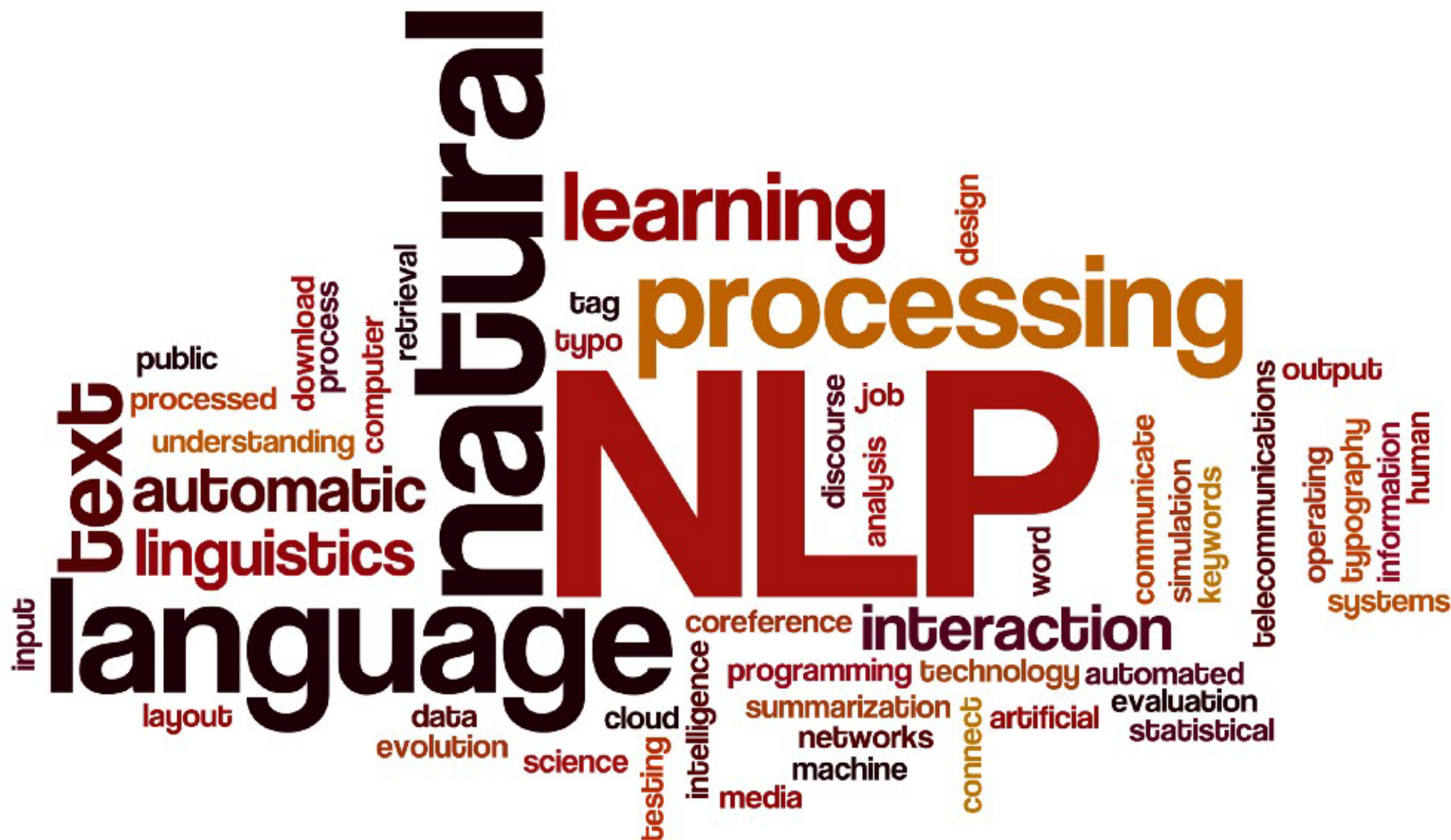




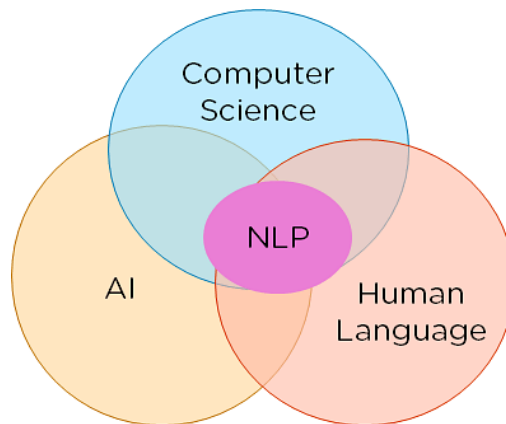
Procesamiento del Lenguaje Natural

¿Qué es el procesamiento del lenguaje natural?

- ✓ Término genérico que abarca todo aquello que permite a las máquinas *procesar* el *lenguaje humano* tanto en forma *escrita, verbal, o visual*.

¿Porqué es importante el procesamiento del lenguaje natural?

- ✓ Componente/Capacidad fundamental de los sistemas de IA.



Capacidades de un sistema de IA

- Percepción
- Aprendizaje
- Representación del conocimiento
- Razonamiento

Procesamiento del Lenguaje Natural

Comprensión de la información

El objetivo final es **comprender** el mensaje codificado en el lenguaje.

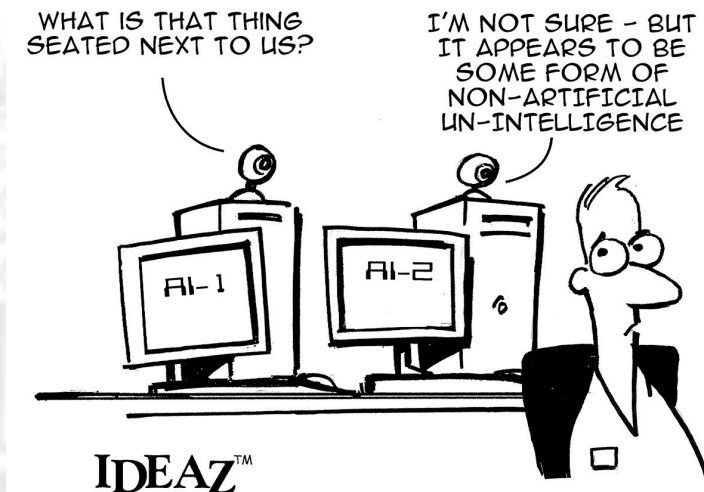
Comprender

Percibir y tener una idea clara de lo que se dice, se hace o sucede o descubrir el sentido profundo de algo.

Implica entender conceptos y procesos para poder explicarlos y describirlos de forma adecuada.



Nos proporciona herramientas para representar el **conocimiento**



Procesamiento del Lenguaje Natural

Representación del conocimiento

formalismos de representación del conocimiento:

Redes Semánticas (relaciones semánticas entre objetos en una red)

Frames(marcos), colección de datos estructurados
slots (propiedades) & fillers (valores) & Métodos

Reglas

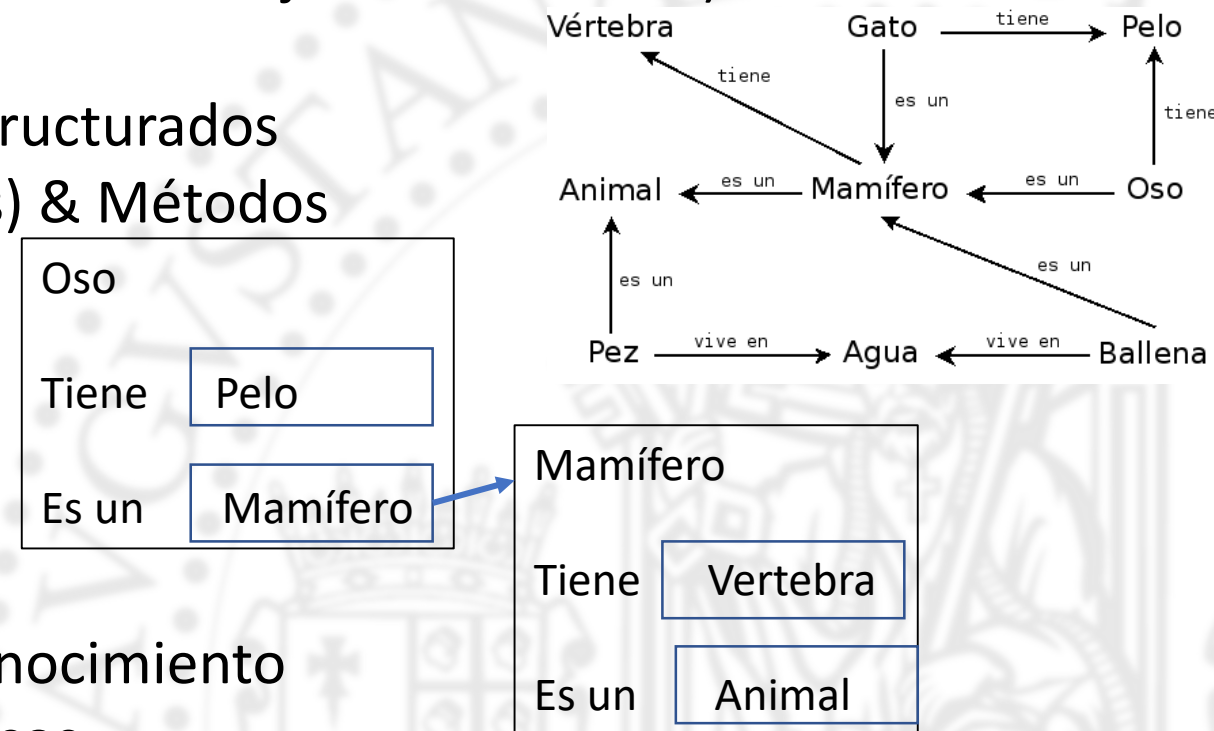
Si <condición>
entonces <conclusión>

Ontologías

esquemas de representación del conocimiento
basados en redes semánticas o marcos.

Espacios semánticos

Representación vectorial capaz de capturar el significado



Procesamiento del Lenguaje Natural

¿Qué tareas podemos hacer?

