Maestría en Inteligencia Artificial Aplicada (MNA)

Breve Introducción NLP

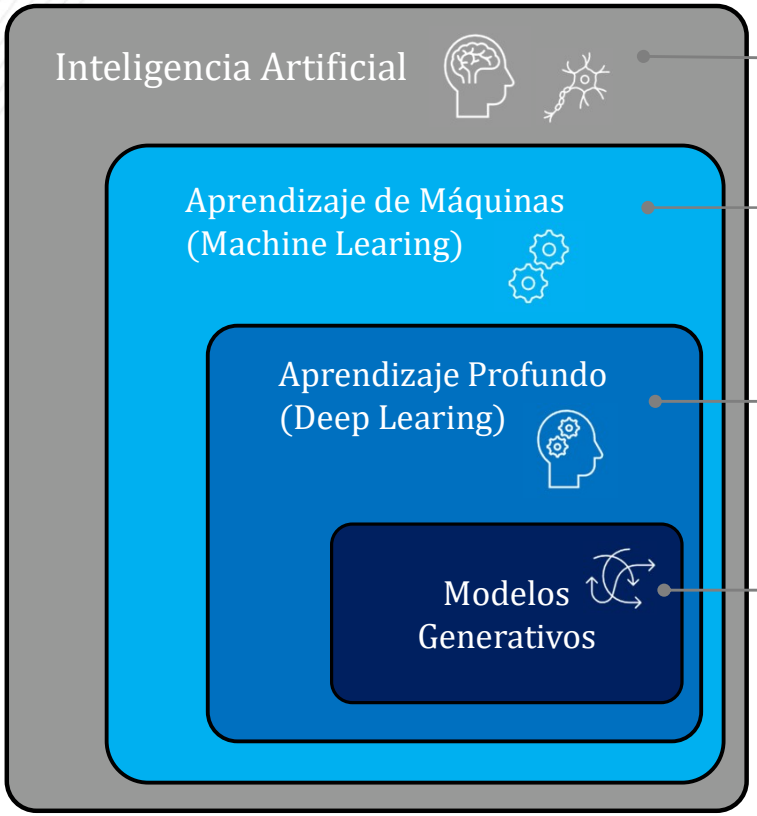
Procesamiento de Lenguaje Natural (PLN)
Natural Language Processing (NLP)

Luis Eduardo Falcón Morales

Introducción NLP

En esta semana repasaremos algunos de los sucesos más importantes que han ido formando lo que actualmente conocemos como el área de Procesamiento de Lenguaje Natural.

Estaremos usando en general las iniciales NLP por las siglas del nombre en inglés, Natural Language Processing. Aún en nuestro idioma es usual hacer referencia al tema con estas iniciales.



$1950's$

Inteligencia Artificial

Cualquier método que permita a una computadora simular el comportamiento humano.

$1980's$

Aprendizaje de Máquinas (Machine Learning-ML)

Área de la IA que aplica técnicas diversas para que las computadoras mejoren su desempeño con base a la experiencia.

$2010's$

Aprendizaje Profundo (Deep Learning-DL)

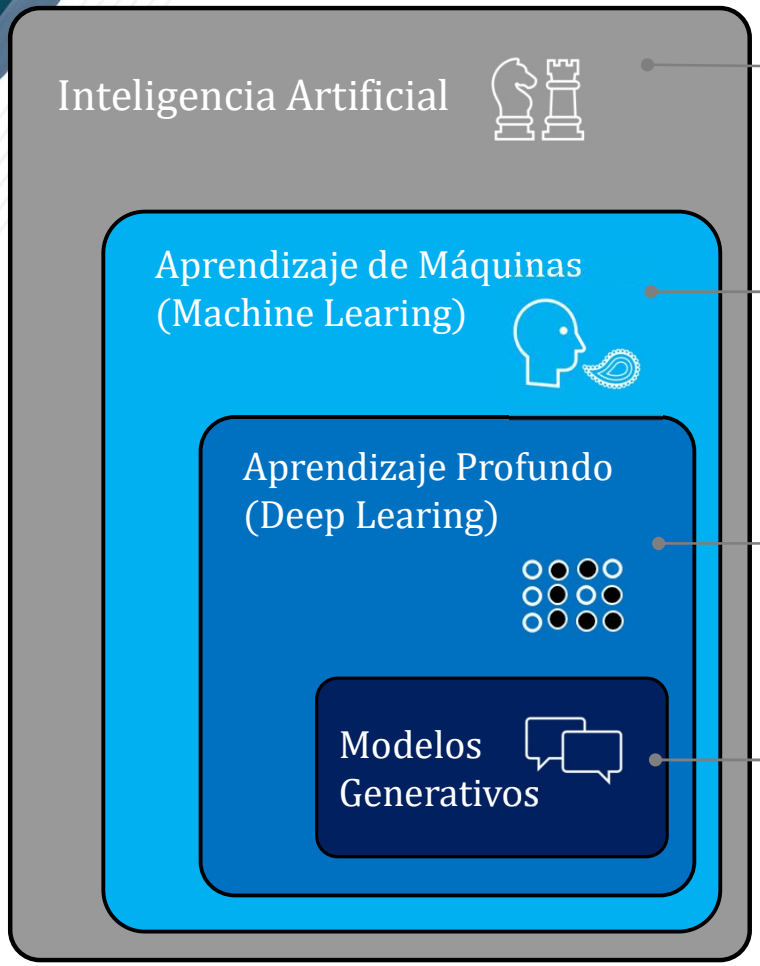
Área de ML que implementa modelos de redes neuronales profundas.

$2020's$

Modelos Generativos (Generative models)

Área de DL que genera datos de manera probabilística con base al conocimiento de la distribución de los datos.

Eventos mediáticos han permitido que el público en general, pero sobre todo instituciones de investigación, empresarios y gobiernos, volteen e inviertan en sistemas de IA:



IBM - Deep Blue 1996 / Deeper Blue 1997
Primera sistema de IA en ganar una partida a un campeón mundial de ajedrez, al ruso Garri Kaspárov y en ganar una serie completa con las reglas oficiales del ajedrez: 3.5-2.5.

