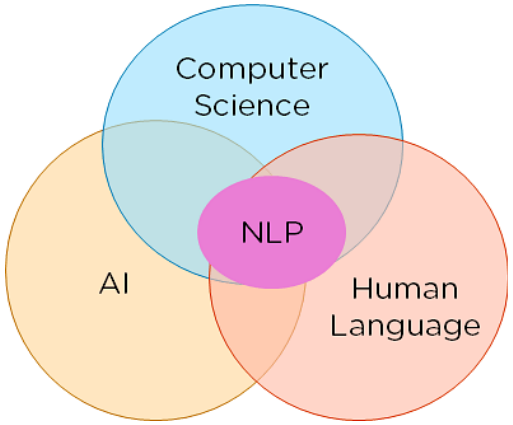
Procesamiento del Lenguaje Natural

¿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.



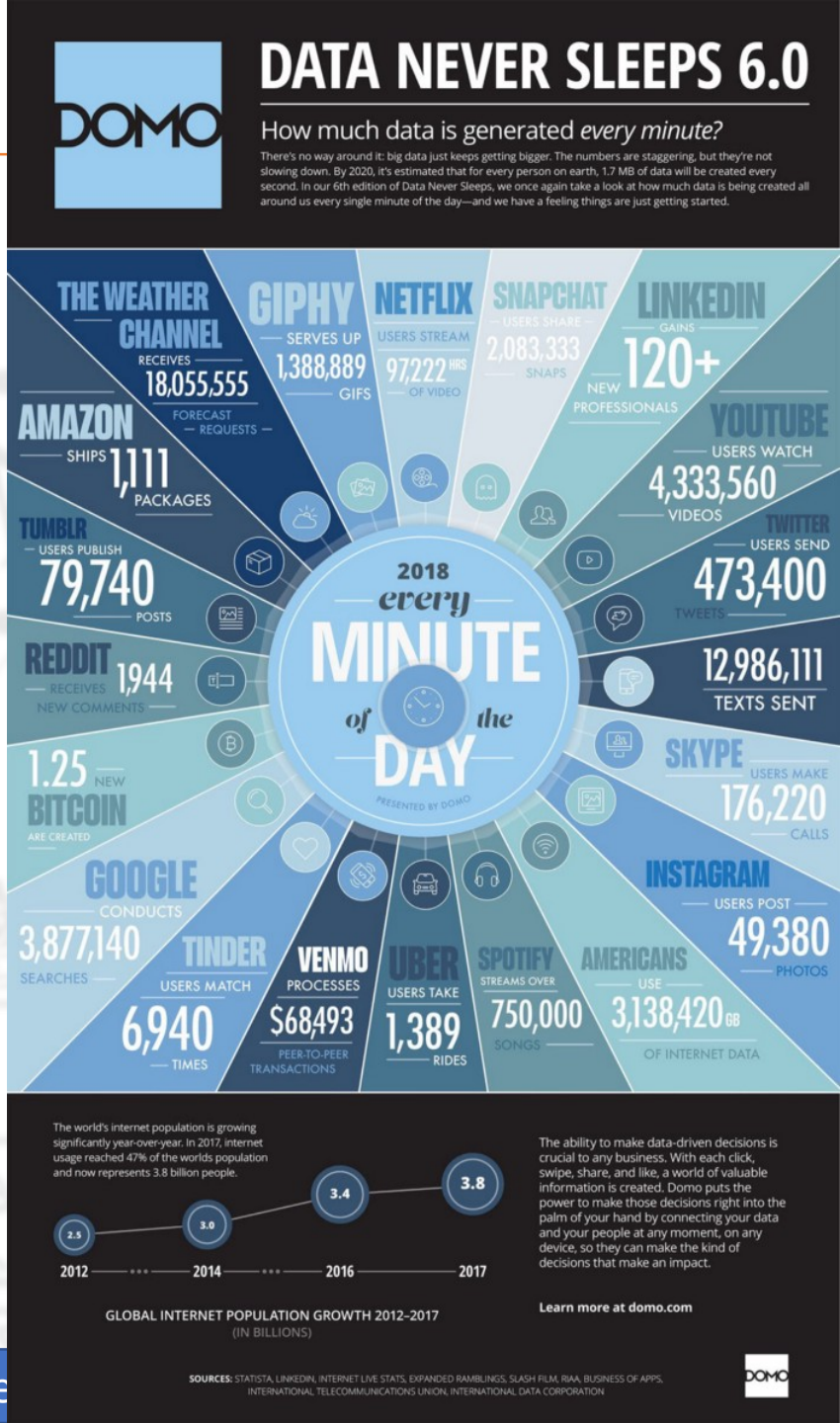
Capacidades de un sistema de IA
- Percepción
- Aprendizaje
- Representación del conocimiento
- Razonamiento

Procesamiento del Lenguaje Natural

Procesado masivo de datos

Cantidades masivas de datos no estructurados (raw data) : texto, audio e imágenes

