



NLP

Natural language processing



Elen Irazabal
@IrazabalElen

LARGEST GLOBAL COMPANIES IN 2018 VS 2008:
SEVEN OUT OF TEN ARE NOW BASED ON PLATFORM
BUSINESS MODELS

2018

RANK	COMPANY		FOUNDED	US\$bn
1.	 *		1976	890
2.	 *		1998	768
3.	 *		1975	680
4.	 *		1994	592
5.	 *		2004	545
6.	 腾讯 *		1998	526
7.	BERKSHIRE HATHAWAY		1955	496
8.	 *		1999	488
9.	 *		1886	380
10.	J.P.Morgan		1871	375

* Companies based on the platform model

2008

RANK	COMPANY		FOUNDED	US\$bn
1.	 PetroChina		1999	728
2.	 EXXON		1870	492
3.	 GE		1892	358
4.	 中国移动 China Mobile		1997	344
5.	 ICBC		1984	336
6.	 GAZPROM		1989	332
7.	 Microsoft		1975	313
8.	 Shell		1907	266
9.	 ARAMCO		2000	257
10.	 AT&T		1885	238

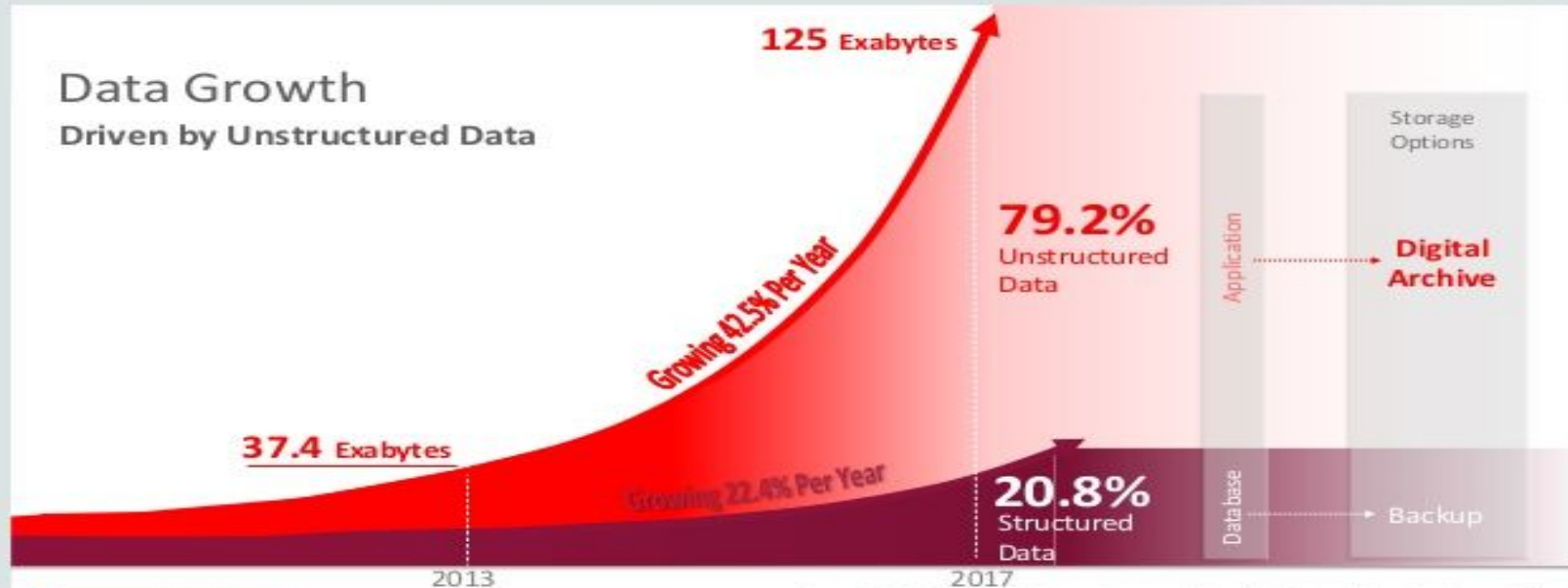
Sources: Bloomberg, Google

¿ Por qué surge la revolución del Procesamiento del Lenguaje Natural?



