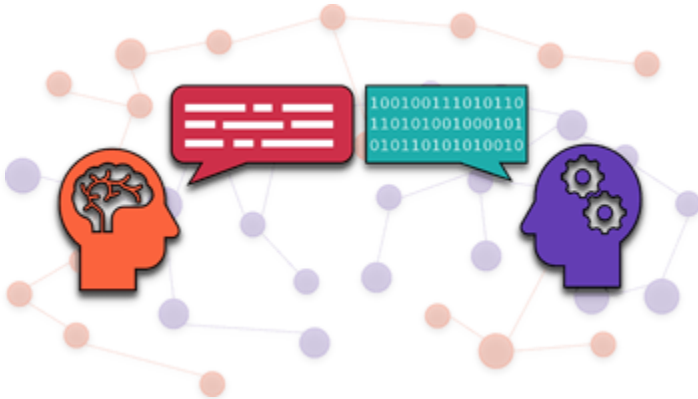
Modelos de Inteligencia Artificial

CIPFP Mislata
Centre Integrat Públic Formació Professional Superior

:: U3 ::

Procesamiento del Lenguaje Natural



Uso de modelos lingüísticos

LLMs y LangChain

Curso 2023-24

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tabla de contenidos

3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Las **incrustaciones de palabras** (*embeddings*) son representaciones vectoriales de una palabra.

- **Objetivo**: tomar toda la información que está almacenada en una palabra, ya sea su significado o su parte gramatical, y convertirla en una forma numérica más comprensible para un ordenador.

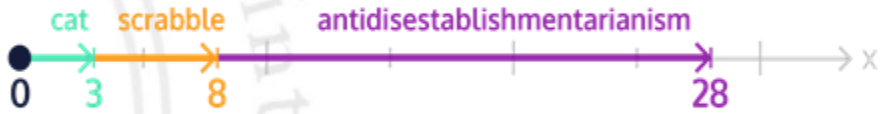
- Video: INTRO al Natural Language Processing (NLP) #2

3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Vectores:** son útiles porque nos ayudan a resumir información sobre un objeto usando números
  - **Ejemplo**:

    Con un vector unidimensional podemos guardar la longitud de una palabra ...



"cat" ----> **[3]**

"scrabble" ----> **[8]**

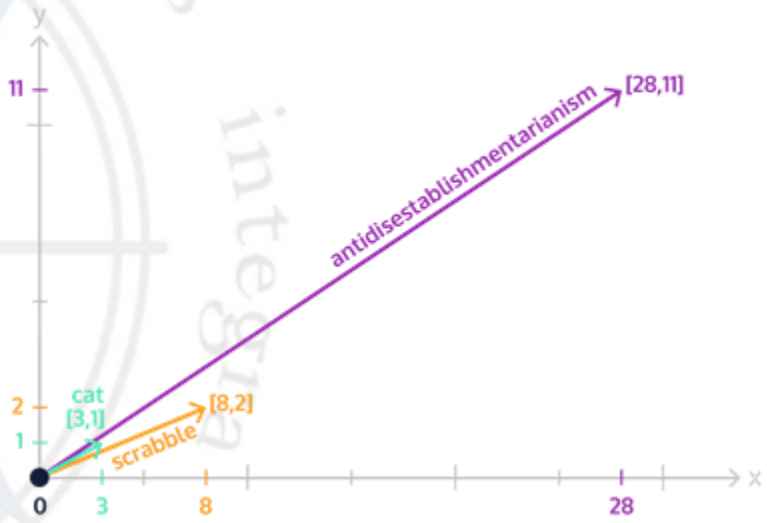
"antidisestablishmentarianism" ----> **[28]**

3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Vectores:** con más dimensiones es mayor la capacidad de guardar información
  - **ejemplo**:

    ... con un vector bidimensional podríamos representar tanto la longitud de una palabra (eje x) como el número de vocales (eje y)



"cat" ----> **[3,1]**

"scrabble" ----> **[8,2]**

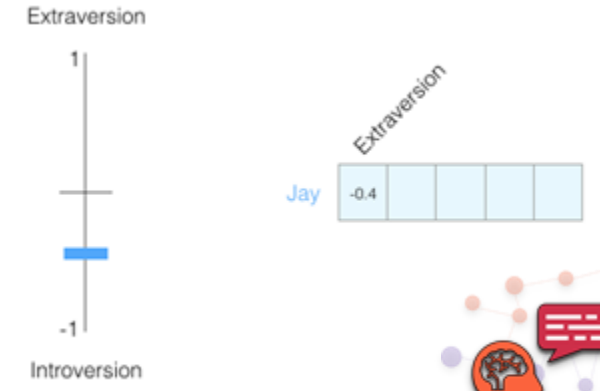
"antidisestablishmentarianism" ----> **[28,11]**

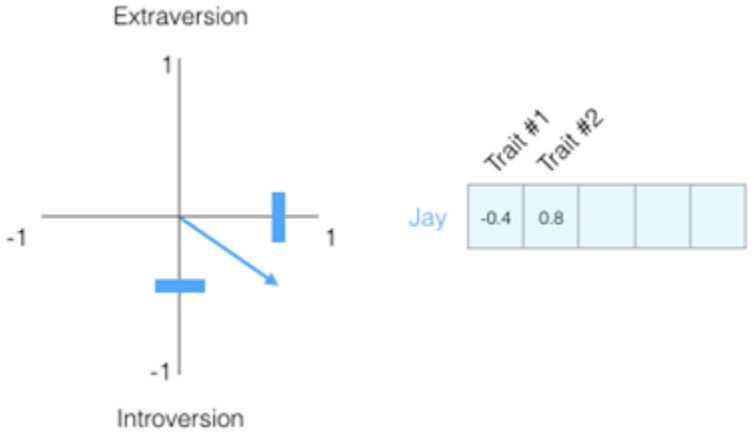
3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Ejemplo: ¿cómo podríamos representar la personalidad?

Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100

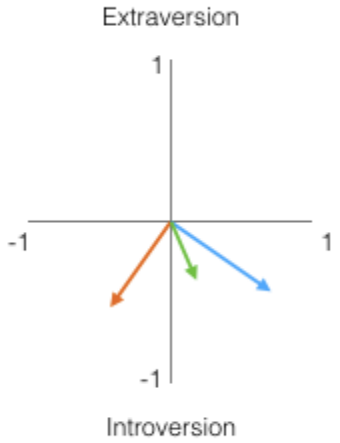


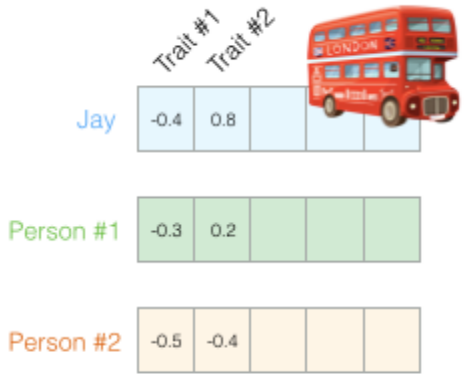


3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Los vectores permiten manejar la idea de distancia ...





