TEMA 0

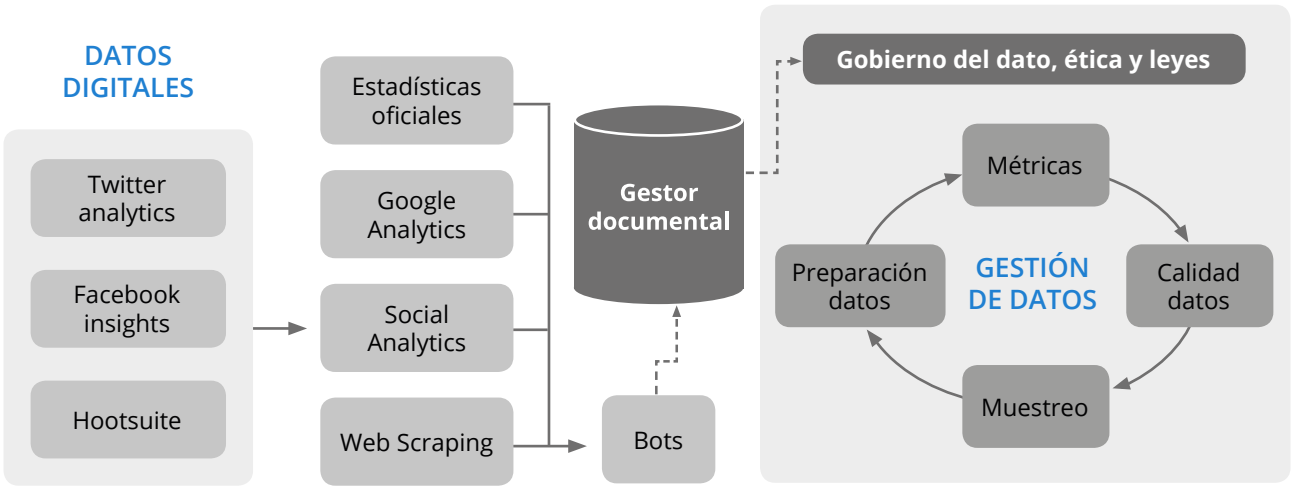
MÓDULO:
DATA MANAGEMENT & DATA DIGITAL

INTRODUCCIÓN AL MÓDULO

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

1. MIND MAP

Para tener una visión general del tema, se presenta el siguiente esquema global:



Fuente: Elaboración propia a partir de imágenes con licencia Pixabay.

El módulo se estructura en dos temas:

1. **Datos digitales o Data digital.**
2. **Gestión de datos o Data Management.**

OBJETIVOS

- Conocer las fuentes principales de datos digitales.
- Obtener datos digitales mediante herramientas analíticas.
- Conocer los principios básicos del gobierno del dato, ética y leyes.
- Preparar los datos, medir la información que aportan y mejorar su calidad.
- Aplicar técnicas de muestreo para el posterior modelado de los datos.

1. DATA MANAGEMENT & DATA DIGITAL

Los datos digitales son una fuente de información muy valiosa. Es interesante esta cita realizada por Joris Toonders en la revista Wired en 2014:

CITA

Los datos son el nuevo petróleo de la economía digital.

Yonego, Joris Toonders (23 de julio de 2014).
Data Is the New Oil of the Digital Economy. Wired

La cita da una imagen muy acertada de la importancia que están teniendo los datos en la nueva era digital. Posteriormente, en 2017, se reproducía una cita similar en "The Economist":

CITA

El recurso más valioso del mundo ya no es el petróleo, sino los datos.

(Mayo del 2017).
The world's most valuable resource is no longer oil, but data. The Economist

Esta segunda cita, confirmaba ya el predominio de las empresas basadas en datos como: Google, Amazon, Apple, Facebook y Microsoft, las 5 empresas mejor valoradas del mundo en 2017.

¿CUÁNTOS DATOS SE GENERAN CADA DÍA EN 2020?

1. 1,7MB de datos se crean cada segundo por persona en 2020. *(Fuente: Domo).*

   Esto supone 2,5 exabytes (10^18 bytes) cada día, en todo el mundo.

2. En los últimos 2 años, se ha generado el 90% de los datos creados por toda la humanidad. *(Fuente: IORG).*

   El impulso de las tecnologías móviles ha dado un gran empuje al crecimiento de los datos.

3. 5 millones de tweets. *(Fuente: Internet Live Stats).*

   Obviamente, muchos menos que los 306.000 millones de emails.

4. 350 millones de fotos se suben a Facebook. *(Fuente: Omni Core Agency).*

   ¿Quién decía que ya se había acabado Facebook...?

5. 4.570 millones de usuarios activos en Internet en todo el mundo. *(Fuente: Statista).*

   Aunque ya sabemos que no todos los datos son generados por personas.