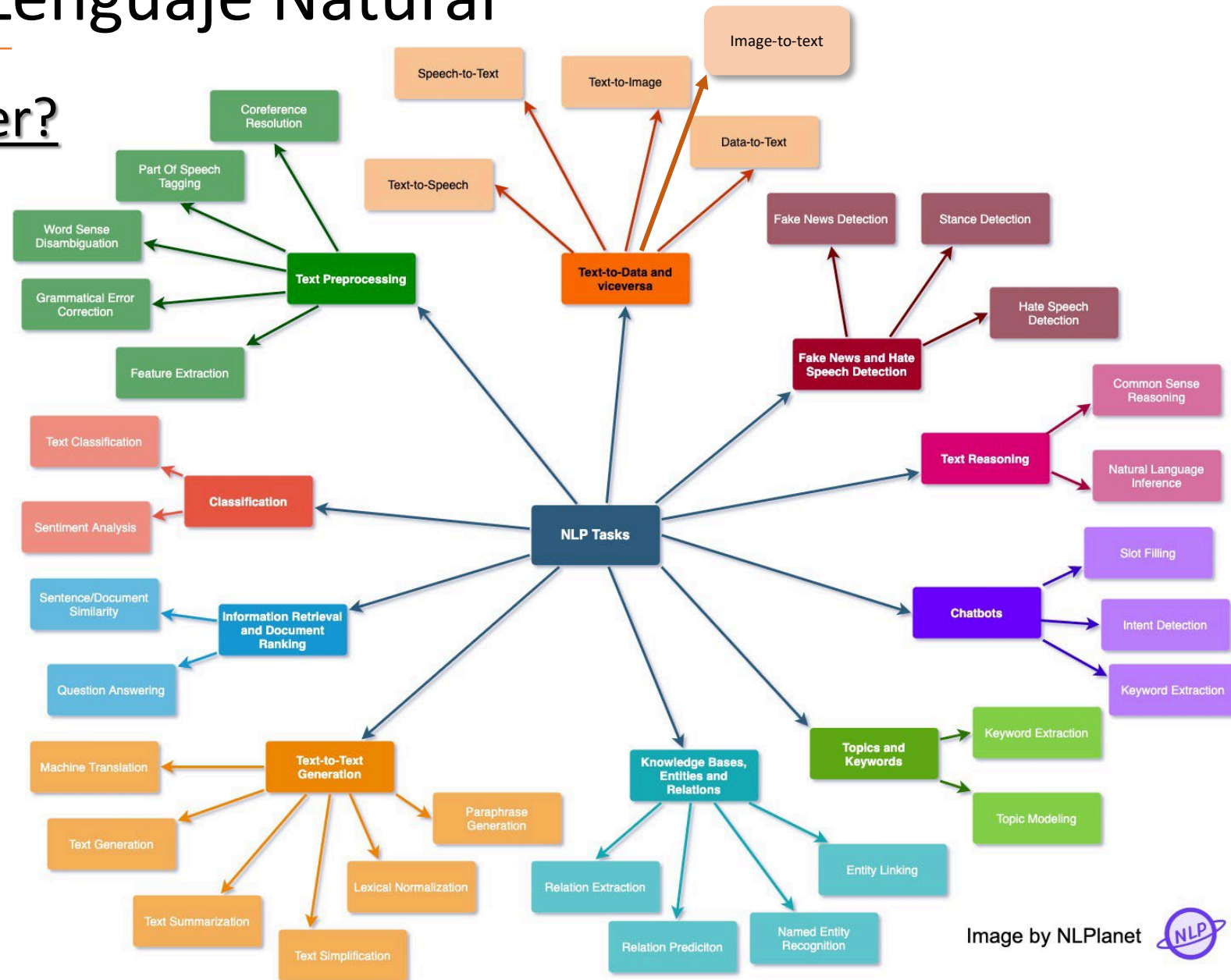
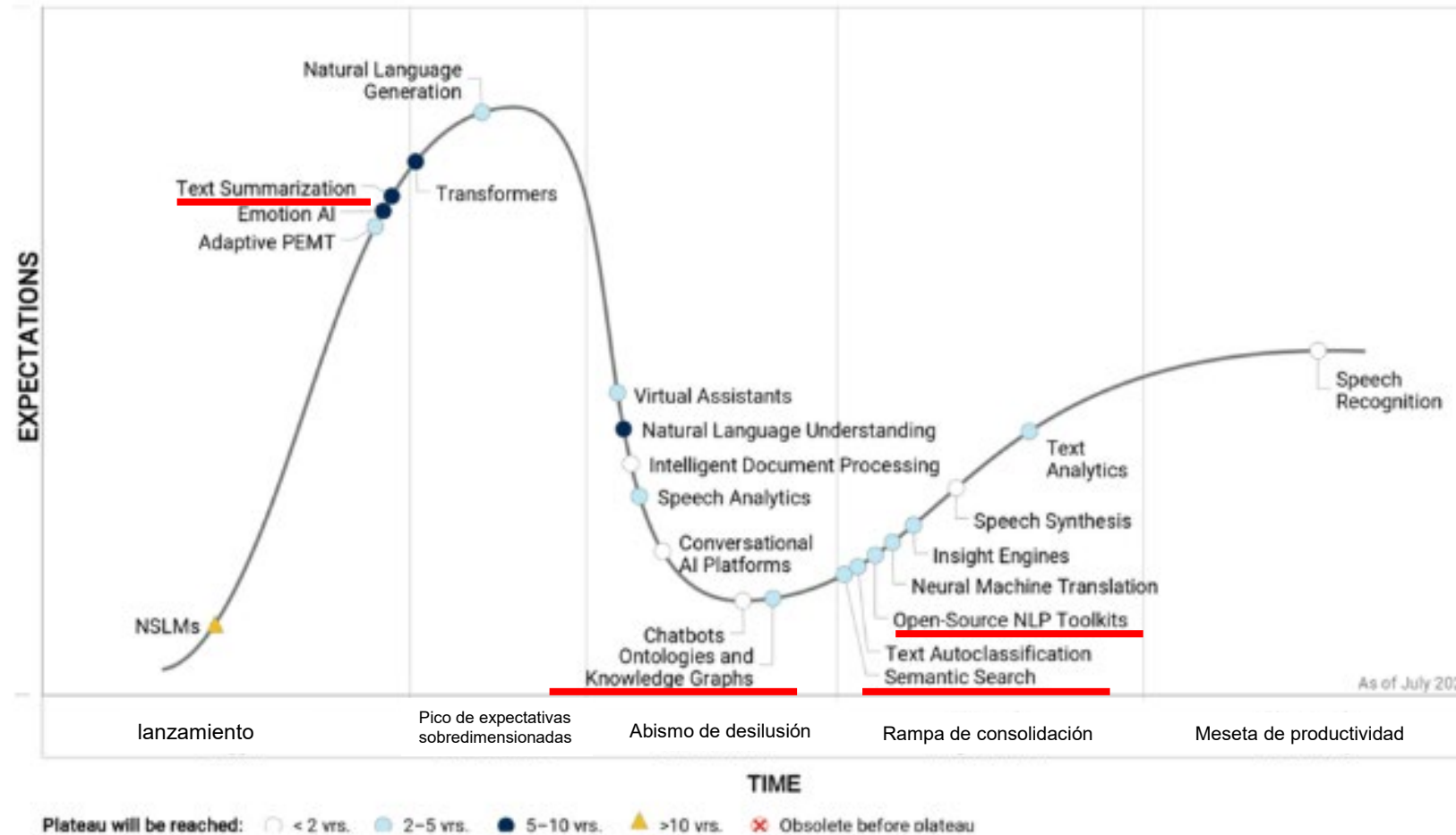


Image by NLPlanet 

Procesamiento del Lenguaje Natural

Hype Cycle for Natural Language Technologies, 2021

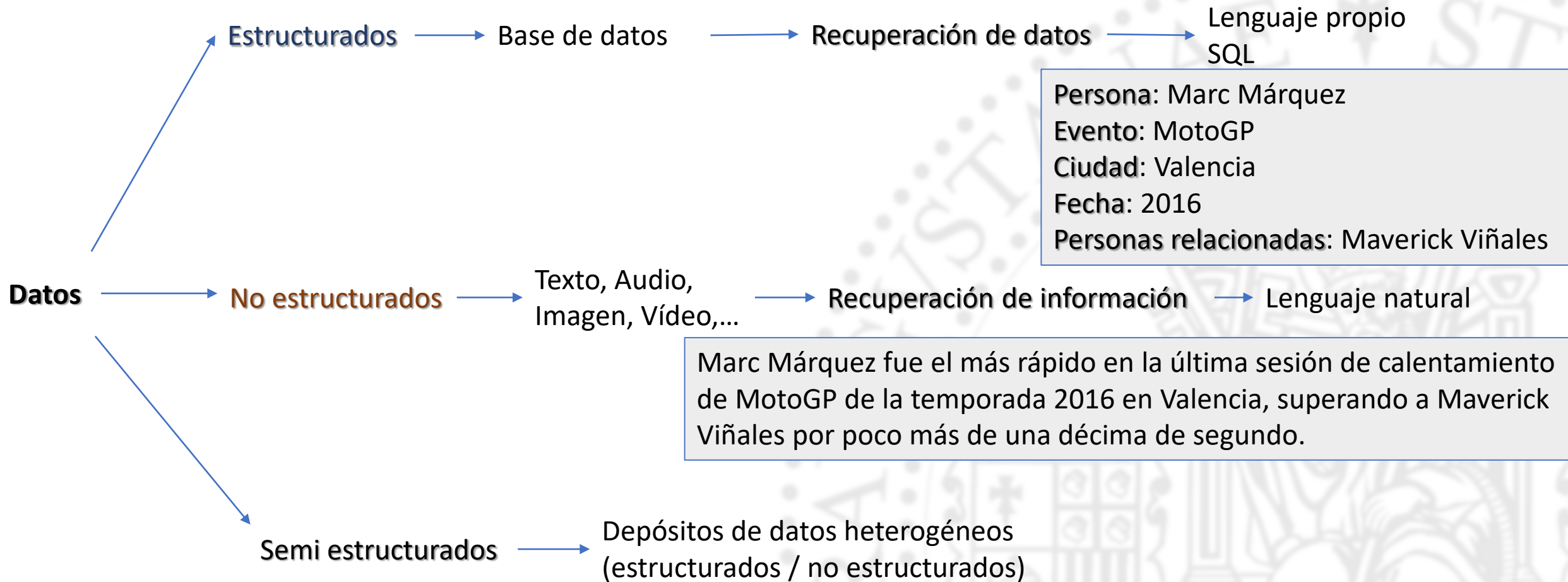


Source: Gartner (July 2021)