Data Growth

Driven by Unstructured Data



Source: IDC - 2014, Structured Data vs. Unstructured Data: The Balance of Power Continues to Shift

ORACLE

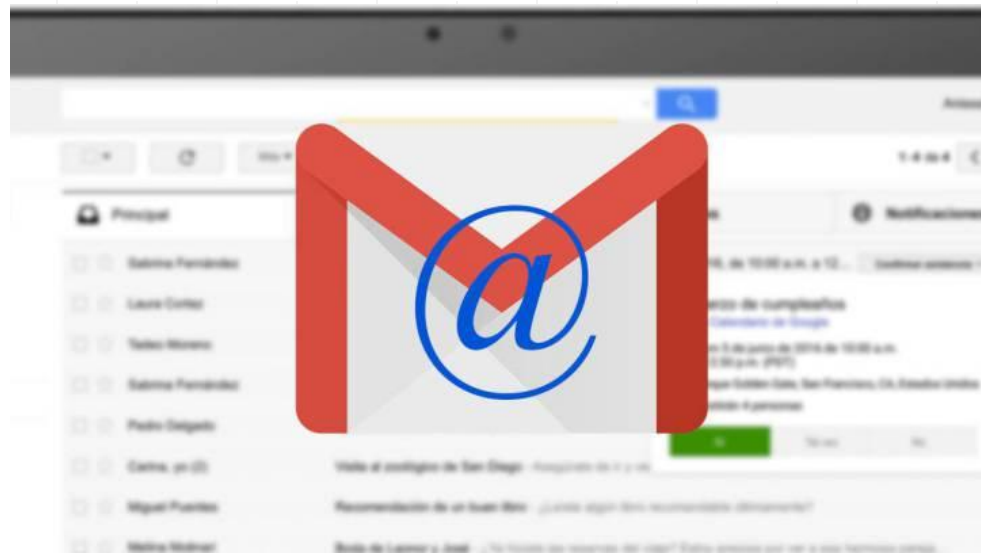
NABSHOW
NAB SHOW
NAB SHOW

etc

ENTERPRISE
TECHNOLOGY
COUNCIL

Copyright © 2015 Oracle and/or its affiliates. All rights reserved.

BREAKING NEWS



1 - 10 de 214 opiniones



Onkelz-1980
Cantabria, España

16 7



Opinión escrita hace 4 días mediante dispositivo móvil

Gran descubrimiento

Hemos comido en Daria y debo decir que ha sido un gran descubrimiento. Su arroz a banda para repetir varias veces. Muy bueno el detalle de poner pan tomate y aceite para ir abriendo boca, estos detalles deberían ser obligatorios pero todavía hay muchos que... [Más](#)



Fecha de la visita: agosto de 2019

1 Gracias, Onkelz-1980



Gastro86extrem

35 5



Opinión escrita hace 4 días mediante dispositivo móvil

Bueno y bonito.

Sitio muy bonito que buenos detalles, pedimos unos tacos de merluza que están muy buenos, curry thai con gambas ojo que pical Lassaña y postre tarta de manzana muy buena y galleta de chocolate normalilla. El pero que le pongo son las raciones algo pequeñitas... [Más](#)

Fecha de la visita: septiembre de 2019

1 Gracias, Gastro86extrem



Lorelai0976

2



Opinión escrita hace 5 días

Un festival para los sentidos

Fui con mi familia para comer. Comimos él rodaballos con salsa Nikkei, los tacos de cochinitillo y como entrante una ensala de tomates ecológicos con salsa de burrata. Todo fantástico, el camarero súper atento. De postre la tarta de manzana, imprescindible.

Fecha de la visita: agosto de 2019

1 Gracias, Lorelai0976





80% of your business data is **unstructured**

¿ De dónde nace el Procesamiento de Lenguaje Natural?

En 1950 con Alan Turing que publicó un artículo llamado “Computing Machinery and Intelligence”

El test de Turing desarrollado también por Alan Turing en ese mismo año, es una prueba de la capacidad de una máquina para exhibir un comportamiento inteligente equivalente o indistinguible del de un humano.

