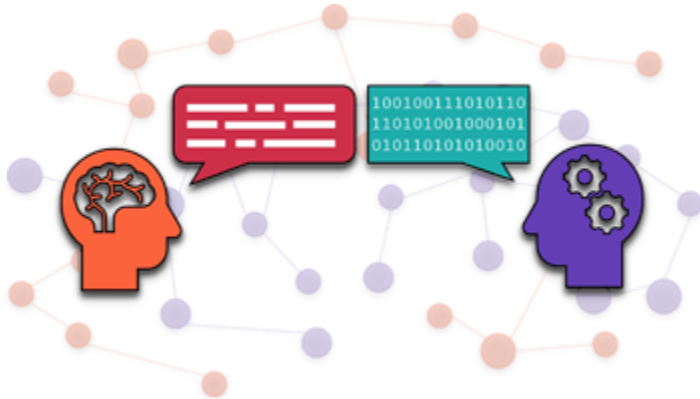
Modelos de Inteligencia Artificial

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

:: U3 ::

Procesamiento del Lenguaje Natural



Introducción al PLN

Curso 2023-24

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Introducción al Procesamiento del Lenguaje Natural (PLN)

1. ¿Qué es el PLN?

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Pensemos en las siguientes tareas ...

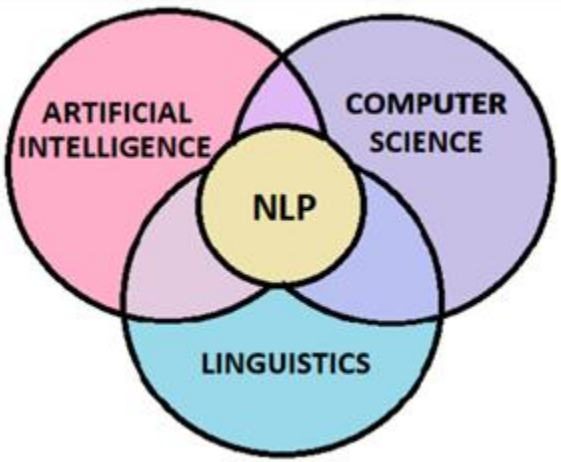
- Corrección ortográfica y autocorrección
- Subtítulos de video generados automáticamente
- Asistentes virtuales como “Alexa” de Amazon
- Detección de SPAM
- Asistentes conversacionales
- ...

... ¿Qué tienen en común?

1. ¿Qué es el PLN?

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- EL **PLN** es un campo que se encuentra en la intersección entre la lingüística, la inteligencia artificial y la informática
- **Objetivo**: permitir que las computadoras interpreten y analicen el lenguaje natural
- **Problema**: entender el lenguaje natural es una tarea no precisamente obvia ...



1. ¿Qué es el PLN?

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Ejemplos de ambigüedad lingüística:

  > Sintáctica:

  "Se citaron en el banco donde se habían conocido"

  > Gramatical:

  "Cuando apoyamos el cuadro sobre la mesa, se rompió"