748656

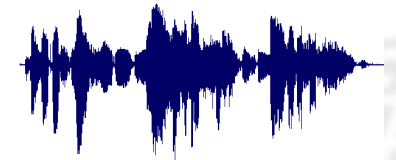
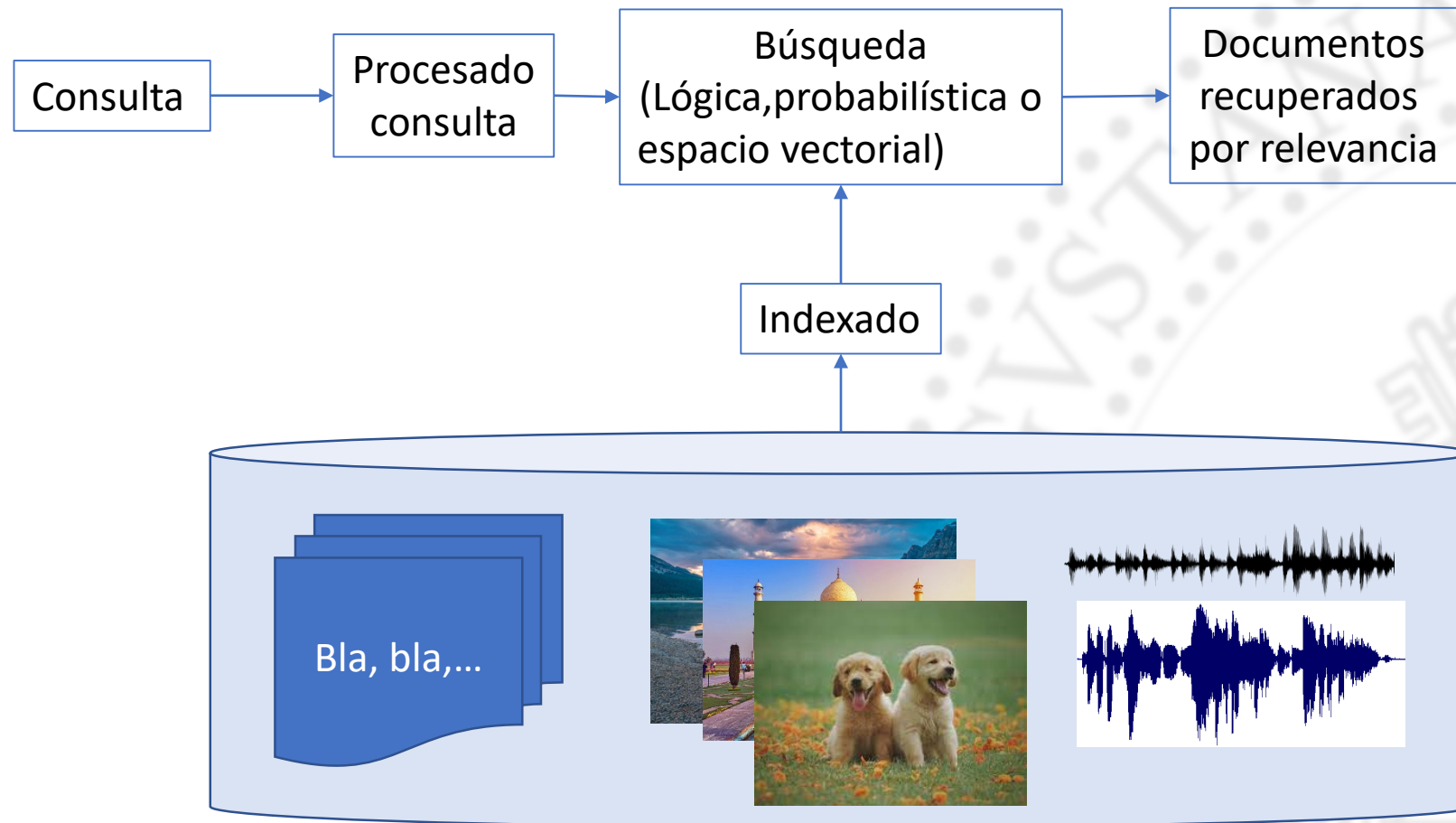
Procesamiento del Lenguaje Natural

Búsqueda y recuperación de información



Procesamiento del Lenguaje Natural

Búsqueda y recuperación de información



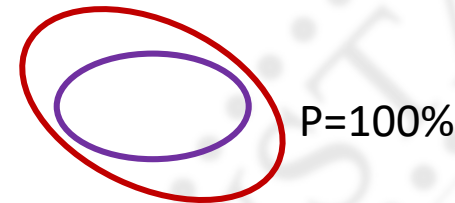
Procesamiento del Lenguaje Natural

Medida de prestaciones en búsquedas y recuperación de información:

Precisión y Exhaustividad/Sensibilidad (Precision/Recall), Valor-F (F-score)

Precisión: Probabilidad de que un documento recuperado sea relevante.

$$P = \frac{|\{\text{documentos relevantes y recuperados}\}|}{|\{\text{documentos recuperados}\}|}$$



Exhaustividad: Probabilidad de que un documento relevante sea recuperado en una búsqueda.

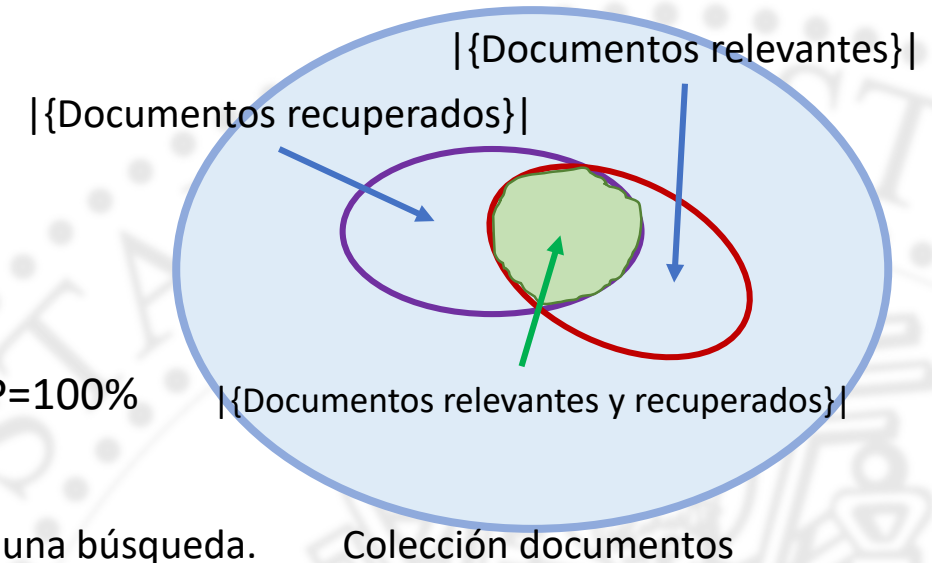
$$R = \frac{|\{\text{documentos relevantes y recuperados}\}|}{|\{\text{documentos relevantes}\}|}$$



Valor-F: Valor único ponderado de la Precisión y la Exhaustividad.

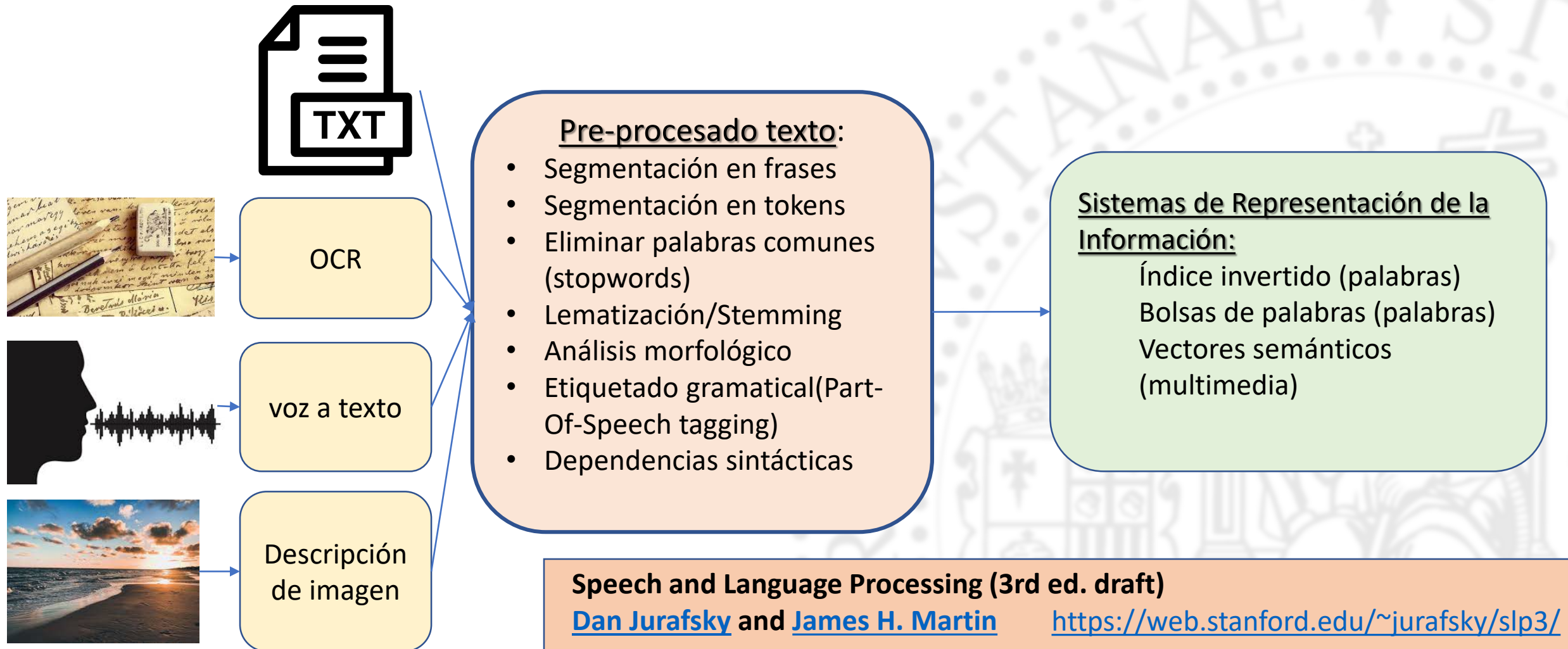
$$F = 2 \frac{\text{Precisión} \times \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precisión} \times \text{Exhaustividad}}{\beta^2 \times \text{Precisión} + \text{Exhaustividad}}$$



Procesamiento del Lenguaje Natural

PLN para la búsqueda y recuperación de información: Procesado consulta



Procesamiento del Lenguaje Natural



voz a texto

Raw data

Segmentación en frases

Tokenización

Lematización

Netflix ha encontrado en el juego del calamar su nuevo fenómeno mundial ni siquiera en la propia plataforma contaban con ello como seguro que tampoco esperaban recibir multitud de quejas por una escena del cuarto episodio sin embargo han estado rápidos para responder a la indignación del público y ha introducido un cambio en el equipo

Netflix ha encontrado en el juego del calamar su nuevo fenómeno mundial. Ni siquiera en la propia plataforma contaban con ello como seguro que tampoco esperaban recibir multitud de quejas por una escena del cuarto episodio. Sin embargo han estado rápidos para responder a la indignación del público y ha introducido un cambio en el equipo.

| netflix | ha | encontrado | en | el | juego | de | el | calamar | su | nuevo |
fenómeno | mundial | . |
| ni | siquiera | en | la | propia | plataforma | contaban | con | ello | ...

| netflix | haber | encontrar | en | el | juego | de | el | calamar | su | nuevo |
fenómeno | mundial | . |
| ni | siquiera | en | el | propio | plataforma | contar | con | ello | ...

