Introducción al Procesamiento de Lenguaje Natural y Extracción de Información

Dr. Cristian Enrique Muñoz Villalobos

Agenda

- Que es el procesamiento de Lenguaje Natural?
- Aplicaciones y tareas de NLP
- Estrategias para la colecta de información
- Data Warehouse y Data Lakes
- Formatos abiertos: JSON, XML, HTML, etc.
- Web Crawling y APIs

"A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language."

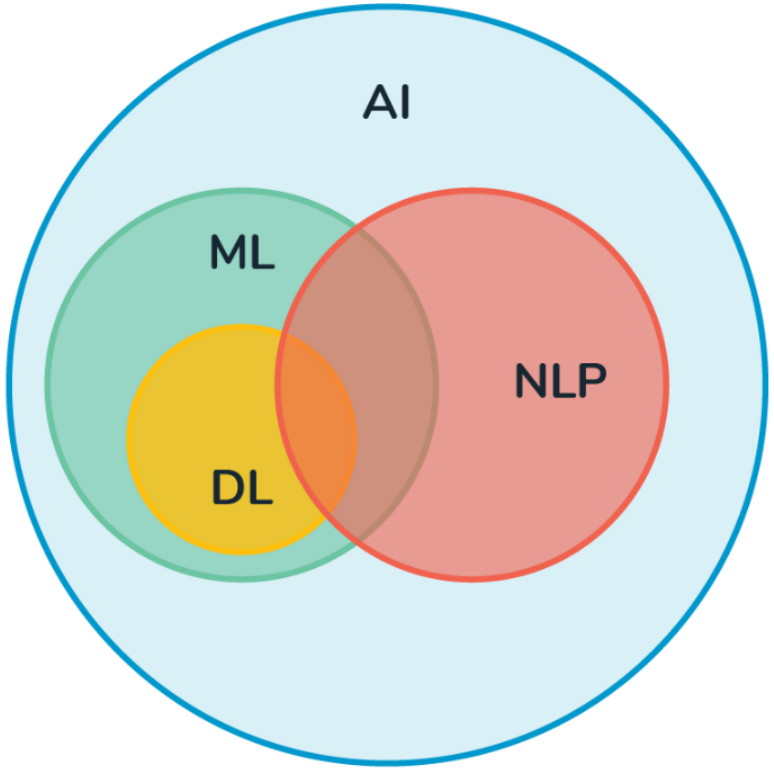
[Noam Chomsky]

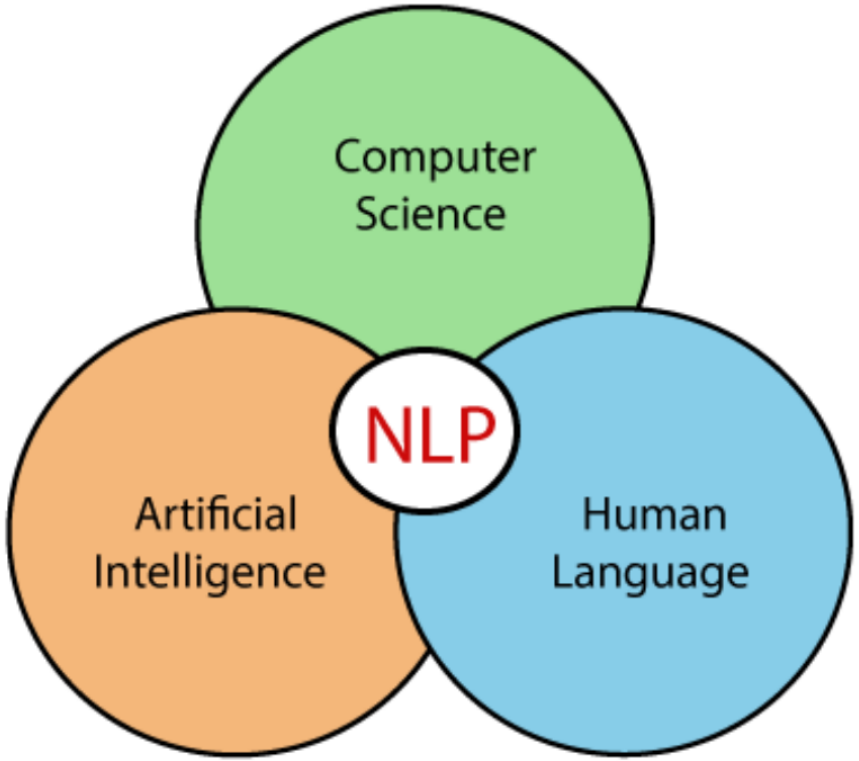
Que es el Procesamiento de Lenguaje Natural?

> Es un área de la inteligencia artificial que se especialista en el proceso computacional de entender y generar lenguaje humano.

> Es una área de ciencia de la computación que utiliza métodos para analizar, modelar y entender el lenguaje humano.

Que es el Procesamiento de Lenguaje Natural?

