

A decorative graphic of circuit lines in a dark blue color, framing the central text. The lines start from the bottom, go up the sides, and then curve back down to the bottom, with small circles at various points along the paths.

presentado en

Pista 2

Lo esencial sobre datos para periodistas que cubren la inteligencia artificial

Lo que veremos en este módulo

- Una breve e incompleta historia de la evolución de los datos
- Los datos en el ámbito de la IA
- Tres ejemplos:
 - Datos estructurados que alimentan algoritmos en la toma de decisiones
 - Historial de comportamiento que alimentan tus feed en las redes sociales
 - Datos de texto para grandes modelos lingüísticos

Comencemos con un ejercicio



¿Qué tipo de datos hay?

Analicemos un tuit.

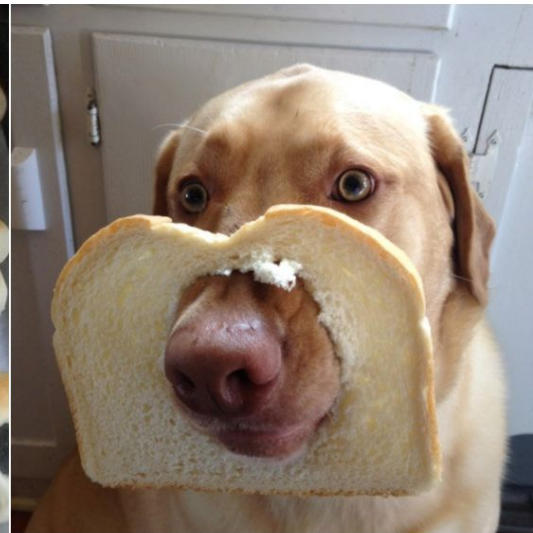
1. Lea esto.
2. Ríanse un poco.
3. Luego díganme qué tipo de datos puede generar este tuit.



Yael
@elle91

Follow

Purebread dogs vs. inbread dogs



RETWEETS
1,446

LIKES
2,758



12:25 PM - 22 Feb 2017

22

1.4K

2.8K

Persona que lo publicó:

- Enlace a la imagen
- Nombre de usuario
- Nombre de usuario de Twitter

Texto

Imágenes

Interacciones:

- Retuits
 - Me gusta
 - Comentarios
- Fecha y hora
(más abajo)

Marca de tiempo



Yael
@elle91

+ Follow

Purebread dogs vs. inbread dogs



RETWEETS
1,446

LIKES
2,758



12:25 PM - 22 Feb 2017



22



1.4K



2.8K

El mismo tuit,
¡formato diferente!
Todo el sitio X.com se
alimenta de flujos de
datos como este.

APIs o interfaz de
programación de
aplicaciones*

 *palabra elegante
para referirse a un flujo
de datos en un formato
que los robots/
computadoras pueden
entender 

```
1  {
2    "contributors": null,
3    "truncated": false,
4    "text": "Purebred dogs vs. inbred dogs https://t.co/e1nC0d7EkR",
5    "is_quote_status": false,
6    "in_reply_to_status_id": null,
7    "id": 834454023260532737,
8    "favorite_count": 2760,
9    "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for
10   * Android</a>",
11    "retweeted": false,
12    "coordinates": null,
13    "entities": {
14      "symbols": [],
15      "user_mentions": [],
16      "hashtags": [],
17      "urls": [],
18      "media": [ {
19        "expanded_url": "https://twitter.com/elle91/status/834454023260532737/photo/1",
20        "display_url": "pic.twitter.com/e1nC0d7EkR",
21        "url": "https://t.co/e1nC0d7EkR",
22        "media_url_https": "https://pbs.twimg.com/media/C5STfk1WQAEzRet.jpg",
23        "id_str": "834454001643044865",
24        "sizes": {
25          "large": {
26            "h": 413, "resize": "fit", "w": 550
27          },
28          "small": {
29            "h": 413, "resize": "fit", "w": 550
30          },
31          "medium": {
32            "h": 413, "resize": "fit", "w": 550
33          }
34        }
35      } ]
36    }
37  }
```