Datos estructurados
Representación numérica adecuada



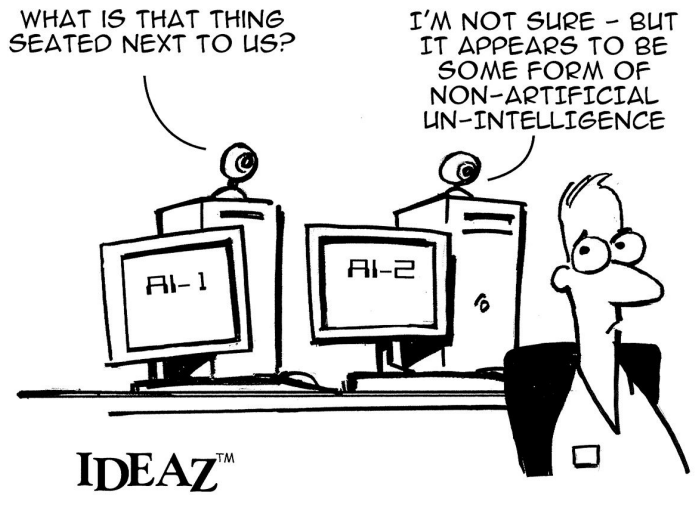
Procesamiento del Lenguaje Natural

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.

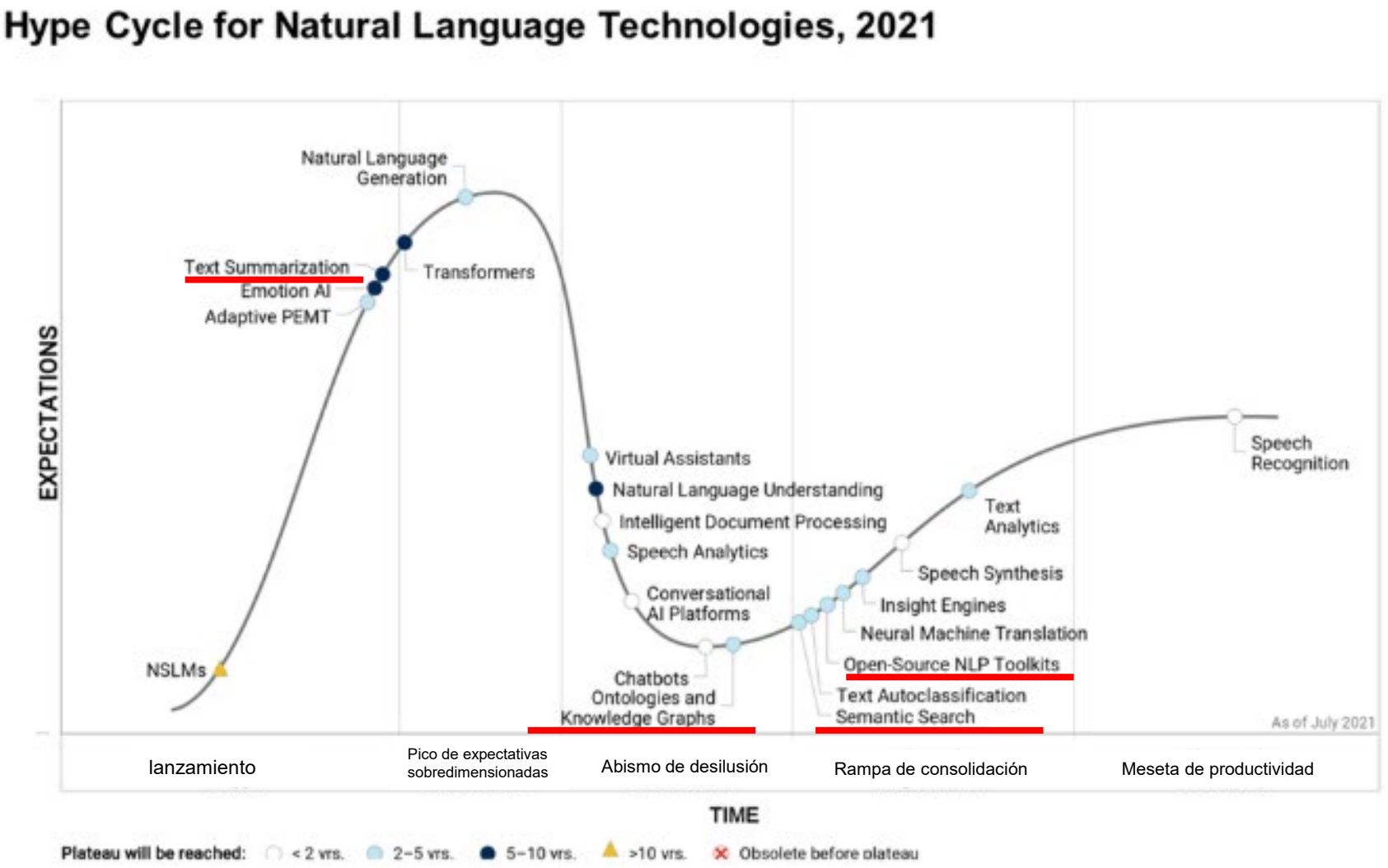


Nos proporciona herramientas para representar el *conocimiento*

Procesamiento del Lenguaje Natural

¿Qué tareas podemos hacer?