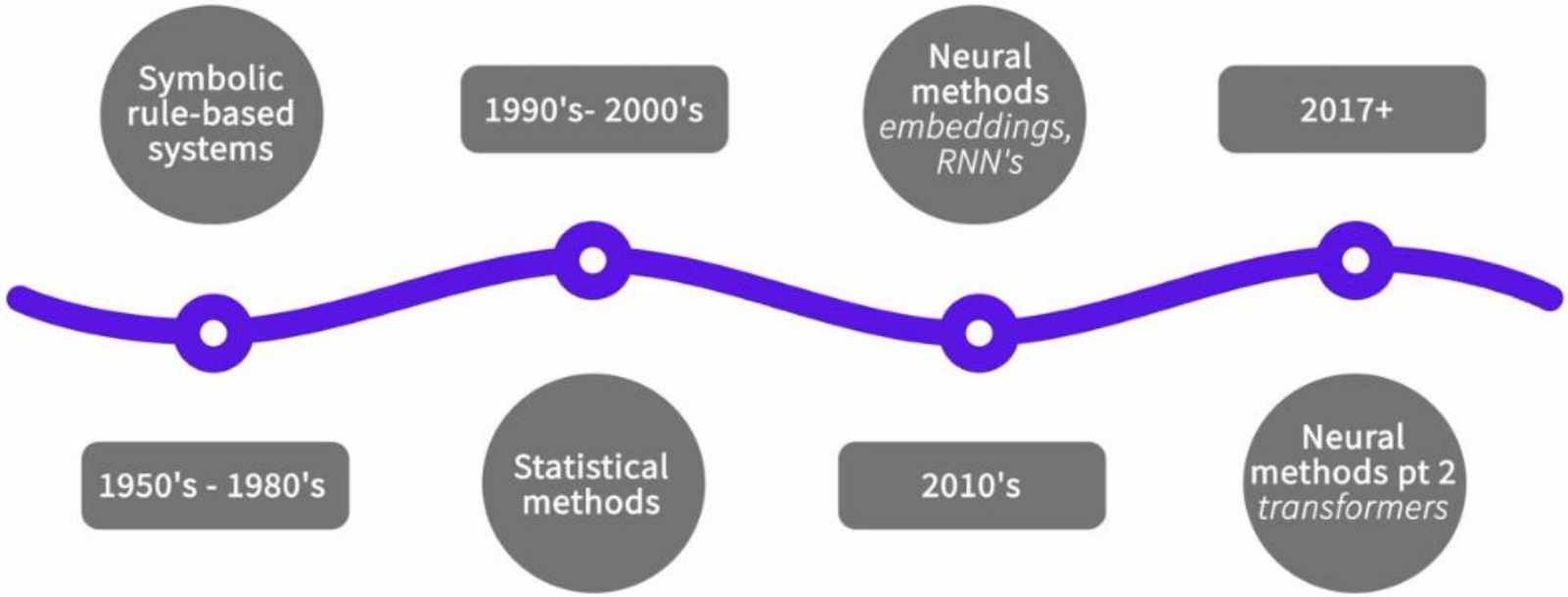
  > Polisémica:

  "Eligió un coche rápido"

1. ¿Qué es el PLN?