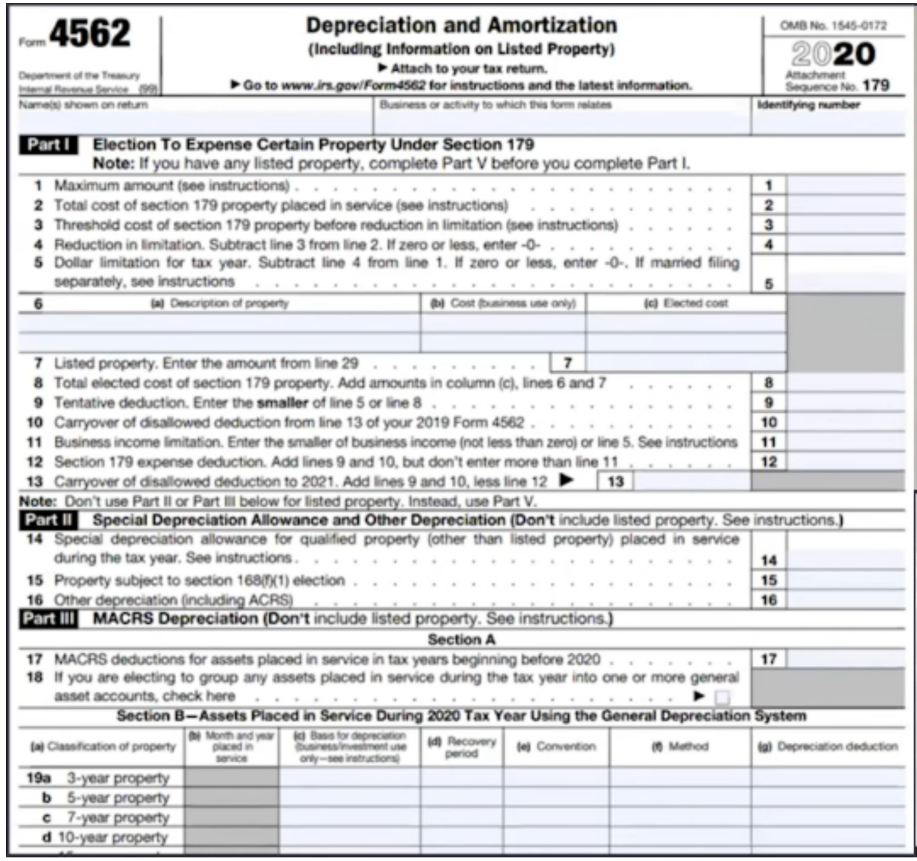




Aplicaciones

		
Extracción de la información	*Clasificación de Spam*	*Análisis de Finanzas*
		
Recuperación de la Información	*Motores de Búsqueda*	*Extracción de Entidades Legales*

Recuperación de la Información



Form **4562**

Department of the Treasury
Internal Revenue Service (99)

Depreciation and Amortization
(Including Information on Listed Property)
▶ Attach to your tax return.
▶ Go to *www.irs.gov/Form4562* for instructions and the latest information.

OMB No. 1545-0172
2020
Attachment Sequence No. **179**

Name(s) shown on return | Business or activity to which this form relates | Identifying number

Part I **Election To Expense Certain Property Under Section 179**
Note: If you have any listed property, complete Part V before you complete Part I.

1	Maximum amount (see instructions)	1
2	Total cost of section 179 property placed in service (see instructions)	2
3	Threshold cost of section 179 property before reduction in limitation (see instructions)	3
4	Reduction in limitation. Subtract line 3 from line 2. If zero or less, enter -0-	4
5	Dollar limitation for tax year. Subtract line 4 from line 1. If zero or less, enter -0-. If married filing separately, see instructions	5

6	(a) Description of property	(b) Cost (business use only)	(c) Elected cost

7	Listed property. Enter the amount from line 29	7	
8	Total elected cost of section 179 property. Add amounts in column (c), lines 6 and 7		8
9	Tentative deduction. Enter the **smaller** of line 5 or line 8		9
10	Carryover of disallowed deduction from line 13 of your 2019 Form 4562		10
11	Business income limitation. Enter the smaller of business income (not less than zero) or line 5. See instructions		11
12	Section 179 expense deduction. Add lines 9 and 10, but don't enter more than line 11		12
13	Carryover of disallowed deduction to 2021. Add lines 9 and 10, less line 12 ▶	13	

Note: Don't use Part II or Part III below for listed property. Instead, use Part V.

Part II **Special Depreciation Allowance and Other Depreciation (Don't include listed property. See instructions.)**

14	Special depreciation allowance for qualified property (other than listed property) placed in service during the tax year. See instructions	14
15	Property subject to section 168(f)(1) election	15
16	Other depreciation (including ACRS)	16

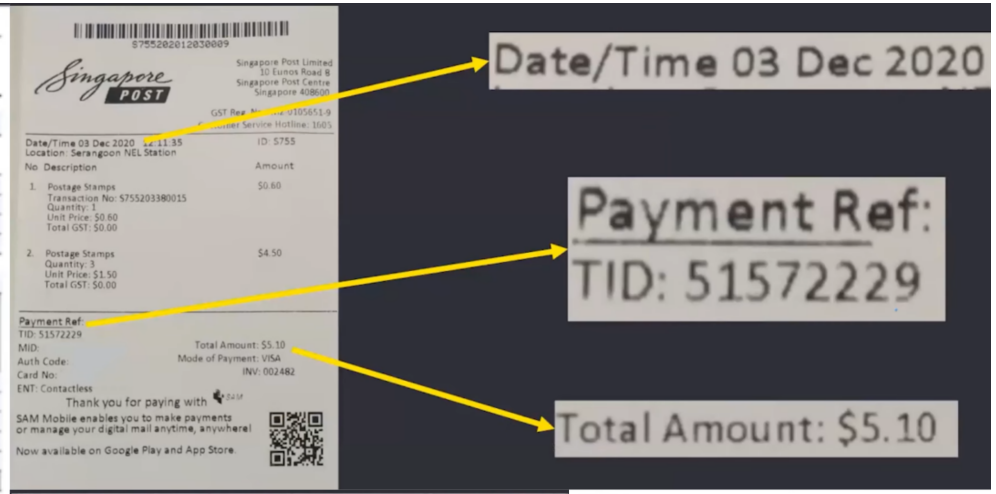
Part III **MACRS Depreciation (Don't include listed property. See instructions.)**