Procesamiento del Lenguaje Natural

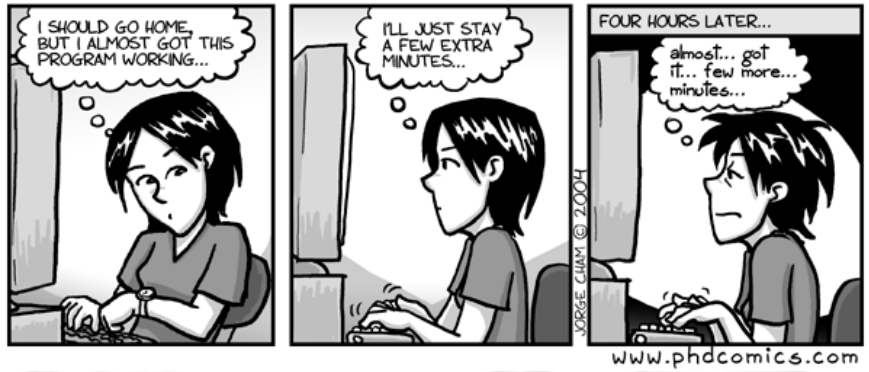


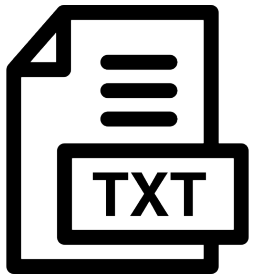
Source: Gartner (July 2021)

748656

Procesamiento del Lenguaje Natural

¿Cómo funciona el procesado de lenguaje natural?





OCR

voz a texto

Descripción de imagen

Pre-procesado texto:

- Segmentación en frases
- Segmentación en tokens
- Eliminar palabras comunes (stopwords)
- Lematización/Stemming
- Análisis morfológico
- Etiquetado gramatical(Part-Of-Speech tagging)

Sistemas/Algoritmos basados en:

- reglas
- datos (aprendizaje automático)
  - ✓ Espacios semánticos
  - ✓ Modelos de lenguaje

Procesamiento del Lenguaje Natural

voz a texto

Raw data

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

Segmentación en frases

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.

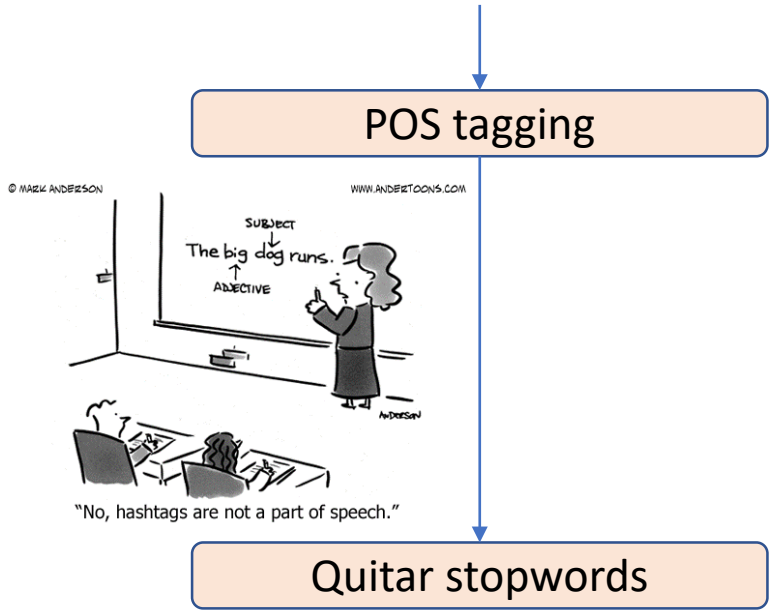
Tokenización

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

Lematización

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

Procesamiento del Lenguaje Natural



| Netflix [NP00SP0] | haber [VAIP3S0] | encontrar [VMP00SM]| en [SP] | el [DA0MS0] | juego [NCMS000] | del [SP] |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] |….

| 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] |….

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

Procesamiento del Lenguaje Natural

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

La *máquina trabaja con números*, …..

…. luego *debemos trasformar las palabras a números*

<u>Opción 1.</u>

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 → 3 dimensiones.