Procesamiento del Lenguaje Natural

POS tagging

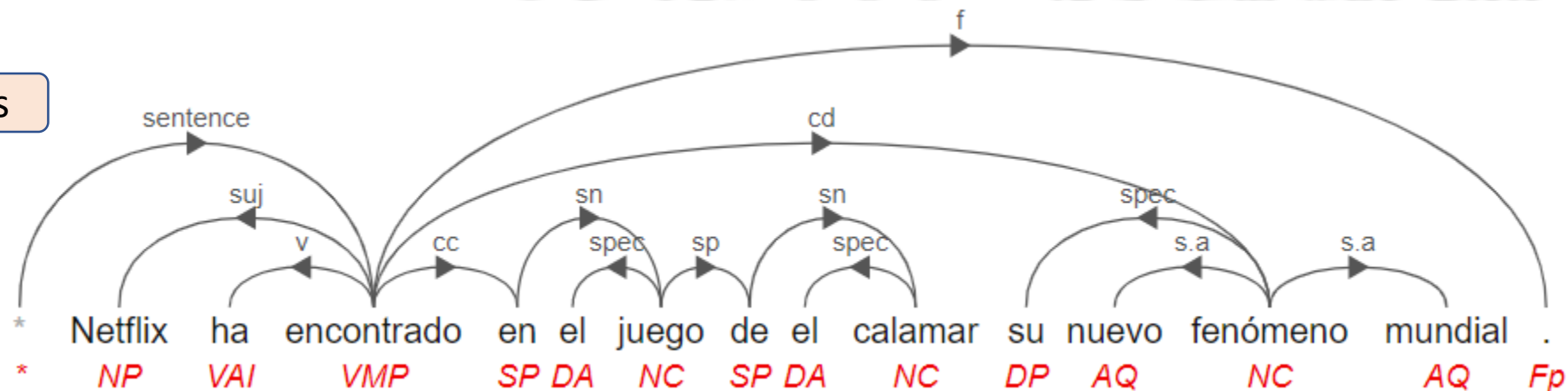


| Netflix [NP00SP0] | haber [VAIP3S0] | encontrar [VMP00SM] | en [SP] | el [DA0MS0] | juego [NCMS000] | de [SP] | el [DA0MS0] | calamar [NCMS000] | su [DP3CSN] | nuevo [AQ0MS00] | fenómeno [NCMS000] | mundial [AQ0CS00] | . [Fp] |
| ni [CC] | siquiera [RG] | en [SP] | el [DA0FS0] | propio [AQ0FS00] | plataforma [NCFS000] | contar [VMII3P0] | con [SP] | ello [PD00S00] |

Quitar stopwords

| Netflix [NP00SP0] encontrar [VMP00SM] | juego [NCMS000] | calamar [NCMS000] | nuevo [AQ0MS00] | fenómeno [NCMS000] | mundial [AQ0CS00] | . [Fp] |
| ni [CC] | siquiera [RG] | propio [AQ0FS00] | plataforma [NCFS000] | contar [VMII3P0] |

Dependencias sintácticas



Procesamiento del Lenguaje Natural

Recursos:

Freeling (<https://nlp.lsi.upc.edu/freeling/index.php/>) permite: análisis morfológico, detección de entidades, POS-tagging, desambiguación del significado de palabras, análisis sintáctico, etiquetado de la función semántica...
Demo on-line:

<https://nlp.lsi.upc.edu/freeling/demo/demo.php>

SPACY (<https://spacy.io/>) toolkit en Python con el estado del arte en técnicas de procesamiento del lenguaje natural
Demos:

<https://spacy.io/universe>

GENSIM (<https://radimrehurek.com/gensim/>) toolkit en Python para el modelado de documentos y recuperación de información
Demos:

https://radimrehurek.com/gensim/auto_examples/index.html#documentation

NLTK (<https://www.nltk.org/>), Natural Language ToolKit *“leading platform for building Python programs to work with human language data”*

PATTERN (<https://github.com/clips/pattern/wiki>), *“web mining module for the Python programming language”*, NLP

TextBlob (<https://textblob.readthedocs.io>), *“API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.”*

Basada en NLTK y PATTERN

Procesamiento del Lenguaje Natural

Pero, ¿cómo representamos los tokens/palabras en una máquina?

La *máquina trabaja con números*,

.... luego *debemos transformar las palabras a números*

Opción 1.

Utilizamos un código numérico, p.e. las numeramos de forma correlativa

netflix	1
encontrar	2
juego	3
calamar	4
nuevo	5
fenómeno	6
mundial	7
ni	8
siquiera	9
propio	10
plataforma	11
contar	12

¿tiene algún significado el valor numérico?

¿Podemos calcular la proximidad semántica?, ¿tiene sentido?

Procesamiento del Lenguaje Natural

Opción 2.

Definimos un espacio matemático donde todas las palabras estén a la misma distancia.

El espacio matemático lo definimos con unos ejes de coordenadas que serán las dimensiones del espacio.

P.e. un vocabulario de 3 palabras $\rightarrow$ 3 dimensiones.

Un punto en un espacio de tres dimensiones estará definido por el valor de las 3 coordenadas

(x,y,z) , que llamaremos **vector**

Palabra A $(1,0,0)$

Palabra B $(0,1,0)$

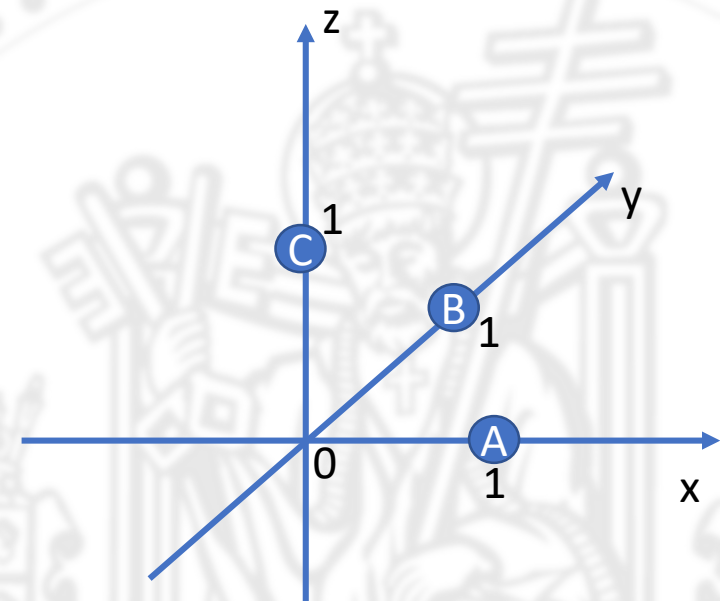
Palabra C $(0,0,1)$

En este espacio las tres palabras están a la misma distancia.

Para un vocabulario de N palabras, necesitaremos N dimensiones.

Representación **one-hot vector**, todos los valores a 0 excepto uno a 1

netflix	(1,0,0,0,0,0,0,0,0,0,0,0)
encontrar	(0,1,0,0,0,0,0,0,0,0,0,0)
juego	(0,0,1,0,0,0,0,0,0,0,0,0)
....	
plataforma	(0,0,0,0,0,0,0,0,0,0,1,0)
contar	(0,0,0,0,0,0,0,0,0,0,0,1)



¿tiene algún significado el valor numérico?

Representación dispersa (muchos 0)

Procesamiento del Lenguaje Natural

Opción 3.

Reflexionemos,

¿qué buscamos? (carta a los reyes magos)

- ✓ Queremos representar el significado de unidades lingüísticas (tokens/palabras)
- ✓ Queremos definir una medida de similitud semántica entre unidades
- ✓ Queremos que sea una representación numérica densa: “embeddings”

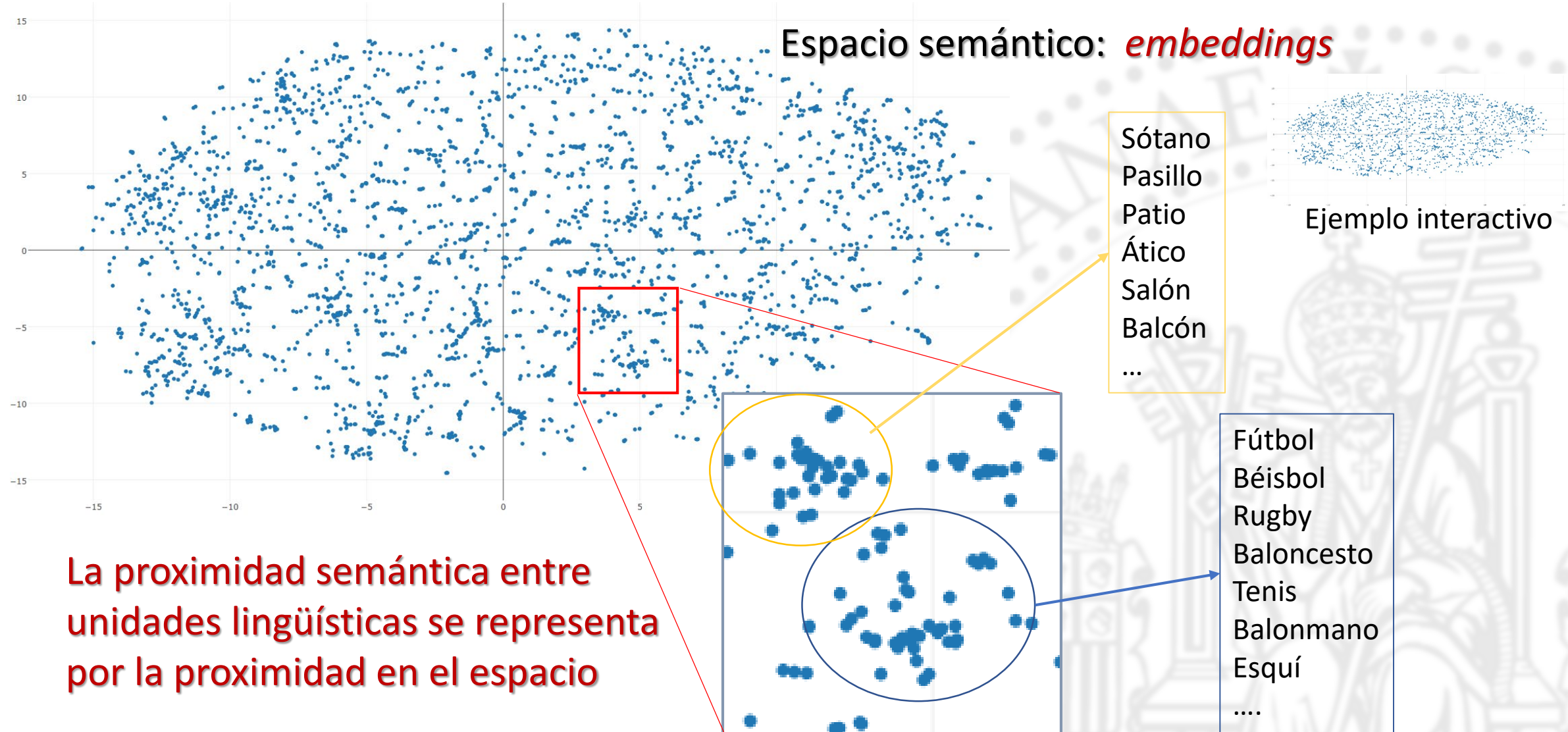
En definitiva:

Un espacio matemático de representación compacto donde la posición de los vectores que me identifican a las unidades contenga información semántica y que llamaremos espacio semántico

¿cómo lo construimos?