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

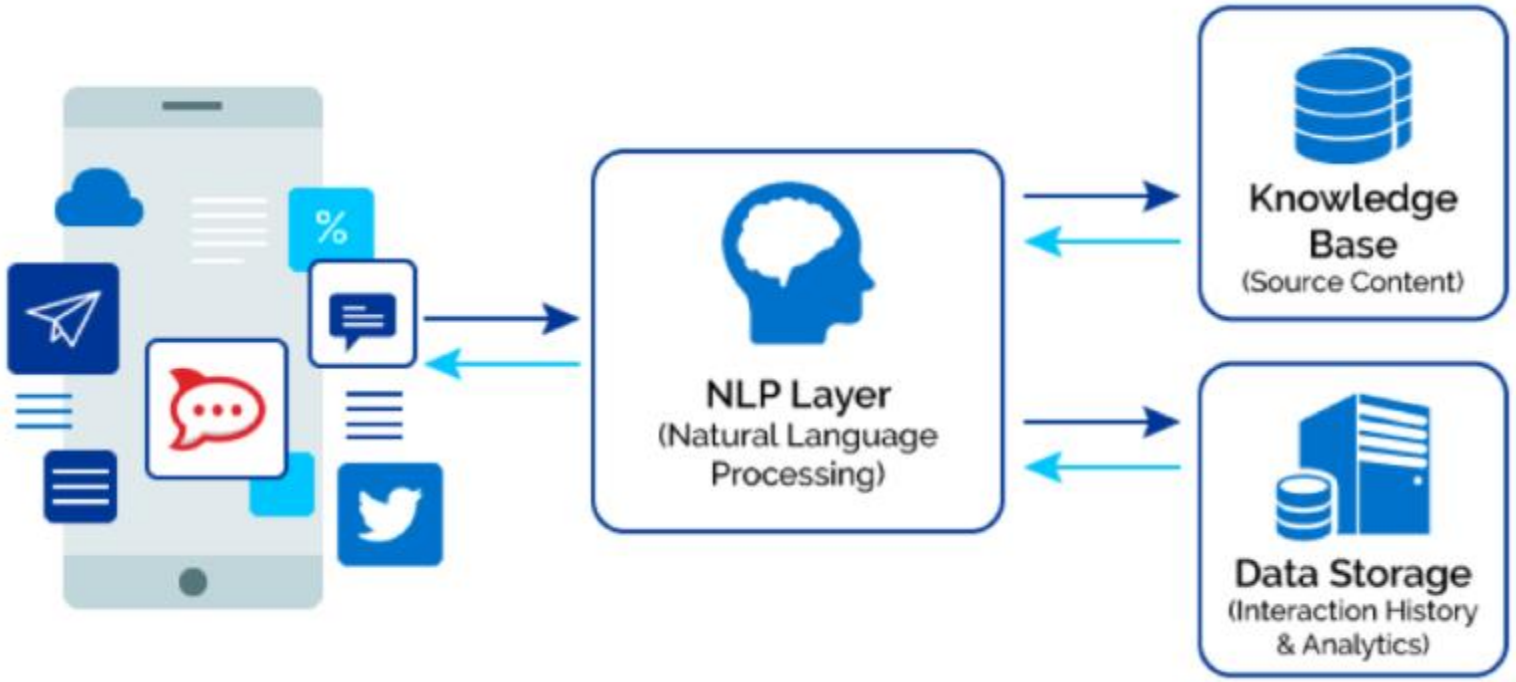
Evolución del PLN



1. ¿Qué es el PLN?

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

El **PLN** es un interfaz entre tipos de datos y soportes ...



2. Aplicaciones del PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Sentiment analysis

Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter...I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mollette brilliant! May the fourth be with you #starwarsday #starwars	Positive
Good morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