6. 31.000 millones de dispositivos IoT (Internet de las cosas). *(Fuente : Statista).*

La aportación de valor del módulo 3, es la de conocer la existencia de estas fuentes de datos digitales, saber extraer estos datos, como si de petróleo se tratara, y procesarlos, mediante herramientas analíticas, para obtener el máximo valor que nos puedan aportar.

Después de la extracción, se aprenderán técnicas para su depuración y uso, es decir, transformaremos ese petróleo en gasolina. Dando valor y calidad, además de aportar ética en la gestión de los datos.

1. DATOS DIGITALES

El primer tema, aporta conocimientos sobre qué datos se pueden encontrar en la red, explica las distintas estrategias para obtener estos datos y muestra, mediante ejemplos, cómo se obtienen con Python.

Los orígenes de los datos digitales que se explican son:

- **Fuentes externas oficiales:** como ejemplos se analizan los datos del Instituto Nacional de Estadística (INE), la oficina estadística de la Comisión Europea (EUROSTAT) y los datos abiertos del Banco Mundial (WORLD DATA BANK).

  En este apartado, se explican los datos más relevantes qué contiene cada oficina estadística. Además, se trabaja, mediante ejemplos, su descarga mediante Python.

- **Google Analytics:** esta herramienta de analítica web de la empresa Google permite, al administrador del sitio web, obtener infinidad de estadísticas del tráfico, facilitando, por ejemplo, captar nuevo tráfico, optimizar la navegación, fidelizar a los clientes, etc.

  Se explica qué datos contiene la herramienta y cómo obtener estos datos mediante Python, facilitando así su análisis.

- **Social Analytics:** como ejemplos, se comentan los informes de Twitter analytics, Facebook insights y Hootsuite Analytics.

  Se trabajan ejemplos de obtención de datos mediante Python a través de sus APIs.

- **Web scraping:** este es el último bloque de datos digitales y consiste en la obtención de datos directamente de la web.

  Se explica cómo diagnosticar el tipo de datos que contiene una web, estructurados o no, datos estáticos/dinámicos, así como las distintas estrategias y herramientas basadas en Python para obtener estos datos.

Otro punto importante aquí es la introducción de las técnicas para procesar texto mediante expresiones regulares. Esta técnica permitirá estructurar la información de una web de forma simple y rápida.

También, se explican algunas dificultades técnicas asociadas a la obtención de estos datos de forma regular.

Por último, se explica un ejemplo sencillo de bot para descargar los datos y almacenarlos mediante un gestor documental. En este caso, se utiliza MongoDB, aplicación que seguro ya has trabajado en el tema de herramientas Big Data del módulo 1.

2. GESTIÓN DE DATOS

La gestión de datos se representa en el diagrama como un ciclo en el sentido que se realiza a partir de ir incorporando tratamientos nuevos que van mejorando la calidad del dato. De esta forma se obtienen mejores métricas que permiten detectar donde hay más o menos información.

Se explican los tratamientos de datos habituales mediante ejemplos de código en R, para estructurar, transformar y codificar los datos necesarios para iniciar cualquier tarea de análisis de datos.

Por otro lado, se dan las métricas generales, tanto supervisadas (cuando hay una variable objetivo definida), como no supervisadas (basadas en la propia distribución de las variables), para poder decidir si una variable aporta más o menos información. Así mismo, estas métricas se utilizan como criterio para aplicar mejoras en el tratamiento de las variables mediante R.

En el apartado de calidad de datos, se profundiza, con el uso de R, en las técnicas avanzadas para detectar valores extremos, así como, analizar e imputar los valores faltantes.

El muestreo se explica como un paso necesario para evaluar si las mejoras obtenidas en las métricas se pueden generalizar al conjunto de la población, o bien, son sólo un caso particular de la muestra de datos utilizada.

Finalmente, el ciclo de la gestión de datos, se enmarca como un proceso desarrollado bajo unos criterios para el gobierno del dato, comunes para todos los miembros de una organización. También, se aportan elementos que faciliten saber que se está cumpliendo con criterios éticos de equidad y justicia en el tratamiento de los datos. Por último, se comentan las principales leyes que regulan la protección de datos en el uso de estos, en el ámbito de los datos de personas físicas.

EVALUACIÓN

Evaluación continua del trabajo realizado en clase mediante la resolución de 3 partes:

1. Prueba **teórica:** al superar los **test** con éxito se alcanzará la posición de ***Initiate Level.***
2. Prueba **individual:** al superar el **trabajo individual** se logrará la posición ***Padawan Level.***
3. Prueba **grupal:** superar el **trabajo colectivo** supondrá conseguir la posición ***Knight Level .***

CRITERIOS MÍNIMOS

El *alumno/a Padawan* para alcanzar el nivel debe superar con éxito los siguientes hitos:

- ***Initiate Level:*** prueba de asentamiento de conceptos teóricos, para superar está parte deberás obtener una calificación superior a 5.