IBM - Watson : Jeopardy 2011 (Machine Learning-ML)
Primer sistema de IA basado en NLP en ganar un concurso hombre vs máquina respondiendo a una serie de preguntas.

Google - Deep Mind : AlphaGo 2015 / AlphaZero 2017
Primera sistema de IA en ganar una partida de Go y en ganar una serie completa 4-1, al jugador profesional 9° dan, el surcoreano Lee Sedol.

OpenAI - Dall-E 2021 / ChatGPT 2022
Dall-E: sistema de IA que genera imágenes muy detalladas a partir de una breve descripción de texto. chatGPT: sistema de IA que genera texto de manera muy similar a la de un ser humano. Es el primer sistema que alcanza 1 millón de usuarios en apenas 5 días.

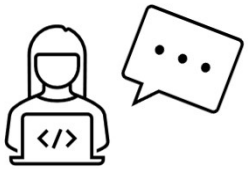
Inteligencia Artificial (IA)

- Aprendizaje Automático
  - Aprendizaje supervisado
  - Aprendizaje no supervisado
  - Aprendizaje por refuerzo
  - Aprendizaje profundo
- Procesamiento de Lenguaje Natural
  - Traducción automática
  - Extracción de contenido
  - Preguntas & Respuestas
  - Generadores de texto
- Visión Computacional
  - Reconocimiento de imágenes
  - Video análisis
  - Segmentación
  - Reconocimiento biométrico
- Audio
  - Audio a texto
  - Texto a audio
- Robótica
- Optimización
- Sistemas Expertos
- ⋮

¿Qué tipo de tareas desearíamos realizar con NLP?

Resumen de un libro:

Chatbots

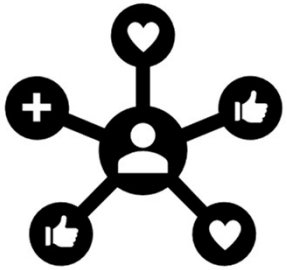


Traducir de un idioma a otro:

早安孙先生

Etiquetado automático de imágenes

Clasificación de noticias:



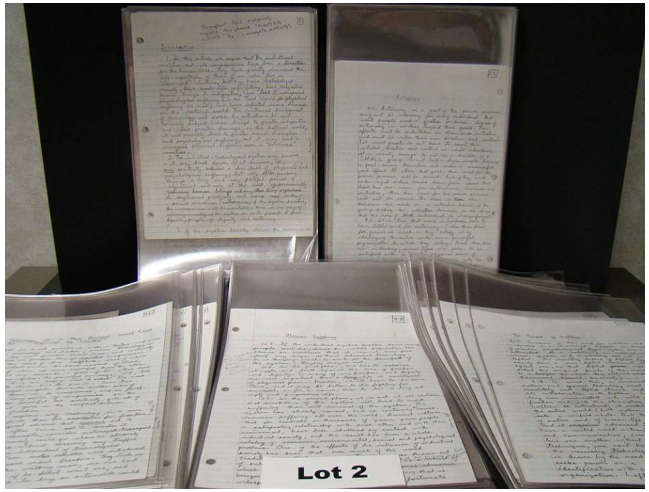
Entender los comentarios escritos en redes sociales

Identificar correos maliciosos

Técnicas de NLP para la identificación de asesinos seriales

A partir del caso del asesino serial conocido como UNABOMBER en EEUU, el FBI (Federal Bureau of Investigation) empezó a utilizar técnicas de análisis lingüístico para identificar/perfilar asesinos seriales a partir de la forma en que escriben.

Es común que asesinos seriales manden de manera incógnita cartas o documentos a la policía a manera de desafío o reto. Mientras más largo sean dichos documentos, mayor posibilidad de encontrar rasgos o características asociadas a la persona que lo escribió y que permitan dar con su identidad o paradero.



https://en.wikipedia.org/wiki/Ted_Kaczynski#/media/File:Unabomber_Auction_03.jpg

How Does Spotify Know You So Well?

Spotify fue de las primeras empresas en proponer utilizar información de sus usuarios más allá de lo que generan dentro de su propia plataforma, para sugerir recomendaciones más personalizadas.

A partir de entonces, los datos que generan los usuarios en sus diferentes redes sociales (Facebook, Instagram, Twitter, Tik-Tok, WhatsApp, etc), será información muy valiosa que se utilizará con diferentes sistemas de recomendación para generar recomendaciones más personalizadas en cualquier aplicación.

¿Qué se entiende por Procesamiento de Lenguaje Natural?

Lee primeramente las siguientes definiciones dadas por estas grandes compañías:

Google:

https://cloud.google.com/learn/what-is-natural-language-processing?hl=es

Amazon:

https://aws.amazon.com/es/what-is/nlp/

IBM:

https://www.ibm.com/mx-es/topics/natural-language-processing

Microsoft:

https://learn.microsoft.com/es-es/azure/architecture/data-guide/technology-choices/natural-language-processing

Oracle:

https://www.oracle.com/mx/artificial-intelligence/what-is-natural-language-processing/

Procesamiento del habla y del lenguaje natural

El entendimiento del lenguaje escrito y hablado no es un problema trivial, son problemáticas que históricamente se han abordado desde diferentes áreas de la ciencia, entre las que podemos mencionar las siguientes:

- **Lingüística**: Ciencia que estudia el lenguaje humano y las lenguas.
  - **Lingüística computacional**: modelos basados en reglas lógicas y de estadística para el modelado del lenguaje, de manera que una computadora pueda comprender su significado. La **Gramática** es el área de la Lingüística que se encarga en general de todas aquellas reglas y normas que nos permiten escribir y hablar correctamente.
- **Ciencias computacionales**:
  - **Procesamiento de lenguaje natural**: uso de técnicas de machine y deep learning para el modelado del lenguaje.
- **Ingeniería eléctrica**:
  - Reconocimiento del habla a través del audio.
- **Psicología**:
  - Psicología del lenguaje computacional o teoría computacional de la mente. Se basa en el supuesto de que existen reglas y procesos mentales internos que pueden ser extraídos, modelados y estructurados para la comprensión del pensamiento humano.

