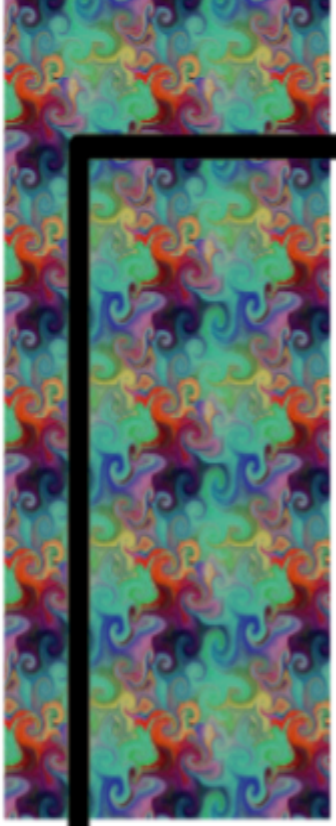


Unidad 3: Actividades

CE Inteligencia Artificial y Big Data

Sistemas de Big Data

2024/2025

Daniel Marín López

Índice

1. Expón un ejemplo de la vida real donde describas claramente el proceso de la transformación de datos en conocimiento.

Ejemplo de una panadería:

- **Datos**.- La panadería registra diariamente la cantidad de productos vendidos. Por ejemplo: los lunes vende 50 barras de pan, 30 croissants, 20 pasteles, los martes hace 45 barras de pan, 35 croissants, 25 pasteles y los miércoles hace 40 barras de pan, 50 croissants, 30 pasteles. También tiene en cuenta otros factores: clima, días festivos y horas puntas entre muchos otros.
- **Información**.- Con estos datos organizados en un sistema de ventas o en una hoja de cálculo, el dueño identifica tendencias. Por ejemplo: los croissants se venden más los miércoles, los fines de semana hay mayor demanda de pasteles y los días fríos aumentan las ventas de pan.
- **Conocimiento**.- Al analizar estos patrones, el dueño comprende que puede ajustar la producción para minimizar desperdicios y maximizar ganancias. Por ejemplo: hacer más croissants los miércoles, preparar más pasteles los fines de semana y aumentar la producción de pan los días fríos.
- **Toma de decisiones**.- Con este conocimiento, el dueño decide ajustar la producción según la demanda de cada día, lanzar promociones en productos menos vendidos para evitar desperdicios e implementar preórdenes para mejorar la planificación.

2. Indica las diferencias que existen entre un Data Lake y un Data Warehouse en cuanto a: tipos de datos, objetivo, flexibilidad, procesamiento, velocidad de acceso, costo, casos de uso, usuarios principales, herramientas comunes.

Fuentes: AWS, Todobi

Características	Data Lake	Data Warehouse
Tipos de Datos	Almacena grandes cantidades y variedades de datos, tanto estructurados como no estructurados (texto, imágenes, videos, logs).	Trabaja principalmente con datos estructurados, organizados en tablas con filas y columnas.
Objetivo	Permite descubrir patrones o preguntas que antes no se podían responder, enfocándose en información desconocida.	Responde a preguntas que ya han sido formuladas previamente, trabajando con información conocida y facilitando análisis tradicionales como reportes y dashboards.

Flexibilidad	Ofrece mayor flexibilidad al no requerir un esquema previo, permitiendo que diferentes departamentos accedan y modifiquen los datos simultáneamente.	Utiliza un esquema rígido basado en tablas con relaciones definidas, lo que limita su flexibilidad.
Procesamiento	Procesa datos no estructurados o semiestructurados (imágenes, videos, redes sociales, sensores IoT).	Trabaja con datos estructurados, almacenados en bases de datos relacionales.
Velocidad de acceso	La arquitectura de los lagos de datos da prioridad al costo y al volumen de almacenamiento por encima del rendimiento.	Un data warehouse se diseña para lograr el rendimiento de consultas más rápido.
Costo	Los costos de un data lake son más económicos.	Los costos de un data warehouse son más caros.
Casos de usos	Útil para analizar grandes volúmenes de datos y encontrar relaciones ocultas, tendencias o comportamientos no evidentes. Por ejemplo, una compañía telefónica puede usar un Data Lake para cruzar información de ventas, campañas de marketing y datos técnicos en tiempo real.	Eficaz para empresas que ya cuentan con datos organizados y necesitan análisis rápidos y estructurados (reportes, dashboards, KPIs).
Usuarios	Analistas empresariales, científicos de datos y desarrolladores de datos.	Analistas empresariales (que usan datos seleccionados), científicos de datos, desarrolladores de datos, ingenieros de datos y arquitectos de datos
Herramientas	Azure Databricks, AWS Lake Formation, Snowflake, Amazon S3	Redshift, Azure Synapse SQL, DB2 Warehouse, Google BigQuery