3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Representación vectorial de la palabra en inglés "**king**"

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 ,
0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 ,
0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 , -
0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -
0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754
, 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -
0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

```python
# Importamos 'SpaCy'
import spacy

nlp = spacy.load("en")
nlp("cat").vector
```

Output:

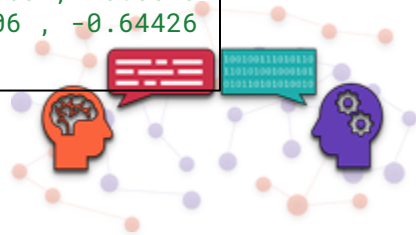
```
array([ 1.4532998 ,  0.47115922,  0.71999407,  1.1403103 ,  3.1126573 ,
        0.6275163 ,  2.1985044 ,  3.1114974 ,  1.0527445 ,  3.2556517 ,
        3.7162375 , -0.10062158,  2.56286   , -3.1346574 ,  1.2517695 ,
       -0.56736076, -2.0715995 ,  1.2972176 , -1.0384768 , -2.8220317 ,
        0.4024135 ,  1.1475914 , -0.9666666 , -1.8358834 ,  0.5618948 ,
       -1.109226  ,  3.1648145 , -3.0343432 ,  1.5009679 ,  1.8909831 ,
        2.1640768 , -0.868277  ,  1.5119176 ,  1.7287943 ,  2.2542849 ,
       -1.7297868 ,  1.8519063 ,  0.29412246, -3.1516666 ,  2.2560802 ,
        2.8821528 ,  0.9065895 , -1.4355452 , -4.937831  ,  0.07267573,
       -1.5819955 ,  0.7202336 , -1.0646656 , -0.19294381, -0.21828318,
       -0.21474707, -0.25848413, -0.41141343, -1.650431  , -1.6427782 ,
        2.0982878 ,  0.8212584 ,  2.458171  , -0.09555507, -0.8922238 ,
        0.7312181 , -1.4605643 , -0.46285537, -1.5074668 , -2.750762  ,
       -3.9613488 ,  1.5867741 , -5.1115217 , -0.21341431,  0.9157888 ,
       -2.3833017 ,  2.126943  ,  2.4008074 ,  0.91968673, -1.3888277 ,
       -1.4603075 ,  2.4146914 , -0.75423056, -1.8709452 ,  0.6487465 ,
       -0.8690753 , -1.9576063 , -1.8812385 , -0.77405465,  2.5276294 ,
        3.4152527 , -1.2771642 , -1.703851  , -0.6723095 , -2.5397158 ,
       -0.1799444 , -0.87516195, -2.0097623 ,  3.7264547 ,  1.2048597 ,
       -1.2384005 ], dtype=float32)
```

3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Representación vectorial de la palabra en inglés "**king**"

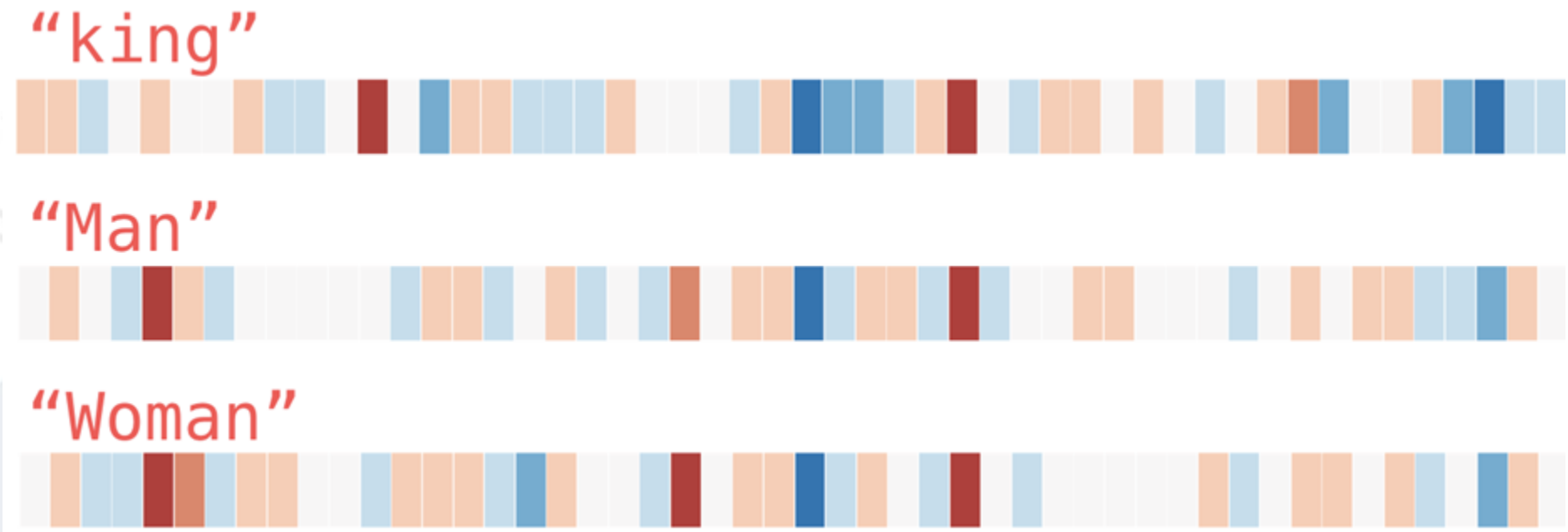

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -
0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809
, 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354
, 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523
, 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426
, -0.51042 ]
```



3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

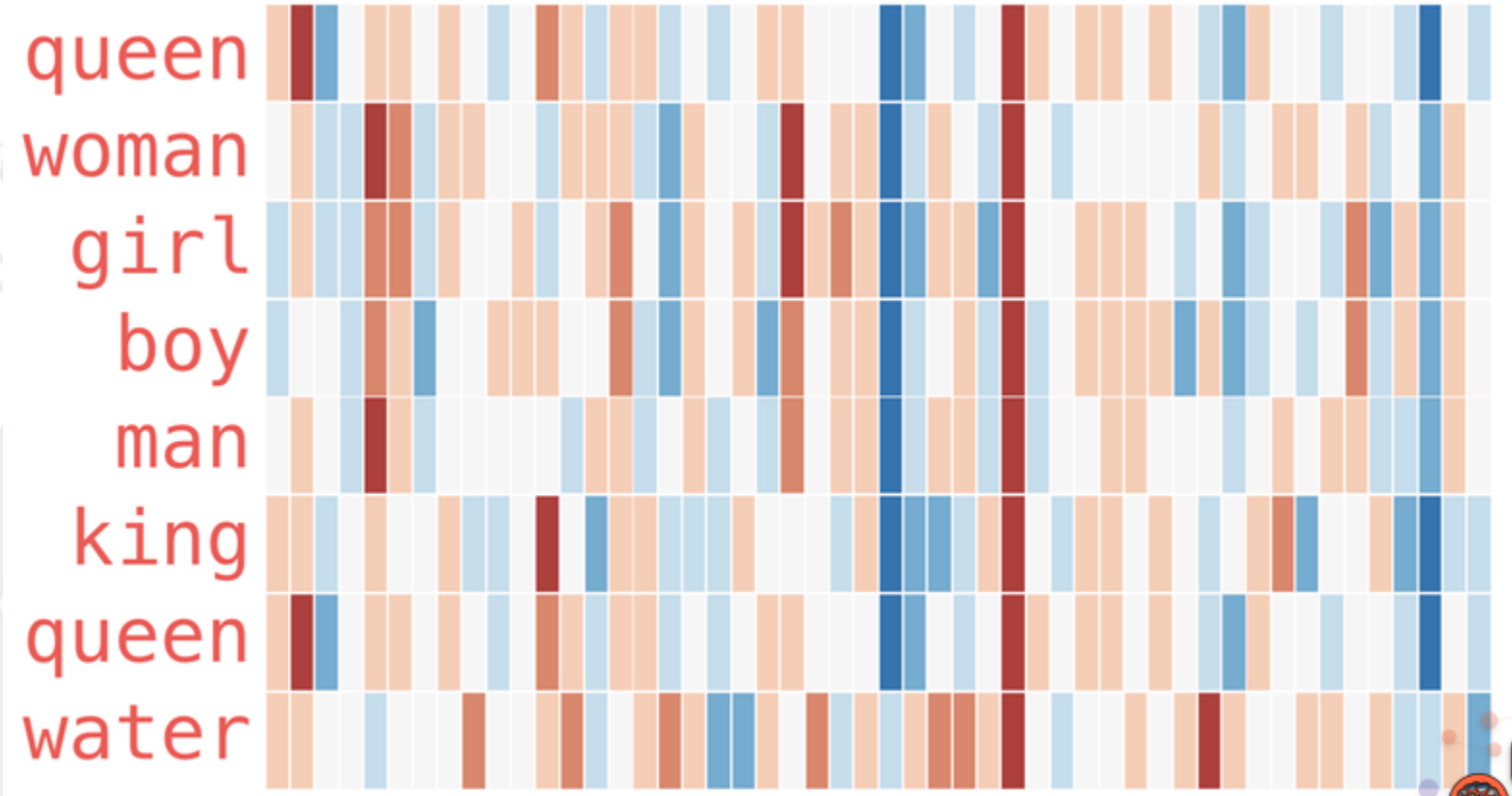
Un término vectorizado puede ser comparado a otros ...



3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Un término vectorizado puede ser comparado a otros ...



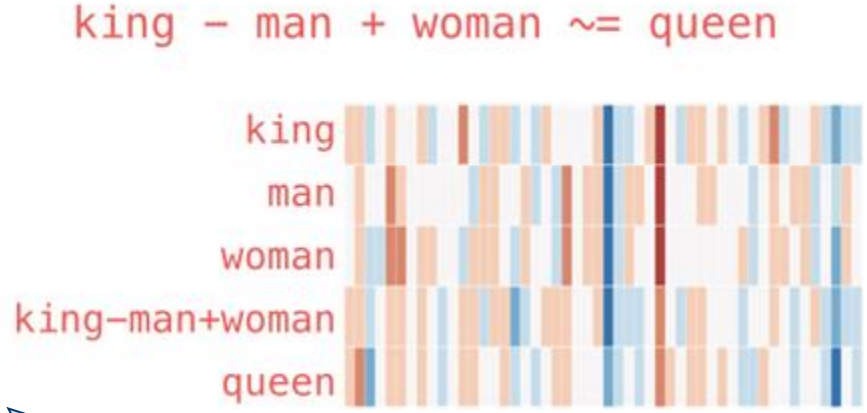
3. Word Embeddings

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

… y se pueden realizar operaciones por **analogía**