El área de NLP puede ser abordado desde diferentes perspectivas:

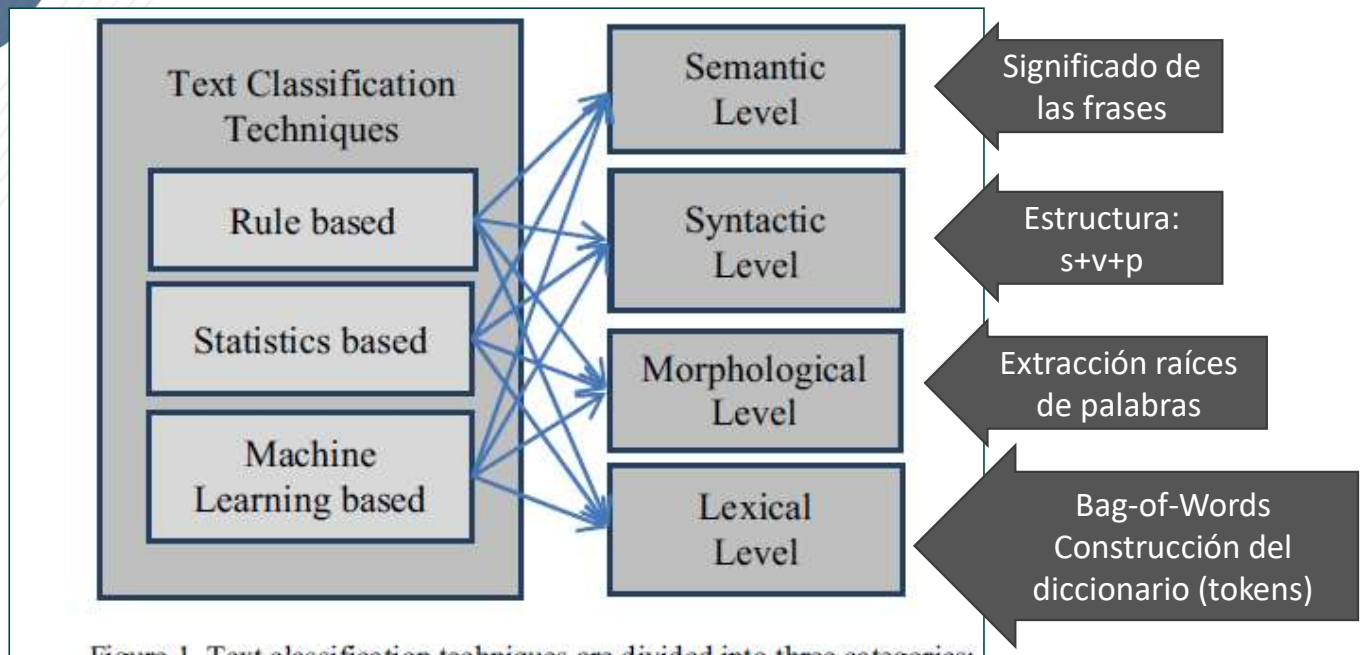


Figure 1. Text classification techniques are divided into three categories: rule based, statistics based and machine learning based. Each of these categories can either be implemented on the syntactic level, morphological level, semantic level or lexical level of the text being analyzed.

Semántico: extrae el significado de las frases.

Sintáctico: Estudia la manera en que se combinan las palabras a través del verbo, sujeto, predicado, etc. Y su relación jerárquica. Es decir, estudia la estructura gramatical de una frase. Árboles sintácticos.

Morfología: estudia la manera en que se forman las palabras. Extrae las raíces.
Por ejemplo:
Tele/visión/transportación/comunicación/…
Gat/o/a/os/as

Léxico: construcción del diccionario (significado) de cada palabra o caracter válido (token), símbolos, signos de puntuación, palabras reservadas, espacios en blanco, etc.

Ver referencia del artículo *Recognising user identity in twitter social networks via text mining*. S. Keretna, A. Hossny and D. Creighton. 2013:

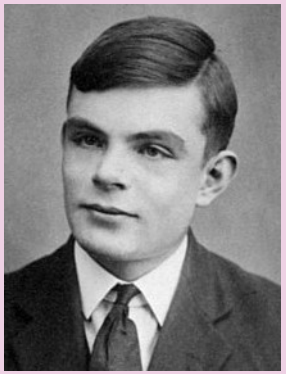
https://www.researchgate.net/publication/256761996_Recognize_User_Identity_in_Twitter_Social_Networks_via_Text_Mining

Primera mitad del siglo XX : Fundamentos

Matemáticas / Lógica / Algoritmia / Informática

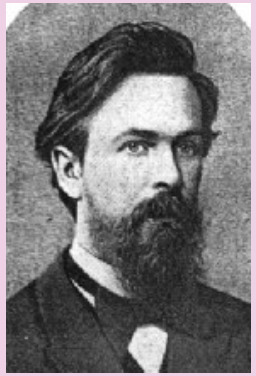
David Hilbert
(1862-1943)
alemán

Kurt Gödel
(1806-1978)
austriaco



Alan M. Turing
(1912-1954)
británico

Modelos probabilísticos



Andréi A. Márkov
(1856-1922)
ruso

Referencia de las imágenes de Wikipedia de la diapositiva anterior:

David Hilbert:

https://es.wikipedia.org/wiki/David_Hilbert#/media/Archivo:Hilbert.jpg

Kurt Gödel:

https://en.wikipedia.org/wiki/Kurt_G%C3%B6del#/media/File:Kurt_g%C3%B6del.jpg

Alan Turing:

https://es.wikipedia.org/wiki/Alan_Turing#/media/Archivo:Alan_Turing_Aged_16.jpg

Andréi Márkov:

https://es.wikipedia.org/wiki/Andr%C3%A9i_M%C3%A1rkov#/media/Archivo:Andrei_Markov.jpg

Prueba de Turing

Turing propone en su artículo de 1950 algunos tipos de diálogo que podrían darse entre el hombre y una máquina, para determinar si esta última "piensa", i.e., es "inteligente".

VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

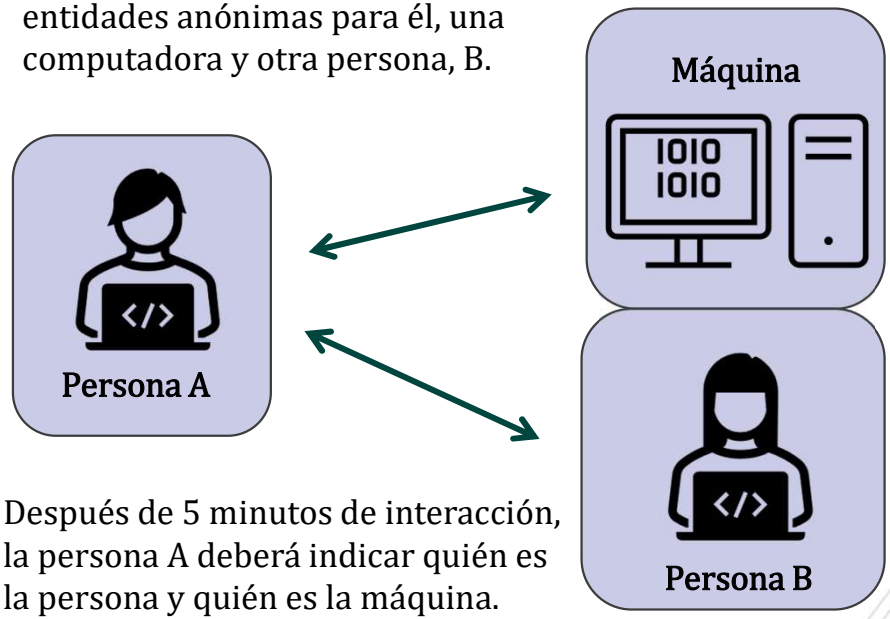
BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'.

https://academic.oup.com/mind/article/LIX/236/433/986238

La persona A interactúa con dos entidades anónimas para él, una computadora y otra persona, B.



Después de 5 minutos de interacción, la persona A deberá indicar quién es la persona y quién es la máquina.

Q : Please write me a sonnet on the subject of the Forth Bridge.
A : Count me out on this one. I never could write poetry.
Q : Add 34957 to 70764
A : (Pause about 30 seconds and then give as answer) 105621.
Q : Do you play chess?
A : Yes.
Q : I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A : (After a pause of 15 seconds) R-R8 mate.

Período de la guerra/posguerra – Décadas 40s y 50s

El estudio de NLP se inició con dos aproximaciones principalmente: la encabezada por Chomsky y su teoría del lenguaje formal, y la basada en la nueva área de estudio llamada inteligencia artificial.

Modelo de la neurona artificial

Walter Pitts
(1923-1969)
&
Warren
McCulloch
(1898-1969)
estadounidenses

Autómatas finitos / Sistemas secuenciales / Expresiones regulares

Stephen Kleene
(1909-1994)
estadounidense

Teoría de la Información / Entropía

Claude E. Shannon
(1916-2001)
estadounidense

En 1950 propone el ajedrez como área de desarrollo para llegar a construir un sistema “inteligente”.

Gramática de estados finitos / Teoría del lenguaje formal / Lingüística

Noam Chomsky
(1928-)
estadounidense

Referencias de imágenes de la diapositiva anterior:

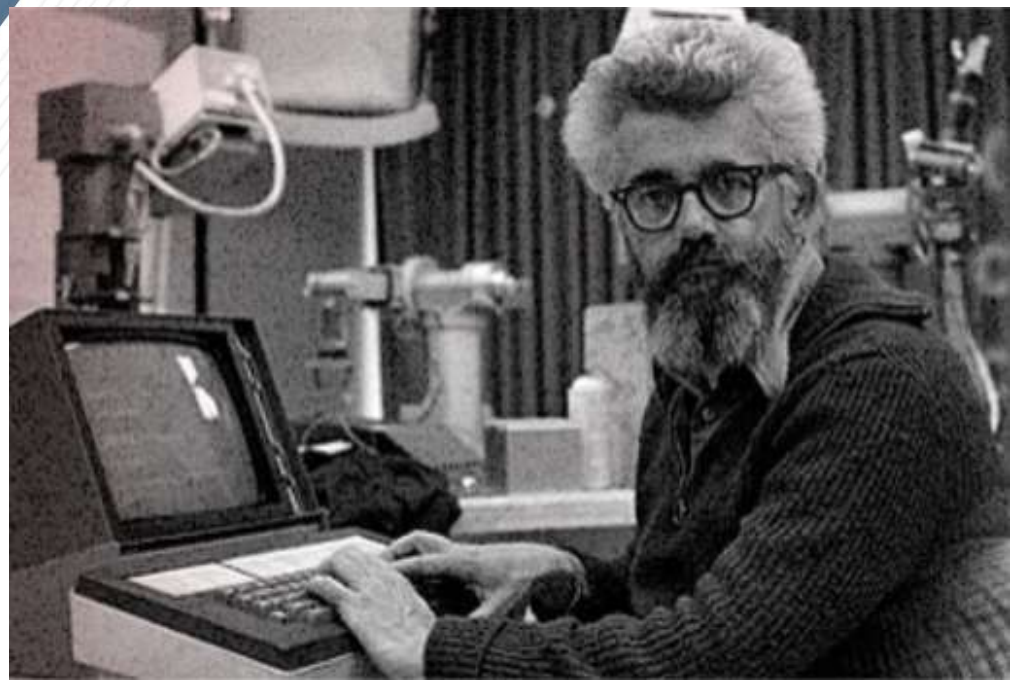
https://en.wikipedia.org/wiki/Walter_Pitts#/media/File:1954_Walter_Pitts_and_a_blackboard.jpg

https://home.csulb.edu/~cwallis/artificialn/warren_mcculloch.html

https://es.wikipedia.org/wiki/Stephen_Kleene#/media/Archivo:Kleene.jpg

https://es.wikipedia.org/wiki/Claude_Shannon#/media/Archivo:ClaudeShannon_MFO3807.jpg

https://es.wikipedia.org/wiki/Noam_Chomsky#/media/Archivo:Noam_Chomsky_portrait_2017.jpg



John McCarthy

Boston, EEUU
(1927 – 2011)

Matemático, informático y ciencias cognitivas

http://jmc.stanford.edu/general/pictures.html

McCarthy acuñó el término de *Inteligencia Artificial* en 1955 en la invitación que hizo para la Conferencia en Dartmouth a realizarse en 1956

A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

- We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire.

