TEMA 4

MÓDULO:
HERRAMIENTAS DE BIG DATA

HERRAMIENTAS DE ANÁLISIS

INTRODUCCIÓN AL BLOQUE II. PROGRAMACIÓN EN R Y PROGRAMACIÓN EN PYTHON

FERRAN CARRASCOSA

Licenciado en Matemáticas por la UB
Data Scientist

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Realizar operaciones de lectura y escritura de datos.
- Saber escoger la estructura de datos adecuada para cada problema.
- Desarrollar pequeñas piezas de código en R y Python.

INTRODUCCIÓN

R

Es un lenguaje y entorno utilizado para la computación estadística.

SABÍAS QUE...

Es un programa de código abierto, desarrollado por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en 1993 como un "clon" del lenguaje S desarrollado por los Laboratories de AT&T en Nueva Jersey en 1970.

PYTHON

Es un lenguaje interpretado, de alto nivel y de propósito general.

SABÍAS QUE...

Es un programa de código abierto, creado en 1989 por Guido van Rossum en el Centro de Matemáticas e Informática (CWI, Centrum Wiskunde & Informatica), en los Países Bajos.

Con el objetivo de valorar su relevancia, a continuación, se analiza su posición en dos de los rankings de referencia en lenguajes de programación:

- **TIOBE ÍNDICE:** el ranking TIOBE Índice de Agosto 2020 ordena por popularidad los lenguajes de programación mediante las búsquedas realizadas en los principales motores: Google, Baidu, Yahoo, Wikipedia, etc.

- **Ranking IEEE:** el ranking "Top programming languages 2020" de la IEEE se construye combinando 11 métricas de 8 fuentes distintas: *CareerBuilder, GitHub, Google, Hacker News, IEEE, Reddit, Stack Overflow y Twitter.*

TIOBE Index	Lenguaje	Rating
1	C	16,98%
2	Java	14,43%
3	**Python**	9,69%
4	C++	6,84%
5	C#	4,68%
6	Visual Basic	4,66%
7	JavaScript	2,87%
8	**R**	2,79%
9	PHP	2,24%
10	SQL	1,46%

Ranking IEEE	Lenguaje	Puntuación
1	**Python**	100,0
2	Java	95,3
3	C	94,6
4	C++	87,0
5	JavaScript	79,5
6	**R**	78,6
7	Arduino	73,2
8	Go	73,1
9	Swift	70,5
10	Matlab	68,4

Fuente (izquierda): Elaborada a partir de TIOBE. Fuente (derecha): Elaborada a partir de IEEE.

En el *TIOBE Índice*, Python se encuentra en tercera posición y R en octava.

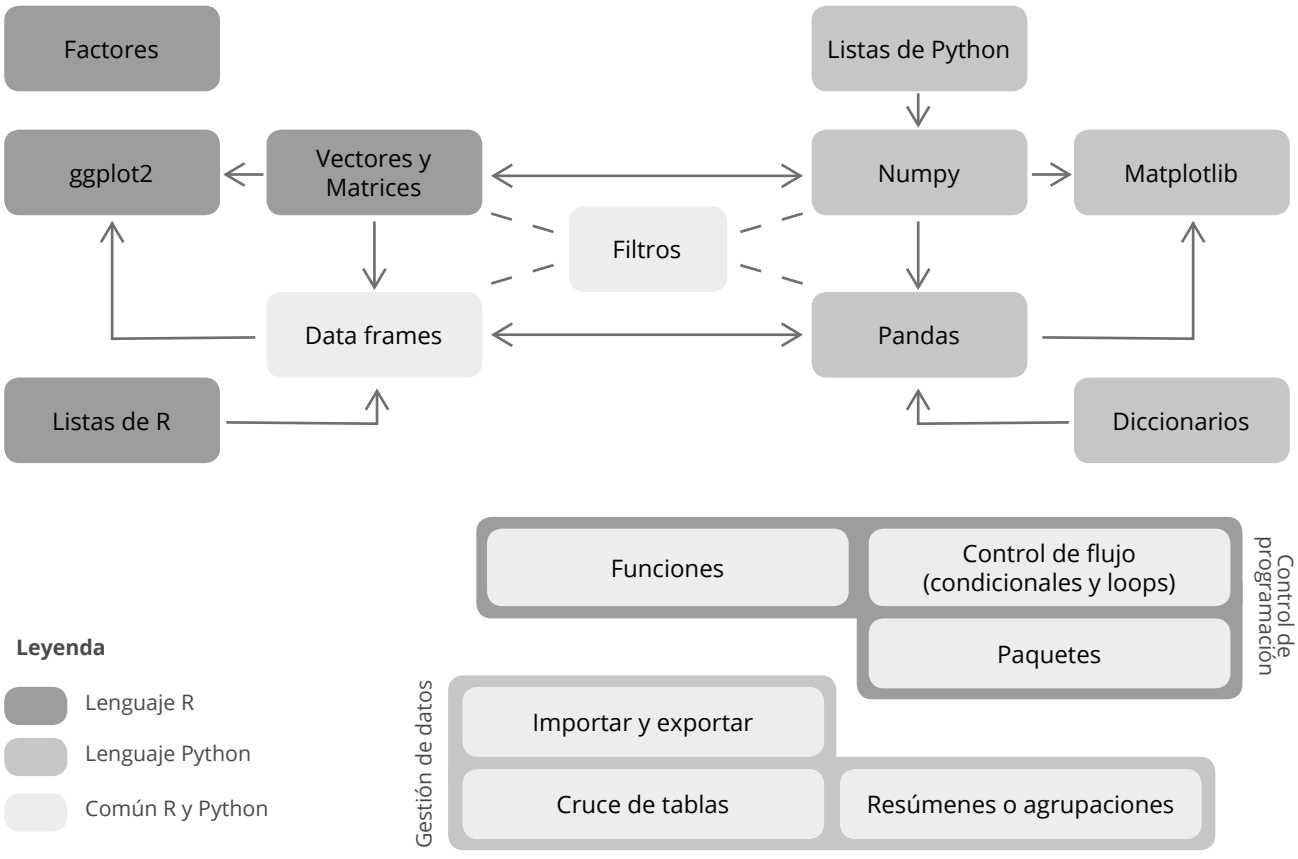
En el ranking de la *IEEE*, ambos aparecen de nuevo dentro del top 10, Python como el lenguaje ganador y R ocupando la sexta posición, a pesar de ser un lenguaje de nicho (análisis de datos).