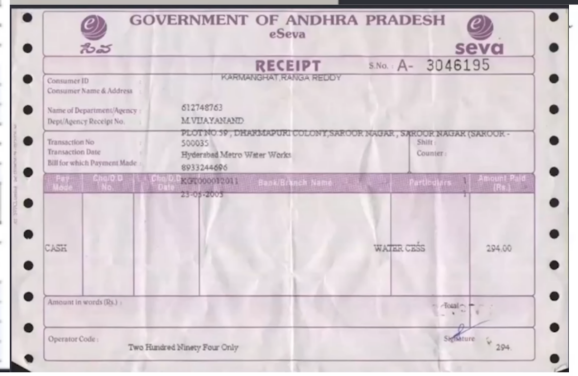
Section A

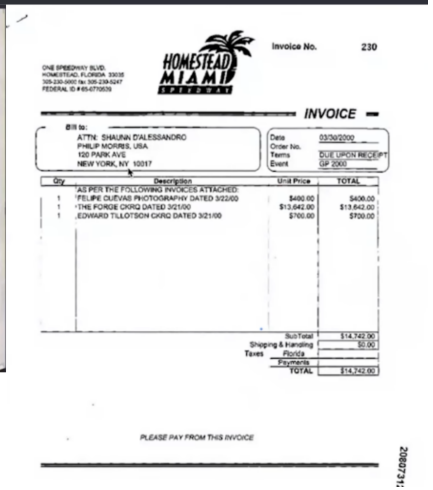
17	MACRS deductions for assets placed in service in tax years beginning before 2020	17
18	If you are electing to group any assets placed in service during the tax year into one or more general asset accounts, check here ▶ ☐	

Section B—Assets Placed in Service During 2020 Tax Year Using the General Depreciation System

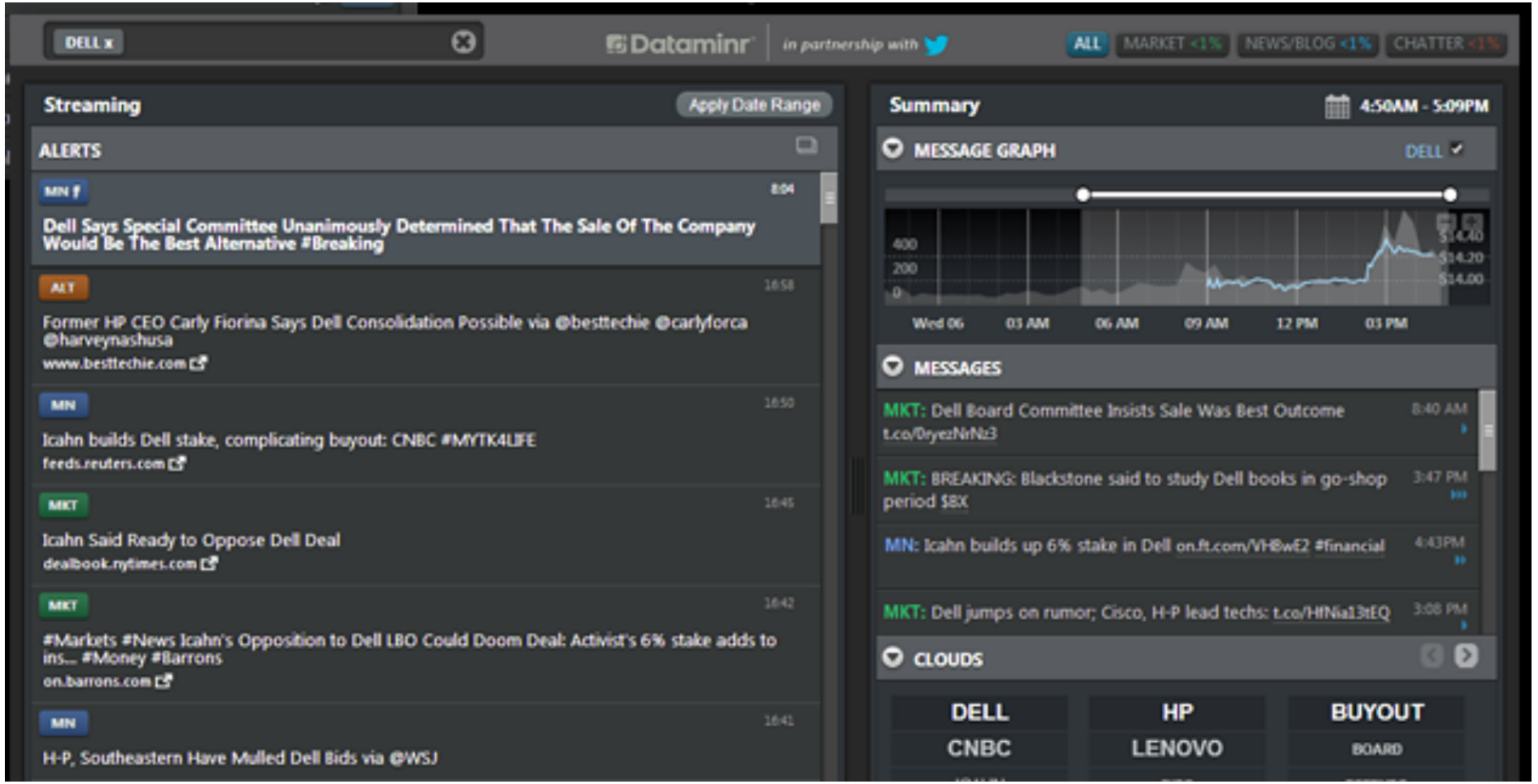
	(a) Classification of property	(b) Month and year placed in service	(c) Basis for depreciation (business/investment use only—see instructions)	(d) Recovery period	(e) Convention	(f) Method	(g) Depreciation deduction
19a	3-year property						
b	5-year property						
c	7-year property						
d	10-year property						