```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.721449077129364),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```



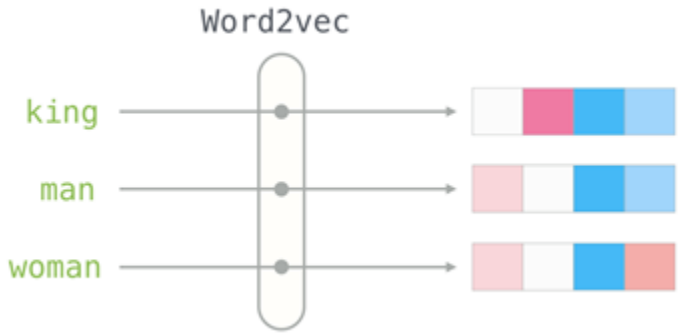
"queen" no equivale a **"king - man - woman"**, pero es lo más aproximado en el *corpus* donde se ubican!!

3. Ejemplo :: Word2Vec

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Qué es Word2Vec?

- Algoritmo de aprendizaje estadístico que desarrolla incrustaciones de palabras, tomando un *corpus* de texto, y generando una representación vectorial de cada palabra.
- Utiliza una **red neuronal superficial de 2 capas** para generar los valores que representan palabras con un contexto similar como vectores cercanos entre sí y palabras con diferentes contextos como vectores muy separados entre sí.



3. Ejemplo :: Word2Vec

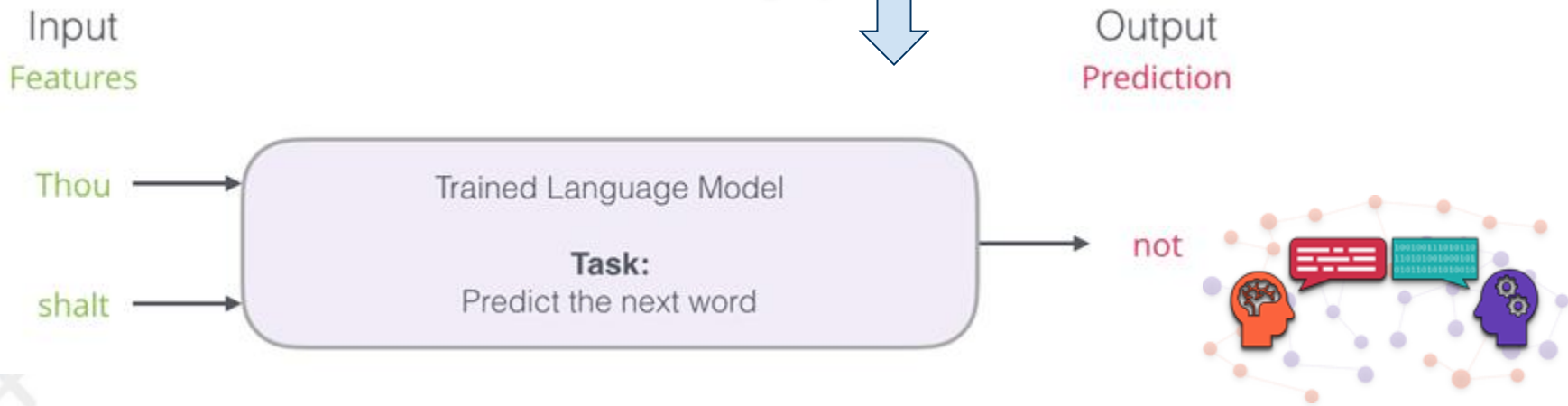
CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Word2vec**

Ejemplo de aplicación:

predicción de la siguiente palabra

modelo de lenguaje



3. Ejemplo :: Word2Vec

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Word2vec**

  Si tienes interés en comprender el enfoque para resolver un problema como el de predicción de la palabra que seguirá a otra en una frase, consulta esta página:

  Fuente:
  https://jalammar.github.io/illustrated-word2vec/

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿ Qué son los Grandes Modelos de Lenguaje (LLMs)?

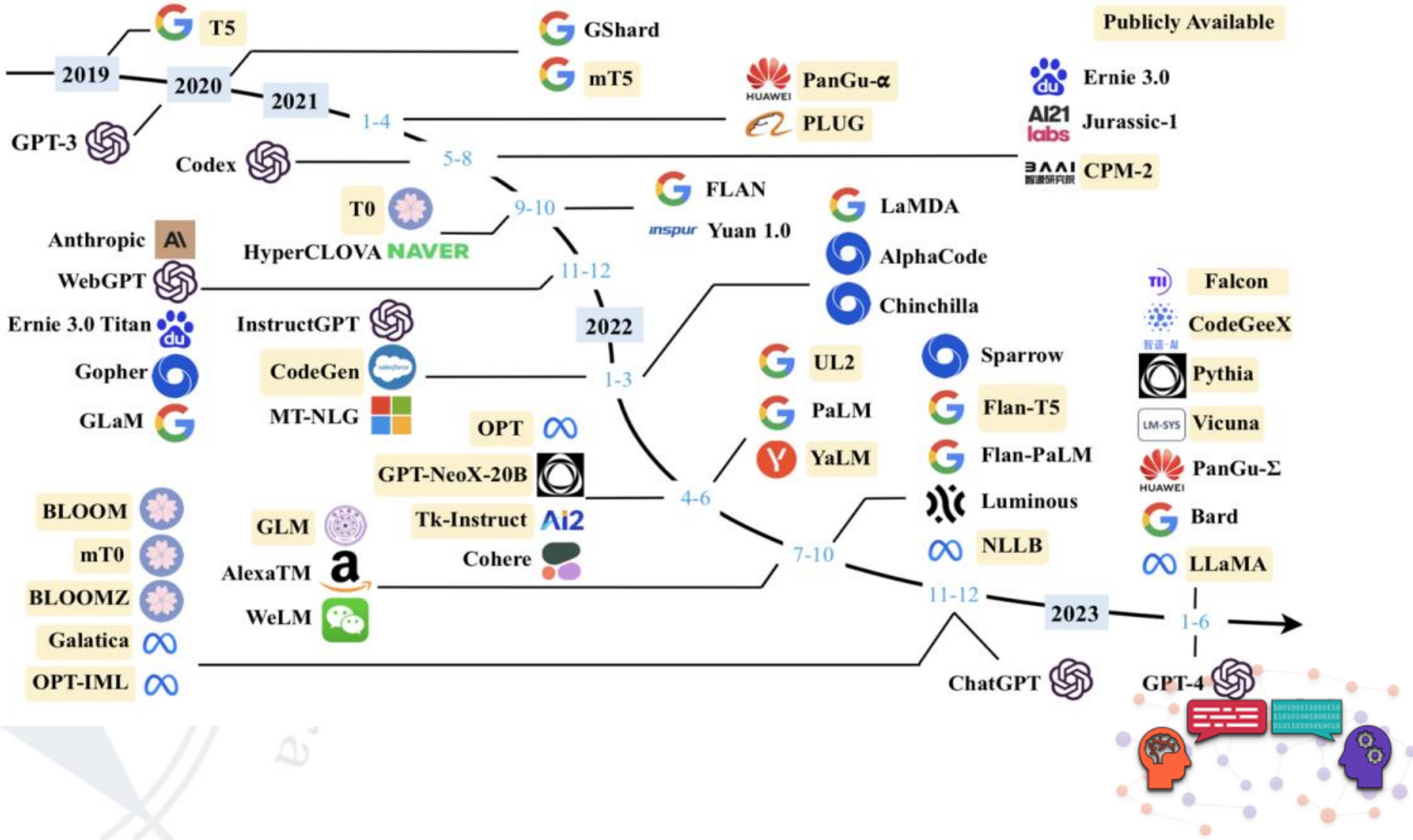
- Los Grandes Modelos de Lenguaje (LLMs) son un tipo de modelo de aprendizaje profundo diseñados para comprender y generar texto en lenguaje natural.