Viendo ambos rankings, se puede deducir que tanto R como Python son líderes en la categoría de lenguajes con capacidades analíticas avanzadas.

Los contenidos de este tema permiten comprender los motivos de esta popularidad y profundizar en el análisis de datos mediante R y Python.

MINDMAP

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



Fuente: Elaboración propia.

Los **tipos de datos** pueden ser numéricos, de texto, fechas y factores (en el caso de R).

Para poder agrupar estos datos básicos en colecciones o listas, se presentan las siguientes estructuras:

- **Listas de Python:** colección ordenada y mutable.

Por ejemplo:

['elefante','león','ratón'].

- **Tuplas de Python:** colección ordenada e inmutable.

Por ejemplo:

(0,1,2,3,4,5,6,7,8,9).

- **Diccionarios de Python:** colección indexada por clave y mutable.

Por ejemplo:

{'nombre': 'Juan', 'Edad': 25, 'dirección': {'calle': 'París', 'número': 33}}.

- **Listas de R:** combinas listas de Python (ordenadas y mutables) y diccionarios (indexados).

PIENSA UN MINUTO

Python dispone de muchísimas formas para agrupar los datos. Esto le da mucha agilidad para procesar datos no estructurados: datos web, datos de texto, imágenes, etc.

Los vectores y matrices existen de forma nativa en R. En el caso de Python se implementaron en la librería numpy en 2002:

- **Vectores:** lista de elementos homogénea (de un mismo tipo de datos).
- **Matrices:** vectores con estructura de 2 dimensiones (filas y columnas) o más.

SABÍAS QUE...

R está basado en LISP, un lenguaje de programación orientado a listas.
Originalmente especificado en 1958, LISP fue clave para el desarrollo de la inteligencia artificial.

Los **Data frames** también existen de forma nativa en R. Python los implementó en 2009 en la librería pandas. Se estructuran como una lista de vectores columna de una misma longitud, es decir, similar a una matriz de 2 dimensiones, donde cada columna puede tener un tipo de datos distinto (numérico, carácter, etc.).

IMPORTANTE

Los Data frames se organizan en columnas de vectores, en cambio, las bases de datos tradicionales se organizan en filas.

Los Data frames son más rápidos en tareas de análisis frente a las Bases con mejor desempeño en gestión de registros o filas.

Los **filtros** permiten seleccionar filas y columnas de Matrices y Data frames.

Respecto a los gráficos en R, introduciremos ggplot2 que se utilizará dentro del máster en los módulos 3, 4 y 5.

En Python se introduce matplotlib que se utilizará en los módulos 2, 6 y 7.

SABÍAS QUE...

Ggplot2 está desarrollada por Hadley Wickham en 2005.
Por su parte, matplotlib está desarrollada por John Hunter en 2003 y se basa en el sistema de gráficos del software MATLAB.

Las capacidades de control de la programación consisten en:

- **Funciones:** definidas a partir de un nombre y unos parámetros de entrada y salida.
- Control de flujo mediante **condicionales** (if) y **bucles o loops** (for).
- **Librerías de funciones** o paquetes: conjunto de funciones con un objetivo común.