Semántica distribucional

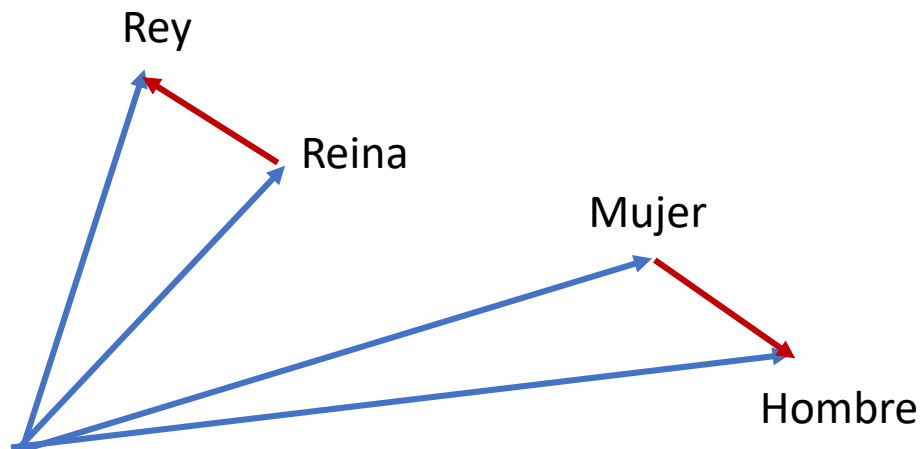
Procesamiento del Lenguaje Natural



Procesamiento del Lenguaje Natural

Relaciones semánticas

$$\text{vector}[\text{Reina}] = \text{vector}[\text{Rey}] - \text{vector}[\text{Hombre}] + \text{vector}[\text{Mujer}]$$



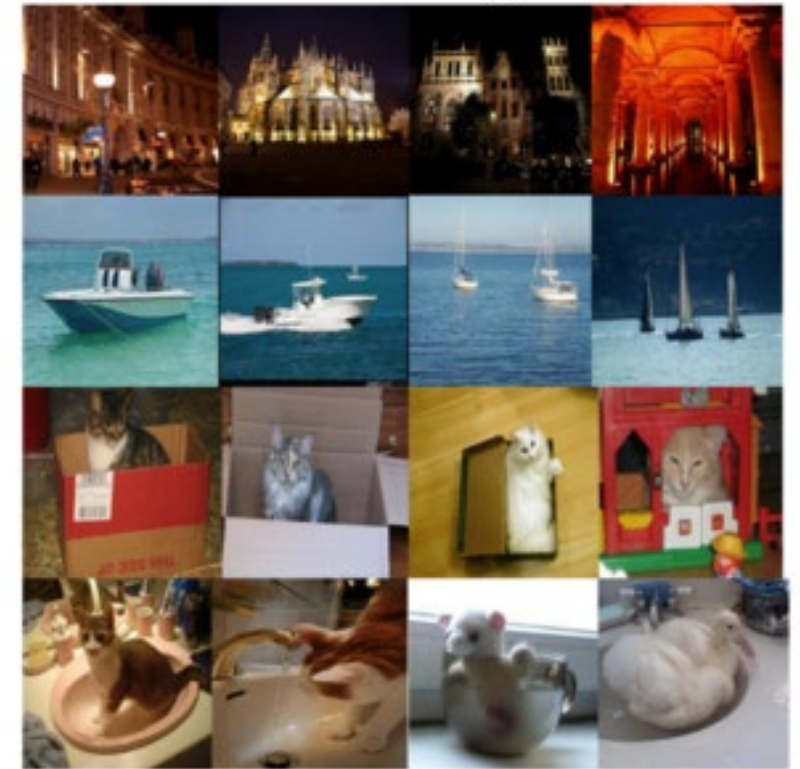
- día + noche =

-volar+navegar =

- taza + caja =

- caja + taza =

Imágenes próximas



(Kiros, Salakhutdinov, Zemel, TACL 2015)

Procesamiento del Lenguaje Natural

Espacios semánticos, ¿cómo los construimos?

Semántica distribucional

Hipótesis

Palabras similares aparecen en contextos similares

Si dos palabras w_1 y w_2 tienen distribuciones similares, podemos asumir que tienen significados similares

Representaciones palabras-contextos (matrices TF-IDF)

	D1	D2	D3	...	DN
P1	30	0	40		80
P2	23	12	0		0
...					
PM	0	67	80		4

Modelos de lenguaje

Predecir una palabra dado un contexto

Predecir un contexto dada una palabra

Word2vec, GloVe, Bert, GTP-3....

..... dotar un presupuesto de la eurozona para invertir en los países.....

↑ ↑ ↙
palabras contexto izquierdo palabra central palabras contexto derecho

Procesamiento del Lenguaje Natural

Semántica distribucional

Matriz términos-documentos: *buscar documentos similares*

Cada fila representa una palabra del vocabulario

Cada columna representa un documento

Cada celda: tf-idf:

$tf_{t,d}$: Frecuencia de aparición de un término t en el documento d

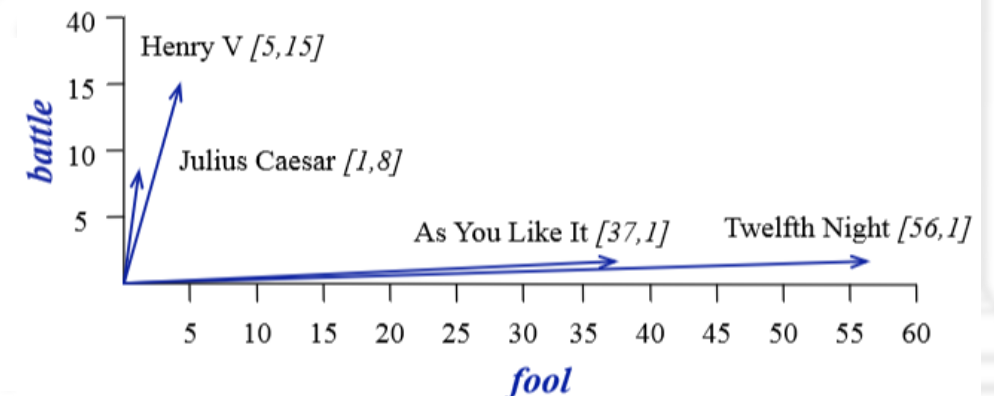
$idf_{t,D}$: inversa frecuencia de aparición término t en el conjunto de documentos D

Cada documento: un vector de cuentas de dimension V (tamaño vocabulario)

Dos documentos son similares si sus vectores son similares

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 15.2 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.



Daniel Jurafsky, James H. Martin

Procesamiento del Lenguaje Natural

Representaciones densas

La mayoría de los elementos del vector son cero

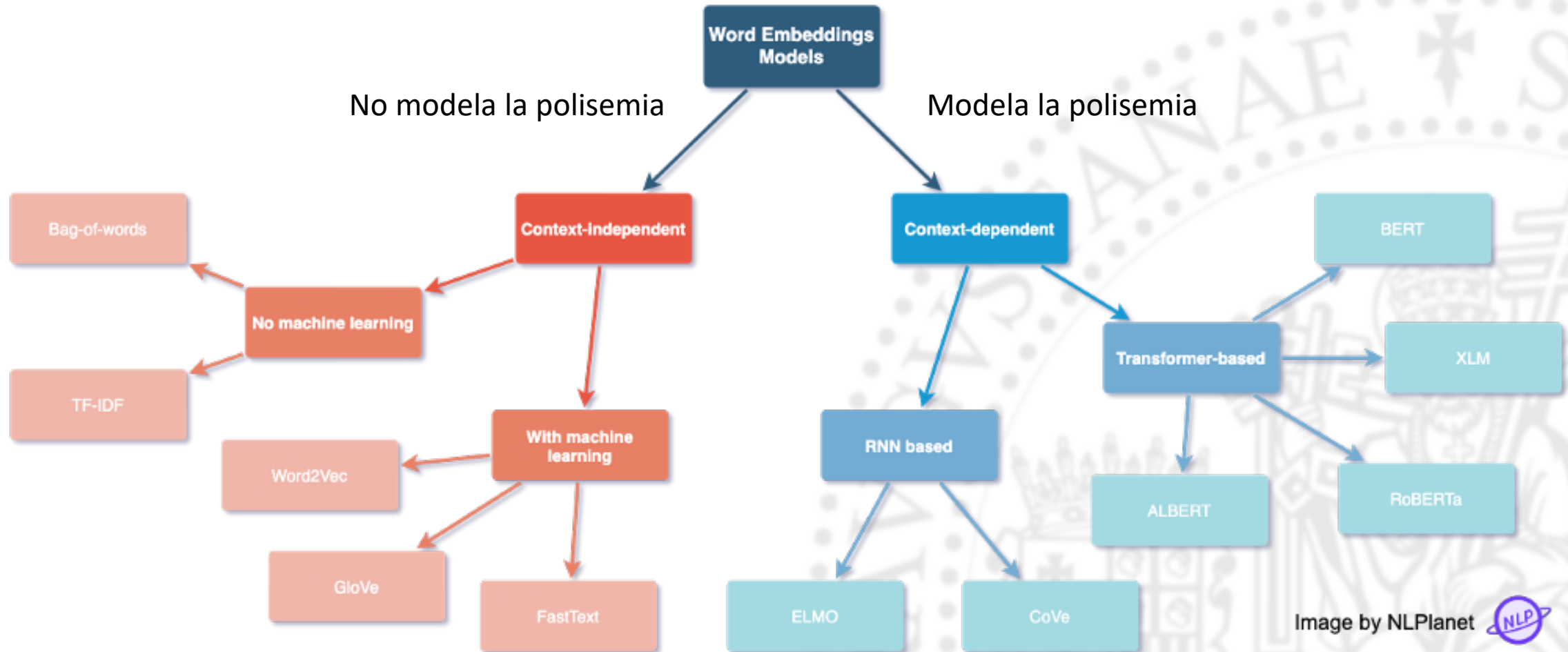
Reducir el tamaño del vector: representaciones densas (embeddings)

- Descomposición en valores singulares
 - LSA: Latent Semantic Analysis
- Word embeddings:
 - Word2vec: cbow y skip-gran (<https://code.google.com/archive/p/word2vec>)
 - Glove: global vectors (<https://nlp.stanford.edu/projects/glove>)
 - fastText: Text representation (<https://fasttext.cc/>)

Representación gráfica de vectores densos

- t-SNE: t-Distributed Stochastic Neighbor Embedding (<https://lvdmaaten.github.io/tsne/>)
- UMAP: Uniform Manifold Approximation and Projection (<https://umap-learn.readthedocs.io>)

Procesamiento del Lenguaje Natural



Procesamiento del Lenguaje Natural

Generalización de los espacios semánticos:

Embeddings de Palabras, Frases, Documentos, Audio, Imagen, Vídeos, ...