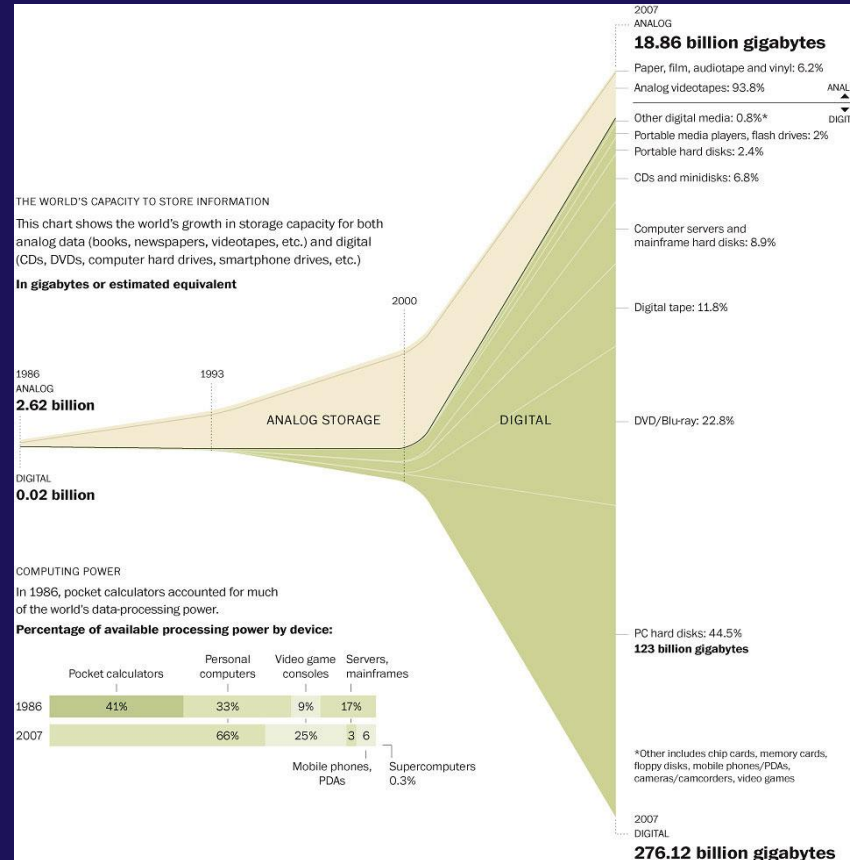
¡Los datos están en todas partes!

Internet se alimenta de datos

A decorative graphic resembling a circuit board or a stylized letter 'U' is drawn in a dark blue color on a yellow background. It features several vertical and horizontal lines with small circles at the ends, suggesting solder points or data nodes. The pattern is symmetrical and frames the central text.

¿Qué está pasando
con los datos?

Cada vez hay más



Fuente: Washington Post (2011)

Ahora hay mucho más que solo datos gubernamentales



Fuente: IBM via Time



Fuente: CNET

Ejemplo de huella digital: WSJ



How Pizza Night Can Cost More in Data Than Dollars

Even a low-key evening at home can mean handing over a trove of personal information to high-tech companies

By Stephanie Stamm, Tripp Mickle and Jessica Kuronen

Published April 10, 2018 at 5:30 a.m. ET

The smartphones, Facebook accounts and other technology products deeply embedded in modern life help people get more things done every day. They also gather more information about us than we often realize.

But Facebook's crisis [over how it handles and protects user data](#) has led some to ask: What data am I giving up?

Imagine "Sally" sets up a pizza-and-movie night with her friend "Kristen." The Wall Street Journal reviewed privacy statements to assess just how much data could be unknowingly shared on top of the price of that pepperoni pie.

El costo en datos

Sally y Kristen posiblemente entregaron al menos 53 datos personales en conjunto. La información descrita refleja los datos que las empresas podrían recopilar según sus declaraciones de privacidad, términos de servicio y documentos relacionados.

Data points collected in this scenario

15 are user-provided (28%)



Fuente: [WSJ](#)