- The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

- An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

- We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

The original seven aspects of AI - 1955

1. Automatic computers.
2. How can a computer be programmed to use a language.
3. Neuron nets.
4. Theory of the size of a calculation.
5. Self-improvement.
6. Abstractions.
7. Randomness and creativity.

En estas décadas se incrementa el área de investigación y desarrollo en NLP desde diferentes áreas de estudio.

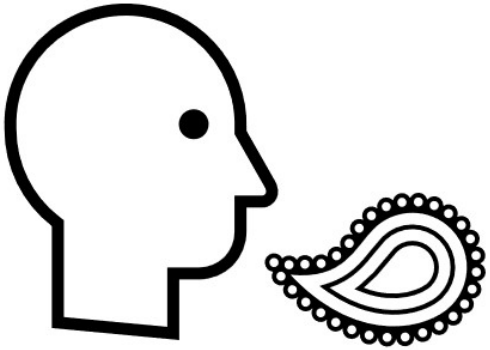
Décadas de los 60s – 70s – 80s

Brown corpus (1963)

- Se construye el primero de los corpus que puede ser leído por una computadora.
- Compilado por investigadores de la Universidad de Brown en EEUU.
- El corpus consiste de 1 millón de palabras extraídas de textos impresos en 1961 en inglés americano.
- Formado por 500 textos de diferentes temáticas, cada uno con ~2000 palabras.
- Los temas están agrupados en 15 categorías.

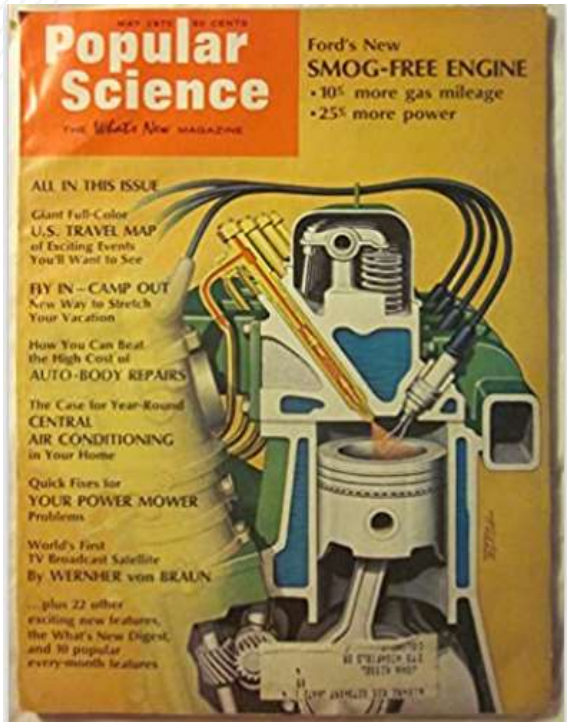
Project DOC (1967)

Diccionario informático para la investigación fonológica



https://www.aclweb.org/anthology/C69-7101.pdf

En ocasiones lo que imagina un escritor de ciencia ficción puede parecer irreal.



mayo 1970

Dr. Wernher von Braun - NASA

Arthur C. Clarke

https://books.google.com.ec/books?id=8QAAAAAAMBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

A look into the future

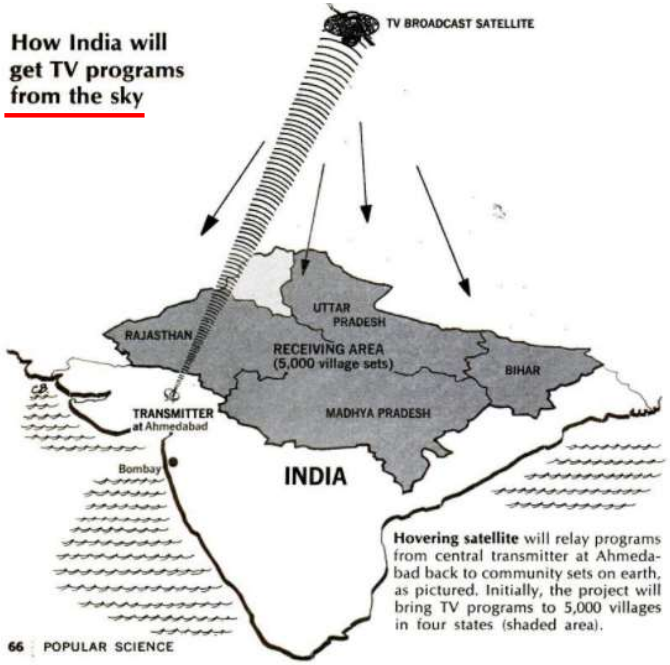
All in all, you are bound to benefit from communications satellites in your lifetime. Arthur C. Clarke, the perceptive British writer and originator of the idea of synchronous communications satellites, sees this prospect ahead:

Imagine a console in your office, he says, combining the features of a Touch-Tone (pushbutton) telephone, a television set, a Xerox machine, and a small electronic computer. Tuned in to a system of synchronous satellites, this console will bring the accumulated knowledge of the world to your fingertips.

By punching a few digits, you can verify a check, get the data on some historical event, or hear an illustrated lecture on any subject you wish. Or you can hold an electronic conference with any group of people scattered all over the world, seeing each other as you talk. And the console will even provide you with a document to seal an agreement.

Once we all have such a console, Clarke foresees, we'll have little need for business trips, and will limit our traveling to the pursuit of fun. A man will be able to run his business just as effectively from a ranch in Wyoming as from an office in New York City or Chicago.