Ejemplo: Búsqueda semántica en periódicos

Buscador de noticias similares <http://signal4.cps.unizar.es:5000/>

- ✓ Cada noticia es un embedding
- ✓ Calcular embedding del texto a buscar
- ✓ Buscar los embeddings de noticias más próximos al del texto

Pero además permite

- ✓ Clasificar las noticias por categorías/temas
- ✓ Reconocer entidades
- ✓ Descubrir estereotipos y sesgos
- ✓ Evolución temporal/espacial de la semántica de palabras
- ✓ Componente principal de los modelos de lenguaje con redes neuronales
- ✓

Procesamiento del Lenguaje Natural

Bases de datos basadas en vectores

MILVUS (<https://milvus.io/>)

Milvus se creó en 2019 con el único objetivo de almacenar, indexar y administrar vectores de embeddings masivos.

Usos:

- Búsqueda masiva de texto, imágenes, audio y vídeo por similitud sin necesidad de estructurar los datos

- Sistemas de recomendación

- Sistemas de pregunta-respuesta

Ejemplos

<https://github.com/milvus-io/bootcamp/tree/master/solutions>

<https://milvusio.medium.com/>

VESPA (<https://vespa.ai/>)

<https://github.com/vespa-engine/vespa>

Procesamiento del Lenguaje Natural

Modelos de Lenguaje

Asignar una probabilidad a una frase o secuencia de palabras

Calcular la probabilidad de una frase

W="la alta comisionada de las naciones unidas para los derechos humanos ha advertido que atacar viviendas supone una violación de las convenciones de ginebra que regulan los conflictos"

$$P(W) = P(w_1, w_2, w_3, w_4, \dots, w_M)$$

Predecir la siguiente palabra en una secuencia

W="la alta comisionada de las naciones..."

$$P(w_n | w_1, w_2 \dots w_{n-1})$$

Modelos estadísticos: N-gramas

Bigramas, Trigramas, ...

Modelos neuronales:

Large Language Models (GPT-3 (175 Gparámetros), PaLM (540 Gparámetros),...)

Fine-tuned Language Modes (OpenAI Codex, orden de Gparámetros)

Edge Language Models (adaptados al hardware local, muchos menos parámetros que LLMs)

Procesamiento del Lenguaje Natural

¿Dónde estamos?

<https://huggingface.co/>

<https://openai.com/>



Completion
Generate or manipulate text and code



Semantic search
Score text based on relevance



Fine-tuning Beta
Train a model for your use case



Classification Beta
Classify text into different categories



Question answering Beta
Generate high-accuracy answers



The AI community building the future.

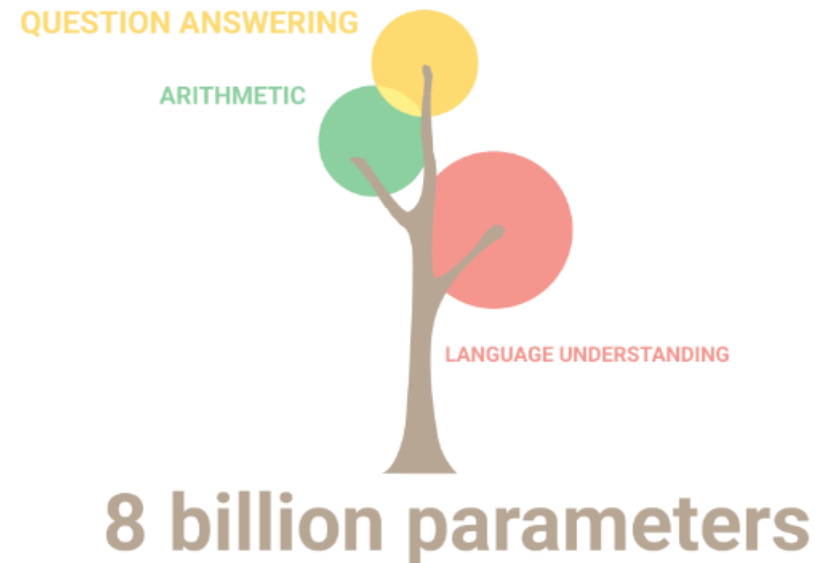
Build, train and deploy state of the art models powered by
the reference open source in machine learning.

GPT-3 Access Without the Wait

We've made improvements to our API and safety
features so developers can get started right away.

<https://beta.openai.com/examples>

Procesamiento del Lenguaje Natural



Pathways Language Models (PaLM) Google AI

<https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

Procesamiento del Lenguaje Natural

GPT-3: engine="text-davinci-001"

Hazme una lista en formato json con la persona, oficio, nacionalidad y año de nacimiento.

Bryan Adams, el fotógrafo encargado de realizar el calendario, es un cantante, guitarrista, compositor, fotógrafo y filántropo canadiense.

Anne Erin Annie Clark, conocida artísticamente como St. Vincent, es una cantautora y multiinstrumentista estadounidense. Es ganadora de tres Premios Grammy por Mejor Canción de Rock.

Kali Uchis, es una cantante, compositora, actriz, directora y diseñadora colomboestadounidense, saltó a la fama internacional en dos mil veintiuno con el gran éxito de su canción Telepatía.

Diamante Quiava Valentin Harper, es una conocida rapera estadounidense.

Cher es una de las figuras más destacadas del este calendario.

```
[{
  "person": "Bryan Adams",
  "oficio": "cantante, guitarrista, compositor, fotógrafo y filántropo",
  "nacionalidad": "canadiense",
  "año de nacimiento": "1958"
},
{
  "person": "Anne Erin Annie Clark",
  "oficio": "cantautora y multiinstrumentista",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1982"
},
{
  "person": "Kali Uchis",
  "oficio": "cantante, compositora, actriz, directora y diseñadora",
  "nacionalidad": "colomboestadounidense",
  "año de nacimiento": "1994"
},
{
  "person": "Diamante Quiava Valentin Harper",
  "oficio": "rapera",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1996"
},
{
  "person": "Cher",
  "oficio": "cantante",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1946"
}]
```


Procesamiento del Lenguaje Natural

Razonamiento: PaLM

Standard Prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

A: The answer is 11.

Prompt

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response

The answer is 50.



Chain of thought prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Prompt

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.

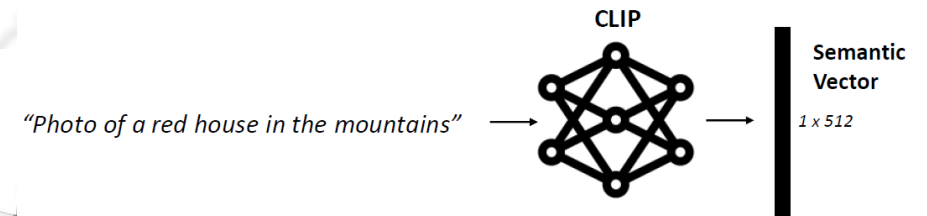
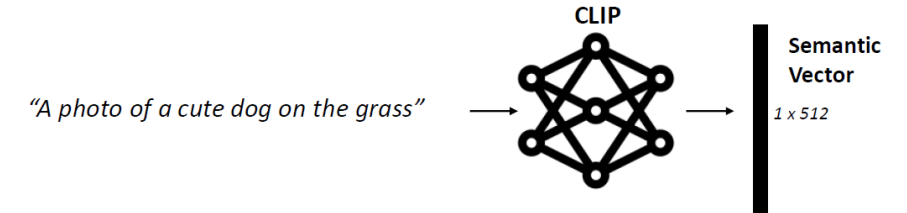
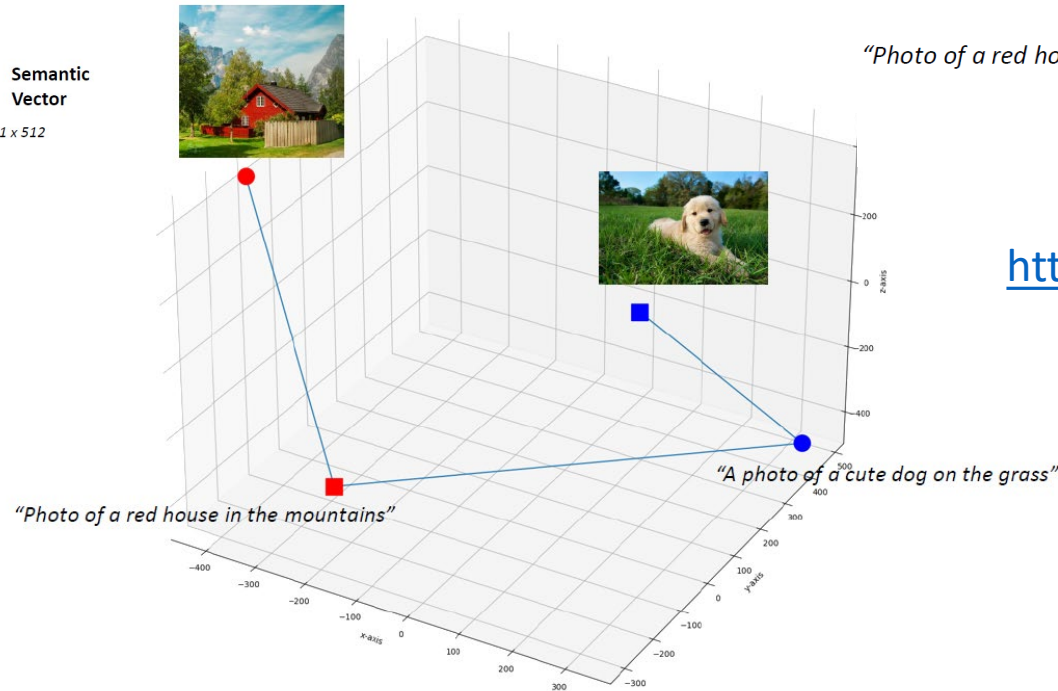
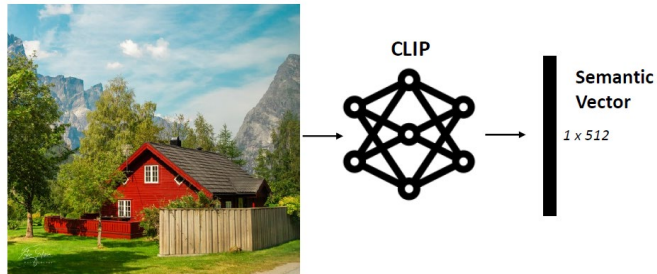
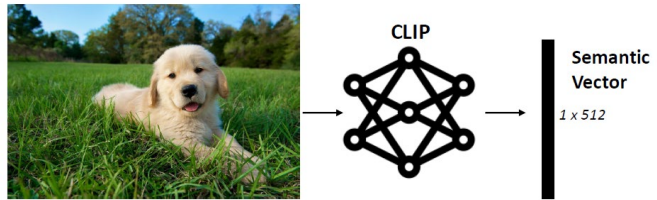


Procesamiento del Lenguaje Natural

CLIP: modelo multimodal de OpenAI

Combina un modelo de lenguaje de inglés con un modelo semántico de conocimiento de imágenes

Entrenado con mas de 400M de pares imagen+texto



<http://catedrartve.unizar.es/jaca2022.html>

<http://signal4.cps.unizar.es:8052/>

Procesamiento del Lenguaje Natural

Algunas aplicaciones

CLIP + GPT2: Descripción de imágenes

Búsqueda en vídeos con lenguaje natural



A couple of people standing next to an elephant.