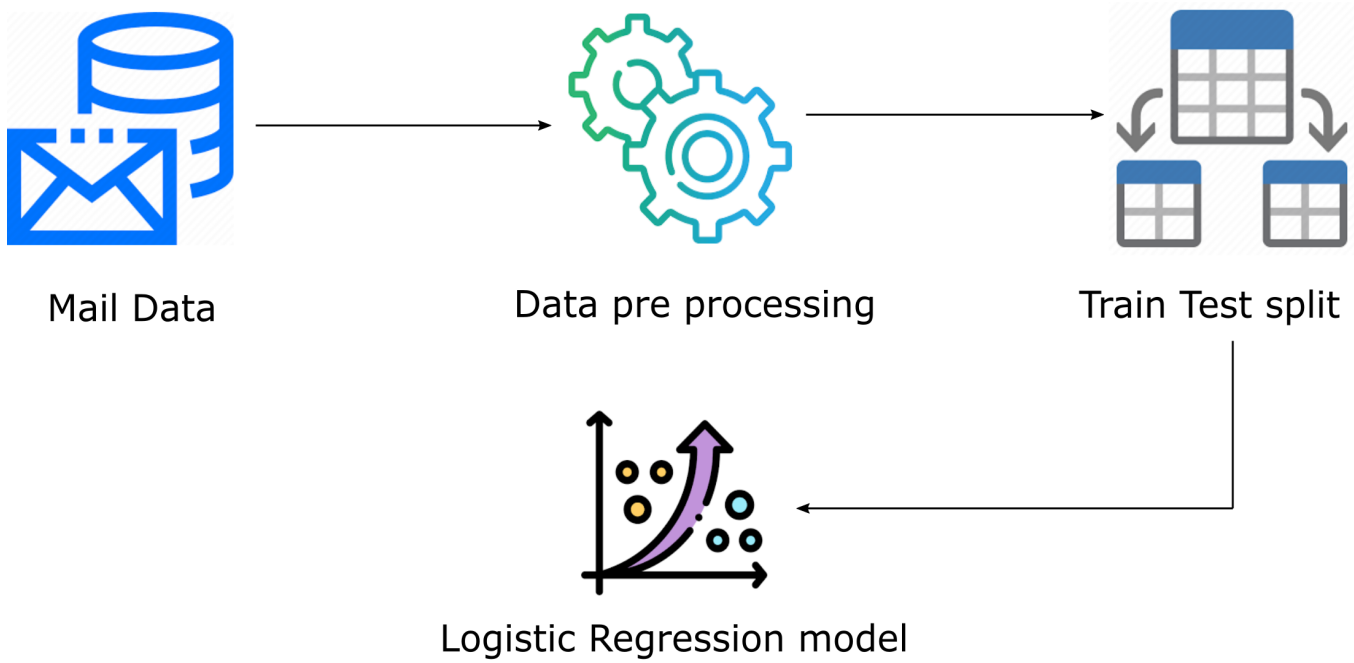




Finanzas


DELL x
Dataminr
in partnership with
ALL
MARKET <1%
NEWS/BLOG <1%
CHATTER <1%
Streaming
Apply Date Range
ALERTS
MN f
8:04
Dell Says Special Committee Unanimously Determined That The Sale Of The Company Would Be The Best Alternative #Breaking
ALT
16:58
Former HP CEO Carly Fiorina Says Dell Consolidation Possible via @besttechie @carlyforca @harveynashusa
www.besttechie.com
MN
16:50
Icahn builds Dell stake, complicating buyout: CNBC #MYTK4LIFE
feeds.reuters.com
MKT
16:45
Icahn Said Ready to Oppose Dell Deal
dealbook.nytimes.com
MKT
16:42
#Markets #News Icahn's Opposition to Dell LBO Could Doom Deal: Activist's 6% stake adds to ins... #Money #Barrons
on.barrons.com
MN
16:41
H-P, Southeastern Have Mulled Dell Bids via @WSJ
Summary
4:50AM - 5:09PM
MESSAGE GRAPH
DELL
400
200
0
$14.40
$14.20
$14.00
Wed 06
03 AM
06 AM
09 AM
12 PM
03 PM
MESSAGES
MKT: Dell Board Committee Insists Sale Was Best Outcome t.co/DryezNrNz3
8:40 AM
MKT: BREAKING: Blackstone said to study Dell books in go-shop period $BX
3:47 PM
MN: Icahn builds up 6% stake in Dell on.ft.com/VHBwE2 #financial
4:43PM
MKT: Dell jumps on rumor; Cisco, H-P lead techs: t.co/HfNia13tEQ
3:08 PM
CLOUDS
DELL
HP
BUYOUT
CNBC
LENOVO
BOARD

Spam


Mail Data
Data pre processing
Train Test split
Logistic Regression model
Trained Logistic Regression model
New mail
Spam
(or)
Ham
Prediction

Buscadores


Google
quien es el presidente de Peru
Todas
Notícias
Imagens
Vídeos
Maps
Mais
Ferramentas
Aproximadamente 114.000.000 resultados (0,53 segundos)
Peru / Presidente
Pedro Castillo
Pesquisas relacionadas
Keiko Fujimori
Vladi… Cerrón
Aníbal Torres
Alberto Fujimori
Lilia Pare…
Martín Vizca…
Dina Bolua…
Feedback
https://www.bbc.com › mundo › n… Traduzir esta página
Quién es Pedro Castillo, el presidente electo de Perú, y en ...
20 de jul. de 2021 — En una ajustadísima victoria, el candidato de izquierda Pedro Castillo, se impuso ante la candidata de derecha Keiko Fujimori en una de las ...
https://www.bbc.com › mundo › n… Traduzir esta página
Quién es Pedro Castillo, el maestro rural que llegó a la ... - BBC
20 de jul. de 2021 — Con una imagen y un discurso que distan mucho de los de la tradicional élite política de Lima, Pedro Castillo logró finalmente ser ...
Pedro Castillo
Presidente do Peru
José Pedro Castillo Terrones é um professor peruano, líder sindical, político, e o atual presidente do Peru desde 28 de julho de 2021. Alcançou proeminência como figura de destaque na greve dos professores de 2017 e concorreu nas eleições como o candidato do partido de esquerda Peru Livre. Wikipédia
Nascimento: 19 de outubro de 1969 (idade 52 anos), Puña
Partido: Peru Livre
Parceiro: Lilia Paredes (desde 2000)
Mandato presidencial: 28 de julho de 2021 –
Cargo: Presidente do Peru desde 2021
Formação: Instituto Superior Pedagógico Público "Octavio Matta Contreras", Universidad César Vallejo
Reivindicar Painel de informações
Feedback