How India will get TV programs from the sky



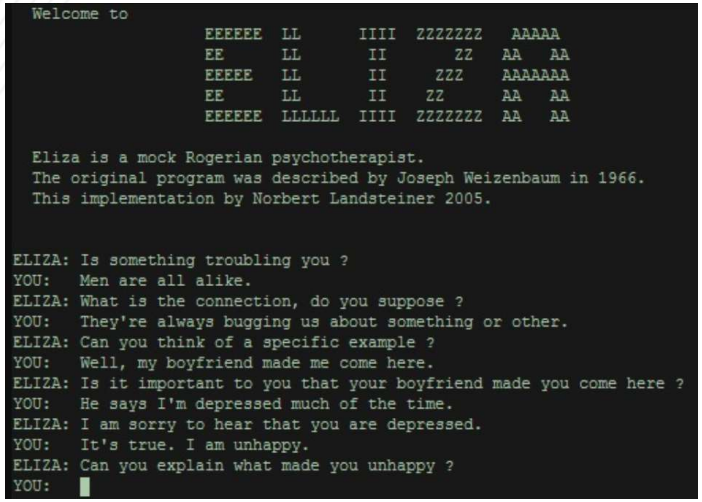
Hovering satellite will relay programs from central transmitter at Ahmedabad back to community sets on earth, as pictured. Initially, the project will bring TV programs to 5,000 villages in four states (shaded area).

66 POPULAR SCIENCE

If Clarke is right, better communication via satellite could stem or reverse America's trend toward the big city—and alleviate problems like urban decay, smog, polluted rivers, clogged traffic, and crowded airports. These effects, of course, would be long-term. But, just as the telephone and airliner helped make the U.S. a great nation, so the communications satellite may yet turn out to be the greatest thing the space program ever contributed to this country. PS

Mediante el uso de reglas lógicas las nuevas propuestas empiezan a aceptar preguntas/comandos de texto en lenguaje natural:

Reconocimiento de patrones & Métodos de sustitución / efecto ELIZA



MIT
(1964-1966)

https://en.wikipedia.org/wiki/ELIZA#/media/File:ELIZA_conversation.png

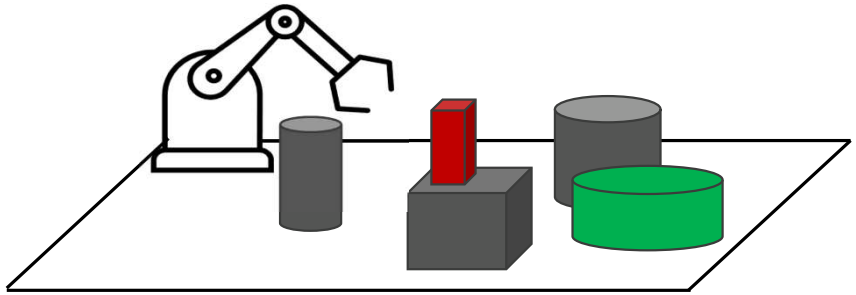
Joseph Weizenbaum
(1923-2008)
alemán - estadounidense

Basado en la búsqueda de patrones y sustitución de expresiones regulares:

```
s/.* YOU ARE (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/
s/.* YOU ARE (depressed|sad) .*/WHY DO YOU THINK YOU ARE \1/
s/.* all .*/IN WHAT WAY/
s/.* always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/
```

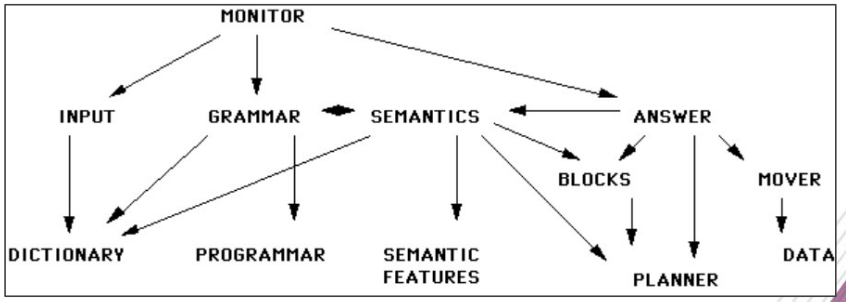
http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

Representación semántica: Natural Language Understanding



PICK UP THE RED BLOCK

SHRDLU (1968-1970) : Interacción virtual
Terry Winograd (1946-)



http://users.cs.cf.ac.uk/Dave.Marshall/AI1/COPY/shrdlu.html

Década de los 90s

- Modelos basados en datos
- Modelos probabilísticos
- Incremento de recursos computacionales: memoria y velocidad
- Telefonía celular & Videojuegos
- Surgimiento de la World-Wide-Web (WWW)
- Inicia el trabajo del "grupo de Toronto": Geoffrey Hinton, Yan LeCun, Yoshua Bengio

1989 – CERN
1990 – 1° navegador

https://es.wikipedia.org/wiki/ViolaWWW

1992 – 1° navegador gráfico

1994

1996 – mensajería instantánea

1984 – Nintendo

IBM – Deep & Deeper Blue

1996 & 1997

Robin Li → Larry Page & Sergey Brin
1996 : RankDex → PageRank algorithms
Matrix Factorization/EigenVectors/Markov chains

Referencias de imágenes de la diapositiva anterior:

https://es.wikipedia.org/wiki/ICQ

https://es.wikipedia.org/wiki/Netscape_Navigator#/media/Archivo:Netscape_logo.svg

https://es.wikipedia.org/wiki/Nintendo#/media/Archivo:NES-Console-Set.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/2/2f/Google_2015_logo.svg/300px-Google_2015_logo.svg.png

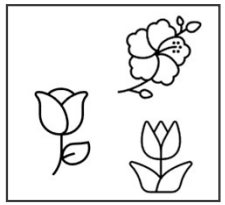
Turing Award – 2018

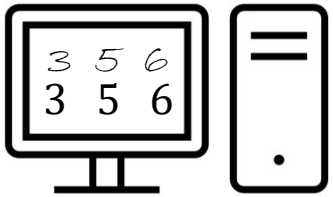
Geoffrey Hinton
británico 1947
Google Brain & UniToronto