- Los grandes modelos de lenguaje (LLM) se llaman "grandes" porque están pre-entrenados con un gran número de parámetros en gran *corpus* de texto.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

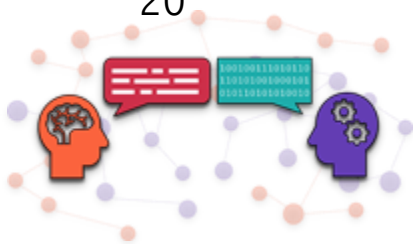
La lista de LLMs empieza a ser inacabable ...



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Modelo	Tamaño del disco (GB)	Uso de memoria (GB)	Parámetros (millones)	Tamaño de los datos de entrenamiento(GB)
BERT-Large	1,3	3,3	340	20
GPT-2 117M	0,5	1,5	117	40
GPT-2 1.5B	6	16	1.500	40
GPT-3 175B	700	2.000	175.000	570
T5-11B	45	40	11.000	750
RoBERTa-Grange	1,5	3,5	355	160
ELECTRA-Large	1,3	3,3	335	20



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿ Qué son los Grandes Modelos de Lenguaje (LLMs)?

- Los LLMs resuelven tareas de PLN relacionadas con la generación de texto, con la sumarización, la traducción y con los procesos pregunta-respuesta (Q&A).

- Estos modelos están basados en arquitecturas de redes neuronales conocida como "**transformers**".

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿ Qué características definen a un **transformador**?

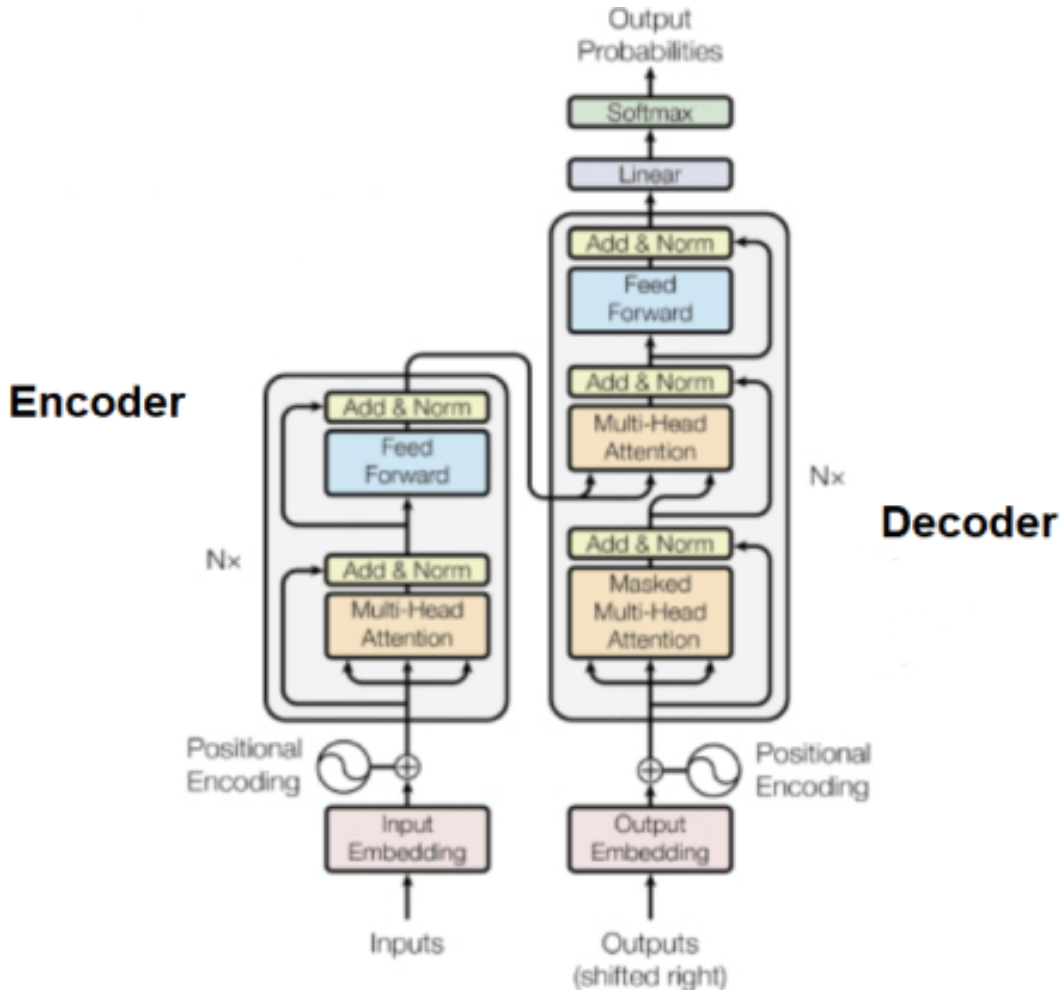
- **Objetivo**: capturar patrones de lenguaje complejos y relaciones entre palabras o frases en conjuntos de datos de texto a gran escala.
- Mecanismo clave: **atención**
  - auto-atención ("self-attention")
  - atención cruzada ("cross-attention")
- La **atención** permite al modelo comprender las palabras de un texto al otorgarles una importancia o peso relativo dentro del contexto del que forman parte.

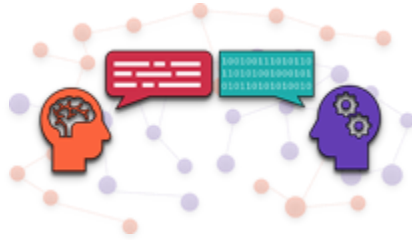
3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Una red neuronal basada en **transformadores** combina 2 elementos:

- **codificadores**
- **decodificadores**





3. Modelos generativos

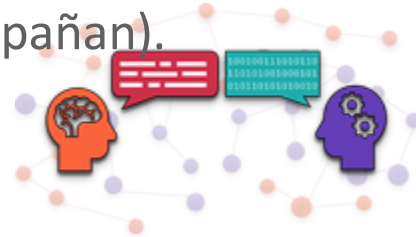
CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿ Qué hace un **codificador** ("encoder")?

- **Objetivo**: una red neuronal de codificación toma una entrada y produce una secuencia de "estados ocultos".