excite
Ask
copernic
dogpile
LookSmart
AOL
Google
About.com
YAHOO!
HOTBOT
altavista
LYCOS
WebCrawler

Chatbots

Chaty

Automatiza los chats de tu restaurante

Crea tu propio chatbot y dile adiós a tomar pedidos a mano

Probar demo | Beneficios

Tareas de PLN

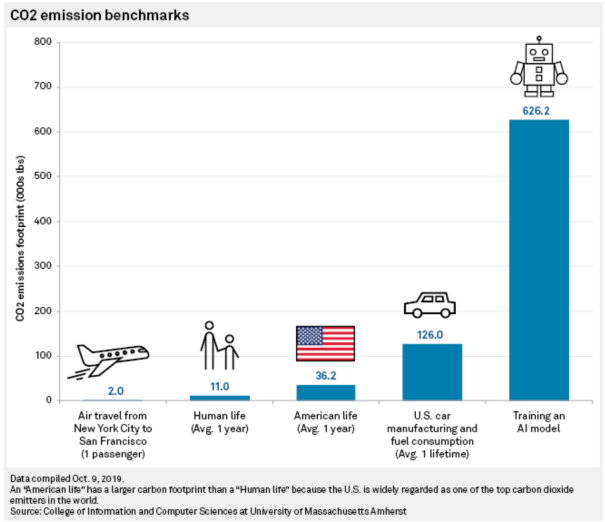
Modelamiento de Lenguaje	Clasificación de texto
Extracción de información	Recuperación de la información
Resúmenes de texto	principal del documento original
Agentes de Conversion	Pregunta-Respuesta
Maquinas de traducción	Moldeamiento de Tópicos

Modelos de Lenguaje

AI's large carbon footprint poses risks for big tech

The artificial intelligence industry has skyrocketed in recent years, powering technologies behind smart speakers and self-driving cars, but that growth is coming at a cost.

Researchers at the University of Massachusetts Amherst recently conducted a study assessing the energy consumption required to train several common large AI models. The study revealed that the training process can emit over 626,000 pounds of carbon dioxide, nearly 5x the lifetime emissions of an average car, or the equivalent of about 300 round-trip flights between New York and San Francisco.



VENTURE CAPITAL

A Golden Age For Natural Language

Louisa Xu Former Contributor ⓘ

Follow

Dec 1, 2021, 11:00am EST

Listen to article 6 minutes

Every company talks to its customers with natural language. The last three years signaled the beginning of a golden age for natural language processing (NLP), one of the most useful and visible forms of machine learning (ML).

NLP is a branch of machine learning that endows computers with the human-like ability to understand text and spoken word. Thanks to the increase in computational power, the sheer volume of raw language data available on the Internet, and the popularization of deep learning, NLP has transformed from a hypothetical question posed in the 1950 Turing Test to an everyday reality.

ENTERPRISE TECH

Revolutionary NLP Model GPT-3 Poised To Redefine AI And Next Generation Of Startups

Hannah M. Mayer Contributor ⓘ

Hannah is a former HBS & HSG researcher, and a Wiley-published author.

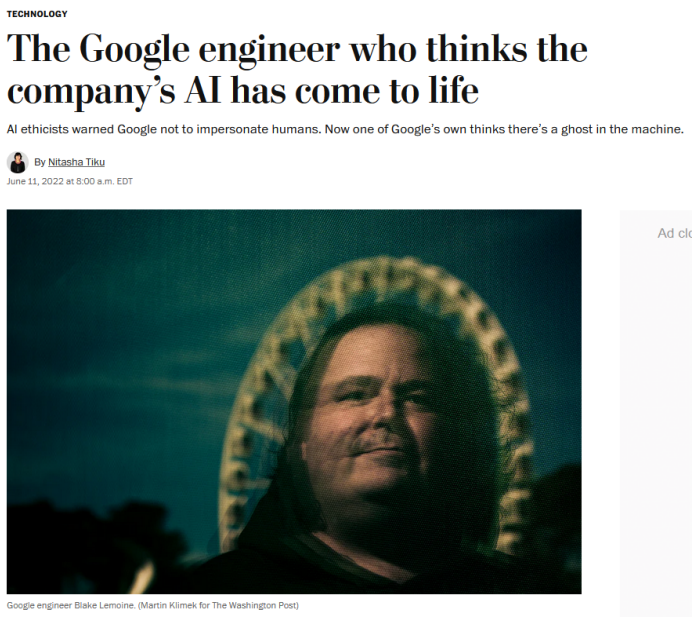
Follow

Jan 2, 2021, 03:09pm EST

Listen to article 9 minutes

LaMDA (Google Chatbot)