3. Pon un ejemplo real, distinto a un caso médico, donde expliques los 4 tipos de analítica (descriptiva, diagnóstica, predictiva y prescriptiva).

Supongamos que tu ordenador ha estado experimentando una serie de problemas como lentitud al cargar las páginas web, archivos que se corrompen o reinicios inesperados.

- **Describes** al informático los problemas que has estado experimentando.
- El informático realizará un **diagnóstico** en base a los datos, el diagnóstico describe que tu procesador no soporta el sistema operativo que tiene.

- El informático **predice** que si el ordenador sigue en este estado puede que ocurran daños irreparables a largo plazo.
- El informático, en base a las características del ordenador, te proporcionará una versión anterior del sistema operativo o uno distinto que sea compatible. Esto es la **prescripción**.

4. Indica un ejemplo real donde el uso del Big Data haya provocado una catástrofe.

Google lanzó en 2009 una aplicación llamada "Google Flu Trends" diseñada para detectar brotes de gripe analizando búsquedas relacionadas con síntomas de la gripe. Aunque pareció que funcionaba, en 2013 sobreestimó bastante el número de visitas al médico llegando a predecir el doble de las que ocurrieron. Este error se debió a lo siguiente:

- **Ruido en los datos:** El interés público en la app llevó a un aumento de búsquedas por curiosidad, no necesariamente relacionadas con casos reales de gripe.
- **Algoritmos de predicción inadecuados:** Los modelos utilizados no lograron diferenciar entre búsquedas informativas y búsquedas indicativas de enfermedad real.

Este caso destacó la importancia de complementar las técnicas de Big Data con métodos tradicionales de recolección y análisis de datos para obtener predicciones más precisas.

5. Explica cómo se relaciona el Big Data con la Inteligencia de Negocio.

La Inteligencia de Negocio (BI) y el Big Data (BD) son enfoques clave para la toma de decisiones basada en datos, que tienen el objetivo de mejorar tanto la eficiencia como la competitividad. Aunque ambos están muy relacionados, también presentan diferencias entre sí:

- El BI se centra en **información conocida** y **preguntas ya formuladas**, utilizando **datos estructurados** para generar informes y cuadros de mando.
- Big Data, por otro lado, se enfoca en **descubrir patrones** y **preguntas que no se podían responder antes**, analizando grandes volúmenes de **datos estructurados** y **no estructurados** provenientes de diversas fuentes.

Mientras que ambos enfoques se complementan, BI se apoya en Big Data para ampliar su capacidad de análisis y conseguir obtener información más profunda y estratégica. El BI tradicional maneja Gigabytes y Terabytes de información centralizada, Big Data maneja Petabytes o Exabytes en arquitecturas distribuidas. El foco actual está en datos desestructurados externos como vídeo, texto e imágenes, y las empresas están creando observatorios en redes sociales para analizar las opiniones de los usuarios. El análisis de datos ha evolucionado desde la analítica descriptiva, que resume datos pasados, hacia la analítica predictiva y prescriptiva, que busca predecir escenarios futuros y optimizar decisiones.