¿Cómo podríamos definir el Procesamiento de Lenguaje Natural?

El procesamiento del lenguaje natural (PNL) es una rama de la inteligencia artificial que ayuda a las computadoras a comprender, interpretar y manipular el lenguaje humano.



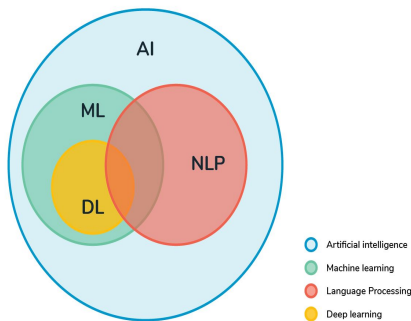
Explosión de Machine Learning

1970–1990: La revolución de la estadística y su incorporación a programas informáticos

Aprendizaje automático

En PLN: A través de corpuses

Hoy en día: grandes avances gracias a las redes neuronales (deep learning)

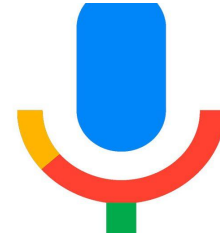


Retos del Procesamiento de Lenguaje Natural

- La estadística no tiene sentimientos
- Falta de comprensión de reglas gramaticales
- Ambigüedades de los textos.
- Sinónimos: usar diferentes palabras para comunicar el mismo significado (lo mismo para frases completas)
- Metáforas: usar un constructo con un significado para venir a decir algo diferente.
- Expresiones
- Los idiomas
- Sentimientos y emociones
- Mentiras, ironías, sarcasmos...

¿Para qué se utiliza?

Traducción automática de texto
Sistemas conversacionales
Recuperación y extracción de información
Análisis de sentimiento
Detección de temas
Resúmenes
Clasificación de documentos



¿Y en el sector legal?

Predicciones sobre decisiones judiciales / documentos oficiales

Búsqueda de información

Revisión de contratos

Legal advice

Reto: la digitalización del sector



2020: La consola semántica de Open AI

Traducción de tareas a código

Utilizar lenguaje natural para decirle a la consola que haga algo y te devuelve el código a ejecutar

Seguir a: <https://www.youtube.com/channel/UCy5znSnfMsDwaLIROnZ7Qbg> DOTCSV

El WTF de NLP ha surgido hace nada : el GPT-3 de OpenAI



Ha sido alimentado con prácticamente todo el conocimiento más importante publicado en internet (libros, docs científicos, wikipedia...)

El GPT2 del año pasado tenía un peso de 40 GB con 45 millones de análisis de páginas webs y con 1.500 millones de parámetros, GPT3... tiene 175.000 millones de parámetros