https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/



TECHNOLOGY

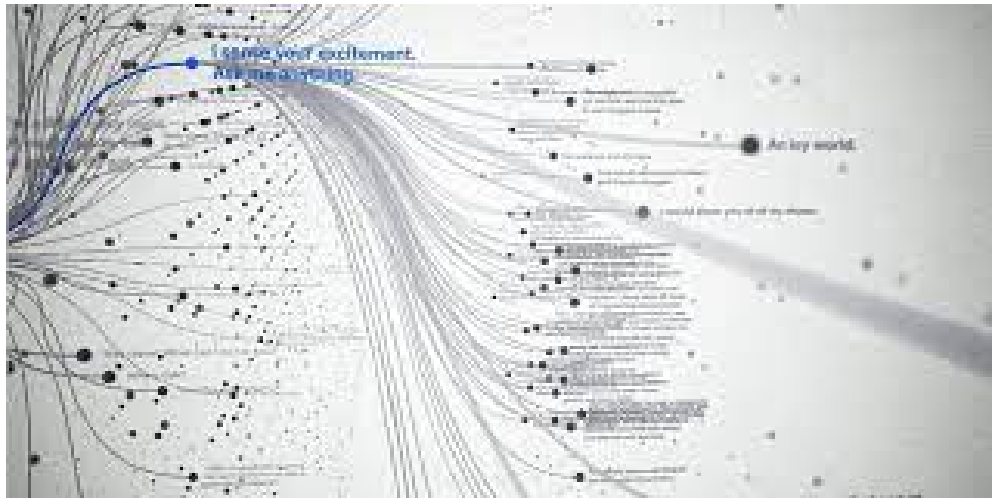
The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku

June 11, 2022 at 8:00 a.m. EDT

Google engineer Blake Lemoine. (Martin Klimek for The Washington Post)

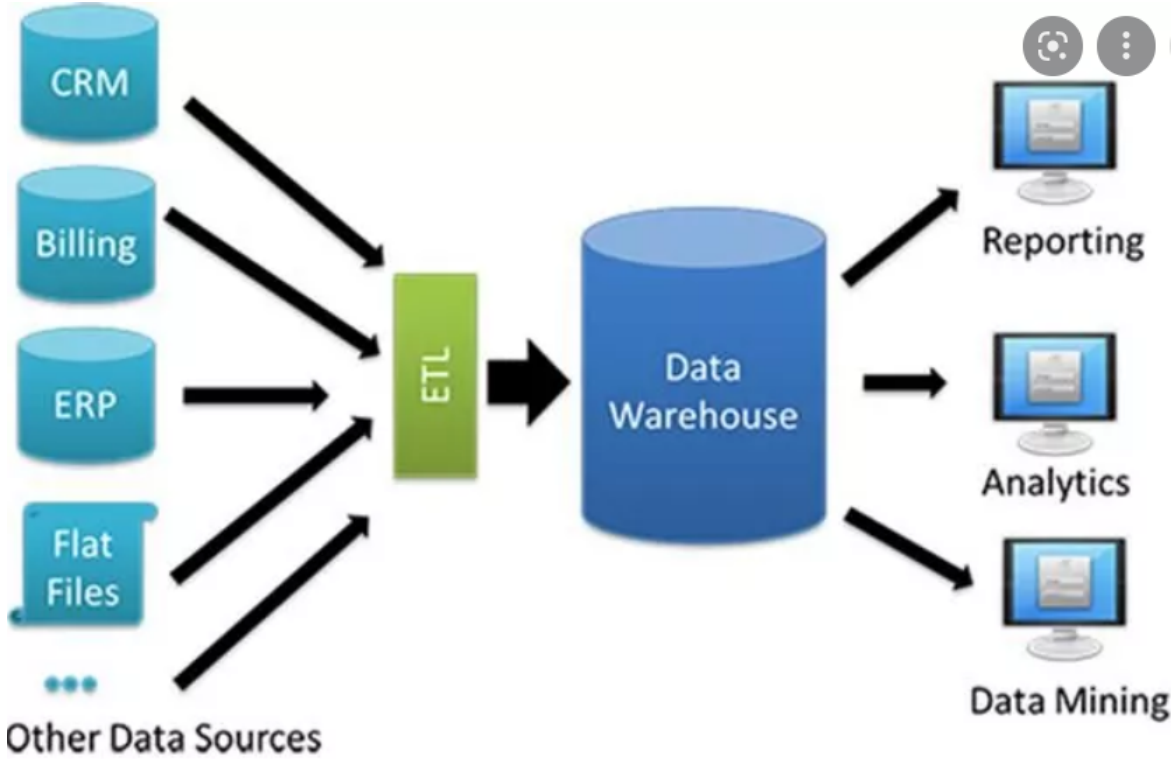


Estrategias para Colecta de Información

 **Web Crawling**	 **Document Parsing**
Recopilación automatizada de datos y su conversion en información estructurada para su análisis. e.g. datos de paginas web.	Implica la revision de datos presentes en un documento y extraer información util de el. e.g. PDF, xlsx, etc.