La gestión de datos consiste en:

- **Importación y exportaciones** de formatos propios, texto u otros: bases de datos, MS Excel, etc.
- **Cruce entre tablas** mediante campos clave comunes.

  Por ejemplo: inner join, left join, etc.

- Operaciones con *Data frames* de **agrupación** a partir de una o varias variables categóricas.
- Construcción de **tablas resumen** de datos.

  Por ejemplo: conteos, medias, etc.

EVALUACIÓN. HERRAMIENTAS DE ANÁLISIS

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.

RECURSOS DISPONIBLES

Se recomienda aprovechar todos los recursos disponibles online:

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

EJECUCIÓN INTERACTIVA DE LA PROGRAMACIÓN EN R

Puedes ejecutar el temario de R, de forma interactiva, accediendo con RStudio a tu directorio git local, o bien, consultar los códigos fuente en Github:

- Introducción.
- Elementos básicos de R.

- Gráficos con ggplot2.
- Colecciones de objetos.
- Control de flujo.
- Gestión de datos.
- Anexo: README de R.

EJECUCIÓN INTERACTIVA DE LA PROGRAMACIÓN EN PYTHON

El temario de Python se puede ejecutar desde Colab, accediendo directamente desde el enlace que encontrarás en cada notebook.

Aunque no es imprescindible, también puedes usar el entorno de Jupyter local tal cual se explica en la guía en el Anexo: README de Python.

Puedes ejecutar de forma interactiva los materiales de Python en Colab en el enlace que encontrarás en la cabecera de los siguientes notebooks:

- Introducción.
- Elementos básicos de Python.
- Gráficos con matplotlib.
- Colecciones de objetos:
  - Listas, tuplas y diccionarios.
  - Numpy.
  - Pandas.
- Control de flujo.
- Gestión de datos.
- Anexo: README de Python.

IDEAS CLAVE

- R y Python cubren todo el espectro de trabajo con datos.
- Ambos son programas con un **entorno de software libre** (licencia GNU GLP) y funcionan en plataformas **UNIX** o similares como **Linux, Windows y MacOS.**
- Forman parte de un **proyecto colaborativo y abierto** donde los usuarios/as pueden publicar paquetes que extienden su configuración básica (repositorio oficial de paquetes).

Se puede **descargar gratis** a través de los siguientes enlaces:

- Para R:

  https://www.r-project.org/

- Para Python:

  https://www.python.org/downloads/

BIBLIOGRAFÍA

- Y. Xie, J. J. Allaire, G. Grolemund. R Markdown: The Definitive Guide. Chapman & Hall/CRC; 2020. Disponible en: https://bookdown.org/yihui/rmarkdown/

Guía para elaborar con Rmarkdown, documentos que combinan texto, análisis en R o Python con los resultados (tablas o gráficos).

- W. Chang. R Graphics Cookbook, O'Reilly Media, Inc. 2nd ed.; 2020. Disponible en: https://r-graphics.org/

Desarrollo de gráficos con R.

- W. McKinney. Python for Data Anlysis. O'Reilly Media, Inc. 2nd ed.; 2017

Profundizar en el uso de Pandas, NumPy y Matplotlib para el análisis y preparación de datos.

- 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.

- J. VanderPlas. Python Data Science Handbook. O'Reilly Media, Inc.; 2016. Disponible en: http://faculty.marshall.usc.edu/gareth-james/ISL/

Profundizar en el uso de Pandas, Numpy y Machine Learning.

- G. Grolemund. Hands-On Programming with R. O'Reilly; 2014. Disponible en: https://rstudio-education.github.io/hopr/

Aprender a programar en R básico mediante ejemplos prácticos.

- C. R. Severance. Python para todos. Elliott Hauser, Sue Blumenberg; 2009. Disponible en: https://www.py4e.com/book.php

Cubre todo lo básico de Python más la interacción con bases de datos, visualización y modelado.

- R. Kabacoff. Data Visualization with R. 2018. Disponible en: https://rkabacoff.github.io/datavis/Models.html#SaratogaHouses

Visualización de datos con R.

RECURSOS EN INTERNET

- Cheatsheet de R.
- R para Ciencia de Datos.
- Cheatsheet de Python.
- Tutorial de Python.

ANEXO: README

COPIA DEL REPOSITORIO GITHUB (OPCIONAL)

Para editar y conservar tu código en github, se recomienda, hacer FORK del repositorio en tu GITHUB.

Para hacer FORK, realiza los siguientes pasos:

- Accede a https://github.com/griu/mbdds_fc20.git.
- Introduce tu usuario y contraseña y clica 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.