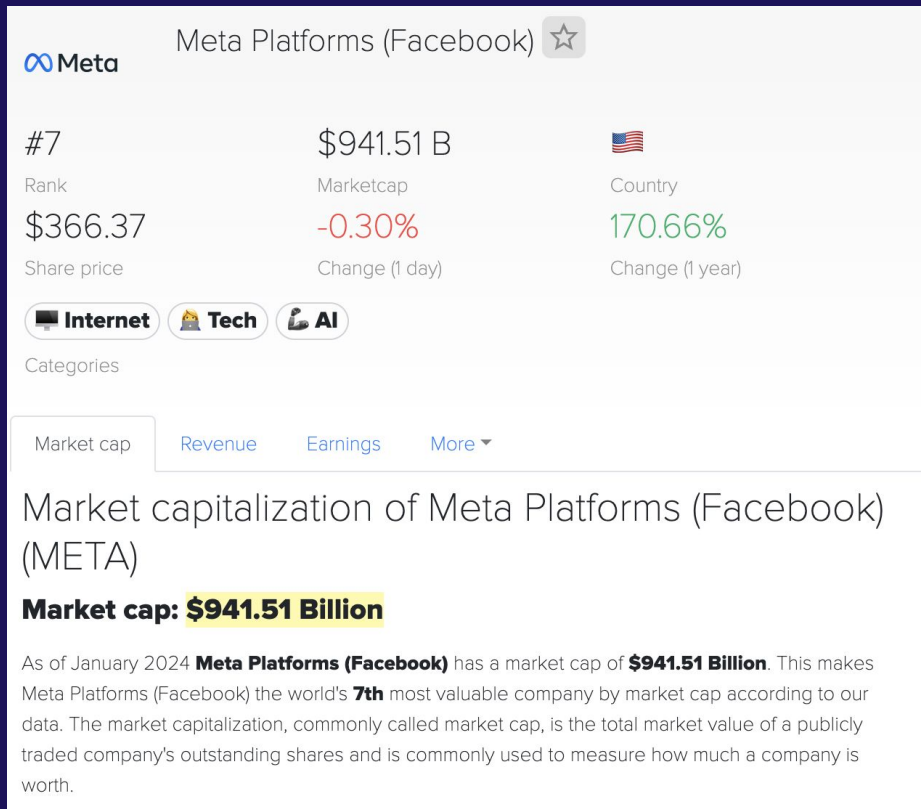
Ejemplo de huella digital: servidores de metadatos