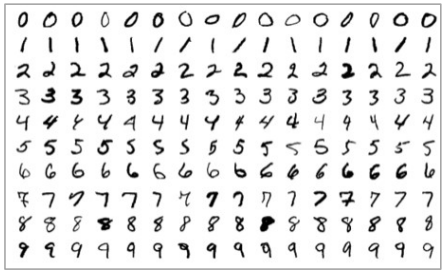
Yann LeCun
francés 1960
Facebook & UniNY

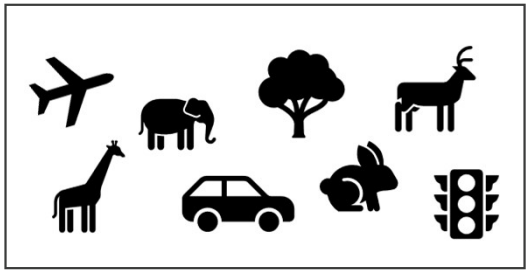
Yoshua Bengio
francés 1964
Uni Montreal & asesor en IA

- Backpropagation (chain rule): 1986
- CNN : 90s
- CNN + GPUs : 2012
- CNN + Hidden Markov Models : 90s
- High-dim Word embeddings : 2000
- GAN: Generative Adversarial Nets : 2010
- Dropout : 2014
- Encoder – Decoder : 2014
- Context vector / Attention concept: 2016
- …









Iris dataset (150 : Ronald Fisher, 1936) → MNIST (60,000 : 1998) → ImageNet (1’000,000 : 2010)

El tamaño de las bases de datos siguen creciendo exponencialmente

Referencias de imágenes de la diapositiva anterior:

https://es.wikipedia.org/wiki/Geoffrey_Hinton#/media/Archivo:Deep_Thinkers_on_Deep_Learning_(cropped_to_Geoffrey_Hinton).jpg

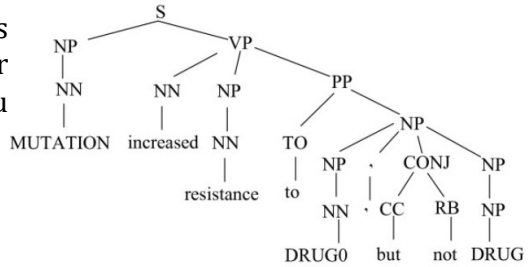
https://es.wikipedia.org/wiki/Yann_LeCun#/media/Archivo:Yann_LeCun_-_2018_(cropped).jpg

https://es.wikipedia.org/wiki/Yoshua_Bengio#/media/Archivo:Yoshua_Bengio_2019_cropped.jpg

Década del 00s

Linguistic Data Consortium

Se incrementan las bases de datos numéricos, texto y voz compartidas por grandes coorporaciones, incluyendo su estructura sintáctica (TreeBank).

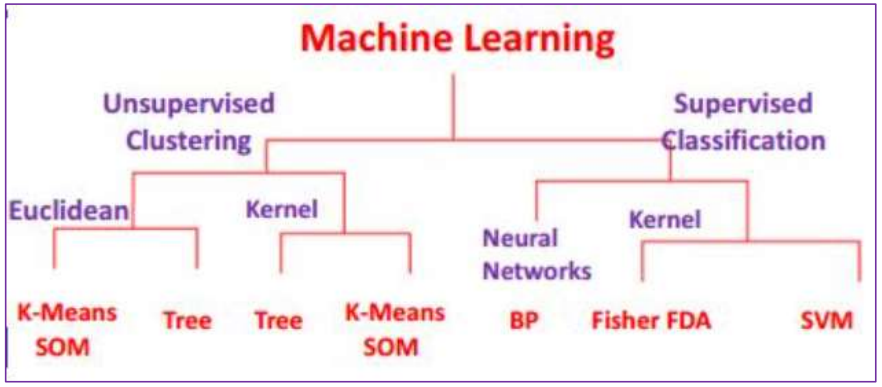


BigData:

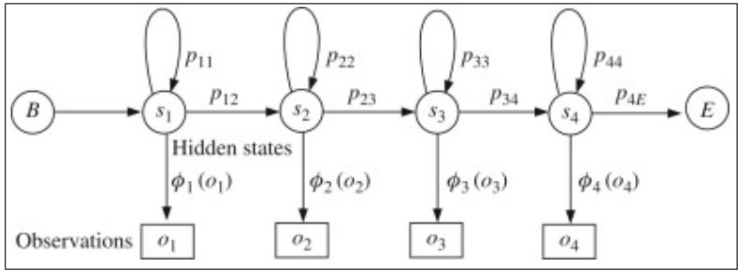
- Volume
- Velocity
- Variety
- Veracity

facebook (2004)

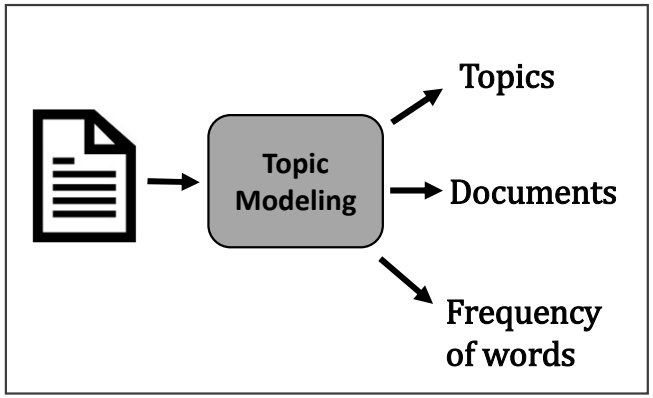
(2006) twitter



Se incrementa el uso, diseño y mejora de los modelos de aprendizaje y Estadística



CNN+HMM+Word Embeddings:
2000/Yoshua Bengio/Geoffrey Hinton



machine translation / chat bots

Década del 2010s

Siri (2011)

(2014)