2. Aplicaciones del PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

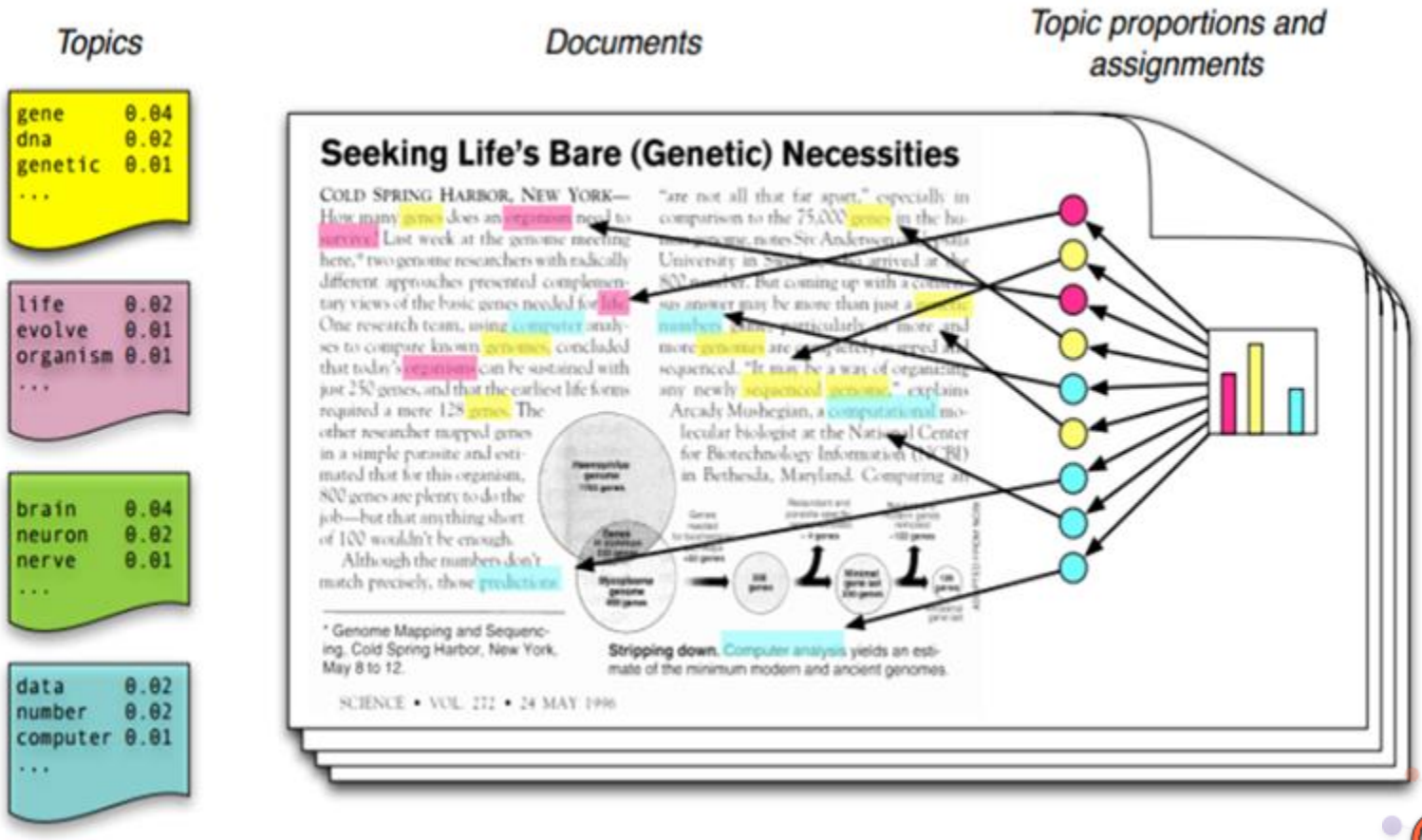
- Generación de resúmenes

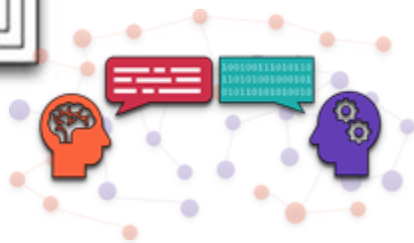
London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium. London's ancient core, the City of London, largely retains its 1.12-square-mile (2.9 km2) medieval boundaries.