Un punto en un espacio de tres dimensionas estará definido por el valor de las 3 coordenadas (x,y,z), que llamaremos **vector**

Palabra A (1,0,0)

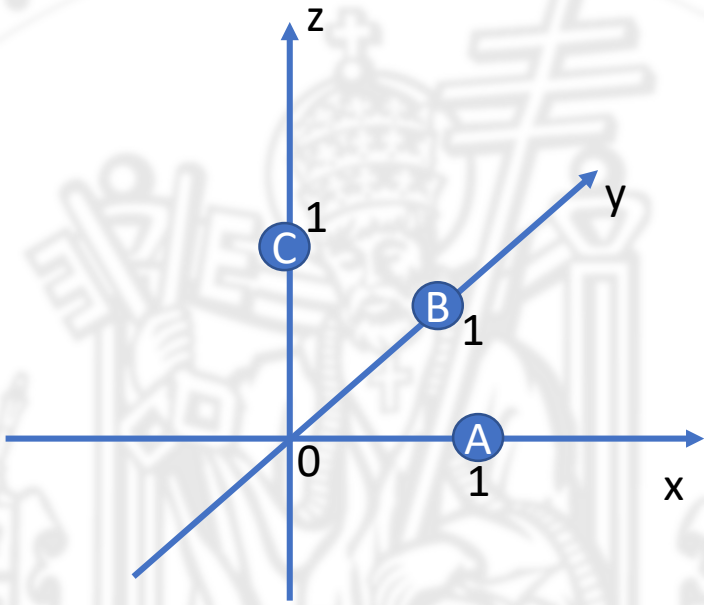
Palabra B (0,1,0)

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

Semántica distribucional

¿Cómo conocemos el significado de una palabra?

John Rupert Firth, "You shall know a word by the Company it keeps"

"Similar words occur in similar contexts"

Ludwig Wittgenstein, "The meaning of a word is its use in language"

Hay una botella de ***Belikin*** sobre la mesa
A todo el mundo le gusta la ***Belikin***
No bebas ***Belikin*** si tienes que conducir
La ***Belikin*** se fabrica con granos de cebada germinada

¿qué podemos deducir sobre la palabra ***Belikin***?

Miramos las palabras que acompañan
Buscamos la similitud semántica con otras palabras ya conocidas
… y deducimos que la ***Belikin*** debe ser una bebida similar a…

Procesamiento del Lenguaje Natural

Aproximaciones para representar palabras

Distribuciones: cuentas

- Usado desde los 90
- Representaciones dispersas de palabras-contexto (matrices TF-IDF)
- Compactación

Predicción: modelos de lenguaje

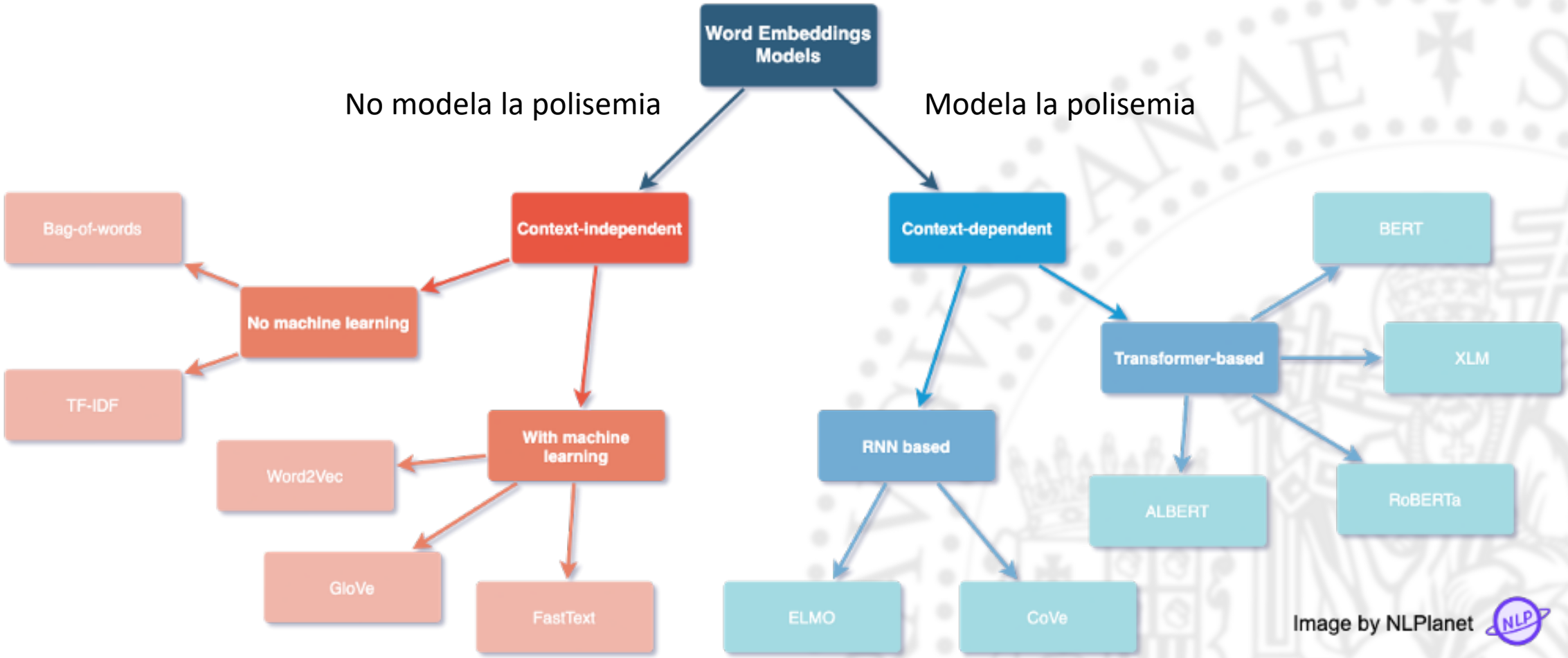
- Inspirado en deep learning
- word2vec *(Mikolov et al., 2013)*
- GloVe *(Pennington et al., 2014)*
- *BERT (Google, 2018)*