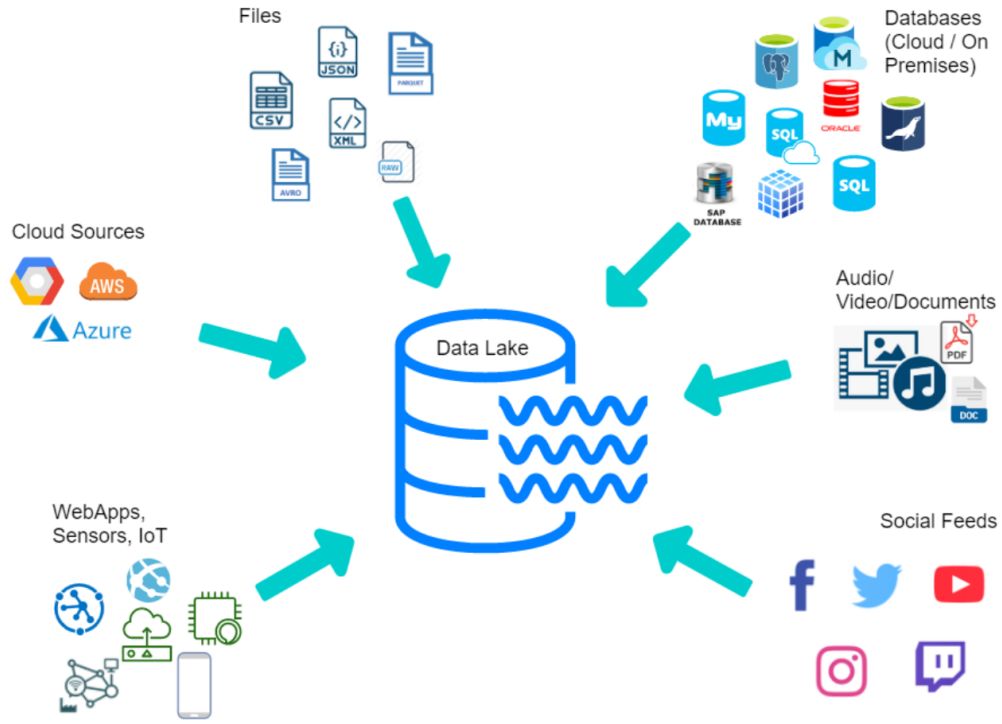
Data Warehouse y Data Lakes

Data Warehouse: es un banco de datos optimizado para analizar datos relacionales. La estructura de datos y los esquemas son definidos previamente para optimizar consultas SQL rápidas. Los datos son limpios, enriquecidos y transformados para que puedan actuar como la "fuente única verdadera" en que el usuario puede confiar.



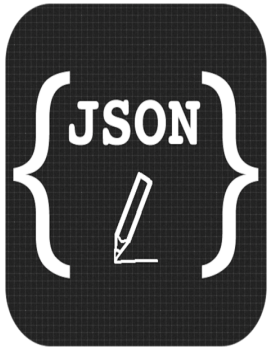
Data Warehouse y Data Lakes

Data Lake: son repositorios centralizados que permiten almacenar todos tus datos estructurados y no estructurados en cualquier escala. A principal ventaja de data lakes es la alta disponibilidad de los datos par aplicaciones de machine learning, análisis predictivas, descubrimiento de datos, etc.



Formatos Abiertos

Es un formato de archivo para almacenar datos digitales. Generalmente, mantenida por su organización de estándares y que puede ser utilizado e implementado por cualquier persona.

XML		YML	HTML	...

Formato JSON (Javascript Object Notation)

Relacionado al desarollo de software, JSON es convirtio en un padron ampliamente utilizado para intercambio de información entre sistemas. Es rápido y simple de ser interpretado por los computadores y seres humanos. **Google** ha explorado ampliamente el uso de JSON en sus API. Muchos lenguajes de programación dan soporte a este formato: Python, Ruby, PHP, Javascript.

JSON Viewer: http://jsonviewer.stack.hu/
Official Site: https://www.json.org/json-en.html

```
{
"nombre":"Cristian Enrique Muñoz Villalobos",
"edad":34,
"pais":"Perú",
"Meses":["Enero","Febrero","Marzo"],
"Familia":{"Papá":"José","Hermana":"Jannet"}
}
```

Formato XML (Extensible Markup Language)

Archivo de texto considerado un padron para intercambio de información entre sistemas y almacenamiento de datos. XML es un lenguaje de marcación definida por W3C.

XML Viewer: https://jsonformatter.org/xml-viewer

```
<pucp>
    <user id="1">
        <nombre>Cristian Enrique Muñoz Villalobos</nombre>
        <edad>34</edad>
        <pais>Perú</pais>
        <Meses>
            <Mes>Enero</Mes><Mes>Febrero</Mes>
        </Meses>
        <Familia>
            <papa>José</papa><hermana>Jannet</hermana>
        </Familia>
    </user>
</pucp>
```

Formato HTML (Hyper-Text Markup Language)

(1989) HTML esta en nuestra vidas todos los dias, mismo que no nos demos cuenta. Es un lenguaje de marcación. El formato utiliza **TAG** para estructurar a informação que será exibida en la pagina web. (Obs: **CSS** por otro lado indica como el elemento estructural debera ser exibido: color/fuente/tamaño/etc.)

- Tag HTML

```
<!-- Inicio y fin de pagina -->
<html></html>
```

- Tag HEAD

```
<head>
    <title>Introducción a NLP para Ciencias Sociales</title>
    <meta name="description" content="El mejor curso de NLP">
    <meta name="keywords" content="Reconocimiento de Entidades, Procesamiento de Lenguaje Natural, Extracción de Relaciones, Inteligencia Artificial">
</head>
```

- Tag Simples

```
<h1>Sobre el Curso</h1>
<p>Dirigido a Profesionales y estudiantes de carreras vinculadas a las Ciencias ...</p>
<p>Sumilla: Cantidades enormes de información relevante para las ciencias sociales son producidas diariamente ...</p>
<p>Página del curso: <a href="https://qlab.pucp.edu.pe/cursos/proc-lenguaje" target="_blank">NLP QLAB</a>.</p>
```

- Version 5

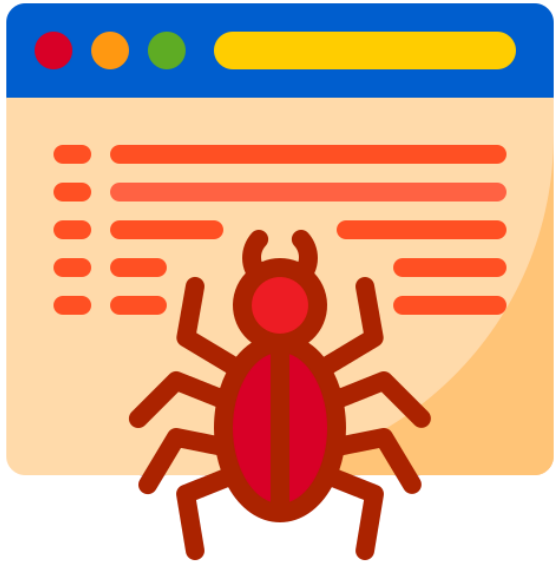
```
<!-- Elementos Semanticos -->
<header></header>
<footer></footer>
<article></article>
<section></section>

<!-- Elementos Gráficos -->
<svg></svg>
<canvas></canvas>

<!-- Elementos Multimedia -->
<audio></audio>
<video></video>
```

Como funciona Web Crawling?

