Mónica Romero Ferrón

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.**

Se ha decidido buscar información sobre tenistas femeninas profesionales. En concreto se quiere estudiar a las tenistas que han sido número uno a lo largo de estos años y obtener información sobre ellas como por ejemplo semanas que han sido número uno del mundo, semanas seguidas cronológicamente, además se puede obtener información por países número uno del mundo en tenis femenino y Grand Slam.

Este tipo de información no se recoge de manera tan detallada y con histórico en su página web oficial wtatennis.com ya que aquí se recogen datos actuales. Buscando información por internet sobre información de la historia del tenis femenino hay un anexo en Wikipedia donde se ha ido recogiendo esta información en tablas con los datos ya limpios y preparados para hacer un análisis de ellos. He elegido esta página por toda la información que contiene y la simplicidad de extracción de tablas ya que otras webs no estaban estructurada la información y sin organización.

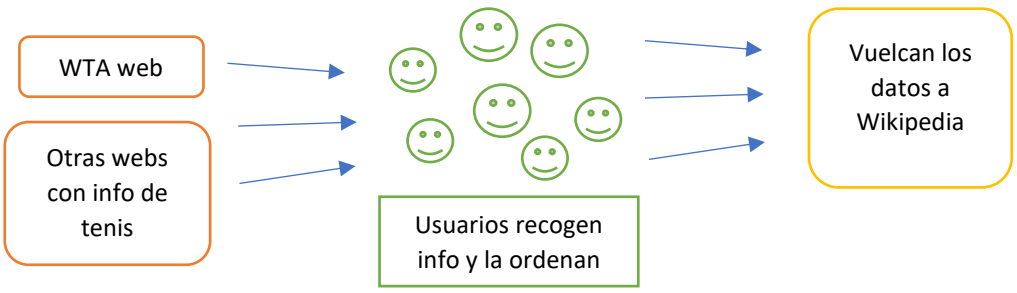
2. **Definir un título para el dataset. Elegir un título que sea descriptivo.**

Información sobre tenistas femeninas número uno en el ranking WTA.

3. **Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

Como se dice en el título elegido en el apartado 2 este dataset está basado en datos sobre mujeres tenistas que han sido número uno en algún momento de la historia información sobre las tenistas, el periodo (fecha de comienzo y fin de estar en el ranking como la primera) y tiempo que estuvieron siendo número uno del mundo también a nivel de país en que años y tiempo fueron número uno sus tenistas femeninas. Estos datos vienen estructurados, pero aun hace falta una pequeña limpieza de ellos para poder tratarlos en los posteriores análisis.

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente**



5. **Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

Las métricas que aparecen en las tablas son:

- Nombre de la jugadora (string).

- Su fecha de inicio y fin de aparición número uno en el ranking esta métrica hay que tener cuidado ya que no son fechas en un formato date si no strings del tipo “3 de Noviembre de 1975”.

- Conteo de semanas tanto consecutivas como semanas totales (integer).

- País al que pertenecen las jugadoras (string).

- Nº de tenistas que acumulan el número 1 en el ranking (integer)

- Año en la que fueron número uno (string)

- Nombre del Grand Slam (string).

El periodo de tiempo que estamos trabajando es desde 1975 que es la fecha donde data la primera tenista hasta el 1 de noviembre del 2020 que es cuando se han actualizado por última vez estos datos.

Estos datos fueron recogidos mediante web scraping en un notebook con lenguaje Python sobre la página de Wikipedia donde apareceré todas estas tablas con esta información. Tras esto se guardan estas tablas cada una un csv.