- **Ejemplo**: en una frase del tipo *"el gato está sentado en la alfombra" ...*

    1. Se procesaría cada palabra, vectorizándola como **incrustaciones** ("embeddings").
    2. El procesamiento implica usar la **auto-atención** para generar una serie de estados ocultos donde, para cada palabra, un vector guarda información de su contexto (otras palabras que le acompañan).



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿ Qué hace un **decodificador** ("decoder")?

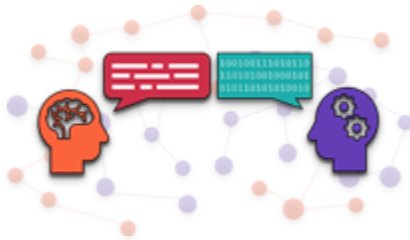
- **Objetivo**: una red neuronal de decodificación considera una secuencia y usa la salida de un codificador para generar una serie de predicciones.
- **Ejemplo**: al traducir al francés una frase del tipo *"el gato está sentado en la alfombra" ...*
    1. Se toma una secuencia de palabras, incorporando los estados ocultos de cada palabra generados en la fase de codificación.
    2. El procesamiento implica usar el mecanismo de **atención cruzada** para, tomando cada palabra y sus estados ocultos, proponer un equivalente al francés en función del contexto.
    3. Sucesivas iteraciones permitan actualizar el estado del decodificador hasta completar la secuencia inicial.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tipos de LLMS

- Modelos autoregresivos
- Modelos autocodificado
- Modelos híbridos



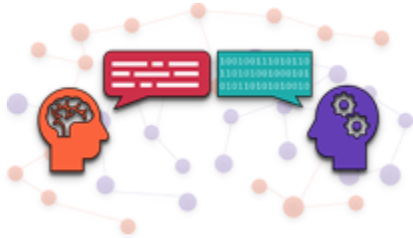
3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tipos de LLMS

Modelos autoregresivos | **Modelos de autocodificado** | **Modelos híbridos**

- Generan texto prediciendo la siguiente palabra dada una secuencia previa de palabras.
- Están entrenados para maximizar la probabilidad de cada palabra en un conjunto de datos para cierto contexto.
- **Ejemplos**: La familia de modelos GPT (*Generative Pre-trained Transformer*) de OpenAI



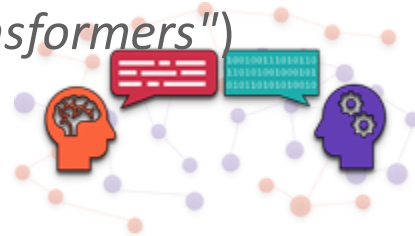
3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tipos de LLMS

Modelos autoregresivos | **Modelos de autocodificado** | **Modelos híbridos**

- Generan representaciones vectoriales de una entrada de texto al reconstruir la entrada original a partir de una versión "incorrecta" de la misma
- Están entrenados para predecir palabras ausentes en una entrada de texto dado cierto contexto.
- **Ejemplos**: BERT (*"Bidirectional Encoder Representations from Transformers"*) de Google



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tipos de LLMS

Modelos autoregresivos

Modelos de autocodificado

Modelos híbridos

- Combinan modelos autoregresivos y de autocodificado.
- **Ejemplos**: T5 (*"Text-to-Text Transformer"*) permite realizar tareas de generacinón y comprensión de texto, con usos variados tales como traducción de texto, realización de resúmenes o respuesta a preguntas.

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Qué es **LangChain**?

- **LangChain** es un potente *framework* para crear aplicaciones que generan texto, responden preguntas, traducen idiomas y otras tantas funciones relacionadas con el PLN.

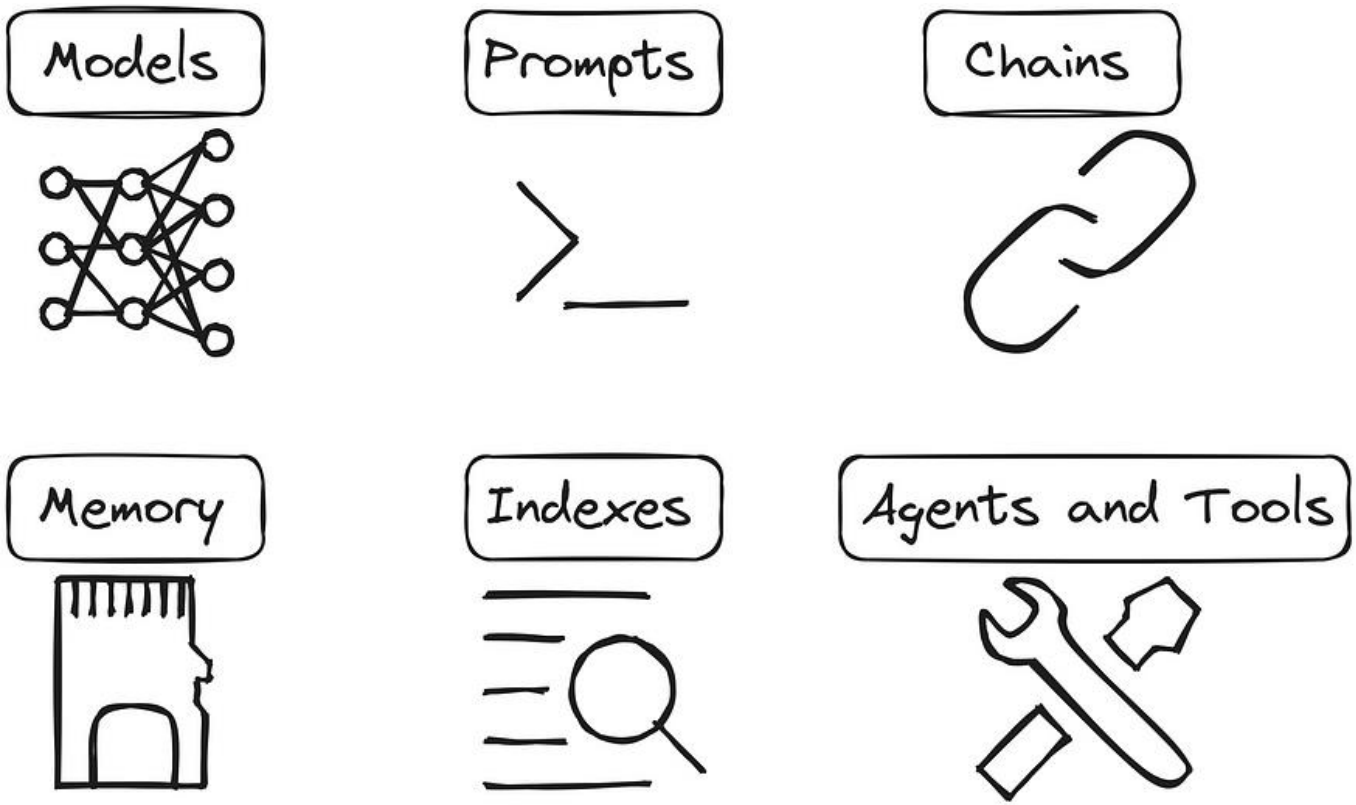
- "**Lang**" significa lenguaje (enfoque principal de LangChain) y "**chain**", se refiere al componente de ***cadena*** (connotación de conectar cosas) utilizado en LangChain.