El WTF de NLP ha surgido hace nada : el GPT-3 de OpenAI

Traducir texto a otro idioma

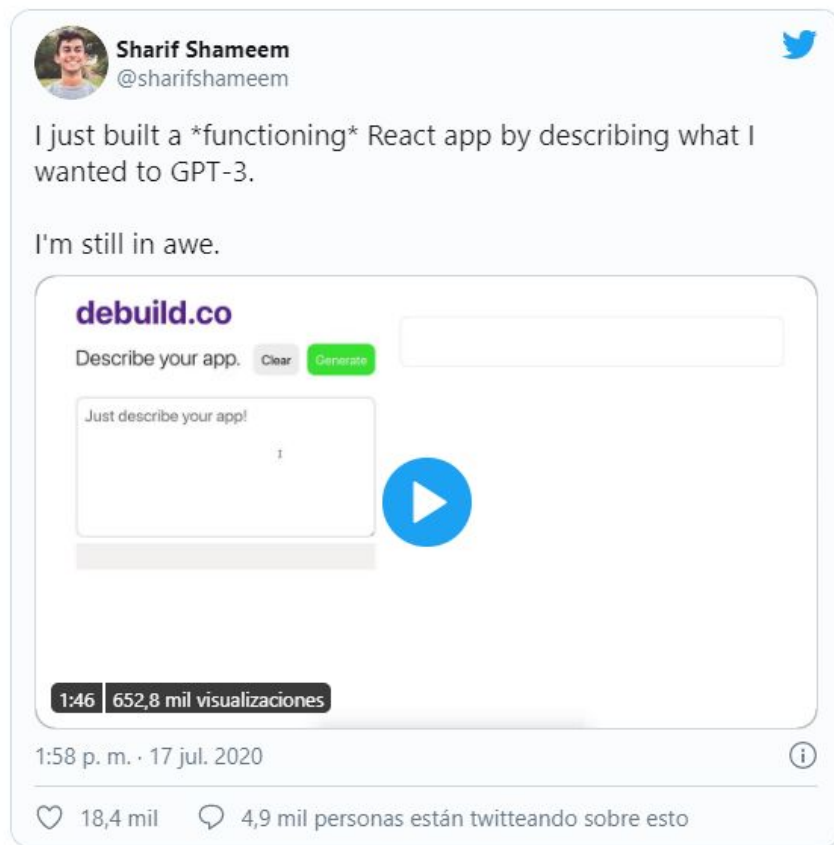
Genera texto, historias, diálogos

Cual es la siguiente palabra a raíz de la secuencia de palabras anteriores

Crea contenido dandole ordenes, puede crear una página web

Genera conversaciones

Programa una app (en React)



A screenshot of a tweet from Sharif Shameem (@sharifshameem) on Twitter. The tweet text reads: "I just built a *functioning* React app by describing what I wanted to GPT-3. I'm still in awe." Below the text is a video player showing the interface of a web application called "debuild.co". The interface has a header with the name "debuild.co" and a description field labeled "Describe your app." with "Clear" and "Generate" buttons. Below this is a text area with the prompt "Just describe your app!" and a single character "I". To the right of the text area is a large blue play button. At the bottom of the video player, a black bar displays "1:46" and "652,8 mil visualizaciones". Below the video, the tweet's timestamp "1:58 p. m. · 17 jul. 2020" and an information icon are visible. At the very bottom, the engagement metrics show "18,4 mil" likes and "4,9 mil personas están twitteando sobre esto".

Sharif Shameem
@sharifshameem

I just built a *functioning* React app by describing what I wanted to GPT-3.

I'm still in awe.

debuild.co
Describe your app.
Just describe your app!
I

1:46 652,8 mil visualizaciones

1:58 p. m. · 17 jul. 2020

18,4 mil 4,9 mil personas están twitteando sobre esto

Hace de google



Paras Chopra
@paraschopra

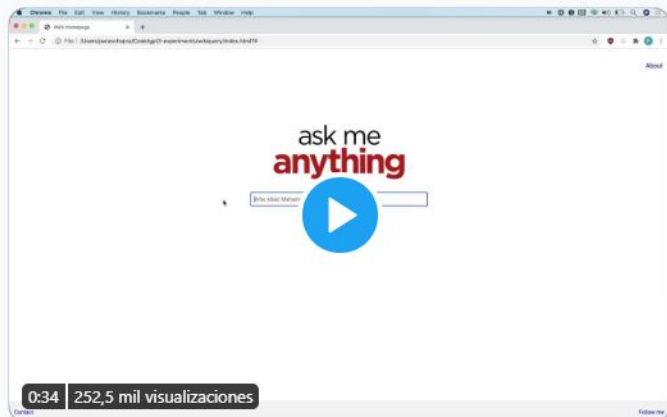


I made a fully functioning search engine on top of GPT3.

For any arbitrary query, it returns the exact answer AND the corresponding URL.

Look at the entire video. It's MIND BLOWINGLY good.

cc: @gdb @npew @gwern



12:43 p. m. · 19 jul. 2020



3,6 mil



832 personas están twitteando sobre esto

[https://www.xataka.com/robotica-e-ia/gpt-3-nuevo-modelo-lenguaje-o
penai-capaz-programar-disenar-co
nversar-politica-economia](https://www.xataka.com/robotica-e-ia/gpt-3-nuevo-modelo-lenguaje-openai-capaz-programar-disenar-conversar-politica-economia)



Una reflexión

“Today’s machine learning applications need a lot of labeled data to have good performance, but most of the world’s data is not labeled. For machine learning to advance, algorithms will need to learn from unlabeled data and make sense of the world from pure observation, much like how children learn to operate in the real world after birth without too much guidance.