  Nota: Las preguntas que no se contesten de forma correcta restará puntos (indicado en cada actividad).

- ***Padawan Level:*** Realizar, al menos una práctica individual, defendiéndola y justificándola adecuadamente.

- ***Knight Level:*** Realizar al menos una práctica colectiva (participación activa en reuniones y discusiones de grupo, así como en la elaboración de informes, etc.), defendiéndola y justificándola adecuadamente.

Los porcentajes de cada hito estarán reflejados en el plan docente y en cada actividad.

Para aprobar el módulo, la media de todos los hitos debe ser superior al 5.

Recuerda que es evaluación continua por lo que cuantas más prácticas realices más posibilidades tendrás de alcanzar el máximo nivel Padawan.

Se recomienda aprovechar todos los recursos disponibles online:

- Materiales teóricos en PDF.
- Vídeos didácticos: dos para cada tema.
- Notebooks de acompañamiento:
  - Rmarkdown en el caso de R.
  - Colab en el caso de Python.

EJECUCIÓN INTERACTIVA

- **Datos digitales**

Puedes ejecutar el temario de forma interactiva accediendo a Colab o Jupyter y RStudio. El código fuente de los materiales se puede descargar desde el aula a tu directorio git local.

  - Introducción.
  - Fuentes externas oficiales.
  - Google Analytics.
  - Social Analytics.
  - Web scraping.
  - Anexo: README de Datos digitales.

- **Gestión de datos**

  - Introducción.
  - Gobierno. Ética y leyes.
  - Preparación de los datos.
  - Métricas.
  - Calidad de los datos.
  - Anexo: README de Gestión de datos.

BIBLIOGRAFÍA

AL SWEIGART. Automate the Boring Stuff with Python. No Starch Press, 2nd Ed; Noviembre 2019. Disponible en: https://automatetheboringstuff.com/

Libro de propósito general. Muestra, de forma muy didáctica y sin asumir conocimientos previos, cómo automatizar tareas con Python: por ejemplo, web Scraping, manipulación de texto o calendarizar tareas.

S. V. BROUCKE, B. BAESENS. Practical Web Scraping for Data Science. Apress.; 2018. Disponible el código en: https://github.com/Apress/practical-web-scraping-for-data-science

Mejores prácticas y ejemplos con Python.

R. MITCHELL. Web Scraping with Python. O'Reilly Media, Inc. 2nd ed.; 2018. Disponible el código en: https://github.com/REMitchell/python-scraping

Guía exhaustiva de web scraping con Python. Cubre todas las modalidades de web scraping.

G. GROLEMUND, H. WICKHAM. R for Data Science. O'Reilly; 2017. Disponible en: https://es.r4ds.hadley.nz/ (Castellano)

Aprender a cargar datos en R, escoger la estructura de datos óptima, transformarlos, visualizarlos y modelarlos.

G. JAMES, D. WITTEN, T. HASTIE, R. TIBSHIRANI. An Introduction to Statistical Learning with applications in R. Springer; 2017. Disponible en: https://www.statlearning.com/

Aporta las métricas más relevantes para medir la calidad de la información, explica los principales problemas de calidad de la información y aporta las técnicas de muestreo más frecuentes.

RECURSOS EN INTERNET

- Cuántos datos se crean cada día en 2020.Disponible en: https://techjury.net/blog/how-much-data-is-created-every-day/#gref
- Cómo usar Python y Selenium para hacer web scraping. Disponible en: https://thenextweb.com/syndication/2020/07/22/how-to-use-python-and-selenium-to-scrape-websites/
- Automatizar Web Scraping. Disponible en: https://analyticsindiamag.com/autoscraper-tutorial-a-python-tool-for-automating-web-scraping/
- Guía paso a paso Web Scraping con Python. Disponible en: https://towardsdatascience.com/a-step-by-step-guide-to-web-scraping-in-python-5c4d9cef76e8
- Cómo aprender a no preocuparse y querer el web scraping. Disponible en: https://www.nature.com/articles/d41586-020-02558-0

ANEXO: README

Copia del repositorio Github

Para editar y conservar tu código en github, te recomendamos hacer FORK del repositorio en tu Github.

Para hacer FORK:

- Accede a https://github.com/griu/mbdds_fc20.git
- Introduce tu usuario y contraseña y haz clic en el botón de FORK.

A partir de este momento, siempre que tengas que clonar el repositorio, tú eliges si trabajar con el común, o bien, con tu propio repositorio:

- https://github.com/griu/mbdds_fc20.git
- https://github.com/TU_USUARIO/mbdds_fc20.git

Revisa los **Readme de R** y **Readme de Python** para completar la plataforma RSTUDIO-JUPYTER-COLAB.