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).**

Agradecer a las personas que contribuyen en Wikipedia que siempre es de carácter anónimo. Para poder contactar se puede hacer mediante su correo info-es@wikipedia.org. Hay posibilidad de realizar web scraping siempre y cuando no se descarguen artículos completos. Se puede analizar el archivo robot.txt donde se ve que no está bloqueado ningún tipo de permisos y se ha leído información sobre esto en https://moz.com/learn/seo/robotstxt

7. **Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.**

El interés sobre este tipo de análisis viene dado por mi interés en el análisis de datos acerca de deportes. En concreto estos datos son interesantes para estudiar como en el paso del tiempo las tenistas número uno duraba más semanas que actualmente que es más volátil además de ser datos actualizados se puede realizar un estudio cada cierto poco tiempo de los datos actualizados de las tenistas número uno. Además, responde preguntas como ¿Cuál es la tenista número uno actualmente y cuantas semanas (consecutivas o no) lleva siéndolo? ¿Cuántas semanas y quien fue la tenista número uno que más tiempo estuvo? ¿Cuál es el país que más semanas seguidas estuvo número uno del mundo? ¿En XXXX año (por ejemplo, 2019) quien fue número uno?

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

La licencia elegida seria *Released Under CC BY-NC-SA 4.0 License* ya que se puede copiar y redistribuir en cualquier formato donde le nombre del creador del dataset estará y se debe comentar cualquier cambio que se realice así se puede tanto ver el trabajo realizado por el creador que seriamos nosotros y el que hagan ellos. No se permitirá un uso comercial ya que este tipo de datos son datos de carácter abierto por Wikipedia y no se podría hacer acopio de ellos para fines comerciales. Y por último se tiene que distribuir bajo el mismo tipo de licencia que estamos usando y así permitir que el autor de este dataset sea siempre reconocido.

9. **Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.**

```
import requests
from selenium import webdriver
import os
from bs4 import BeautifulSoup
import urllib.request
import requests
import pandas as pd


#leemos el codigo robot.txt creamos una funcion
def robot_txt(url):
    if url.endswith('/'):
        path = url
    else:
        path = url + '/'
    req = requests.get(path + "robots.txt", data=None)
    return req.text
print(robot_txt("https://es.wikipedia.org"))

#elegimos la opcion del webdriver como abrirlo en modo incognito y
usar el academic crawler
opcion = webdriver.ChromeOptions()
opcion.add_argument(" — incognito")
opcion.add_argument("user-agent=AcademicCrawler")
#archivo path
My_path = os.path.dirname(os.path.abspath("__file__"))

browser = webdriver.Chrome(executable_path=My_path +'/chromedriver')

url =
"https://es.wikipedia.org/wiki/Anexo:Tenistas_n%C3%BAmero_1_de_la_WTA"
browser.get(url)
```

```
#revisamos el user agent
agent = browser.execute_script("return navigator.userAgent")
print(agent)

#buscamos dentro del enlace las tablas que queremos descargarnos
page = requests.get(url)

soup = BeautifulSoup(page.content, "lxml")

table = soup.find('table',{'class':'wikitable'})
links = table.findAll('a')
links

table1 = pd.read_html(url,header=0)[0]
table1.head()
table2 = pd.read_html(url,header=0)[2]
table2.head()
table3 = pd.read_html(url,header=0)[4]
table3.head()
table4 = pd.read_html(url,header=0)[6]
table4.head()
table5 = pd.read_html(url,header=0)[7]
table5.head()
table6 = pd.read_html(url,header=0)[8]
table6.head()
table7 = pd.read_html(url,header=0)[9]
table7.head()
table8 = pd.read_html(url,header=0)[10]
table8.head()

#guardamos cada tabla en un csv distinto
table1.to_csv('table_1.csv', index=False)
table2.to_csv('table_2.csv', index=False)
table3.to_csv('table_3.csv', index=False)
table4.to_csv('table_4.csv', index=False)
table5.to_csv('table_5.csv', index=False)
table6.to_csv('table_6.csv', index=False)
table7.to_csv('table_7.csv', index=False)
table8.to_csv('table_8.csv', index=False)
```

10. **Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.**

Se encuentra en Github. https://github.com/monirome/PRAC1

En Zenodo: https://zenodo.org/record/4264752#.X6ldTGhKg2w

DOI : 10.5281/zenodo.4264752