2. Aplicaciones del PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Identificación y clasificación de temas

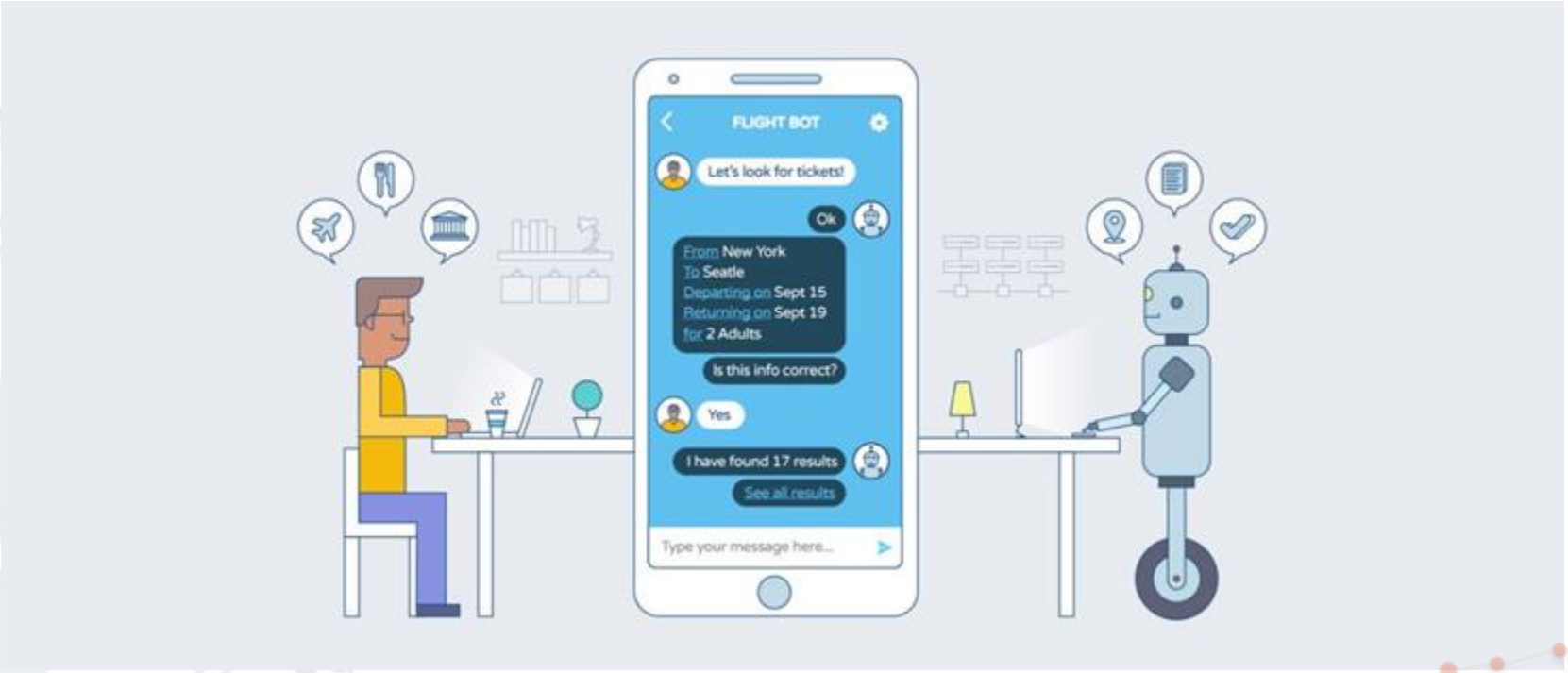




2. Aplicaciones del PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

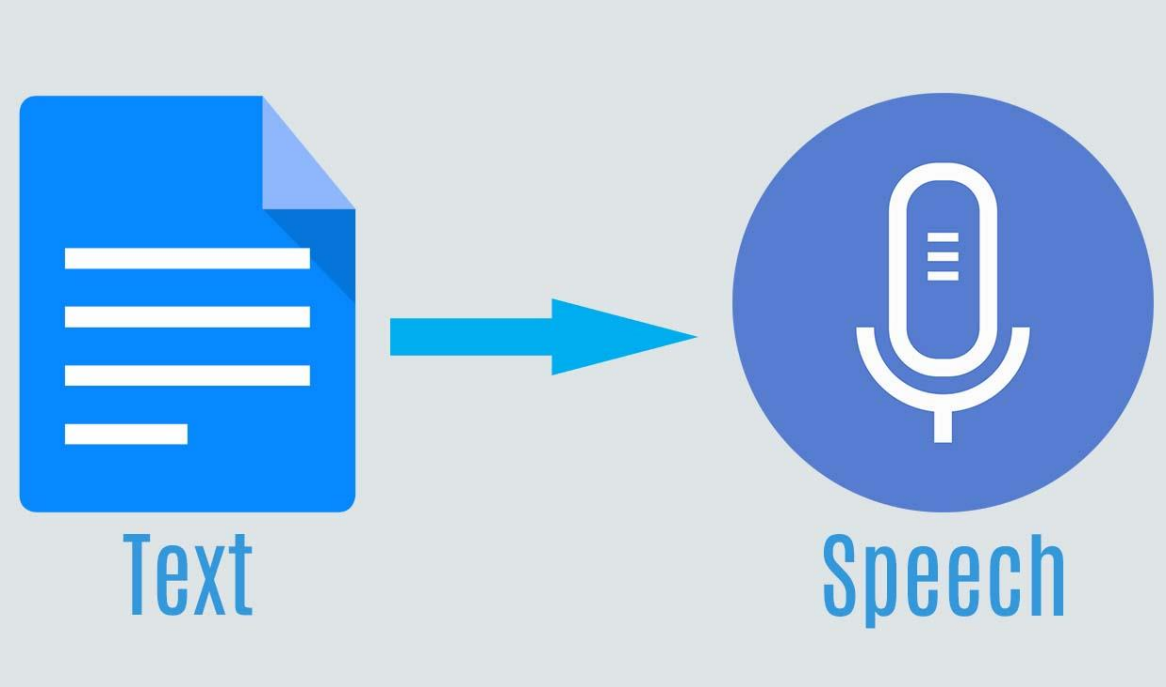
- Creación de chatbots conversacionales



2. Aplicaciones del PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Reconocimiento y síntesis texto a voz



3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

a) Definición problema (pregunta)

b) Exploración de datos (EPA)

c) Aplicación de técnicas de análisis

d) Comunicación de los resultados

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

a) **Definición problema (pregunta)**

b) Exploración de datos (EPA)

c) Aplicación de técnicas de análisis

d) Comunicación de los resultados

"¿Cómo podemos utilizar el procesamiento del lenguaje natural para optimizar la clasificación y asignación de tickets de soporte?"

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

a) Definición problema (pregunta)

b) **Exploración de datos (EPA)**

c) Aplicación de técnicas de análisis

d) Comunicación de los resultados

- **datos**: historiales de conversaciones de clientes y agentes de soporte

- **objetivo**: detectar patrones, analizar variabilidad del lenguaje, actitudes, ...

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

a) Definición problema (pregunta)

b) Exploración de datos (EPA)

c) **Aplicación de técnicas de análisis**

d) Comunicación de los resultados

- uso de **técnicas** de procesamiento del lenguaje natural

- uso de un **modelo** que permita reconocer categorías, niveles de urgencia, etc.

