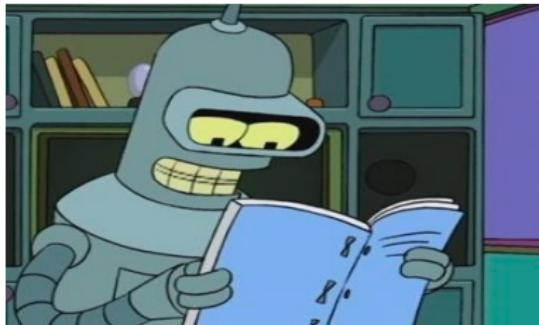
Oraciones que no implican presencia

- ▶ “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)

Oraciones que implican 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 marina detecto la presencia de Arturo Beltrán Leyva en la ciudad de Cuernavaca” (El Universal 2010)

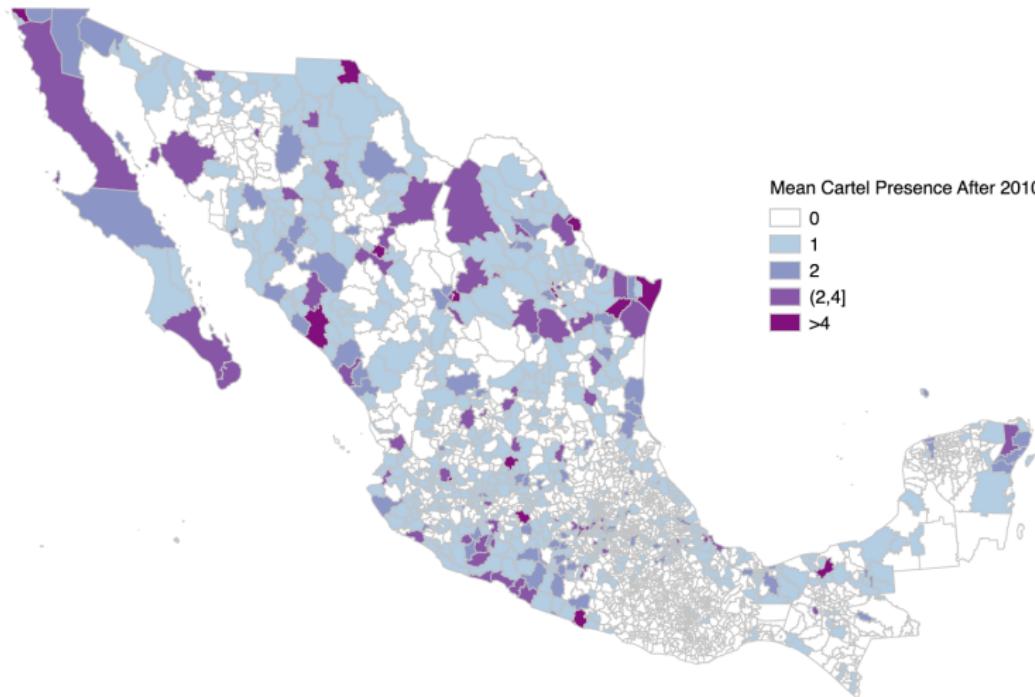
Procesamiento natural del lenguaje



Enseñarle a leer a la máquina

- ▶ Transformar la oración en algo que la maquina pueda leer
- ▶ Escoger un algoritmo para la tarea específica
- ▶ Clasificación manual de algunas oraciones
- ▶ Entrenar el algoritmo
- ▶ Probar y evaluar como funciona

Resultado



Descripción del curso

Nuevo outline

1. Redes Neuronales
2. Introducción al Procesamiento Natural del Lenguaje
 - ▶ texto como data
 - ▶ palabras a vectores básico
3. Web Scraping
4. Aprendizaje no supervisado
5. Método de NLP mas avanzados
 - ▶ LDA
 - ▶ modelos del lenguaje
 - ▶ state of the art models