- LangChain permite *llegar* donde un LLM de por sí no llega: añade funcionalidad y lo complementa con datos extra.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

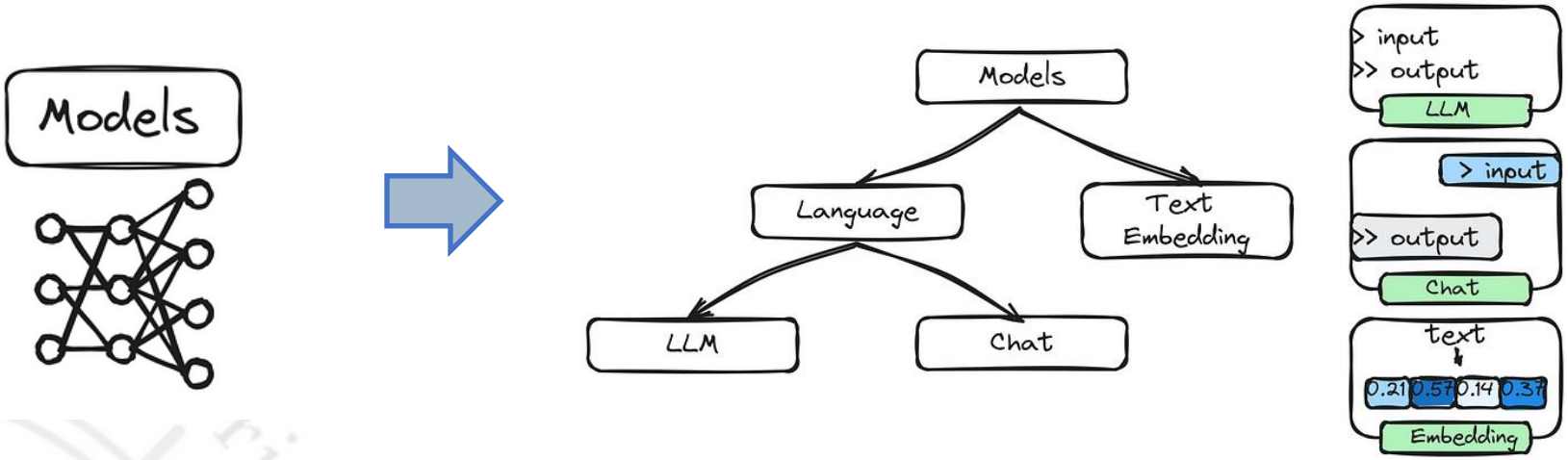
Componentes de LangChain



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

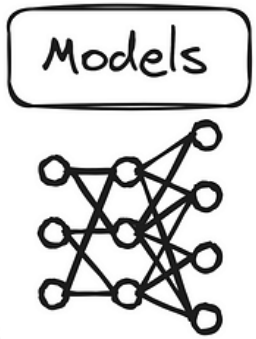


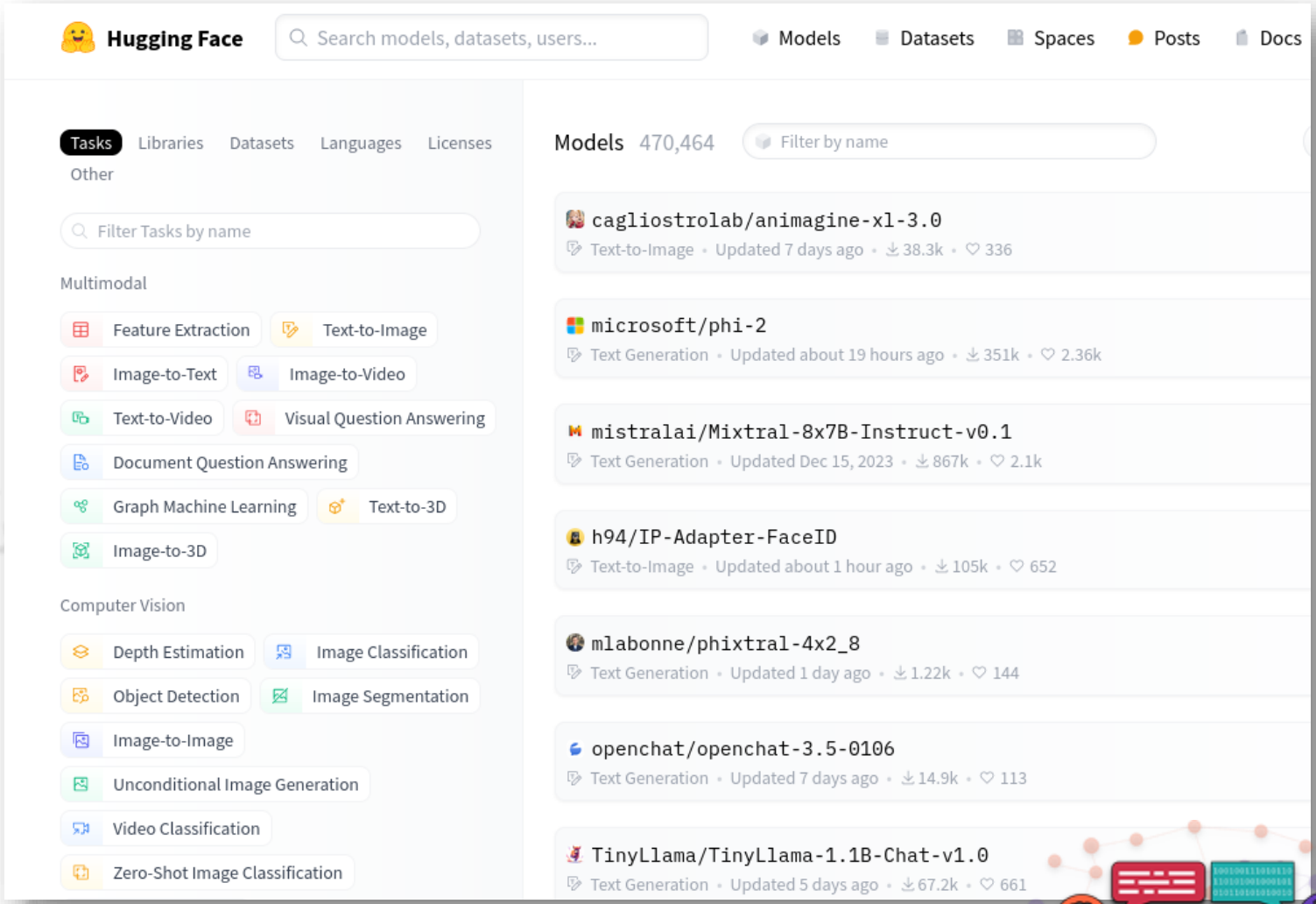
- Los **modelos** en LangChain son LLMs entrenados en enormes cantidades de conjuntos de datos masivos de texto y código.

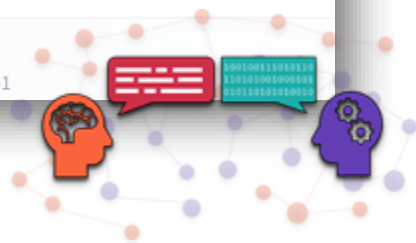
- Los modelos se utilizan en LangChain para generar texto, responder preguntas, traducir idiomas, etc.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior



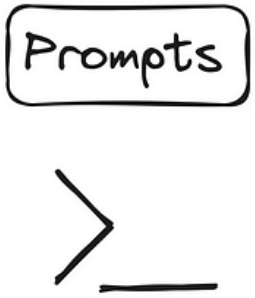


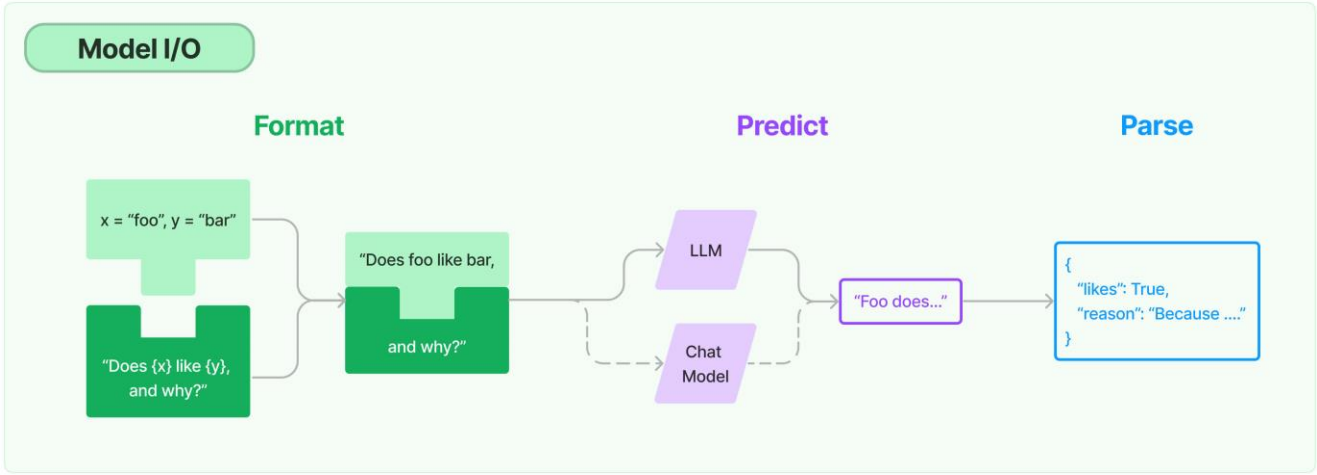


3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain





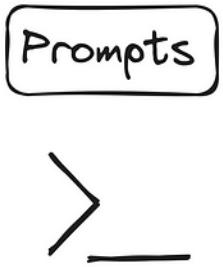
- Los **prompts** son fragmentos de texto que guían al LLM para generar el resultado deseado.