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Cómo afrontamos la resolución de un problema mediante aprendizaje automático?

a) Definición problema (pregunta)

b) Exploración de datos (EPA)

c) Aplicación de técnicas de análisis

d) **Comunicación de los resultados**

- uso de métricas que muestran la fiabilidad del modelo

- elementos visuales par auna comunicación eficaz

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- El proceso de resolución de problemas en el PLN se asemeja al esquema de trabajo seguido en el aprendizaje automático.

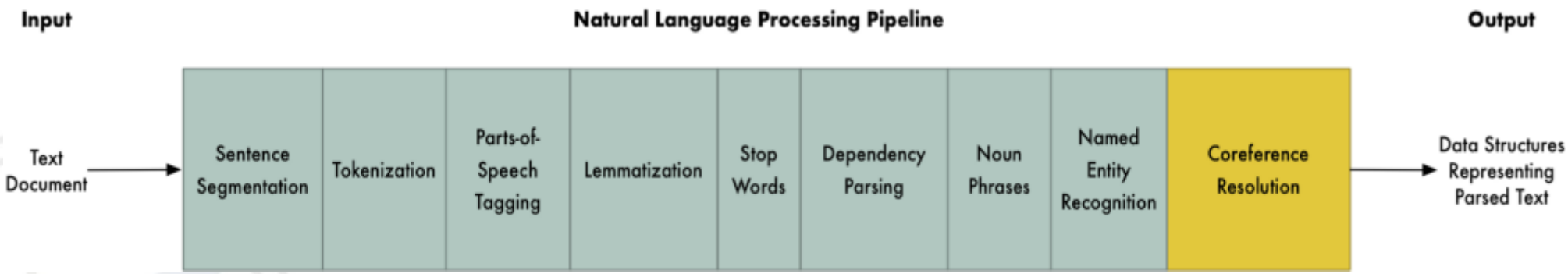
3. El flujo de trabajo en el PLN

- El proceso de resolución de problemas en el PLN se asemeja al esquema de trabajo seguido en el aprendizaje automático.
    - se organiza como una tubería (“**pipeline**”)
    - una tubería comprende un conjunto de pasos que definen un “flujo de trabajo” (“**workflow**”)

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Ejemplo de “pipeline” en PLN



3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Un “pipeline” en PLN puede incluir 3 apartados generales:

    a. **Preprocesamiento del texto**

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Un “pipeline” en PLN puede incluir 3 apartados generales:
    a. Preprocesamiento del texto
    b. Análisis del texto (“parsing”)

3. El flujo de trabajo en el PLN

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Un “pipeline” en PLN puede incluir 3 apartados generales:
    a. Preprocesamiento del texto
    b. Análisis del texto (“parsing”)
    c. Uso de modelos estadísticos

3a. Preprocesamiento del texto

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Supone limpiar y preparar el texto para su procesamiento posterior:
- **Técnicas**:
  - Eliminación de ruido (“Noise removal”)
  - Tokenización
  - Normalización
    - Derivación (“Stemming”)
    - Lematización

3b. Análisis del texto (“parsing”)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- El análisis es un proceso del PLN relacionado con la segmentación de texto según su **sintaxis**.
- Este tipo de análisis puede resultar útil para saber cómo se relacionan las palabras entre sí (es decir, cuál es la sintaxis subyacente)

3b. Análisis del texto (“parsing”)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Técnicas**:
  - Etiquetado de parte del discurso (“**Part-of-speech tagging**”)
  - Reconocimiento de entidades nombradas (“**Named Entity Recognition**”)
  - Árboles de dependencia gramatical
  - Otras tareas: conversión a minúsculas, eliminación de palabras vacías (“Stopwords”), corrección ortográfica, etc.

3c. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Los modelos lingüísticos son modelos **probabilísticos** del lenguaje basados en el procesamiento de datos vectoriales.
- Estos modelos se construyen para calcular la probabilidad de aparición o uso de un sonido, letra, palabra o frase determinados en cierto contexto.
- Una vez que se ha entrenado un modelo, se puede probar en nuevos textos → “aprendizaje automático”.

3c. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Modelos de uso común:
  - "Saco de palabras" ("**Bag-of-words**")
  - Frecuencia de términos / frecuencia inversa en documentos (**tf-idf**)
  - Incrustaciones de palabras ("**Word Embeddings**")
  - Modelos temáticos ("**Topic modelling**")
    - K-Means
    - asignación de *Dirichlet* latente (LDA)
  - Modelos "grandes" (**GPT-4, LaMDA, Bert, GPT4All, ...**)