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

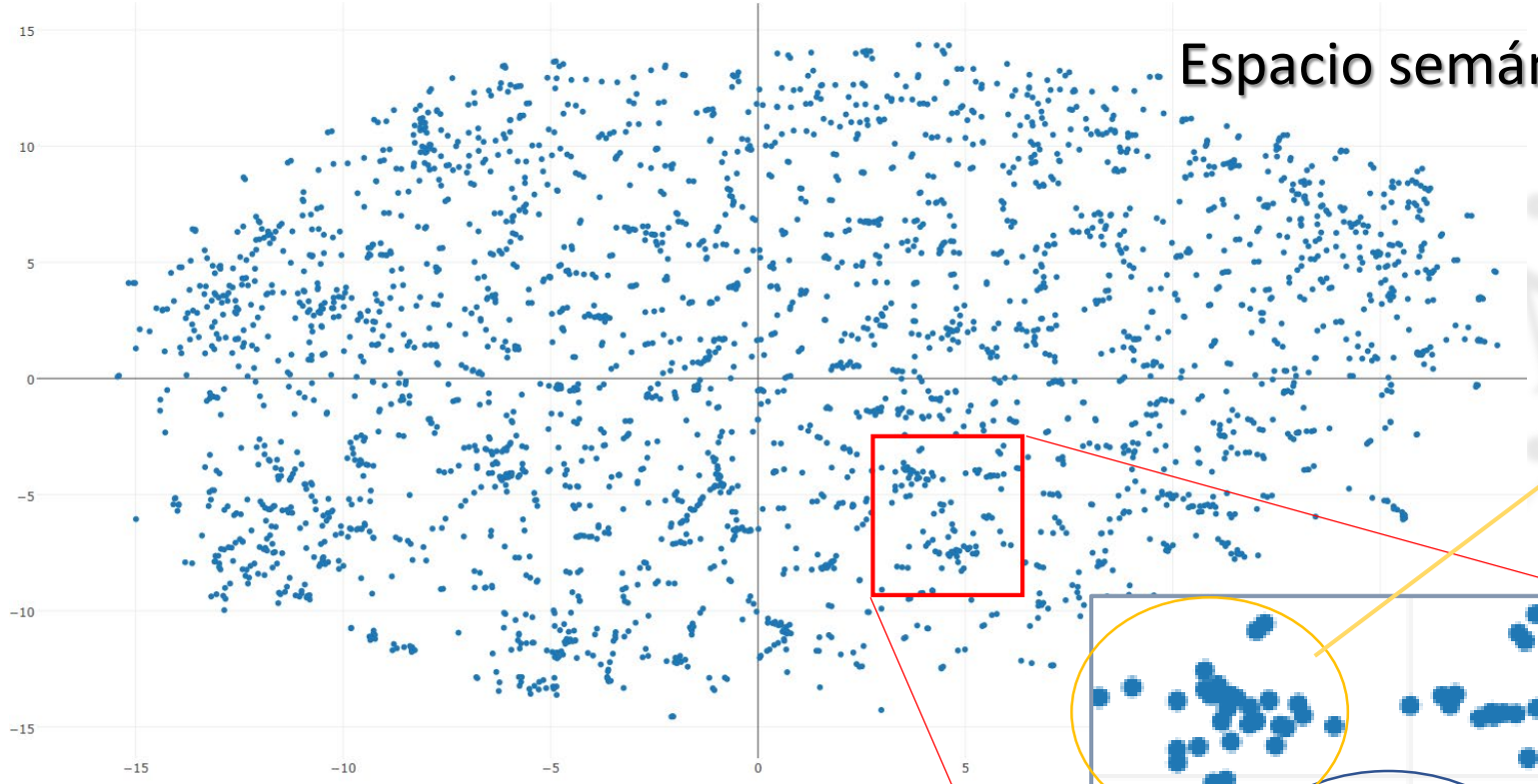
….. dotar un presupuesto de la eurozona para invertir en los países….