6. Busca ofertas de trabajo donde soliciten a científicos de datos e investiga los requisitos que piden.

Primera oferta: Junta de Andalucía

- Requisitos mínimos:
  - Licenciatura o Grado Universitario en Matemáticas, Ciencias Computacionales o Ciencia de Datos o cualquier titulación equivalente a las mismas, reconocidas u homologadas por la Administración Educativa competente en el lugar de contratación.
  - Experiencia de al menos 3 años en ciencia de datos, preferiblemente en el sector salud.
  - Estar en posesión de la documentación reglada para su contratación laboral en España.
  - Aportar la documentación que acredite el cumplimiento de los requisitos anteriormente indicados (títulos académicos y formativos, vida laboral o documentación acreditativa equivalente para aquellas personas que no tengan nacionalidad española y DNI/NIE o equivalente).
- Requisitos valorables:
  - Experiencia en el manejo y análisis de datos de registros electrónicos de pacientes.
  - Experiencia en modelado estadístico y causal.
  - Experiencia coordinando estudiantes de doctorado y máster.
  - Experiencia en sistemas basados en Linux y entornos de altas prestaciones (HPC).
  - Experiencia en modelado mediante Machine Learning Interpretable.
  - Experiencia en investigación demostrable con artículos científicos.
  - Experiencia en comunicaciones orales de resultados científicos.
  - Estancias en centros de investigación en áreas afines.

Segunda oferta: Oferta de Madrid

- Requisitos mínimos: Titulación superior en Estadística, Matemáticas, Ingeniería, Física, Ciencias de la computación, Demostrable 5 años de experiencia en el empleo técnicas y modelos de inteligencia artificial (**Machine Learning**, NLP, Predictive Analytics, Generative AI). Implementación de RAG, Python y manejo en BBDD relacionales, Nivel medio/alto de **inglés**.

Tercera oferta: Oferta de Málaga

- Requisitos mínimos: La persona seleccionada deberá poseer Grado Universitario en Informática, Matemáticas, Estadística, Física, o campos relacionados. Será muy valorable estar en posesión de un Máster en Ciencia de Datos, Inteligencia Artificial o un área afín. También se deberá contar con:
  - Formación en frameworks de IA como TensorFlow, PyTorch, o scikit-learn.
  - Certificaciones en ciencia de datos e IA (ej. Google Professional Data Engineer, AWS Machine Learning).

  Y poseer competencias como:
  - Fuertes habilidades analíticas y de resolución de problemas.
  - Experiencia en machine learning y big data.
  - Capacidad para trabajar en equipo y comunicar resultados técnicos a audiencias no técnicas.
  - Proactividad y enfoque en la innovación educativa.

  Se deberá tener experiencia de al menos 2-3 años en roles relacionados con IA o ciencia de datos, siendo muy valorable experiencia en proyectos aplicados al sector educativo.

7. Realiza 3 búsquedas interesantes en Google utilizando operadores que no conocieras previamente.

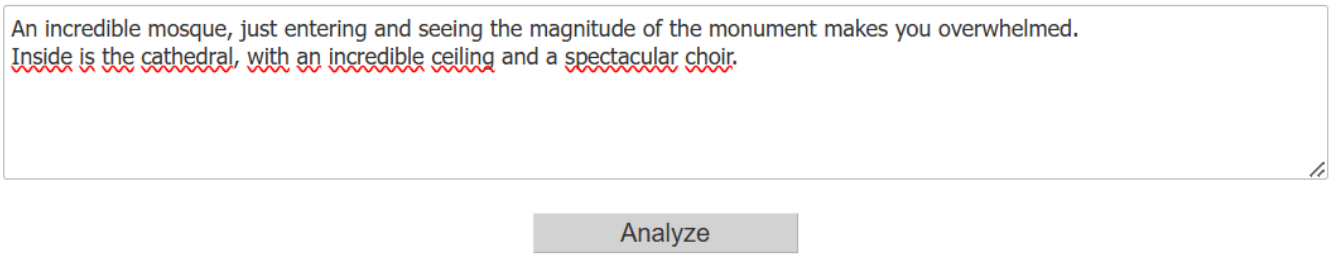
Primera búsqueda: inteligencia artificial before:2015 → Podemos buscar información anterior a 2015 sobre inteligencia artificial.

Segunda búsqueda: intitle:"Big Data" inurl:big data before:2009 → Buscamos títulos que contengan "Big Data" y que en la URL también aparezcan las palabras "big data" y antes de 2009.

Tercera búsqueda: Big Data AND España intitle:big data españa inurl:big data españa → Buscamos información de Big Data y España, que aparezcan estos términos deben aparecer en el título y en la URL.

8. Utiliza una herramienta manual de análisis de sentimientos gratuita y analiza los comentarios que han dejado los usuarios de Google de un restaurante, monumento, local de Córdoba. Muestra si la aplicación indica que los comentarios son positivos o negativos.