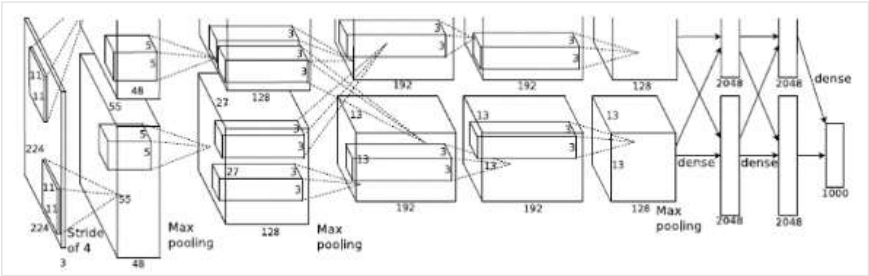
amazon alexa

Jeopardy!
The IBM challenge

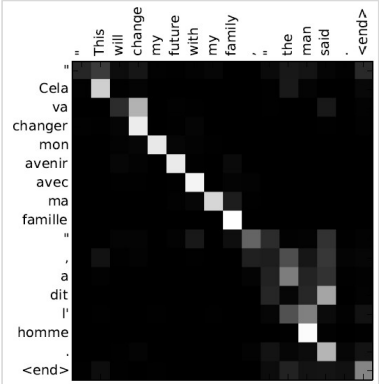
(2011)

AlphaGo (2016)

Google
Deep Mind

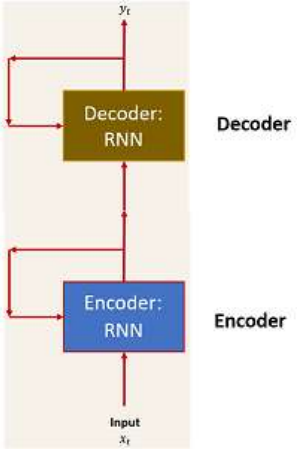


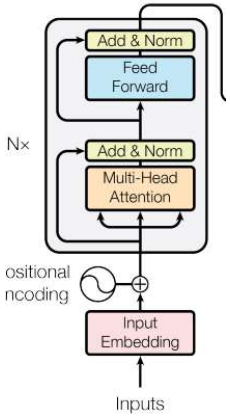
GPUs + Deep Learning (CNN)
1979/K. Fukushima → 1989/Yann LeCun →
2012/Alex Krizhevsky/Geoffrey Hinton



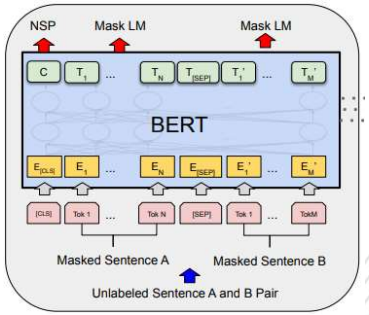
Sequence2Sequence (2014):
Context vector (2016)

1982/Hopfield → 1993/RNN →
1997/LSTM → 2015/Speech Recognition





Attention &
Transofrmer: 2017



2018
BERT - Google

Referencias de imágenes de la diapositiva anterior:

https://en.wikipedia.org/wiki/Siri#/media/File:Siri_Logo_in_2022.png

https://en.wikipedia.org/wiki/Amazon_Alexa

https://es.wikipedia.org/wiki/AlphaGo#/media/Archivo:Alphago_logo_Reversed.svg

https://arxiv.org/abs/1706.03762 Attention

https://arxiv.org/abs/1810.04805 BERT

https://arxiv.org/abs/1409.0473 Seq2Seq

https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html AlexNet

Década del 2020 ...

Modelos Generativos : LLM (Large Language Models)

OpenAI:
chatGPT
DALL-E

+

Microsoft

Bing
Office:
- Word
- Power Point
- Outlook

Google

Bard AI

+

Alphabet

Referencias de imágenes de la diapositiva anterior:

https://upload.wikimedia.org/wikipedia/commons/thumb/0/04/ChatGPT_logo.svg/800px-ChatGPT_logo.svg.png

https://upload.wikimedia.org/wikipedia/commons/thumb/9/96/Microsoft_logo_%282012%29.svg/300px-Microsoft_logo_%282012%29.svg.png

https://upload.wikimedia.org/wikipedia/commons/thumb/2/2f/Google_2015_logo.svg/300px-Google_2015_logo.svg.png

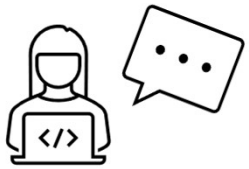
https://upload.wikimedia.org/wikipedia/commons/thumb/7/7a/Alphabet_Inc_Logo_2015.svg/480px-Alphabet_Inc_Logo_2015.svg.png

Procesamiento de Lenguaje Natural
Natural Language Processing (NLP)

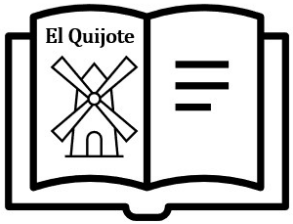
En resumen, podemos decir que NLP es el área de las ciencias computacionales que se encarga del estudio del lenguaje y la lingüística computacional, dentro del área de Inteligencia Artificial.

El objetivo de NLP es el diseño de aplicaciones que faciliten la interacción de las personas con las computadoras, a través del lenguaje natural.

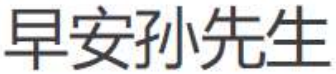
Principales áreas de NLP:



Question & Answer system

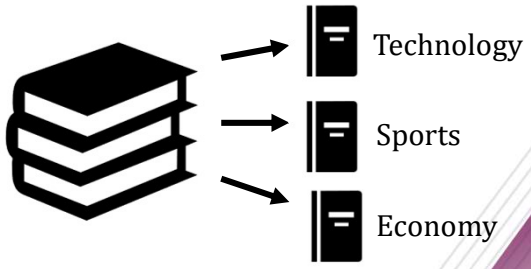


Summary



Machine Translation

Speech recognition



Document classification

Tecnológico de Monterrey

D.R.© Tecnológico de Monterrey, México, 2022.
Prohibida la reproducción total o parcial
de esta obra sin expresa autorización del
Tecnológico de Monterrey.