palabras contexto izquierdo palabra central palabras contexto derecho

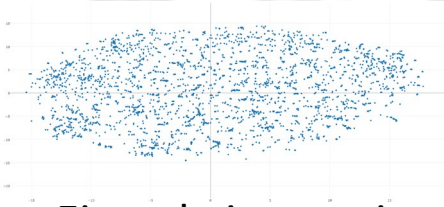
Procesamiento del Lenguaje Natural



Procesamiento del Lenguaje Natural

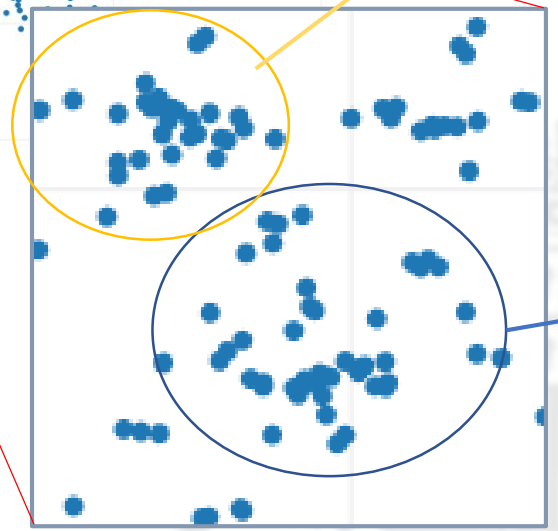


Espacio semántico: *embeddings*



Ejemplo interactivo

Sótano
Pasillo
Patio
Ático
Salón
Balcón
…



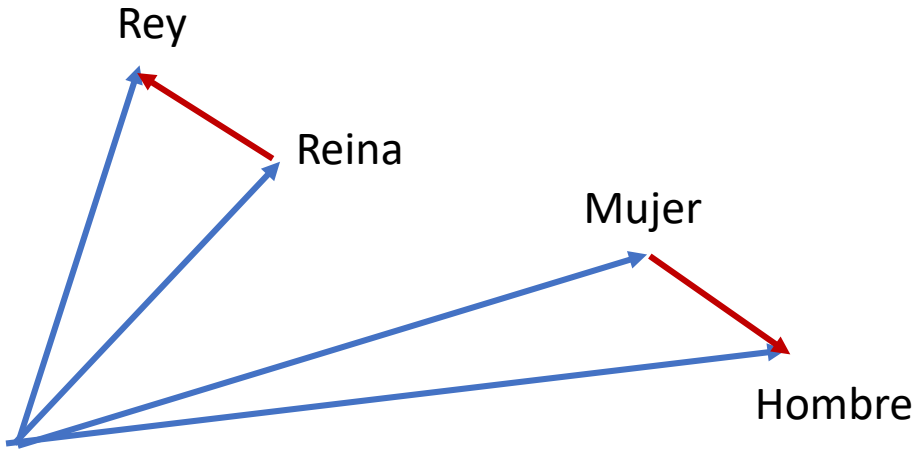
Fútbol
Béisbol
Rugby
Baloncesto
Tenis
Balonmano
Esquí
….

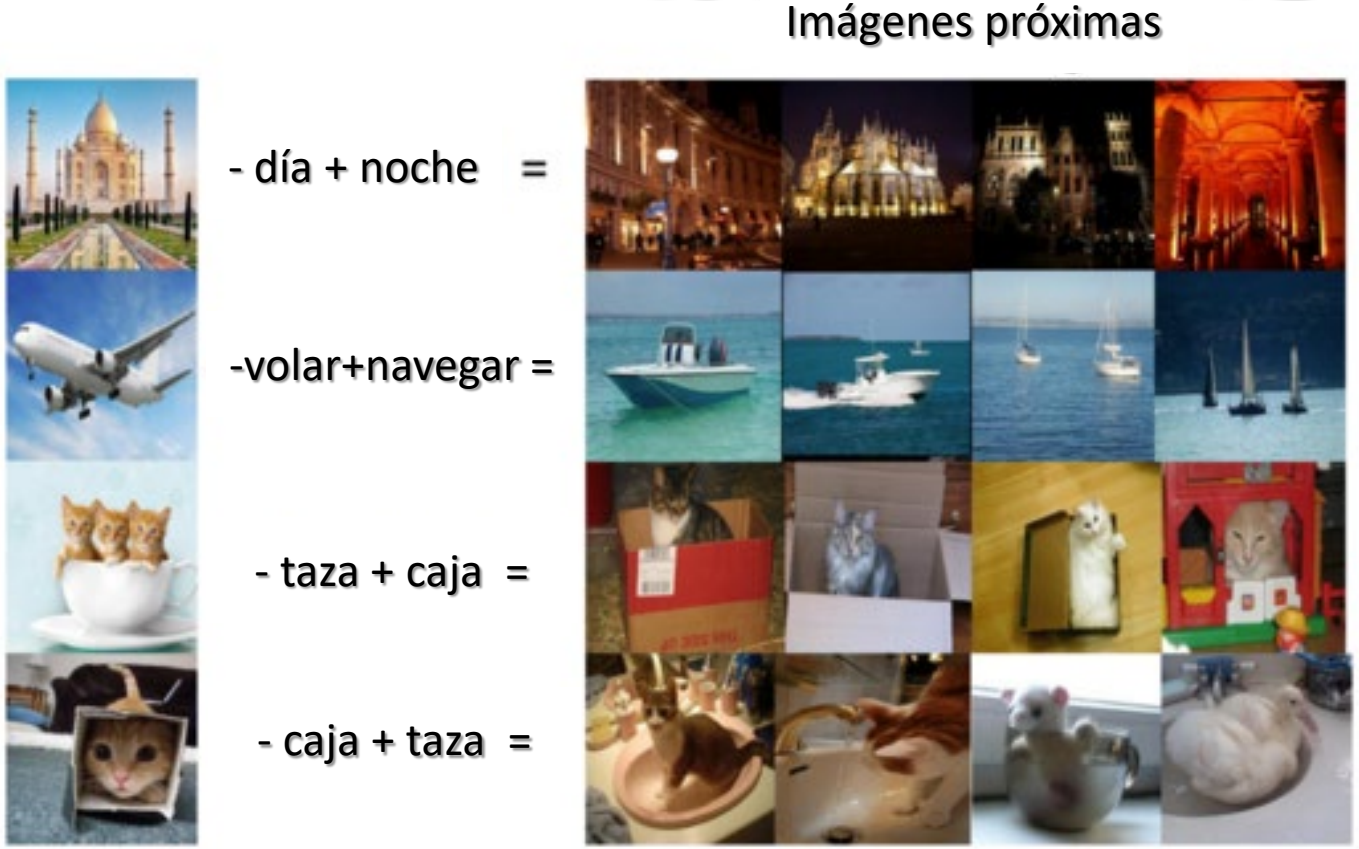
La proximidad semántica entre unidades lingüísticas se representa por la proximidad en el espacio