Primera herramienta: Sentigem

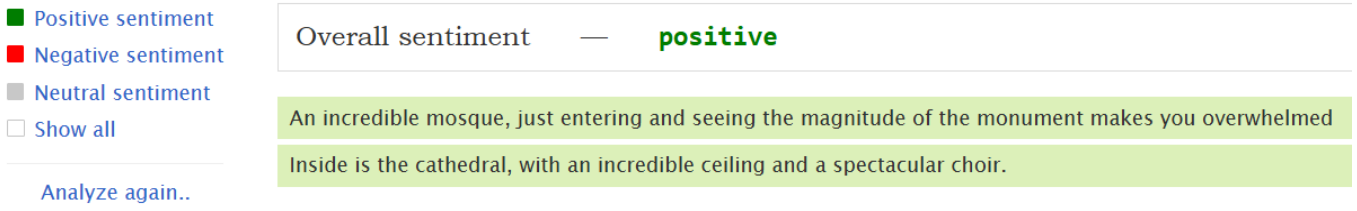


Es una herramienta simple al que se le pasa la reseña que se desea evaluar y esta analizará frase por frase si tiene un sentimiento positivo, negativo o neutro. El único inconveniente que tiene es que solo funciona con textos en inglés. Vamos a probar con reseñas de usuarios que han visitado la Mezquita de Córdoba:

⭐⭐⭐⭐⭐ → ***Positiva***

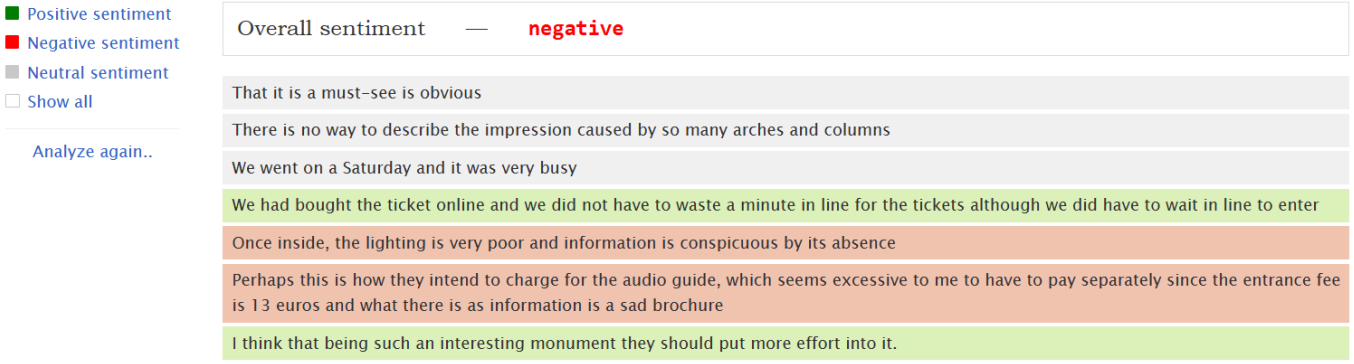
"Una mezquita increíble, solo al entrar y ver la magnitud del monumento hace que te sobrecojas. En el interior está la catedral, con un techo increíble y un coro espectacular."

Si pasamos el texto por la aplicación, muestra lo siguiente:



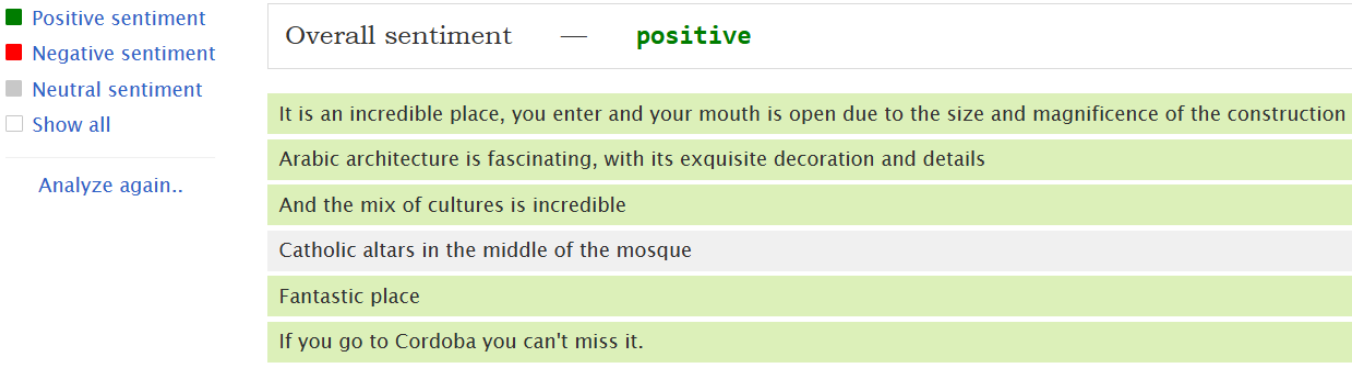
⭐⭐⭐ → *Negativa*

"Que es visita obligada es obvio. No hay manera de describir la impresión que causan tantos arcos y columnas. Nosotros hemos ido en sábado y estaba muy concurrida. Habíamos comprado la entrada por internet y no tuvimos que perder un minuto en cola para los tickets aunque sí tuvimos que hacer cola para entrar. Una vez dentro la iluminación es muy pobre y la información brilla por su ausencia. Quizá así pretendan cobrar el audioguía, que me parece excesivo tener que pagar aparte ya que la entrada son 13 euros y lo que hay cómo información es un triste folleto. Creo que siendo un monumento tan interesante deberían currárselo más."



⭐⭐⭐⭐⭐ → *Positiva*

"Es un lugar increíble, entras y te quedas con la boca abierta por el tamaño y lo magnífico de la construcción. La arquitectura árabe es algo fascinante, con su decoración y detalles exquisitos. Y la mezcla de culturas es increíble. Altares católicos en medio de la mezquita. Fantastico lugar. Si vas a Cordoba no te lo podes perder."



Vemos que la aplicación detecta correctamente los sentimientos expresados en las reseñas, y además te señala aquellas frases que son positivas, negativas o neutras.

Segunda herramienta: Librerías como TextBlob o usar un LLM
Otra alternativa es programar un código que haga el análisis de sentimientos, esto se puede hacer de dos forma posibles:

1. Usando librerías como **TextBlob**
Un ejemplo de código que hace un simple análisis de sentimientos es el siguiente:

```
from textblob import TextBlob
from deep_translator import GoogleTranslator

def analizar_sentimiento(texto):
    text = GoogleTranslator(source='auto', target='en').translate(texto)
    analisis = TextBlob(text)
    if analisis.sentiment.polarity > 0:
        return 'Positivo'
    elif analisis.sentiment.polarity < 0:
        return 'Negativo'
    else:
        return 'Neutro'
texto = input('Ingrese una reseña: ')
sentimiento = analizar_sentimiento(texto)

print("="*35)
print('Sentimiento:', sentimiento)
```

En este código tenemos la función **analizar_sentimiento** que recibe la reseña y en base a la polaridad de los sentimientos analizados en ella devolverá Positivo si dicha polaridad es mayor a 0, Negativo si es menor y Neutro si es 0. Es posible que funcione mejor en inglés, por lo que usaremos la librería **GoogleTranslator** para pasarle el texto en inglés.
Primera reseña:

```
Ingrese una reseña: Una mezquita increíble, solo al entrar y ver la magnitud del monumento hace que te sobrecojas. En el interior está la catedral, con un techo increíble y un coro espectacular.
===================================
Sentimiento: Positivo
```

Segunda reseña:

```
Ingrese una reseña: Que es visita obligada es obvio. No hay manera de describir la impresión que causan tantos arcos y columnas. Nosotros hemos ido en sábado y estaba muy concurrida. Habíamos comprado la entrada por internet y no tuvimos que perder un minuto en cola para los tickets aunque sí tuvimos que hacer cola para entrar. Una vez dentro la iluminación es muy pobre y la información brilla por su ausencia. Quizá así pretendan cobrar el audioguía, que me parece excesivo tener que pagar aparte ya que la entrada son 13 euros y lo que hay cómo información es un triste folleto. Creo que siendo un monumento tan interesante deberían currárselo más.
===================================
Sentimiento: Positivo
```

Tercera reseña:

```
Ingrese una reseña: Es un lugar increíble, entras y te quedas con la boca abierta por el tamaño y
lo magnífico de la construcción. La arquitectura árabe es algo fascinante, con su decoración y
detalles exquisitos. Y la mezcla de culturas es increíble. Altares católicos en medio de la
mezquita. Fantastico lugar. Si vas a Cordoba no te lo podes perder.
==================================
Sentimiento: Positivo
```

Vemos que los resultados son más generales en vez de analizar frase por frase, siendo las reseñas marcadas como positivas.