- LangChain permite definir plantillas para dar forma a los prompts.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain



USOS

1) Especificar el formato de salida deseado

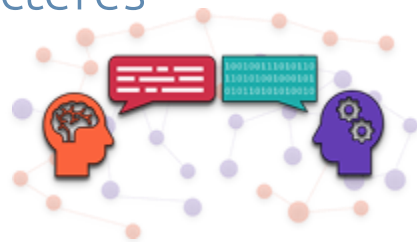
   Ejemplo: "Traduce cierta entrada al chino mandarín"

2) Proporcionar contexto

   Ejemplo: "Asume el rol de un profesor"

3) Limitar la salida

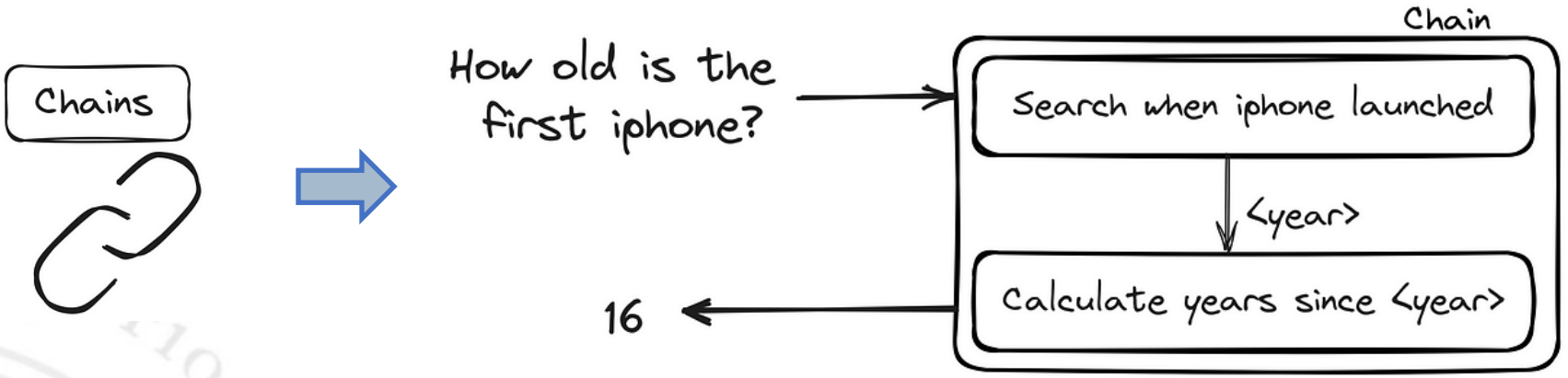
   Ejemplo: "Genera un texto de 140 caracteres"



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

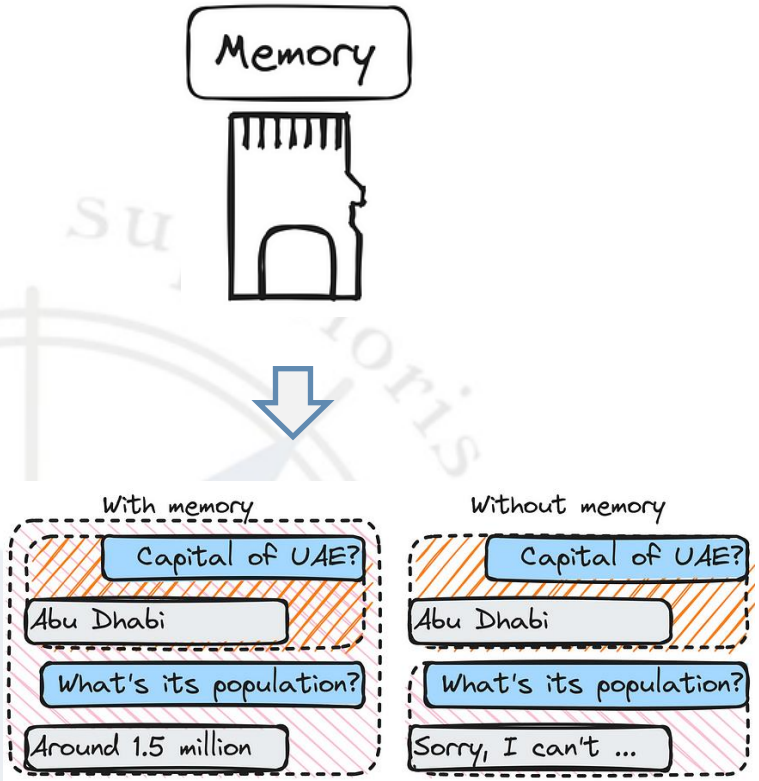


- Las **cadenas** son secuencias de instrucciones que el LangChain ejecuta para realizar una tarea.
- Pueden conectar otros componentes de LangChain en función de los requisitos de la aplicación.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

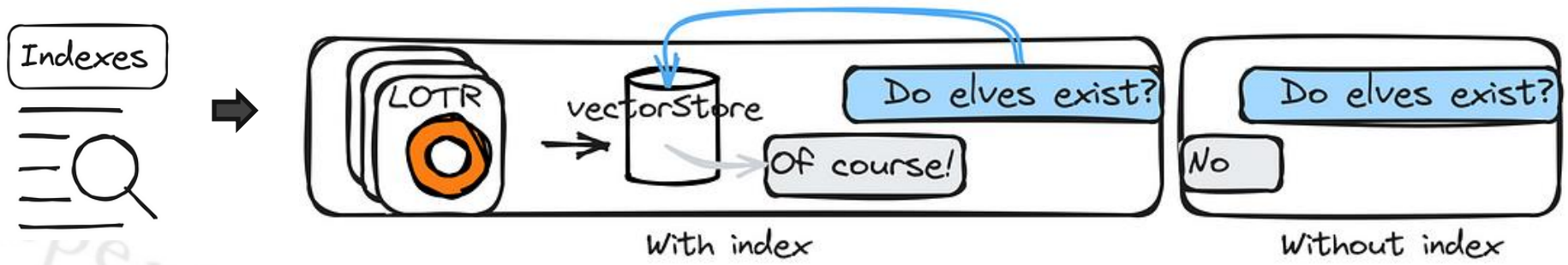


- La **memoria** en LangChain es un método de almacenamiento de datos al que un LLM puede acceder posteriormente.
- Esta información puede incluir los resultados de la cadena anterior, el contexto de la cadena actual y cualquier otra información que requiera el LLM.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain



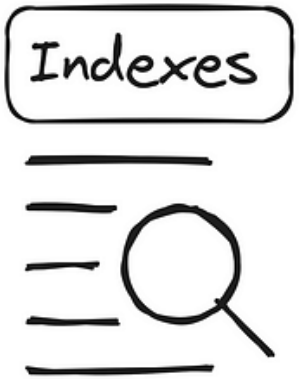
- Para realizar ciertas tareas, los LLM necesitan **acceso a fuentes de datos externas** específicas no incluidas en su conjunto de datos de entrenamiento, como documentos internos, correos electrónicos o conjuntos de datos.
- LangChain hace referencia colectivamente a dicha *documentación externa* como **índices**.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

Cargadores de documentos



Base de datos vectoriales (*store_vectors*)

Divisores de texto

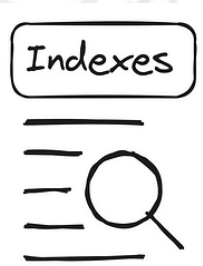
Recuperadores

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

Cargadores de documentos



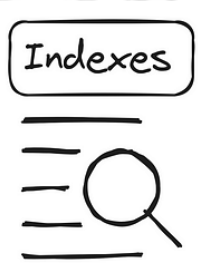
- LangChain ofrece una amplia variedad de *cargadores de documentos* para aplicaciones de terceros.
  - **Servicios de almacenamiento de archivos** (como Dropbox, Google Drive y Microsoft OneDrive)
  - **Contenido web** (como YouTube, PubMed o URL específicas)
  - **Herramientas de colaboración** (como Airtable, Trello, Figma y Notion), **bases de datos** (como Pandas, MongoDB y Microsoft), etc.