A wooden table sitting in front of a window.



A bunch of bananas sitting on top of a table.



A woman holding a plate with a piece of cake in front of her face.



A wooden table topped with lots of wooden utensils.



A red motorcycle parked on top of a dirt field.

Procesamiento del Lenguaje Natural

Casos de uso con CLIP

Detección de imágenes con contenido inadecuado (NSFW-Not safe/suitable for work):

Similitud entre la interpretación de CLIP del texto y la interpretación de CLIP de la imagen.



```
"matches": [ {"text": "the blue car is on the left, the red car is on the right"}, {"text": "the blue car is on the right, the red car is on the left"}, {"text": "the blue car is on top of the red car"}, {"text": "the blue car is below the red car"} ]]
```

```
"the blue car is on the left, the red car is on the right" 0.5232442617416382  
"the blue car is on the right, the red car is on the left" 0.32878655195236206  
"the blue car is below the red car" 0.11064132302999496  
"the blue car is on top of the red car" 0.03732786327600479
```


Procesamiento del Lenguaje Natural

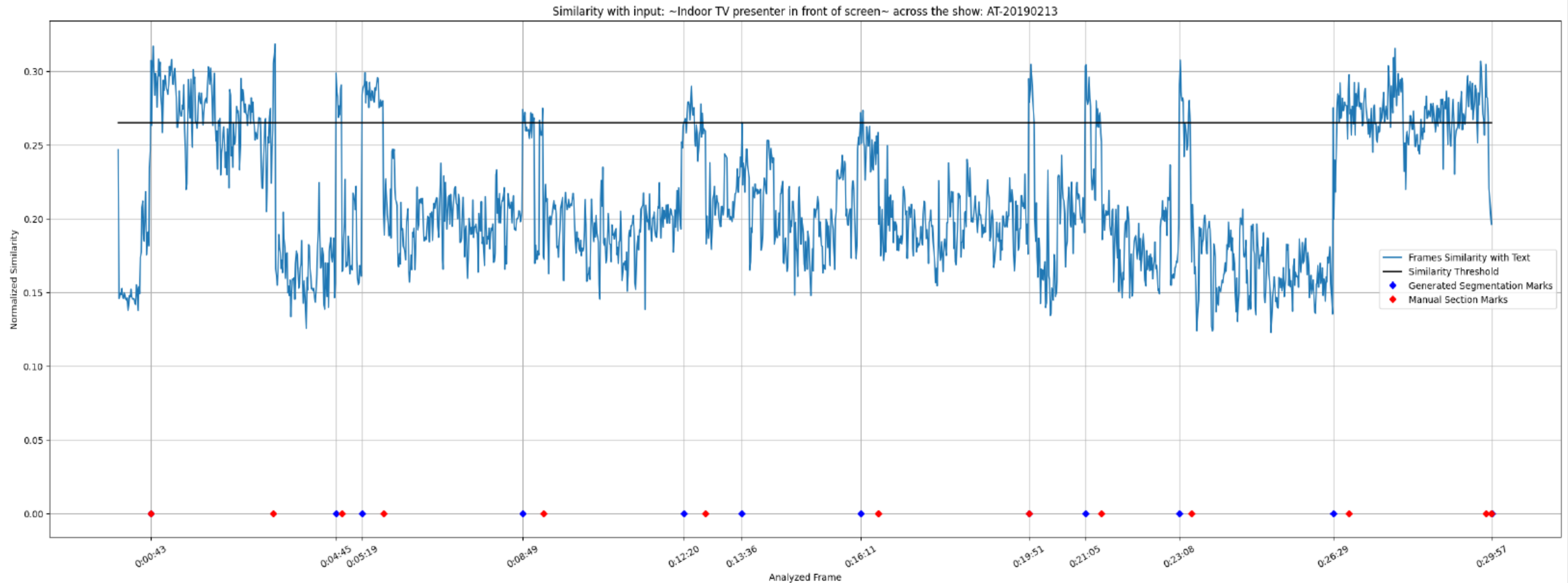
Descripción de segmentos de vídeo



“Chef and his partner were seen chatting to one another as they prepared the meal . Chef was joined by his wife, who was also in the show . The video shows the chef slicing the meat with a knife”

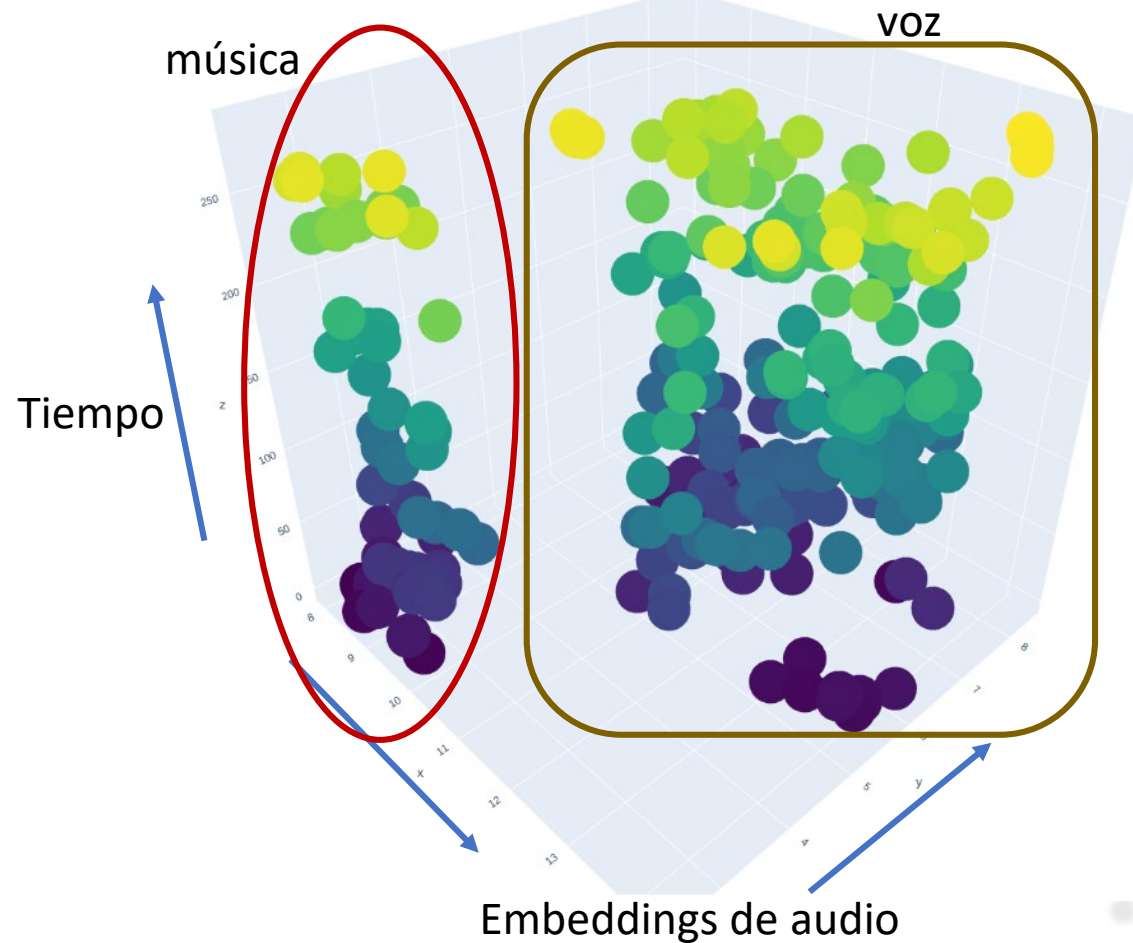
Procesamiento del Lenguaje Natural

Recuperación de segmentos de vídeo



Procesamiento del Lenguaje Natural

AudioCLIP: espacio semántico Audio-Texto-Vídeo



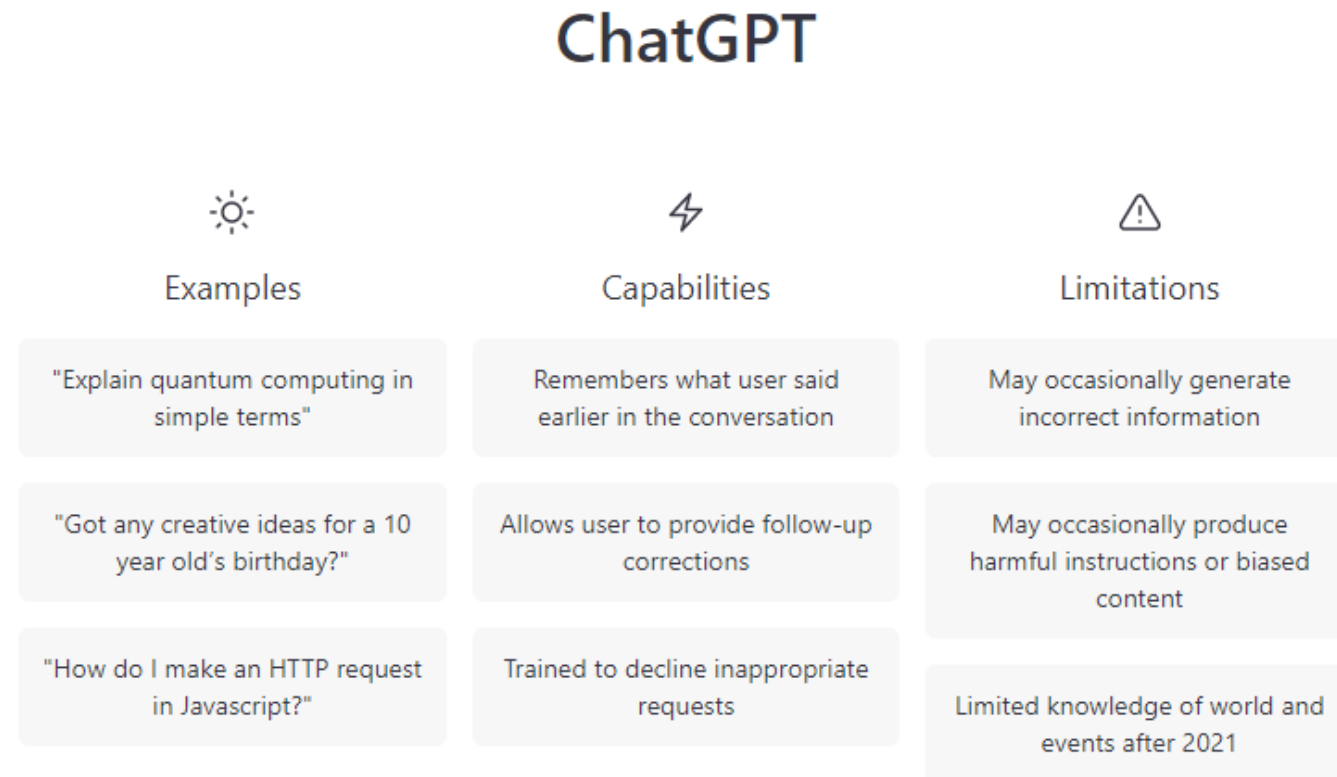
- ✓ Programas de televisión:
Subtítulos / Audio / Vídeo
- ✓ AudioCLIP permite consultas-respuestas multimodales