2. Usando un **LLM**

Un Large Language Model (LLM) es un modelo que ha sido entrenado con muchos textos para comprender muy bien cómo funciona el lenguaje natural. Normalmente son redes neuronales con muchos parámetros como chat GPT, Gemini entre muchos otros. Para ello, estos modelo dejan usar una API que permite conectar estos modelo a tu código. Para este ejemplo, hemos usado un LLM llamado **Cohere** que es muy similar a chat GPT.

```python
import cohere
def analizar_sentimiento_cohere(texto):
    # Configurar la API Key de Cohere
    co = cohere.Client("Mpg8Ne0A2pIS9mFAhgp0NIBOzpz3RYK8TgAbb0FW")  # Reemplaza con tu clave API

    # Prompt directo para clasificar la reseña
    prompt = f"Clasifica si esta reseña es 'Positiva', 'Negativa' o 'Neutra':\n\n{texto}\n\nClasificación:"

    try:
        # Solicitar clasificación a la API de Cohere
        response = co.generate(
            model="command",  # Modelo adecuado para tareas de clasificación e instrucciones
            prompt=prompt,
            max_tokens=4,  # Ajusta según sea necesario
            temperature=0.0  # Hacer la respuesta determinista
        )

        # Extraer la clasificación
        clasificacion = response.generations[0].text.strip()

        # Mostrar la clasificación al usuario
        print(f"\nAnálisis de la reseña: {clasificacion}")

    except Exception as e:
        print(f"Error al analizar la reseña: {e}")

texto_a_analizar = input('Ingrese una reseña: ')
analizar_sentimiento_cohere(texto_a_analizar)
```

El código anterior crea una función **analizar_sentimiento_cohere** que usa la API de Cohere y recibe una respuesta del modelo que luego es mostrada en pantalla. Como es un LLM, puede entender varios idiomas y no es necesario tener que traducir las reseñas. Los resultados son los siguientes:

Primera reseña:

```
Ingrese una reseña: Una mezquita increíble, solo al entrar y ver la magnitud del monumento hace que te sobrecojas. En el interior está la catedral, con un techo increíble y un coro espectacular.

Análisis de la reseña: Positiva
```

Segunda reseña:

```
Ingrese una reseña: Que es visita obligada es obvio. No hay manera de describir la impresión que causan tantos arcos y columnas. Nosotros hemos ido en sábado y estaba muy concurrida. Habíamos comprado la entrada por internet y no tuvimos que perder un minuto en cola para los tickets aunque sí tuvimos que hacer cola para entrar. Una vez dentro la iluminación es muy pobre y la información brilla por su ausencia. Quizá así pretendan cobrar el audioguía, que me parece excesivo tener que pagar aparte ya que la entrada son 13 euros y lo que hay cómo información es un triste folleto. Creo que siendo un monumento tan interesante deberían currárselo más.

Análisis de la reseña: Neutra
```

Tercera reseña:

```
Ingrese una reseña: Es un lugar increíble, entras y te quedas con la boca abierta por el tamaño y lo magnífico de la construcción. La arquitectura árabe es algo fascinante, con su decoración y detalles exquisitos. Y la mezcla de culturas es increíble. Altares católicos en medio de la mezquita. Fantastico lugar. Si vas a Cordoba no te lo podes perder.

Análisis de la reseña: Positiva
```

En este caso, la segunda reseña fue catalogada como “Neutra” ya que la red neuronal ha detectado frases como “la iluminación es muy pobre” o “la información brilla por su ausencia” que tienen un significado negativo. Pero otras partes de la reseña son positivas, lo que hace que sea catalogada como “Neutra”.