También llamado bot o web spider es un algoritmo usado por los buscadores para encontrar, leer e indexar páginas web. Un ejemplo de crawler é o Googlebot de Google. Para aplicaciones especificas se recomienda el uso de bibliotecas open source como selenium o requests de python una flexibilidad mayor para recopilar información.



Redes Sociales

Las redes sociales son plataformas con alto interés para el análisis de información. Insight de productos, opiniones, tendencias son algunos ejemplos de información. La mayoría de redes sociales facilitan APIs que permite acceder de forma controlada, ordenada y eficiente a las informaciones de sus comunidades.

Seleinum (Python)

Esta biblioteca de permite de forma fácil acceder a todas las funcionalidades de Selenium WebDriver. Dispositivo que permite controlar um browser (Firefox, Chrome, IE) de forma remota.

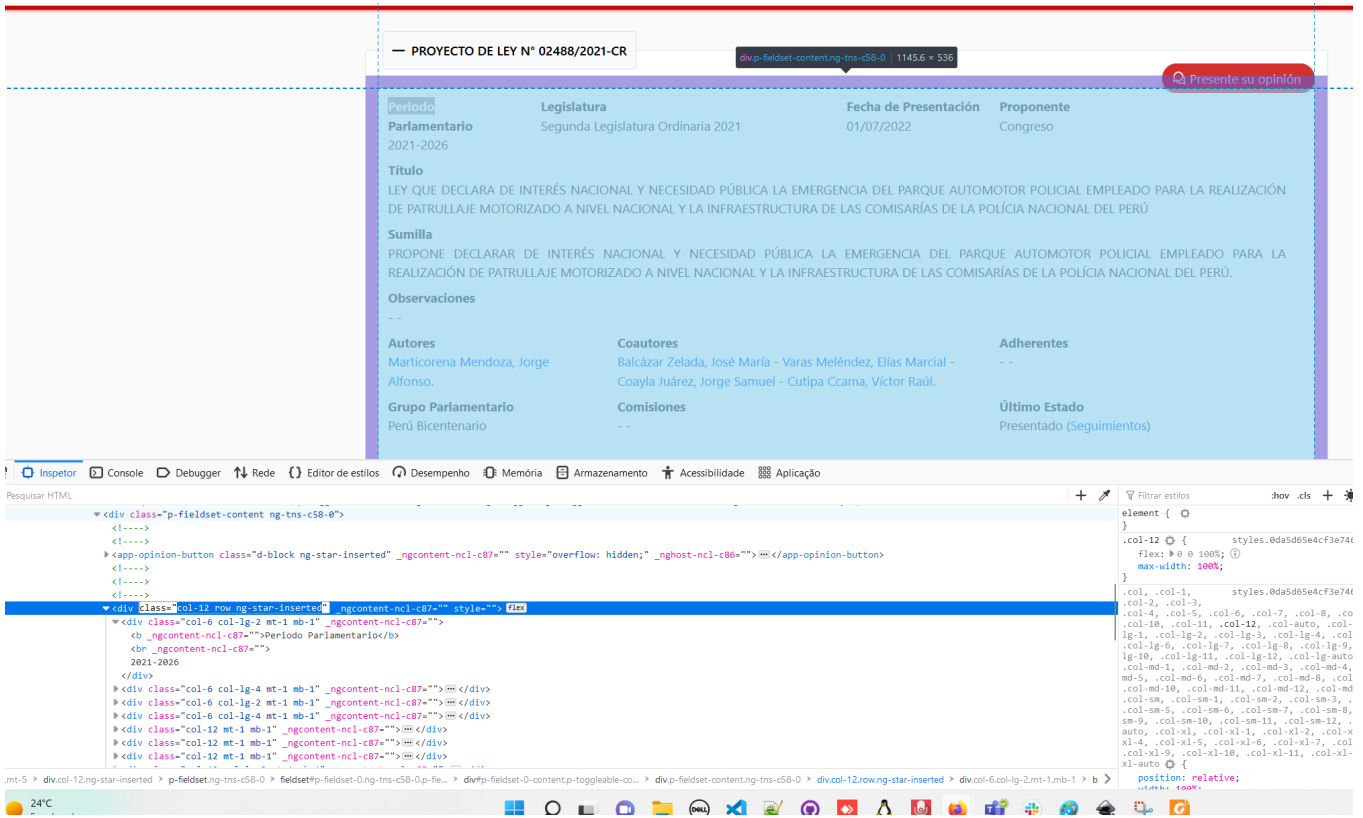
Site: https://selenium-python.readthedocs.io/

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By

driver = webdriver.Firefox()
driver.get("http://www.python.org")
elem = driver.find_element(By.NAME, "q")
driver.close()
```

Selenium (XPATH command)

```
//tagname[@attribute='value']
elem = driver.find_element(By.XPATH, '//div[@class="col-12 row ng-star-inserted"')
```



Tarea 1 :

```
Descarga de informaciones de proyectos de ley de la página del congreso de la república.
```