Procesamiento del Lenguaje Natural

Lo penúltimo

OpenAI: ChatGPT

<https://chat.openai.com/chat>



Procesamiento del Lenguaje Natural

E código en python para dibujar un dendograma



E ¿se puede utilizar la librería plotly?



Para dibujar un dendograma en Python, puedes usar la biblioteca `matplotlib` y su función `dendrogram`. Primero, debes importar la biblioteca y los datos que quieres representar en el dendograma. Luego, puedes usar la función `dendrogram` y pasarle los datos como argumento. Por último, puedes usar la función `show` de matplotlib para mostrar el dendograma en pantalla.



Copy code

```
import matplotlib.pyplot as plt
from scipy.cluster import hierarchy

# Carga los datos que quieres representar en el dendograma
data = ...

# Crea el dendograma usando la función dendrogram de matplotlib
dendrogram = hierarchy.dendrogram(data)

# Muestra el dendograma en pantalla
plt.show()
```

Espero que esto te ayude a empezar a dibujar dendogramas en Python. Si necesitas más ayuda o información, no dudes en preguntar.



Sí, también puedes usar la biblioteca `plotly` para dibujar dendogramas en Python. `plotly` es una biblioteca de visualización de datos que proporciona una interfaz de usuario amigable y herramientas avanzadas para crear gráficos y visualizaciones interactivas.



Para usar `plotly` para dibujar un dendograma, primero debes instalar la biblioteca y importarla en tu script de Python. Luego, puedes usar la función `dendrogram` de `plotly.express` para crear el dendograma y mostrarlo en pantalla.

Copy code

```
# Instala la biblioteca plotly
!pip install plotly

# Importa las bibliotecas necesarias
import plotly.express as px

# Carga los datos que quieres representar en el dendograma
data = ...

# Crea el dendograma usando la función dendrogram de plotly.express
fig = px.dendrogram(data)

# Muestra el dendograma en pantalla
fig.show()
```

Espero que esto te ayude a empezar a usar `plotly` para dibujar dendogramas en Python. Si necesitas más ayuda o información, no dudes en preguntar.

Procesamiento del Lenguaje Natural

E ¿cuántos años tendría John Lennon si viviera todavía?



John Lennon fue un músico y compositor británico, conocido principalmente por ser uno de los fundadores de la banda de rock The Beatles. Lennon fue asesinado el 8 de diciembre de 1980, a la edad de 40 años. Si estuviera vivo en la actualidad, en diciembre de 2022, tendría 87 años.



Lennon fue uno de los principales compositores y letristas de The Beatles, y es considerado uno de los músicos más influyentes y exitosos de la historia. Después de la disolución de The Beatles en 1970, Lennon continuó su carrera como solista y lanzó varios álbumes exitosos, como "Imagine" y "Double Fantasy". Además de su trabajo en la música, Lennon también fue un activista político y pacifista.

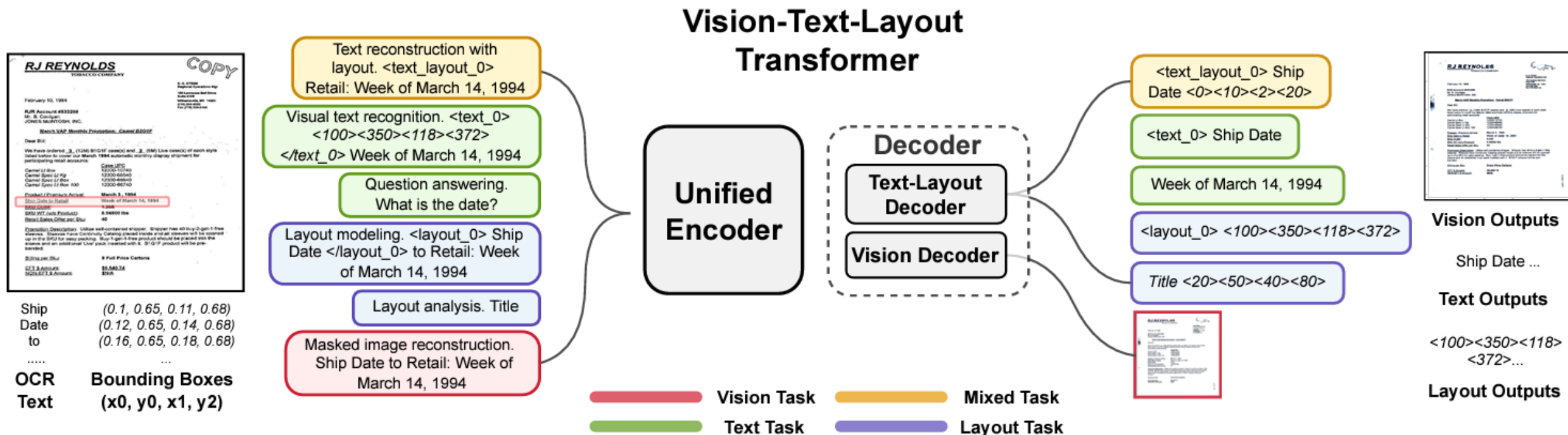
Si deseas conocer más sobre la vida y obra de John Lennon, te recomendaría que consultes biografías y libros sobre The Beatles o sobre la historia de la música popular del siglo XX. También puedes buscar en internet videos y grabaciones de sus presentaciones y entrevistas para conocer más sobre su estilo y personalidad.

Procesamiento del Lenguaje Natural

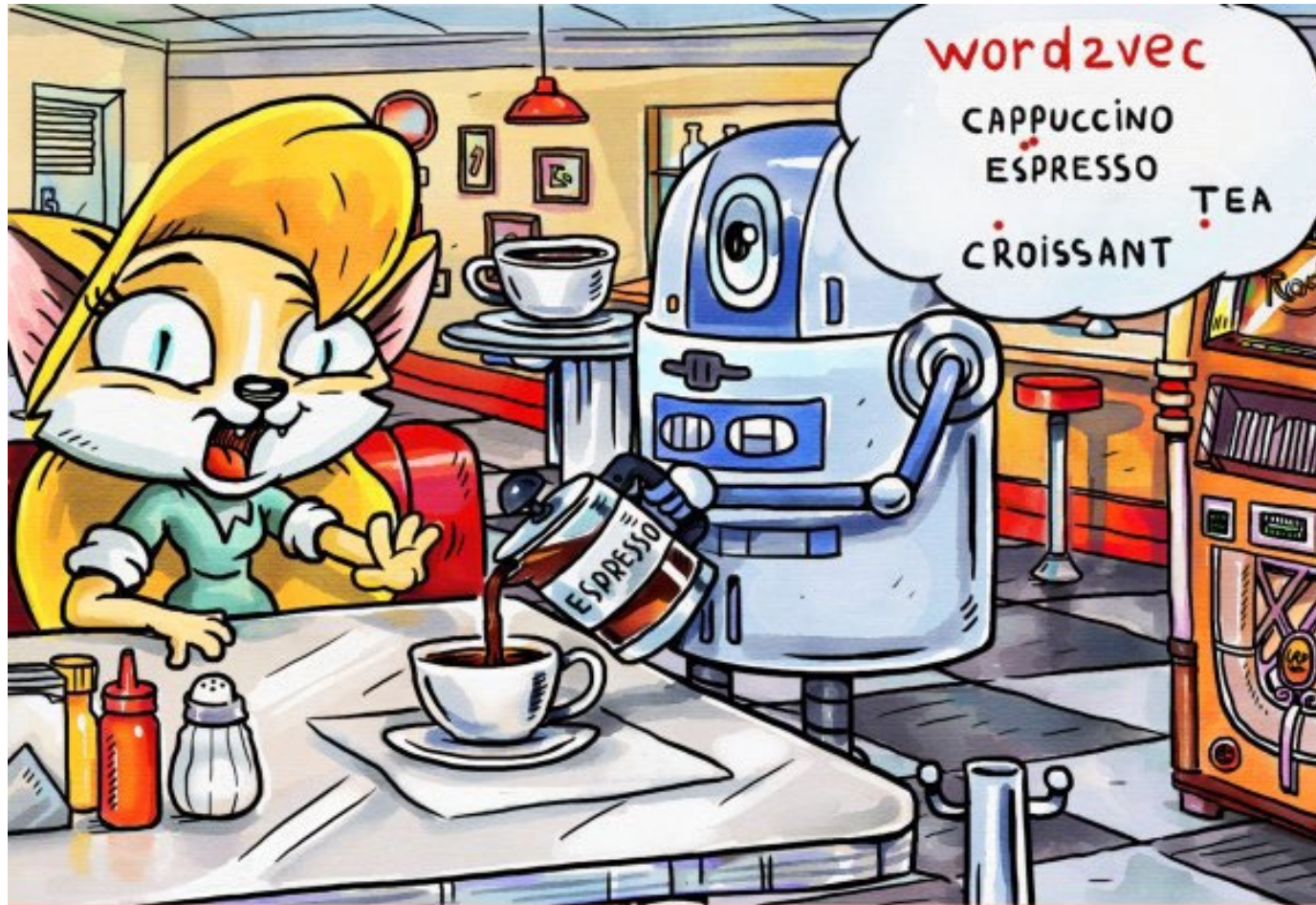
¿Lo ultimo?

Microsoft: Unifying Vision, Text, and Layout for Universal Document Processing (UDOP)

<https://github.com/microsoft/UDOP>



Procesamiento del Lenguaje Natural



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.