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

Base de Datos Vectoriales



- Las **bases de datos vectoriales** (**VectorStore**) representan conjuntos de datos desestructurados, que se almacenan en forma de vectores ("embeddings") indexados.

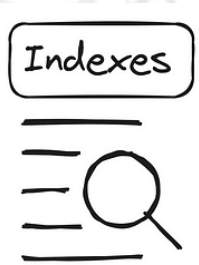
- Este tipo de almacenes de vectores permiten extraer términos (o fragmentos) similares dada cierta entrada de texto

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

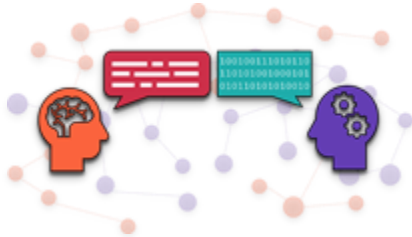
Componentes de LangChain

Divisores de texto



- Para aumentar la velocidad y reducir las exigencias computacionales, Langchain facilita la división de grandes documentos de texto en partes más pequeñas.

- Mediante los ***TextSplitters***, por ejemplo, se puede dividir el texto en pequeños fragmentos semánticamente significativos que luego se pueden recombinar para diferentes usos.

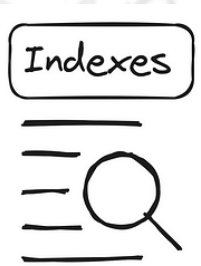


3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain

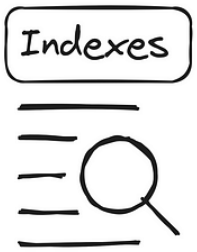
Recuperadores



- Una vez que se han conectado las fuentes externas de datos, el modelo usado debe ser capaz de recuperar e integrar rápidamente la información.
- Los **recuperadores** (**retrievals**) articulan los elementos previos como parte de la cadena definida.

3. Modelos generativos

Ejemplo de cadena con índices



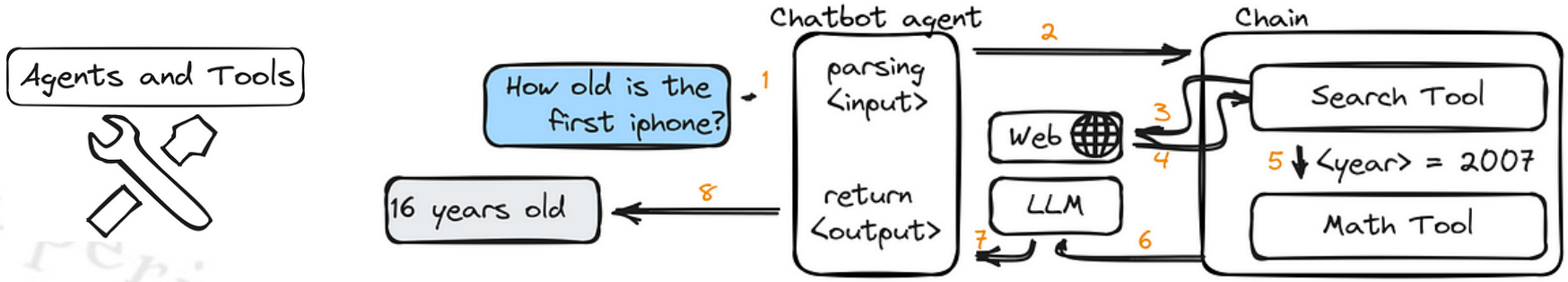
Objetivo: responder preguntas de tipo **financiero.**

1. Se crea una **cadena** que incluya un **índice** para buscar todos los documentos que contegan la palabra "**finanzas**".
2. La cadena consultará una ***base de datos vectorial*** en busca de ese u otros términos que sean similares a "**finanzas**" ("**dinero**", "**inversiones",** etc.).
3. La cadena utilizará un ***recuperador*** para extraer los documentos más útiles para una consulta del tipo: "**¿Cuáles son las formas de invertir?".**

3. Modelos generativos

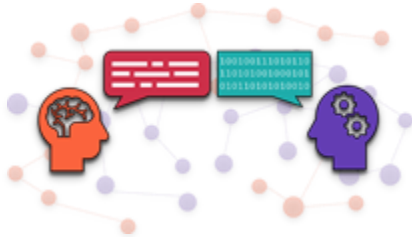
CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain



- Los **agentes** utilizan un modelo de lenguaje determinado como un "motor de razonamiento" para determinar qué acciones realizar.

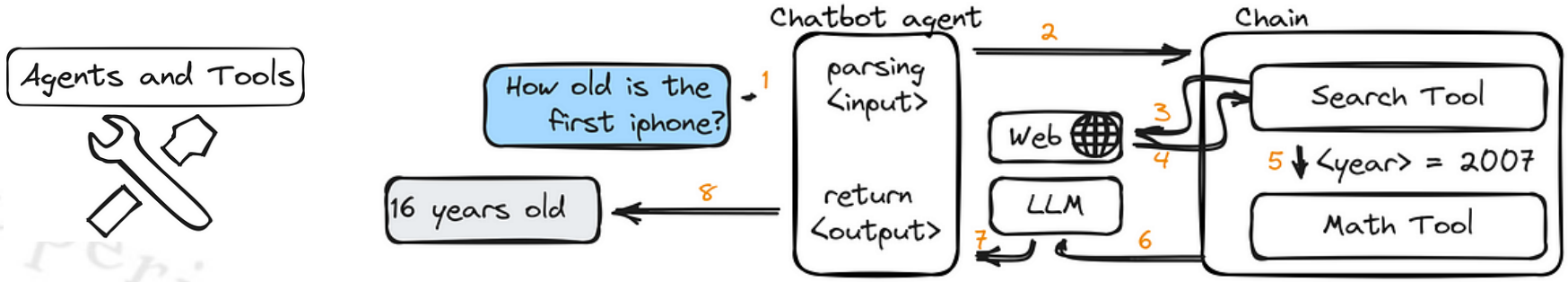
- Al crear una *cadena* para un *agente*, las entradas pueden incluir:
  - una lista de las herramientas disponibles que se pueden aprovechar.
  - entrada del usuario (como prompts y consultas).
  - cualquier paso relevante previamente ejecutado.



3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Componentes de LangChain



- Las **herramientas** son un conjunto de funciones que permiten a los agentes de LangChain interactuar con información del mundo real con el fin de ampliar o mejorar los servicios que puede proporcionar.
- **Ejemplos**: Google Search, Wikipedia, OpenWeather, …

3. Modelos generativos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Práctica 3.5**
  - Realiza la práctica **"PLN - P3.5 :: Creación de un asistente con Langchain"**, cuyo enunciado encontrarás en "Aules".
- **Práctica 3.6**
  - Realiza a continuación la práctica **"PLN – P3.6 :: Chatbot para consultas a documentos PDF"**, cuyo enunciado encontrarás en "Aules".