Vista aérea de algunos de los centros de datos de Facebook

Big Data

Está haciendo ganar mucho dinero a mucha gente

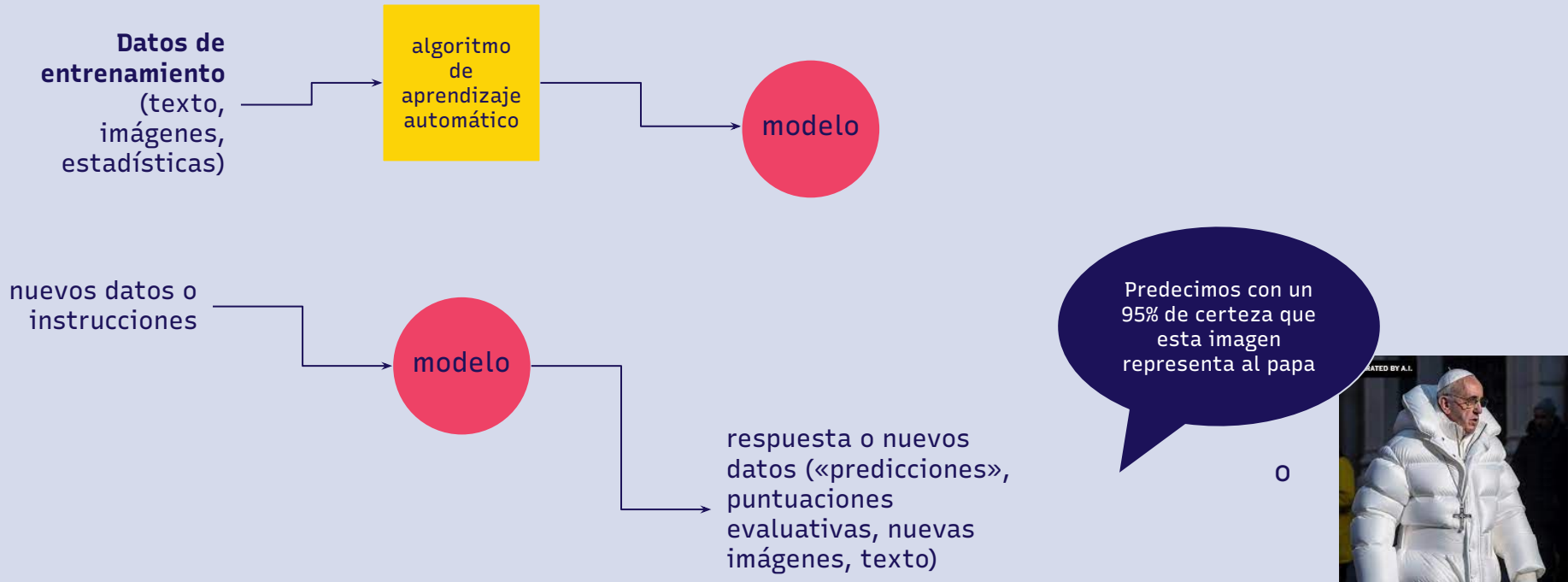


Datos en el ámbito de la IA

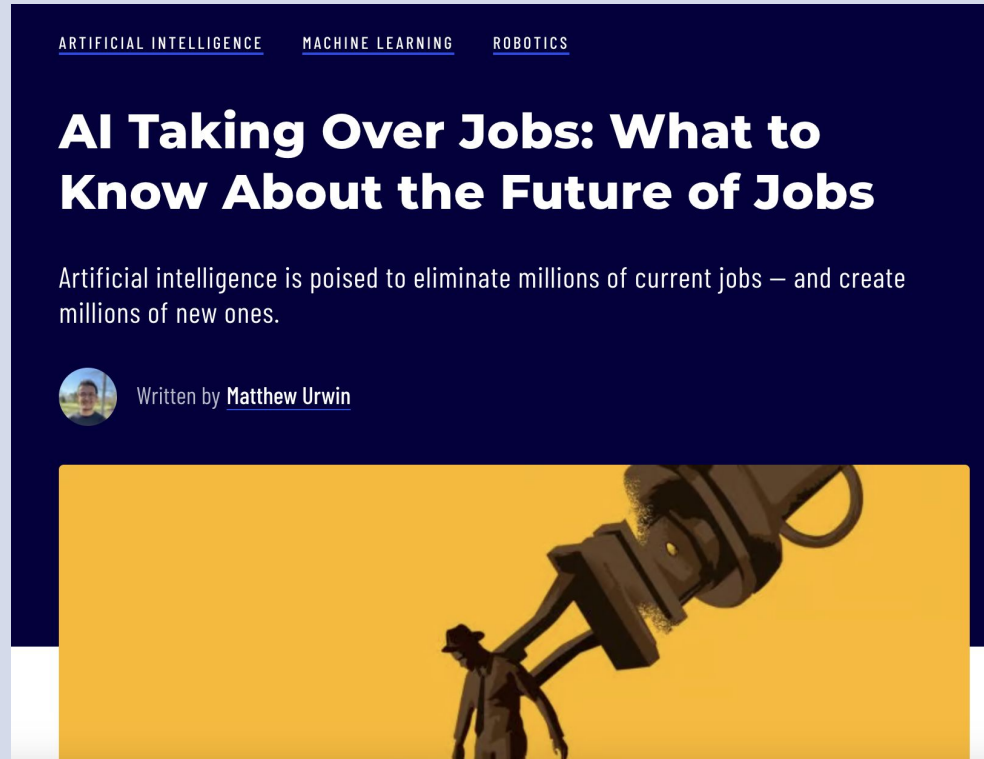


Ciclo de vida de la IA

¿En qué parte del ciclo de vida se encuentran los datos de entrenamiento?



IA: Los datos están cambiando nuestra sociedad



Source: <https://builtin.com/artificial-intelligence/ai-replacing-jobs-creating-jobs>