Procesamiento del Lenguaje Natural

Relaciones semánticas

vector[Reina] = vector[Rey] - vector[Hombre] + vector[Mujer]





(Kiros, Salakhutdinov, Zemel, TACL 2015)

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/

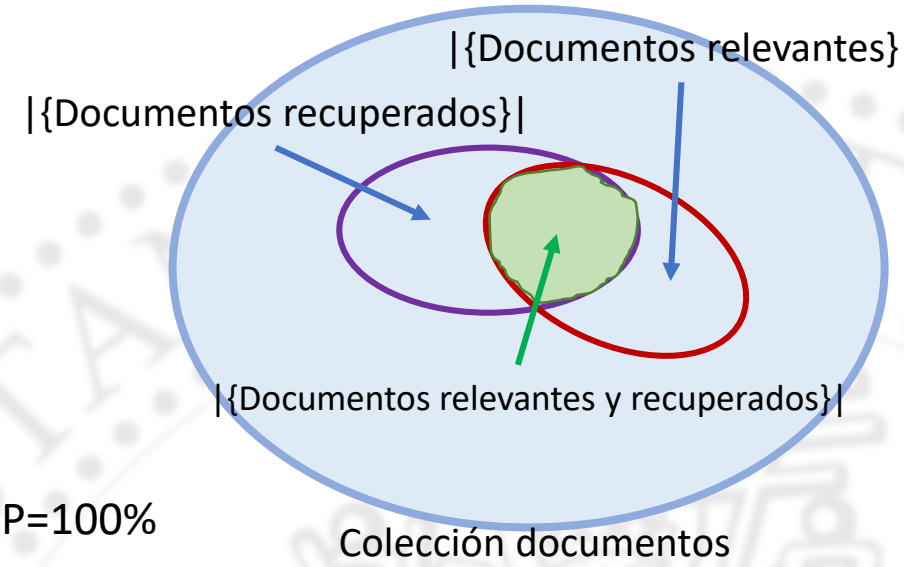
- ✓ Cada noticia es un embedding
- ✓ Calcular embedding del texto a buscar
- ✓ Buscar los embeddings de noticias más próximos al del texto

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

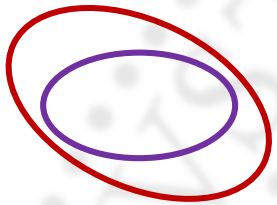
Medida de prestaciones en búsquedas y recuperación de información:



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

Precisión: Probabilidad de que un documento recuperado sea relevante.

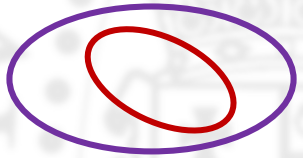
$$P = \frac{|\{documentos\ relevantes\ y\ recuperados\}|}{|\{documentos\ recuperados\}|}$$



P=100%

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

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



R=100%

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

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

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

Procesamiento del Lenguaje Natural

¿Dónde estamos?

https://huggingface.co/

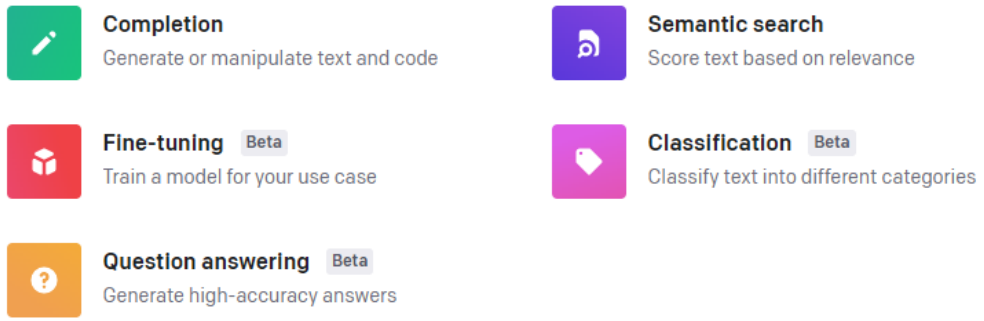
The AI community building the future.