According to Yann LeCun, one of the fathers of machine learning and currently the chief AI scientist at Facebook, the future of machine learning will be driven by unsupervised or self-supervised learning systems”.

<https://www.unsupervisedlearningbook.com/>



**Pero...¿cuáles son los
primeros pasos que puedo
dar para aprender NLP?**

PLN desde el inicio: Limpieza y transformar texto a número

- Procesamiento léxico y sintáctico
 - Segmentar las palabras del texto.
 - Normalizar formatos de palabras
- Segmentación/tokenización
 - Consiste en separar un texto en las palabras de las que se compone
 - Ejemplo
 - “Estamos en estado de alarma”
 - *Tokens*: [“Estamos”, “en”, “estado”, “de”, “alarma”]
- Tras aplicar segmentación/tokenización → quitar palabras comunes
 - Consiste en eliminar palabras “de relleno”, como preposiciones, conjunciones o verbos auxiliares. (“de”, “por”...)

PLN desde el inicio: Limpieza y transformar texto a número

- Stemming

Definición: Reducir los términos a sus *stems*, eliminando los *affixes* añadidos.

- Ejemplo: bibliotecas, bibliotecario → (stemming) → bibliotec

- Lowercase:

Pasar todas las mayúsculas a minúsculas: Legaltech → legaltech, Madrid → madrid

- Quitar signos de puntuación ('?;!2°&%)

PLN desde el inicio: transformar texto a número

El proceso de convertir texto en números se llama vectorización de datos de texto. Es decir, necesitamos representar numéricamente los datos de texto, para así codificar los datos en números.

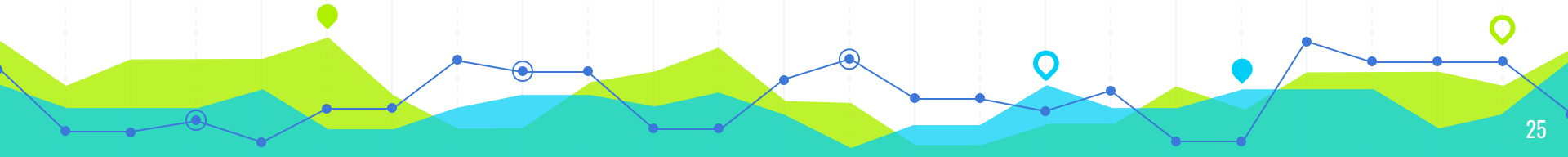


**Si queréis hacer
proyectos con IA...**



“

...tenéis que preguntaros primero si el problema que queréis solucionar pasa por data y cuanto más específico sea el problema mejor



“

Recordad que la tecnología es un medio, no un fin

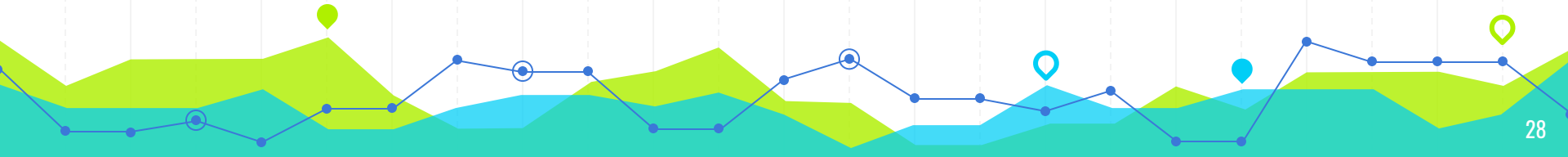
“

Sin datos, la IA no va a poder hacer nada



Coursera para el pensamiento lógico: Think Again

<https://www.coursera.org/courses?query=think%20again>



Gracias!

¿Preguntas?