IA: los datos están cambiando nuestra sociedad



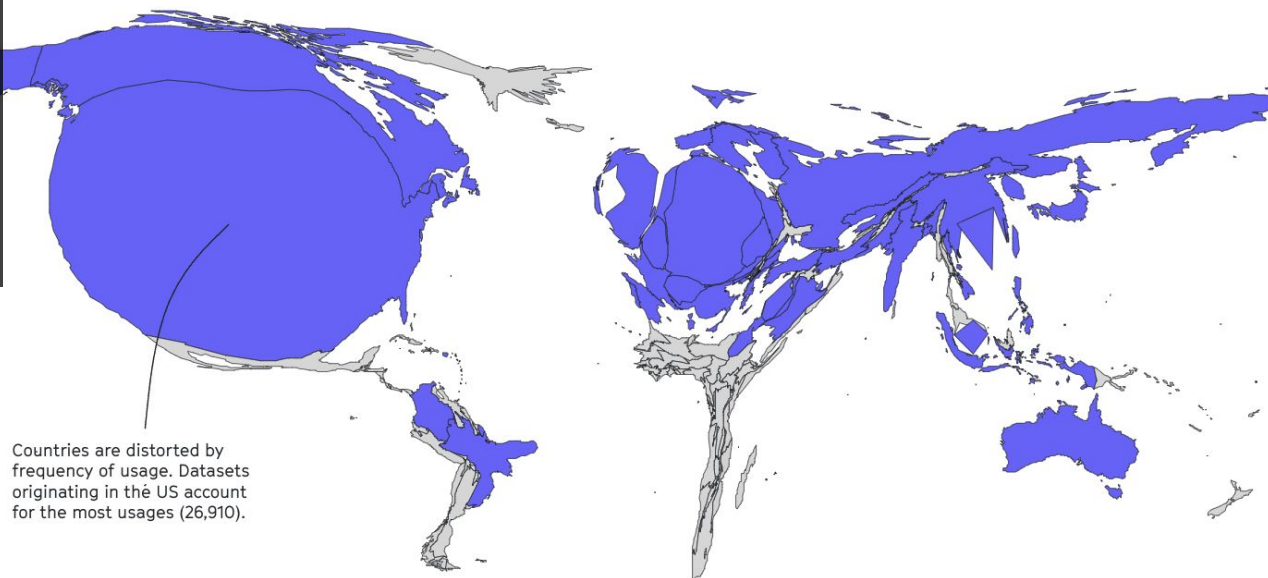
Los gobiernos de todo el mundo están experimentando con algoritmos predictivos de formas que son en gran medida invisibles para el público.

Gran parte de los datos reproducen sesgos

En miles de artículos, se utilizan con mayor frecuencia los mismos conjuntos de datos de unos pocos países para evaluar el rendimiento de los modelos de aprendizaje automático en todas partes.

Frequency of dataset usage by country

● Usage of datasets from here ● No usage of datasets from here



Datos estructurados que
alimentan los algoritmos en
la toma de decisiones



En el programa Old Age Samman Allowance de Haryana, las personas mayores de 60 años cuyos ingresos, sumados a los de su cónyuge, no superen las 300 000 rupias (3600 dólares) al año, tienen derecho a una pensión mensual de 2750 rupias (33 dólares).

En junio de 2020, el estado comenzó a utilizar un nuevo sistema algorítmico, el Repositorio de Datos de Identidad Familiar o la base de datos Parivar Pehchan Patra (PPP), para determinar la elegibilidad de los solicitantes de prestaciones sociales.



Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

El PPP es un identificador único de ocho dígitos que se proporciona a cada familia del estado y contiene datos sobre nacimientos y defunciones, matrimonios, empleo, propiedades e impuestos sobre la renta, entre otros datos, de los miembros de la familia. Recopila la información demográfica y socioeconómica de cada familia mediante la vinculación de varias bases de datos gubernamentales para comprobar su elegibilidad para los programas de asistencia social.

El estado afirmó que el PPP creaba «datos auténticos, verificados y fiables de todas las familias» y lo hizo obligatorio para que los ciudadanos pudieran acceder a todos los programas de asistencia social.

Pero, en la práctica, el PPP marcó erróneamente a Chand como «fallecido», negándole su pensión durante varios meses. Peor aún, las autoridades no cambiaron su estado de «fallecido» incluso cuando se reunió con ellas en persona en repetidas ocasiones.



Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

Historial de comportamiento
que alimentan tu feed en las
redes sociales



Datos utilizados por Facebook

Carpetas incluidas
en un archivo de
datos descargable de
Facebook

El algoritmo de Facebook es un sistema de clasificación que utiliza el aprendizaje automático para organizar el contenido de las noticias de los usuarios. Datos utilizados:

- Datos que introduces en Facebook
- Datos que generas al utilizar Facebook
- Datos que recopilan las páginas con botones de Facebook

Para hacerse una idea de la profundidad de estos datos, los usuarios pueden descargarlos siguiendo las instrucciones de Facebook:

https://www.facebook.com/help/1701730696756992?helpref=hc_global_nav

Más información sobre el tema aquí:

Cómo Facebook te vigila en Internet ([BuzzFeed News](#))

3 formas sencillas en las que entregamos una gran cantidad de información muy personal a Facebook y a aplicaciones aleatorias ([BuzzFeed News](#))

- about_you
- ads
- apps_and_websites
- calls_and_messages
- comments
- events
- following_and_followers
- friends
- groups
- https/
- likes_and_reactions
- location_history
- marketplace
- messages
- other_activity
- pages
- payment_history
- photos_and_videos
- posts
- profile_information
- saved_items
- search_history
- security_an...information
- your_places



Post title

Ejemplo: algoritmo de Facebook

The top post on a Facebook user's news feed, shown as the biggest box, is a prized position based on thousands of data points related to the user and post itself, such as the poster, reactions and comments.