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

https://openai.com/

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

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

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

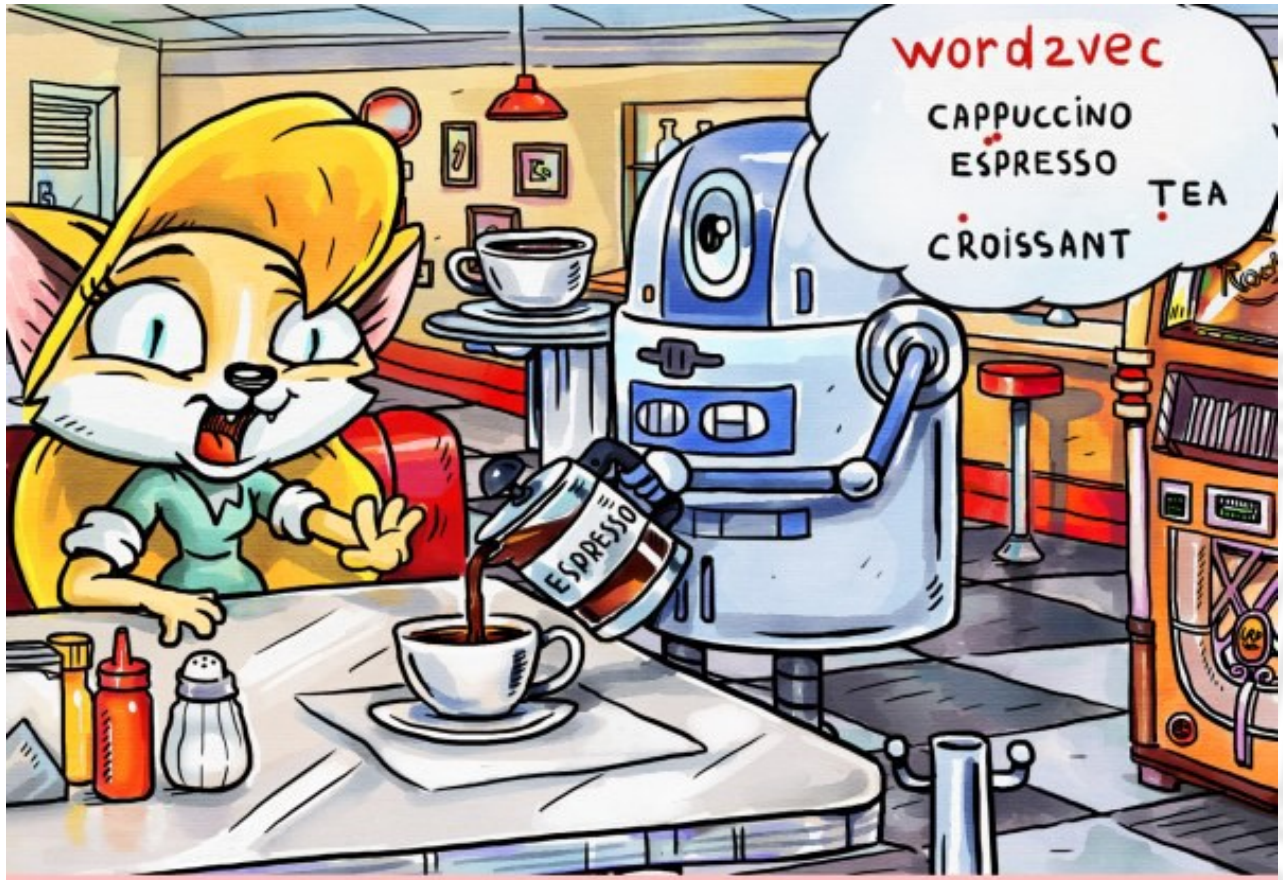
Bryan Adams, el fotógrafo encargado de realizar el calendario, es un cantante, guitarrista, compositor, fotógrafo y filántropo canadiense.
Anne Erin Annie Clark, conocida artísticamente como St. Vincent, es una cantautora y multiinstrumentista estadounidense. Es ganadora de tres Premios Grammy por Mejor 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.
Diamonte Quiava Valentin Harper, es una cocnocida rapera estadounidense.
Cher es una de las figuras más destacas 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": "Diamonte 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

```python
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")
prompt=input("¿Qué deseas saber? ")
response = openai.Completion.create(
  engine="text-davinci-001",
  prompt=prompt,
  temperature=0.3,
  max_tokens=350,
  top_p=1.0,
  frequency_penalty=0.0,
  presence_penalty=0.0
)
print(response["choices"][0]["text"])
```



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