Fuente: [The Washington Post](#)

Show your age

Ejemplo: algoritmo de Facebook

Trip photos

Since 2018, the algorithm has elevated posts that **encourage interaction**, such as ones popular with friends. This broadly prioritizes posts by **friends and family** and **viral memes**, but also **divisive content**.

Zoe A. and 60.8K others

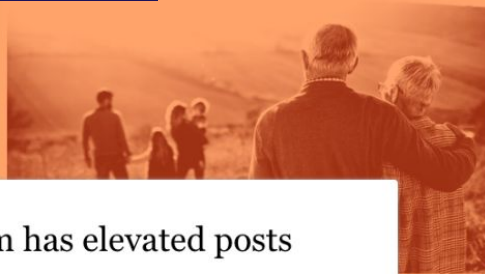
10.7K comments



Like



Comment



Ray M. and 79 others



Like



Comment

Engaged!



Rap name



Kai J. and 45.2K others



Like



Comment

Fuente: [The Washington Post](#)

Player takes political stand

Spicy opinion piece

Ejemplo: algoritmo de Facebook

Each user's feed reflects their expressed interests. For a subset of **extremely partisan** users, today's algorithm can turn their feeds into echo chambers of **divisive content** and **news**, of varying reputability, that support their outlook.

Fuente: [The Washington Post](#)

Datos de texto para el procesamiento del lenguaje natural



Ejemplo: Google

La predicción de texto en la búsqueda de Google se basa en innumerables búsquedas.

¿Cuántas personas con el corazón roto buscaron en Google el nombre de Ryan Gosling para averiguar si estaba casado? (Además, es probable que mucha gente también buscara en Google el nombre de un famoso + esposa).



Algunos conceptos

- 1. Corpus**
- 2. Tokenización**
- 3. Derivación o lematización**

Algunos conceptos

- 1. Corpus**
- 2. Tokenización**
- 3. Derivación o lematización**



Algunos conceptos

1. Corpus
2. Tokenización
3. Derivación o lematización



Algunos conceptos

1. **Corpus**
2. **Tokenización**
3. **Derivación o lematización**

El objetivo tanto de la derivación o lematización es reducir las formas flexivas y, en ocasiones, las formas derivadas de una palabra a una forma base común.

Por ejemplo:

am, are, is → be

car, cars, car's, cars' → car

El resultado de esta correspondencia del texto será algo así como:

the boy's cars are different colors

the boy car be differ color

Algunos conceptos

1. **Corpus**
2. **Tokenización**
3. **Derivación o lematización**

Derivación vs. lematización

Ejemplo: «saw»

(Como en «I saw a bird, or was it a plane? I think it may have been Supergirl.»).

Raíz:

saw → s

Lema:

saw → see*

*si el token es un verbo

Análisis de texto

Una vez que hayas preparado tu corpus en tokens (ya sea como raíces o lemas), puedes analizar el texto. Esto podría incluir:

- Creación de listas de palabras más utilizadas
- Predecir qué palabras son más propensas a aparecer juntas (¡igual que en una buena búsqueda!)
- Detectar el sentimiento en un texto



ryan goslong



Ryan Gosling
Canadian actor



ryan **gosling** movies



ryan **gosling** wife



ryan **gosling** meme



ryan **gosling** height



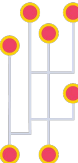
ryan **gosling** instagram



ryan **gosling** eva mendes

A graphic for the AI Spotlight Series. It features a vertical line of five circles on the left, with the top one yellow and the others light blue. A horizontal line connects the top yellow circle to the word "AI".

AI Spotlight Series

A logo for The AI Accountability Network, consisting of a network of interconnected nodes and lines, with nodes colored yellow and red.

**The AI
Accountability
Network**

The Pulitzer Center logo, featuring a stylized circular icon with a white 'P' and 'C' inside a dark blue circle.

Pulitzer Center