# Projeto PIBIC

2025-03-13

## Tarefas estabelecidas em 14/out/2024:

### Escrever as seguintes funções Python em Jupyter Notebook no Jupyterlab:

Dado um arquivo Pandas DataFrame de nome "Studies" com as seguintes colunas, todas elas de formato string, podendo estar vazias: "StudyId", "Authors", "Title", "Reference", "Year_Pub", escrever as seguintes funções, aplicadas às linhas deste DataFrame:

1. Procura por trechos em "Reference" que pareçam ser um link DOI (p.ex., https://doi.org/10.3390/f11030326), copia e coloca numa coluna "doi_link"
2. Procura por trechos em "Reference" de "Studies" que pareçam ser um código DOI, sem o http (p.ex., doi:10.2134/agronj14.0597), e, se achar, copia e coloca numa coluna "doi_code".
3. Procura por trechos em "Reference" que pareçam ser um link que não seja DOI, e, se achar, copia e coloca numa coluna "URL".
4. Busca na web o artigo informado pela coluna "doi_link", indicando se o artigo foi encontrado com sucesso e extrai novos valores para "Authors_found", "Title_found", "Reference_found", e "Year_Pub_found"
5. Idem para "doi_code"

6. Idem para “URL”
7. Idem para busca usando qualquer combinação dos campos originais de “Studies”.

### Dicas:

1. Desenvolva as funções acima na ordem em que estão listadas, pois vão da mais simples para a mais difícil, e as primeiras provavelmente ajudam nas seguintes.
2. Teste bem cada uma antes de passar para a próxima.
3. Cuidado com o caps (maiúscula se minúsculas). Talvez o mais simples seja converter tudo para minúsculas logo no início.
4. Cuidado com “/”, “ ” e caracteres especiais em geral.
5. Não tente inventar nada que já exista: Comece pedindo ajuda ao Gemini, ChatGPT , StackExchange, colega, professor, livro, quem você quiser.

## Solução:

```
import pandas as pd
import re
import requests
from bs4 import BeautifulSoup
```

```
def extract_doi_link(reference):
  doi_link_pattern = r'https?://doi\.org/10\.\S+'
  match = re.search(doi_link_pattern, reference)
  return match.group(0) if match else None
```

```
def extract_doi_code(reference):
  doi_code_pattern = r'doi:10\.\S+'
  match = re.search(doi_code_pattern, reference)
  return match.group(0) if match else None
```

```
def extract_url(reference):
  url_pattern = r'https?://(?:(?!doi\.org).)+\S+'
  match = re.search(url_pattern, reference)
  return match.group(0) if match else None
```

```
def search_article_by_doi_link(doi_link):
  try:
    response = requests.get(doi_link)
    response.raise_for_status()  # Raises exception for error status (4xx or 5xx)

    soup = BeautifulSoup(response.content, 'html.parser')

    title = soup.find('title').text if soup.find('title') else None
    authors = soup.find(
      'meta', {'name': 'citation_author'})['content'] if soup.find(
        'meta', {'name': 'citation_author'}) else None
    year_pub = soup.find(
      'meta', {'name': 'citation_publication_date'})['content'][:4] if soup.find(
        'meta', {'name': 'citation_publication_date'}) else None
    reference = doi_link

    return True, authors, title, reference, year_pub

  except requests.exceptions.RequestException as e:
    return False, f"Error searching article with DOI link {doi_link}: {e}", None, None, None
```

```
  except (AttributeError, KeyError) as e:
    return False, f"Error parsing article page with DOI link {doi_link}: {e}", None, None, None
```

```
def search_article_by_doi_code(doi_code):
  if doi_code:
    doi_link = f"https://doi.org/{doi_code[4:]}"
    return search_article_by_doi_link(doi_link)
  else:
    return False, None, None, None, None
```

```
def search_article_by_url(url):
  try:
    response = requests.get(url)
    response.raise_for_status()

    soup = BeautifulSoup(response.content, 'html.parser')

    title = soup.find('title').text if soup.find('title') else None
    authors = soup.find(
      'meta', {'name': 'citation_author'})['content'] if soup.find(
        'meta', {'name': 'citation_author'}) else None
    year_pub = soup.find(
      'meta', {'name': 'citation_publication_date'})['content'][:4] if soup.find(
        'meta', {'name': 'citation_publication_date'}) else None
    reference = url

    return True, authors, title, reference, year_pub

  except requests.exceptions.RequestException as e:
    return False, f"Error searching article with URL {url}: {e}", None, None, None
```

```
  except (AttributeError, KeyError) as e:
    return False, f"Error parsing article page with URL {url}: {e}", None, None, None
```

```
def search_article_by_fields(row):
  for field in [
    'doi_link', 'doi_code', 'URL', 'Title', 'Authors', 'Year_Pub']:
    value = row[field]
    if pd.notna(value):
      if field == 'doi_link':
        result = search_article_by_doi_link(value)
        if result:
          return result
        else:
          return (False, None, None, None, None)
      elif field == 'doi_code':
        result = search_article_by_doi_code(value)
        if result:
          return result
        else:
          return (False, None, None, None, None)
      elif field == 'URL':
        result = search_article_by_url(value)
        if result:
          return result
        else:
          return (False, None, None, None, None)
      else:
        query = f"{row['Title']} {row['Authors']} {row['Year_Pub']}"
        try:
          response = requests.get(
            f"https://www.google.com/search?q={query}")
```

```
            response.raise_for_status()

            soup = BeautifulSoup(response.content, 'html.parser')

            first_link = soup.find('a')["href"] if soup.find('a') else None
            if first_link:
              result = search_article_by_url(first_link)
              return result if result else (False, None, None, None, None)
            else:
              return False, "No results found on Google.", None, None, None

          except requests.exceptions.RequestException as e:
            return False, f"Error searching article with query '{query}': {e}", None, None, None
          except (AttributeError, KeyError) as e:
            return False, f"Error parsing search page with query '{query}': {e}", None, None, None

    return False, None, None, None, None
```

```
def search_studies(Studies):

    # Applies the functions to the DataFrame
    Studies['doi_link'] = Studies['Reference'].apply(extract_doi_link)
    Studies['doi_code'] = Studies['Reference'].apply(extract_doi_code)
    Studies['URL'] = Studies['Reference'].apply(extract_url)

    # Searches for articles
    Studies[[
      'found_doi_link',
      'Authors_found_doi_link',
      'Title_found_doi_link',
      'Reference_found_doi_link', 'Year_Pub_found_doi_link']] = Studies.apply(
```

```
    lambda row: pd.Series(
      search_article_by_doi_link(row['doi_link'])), axis=1)

Studies[[
  'found_doi_code',
  'Authors_found_doi_code',
  'Title_found_doi_code',
  'Reference_found_doi_code',
  'Year_Pub_found_doi_code']] = Studies.apply(
    lambda row: pd.Series(
      search_article_by_doi_code(row['doi_code'])), axis=1)

Studies[[
  'found_URL',
  'Authors_found_URL',
  'Title_found_URL',
  'Reference_found_URL', 'Year_Pub_found_URL']] = Studies.apply(
    lambda row: pd.Series(
      search_article_by_url(row['URL'])), axis=1)

Studies[[
  'found_columns',
  'Authors_found_columns',
  'Title_found_columns',
  'Reference_found_columns',
  'Year_Pub_found_columns']] = Studies.apply(
    lambda row: pd.Series(
      search_article_by_fields(row)), axis=1)
```

**Exemplo para teste:**

```
Studies = pd.DataFrame({
  'StudyId': ['1', '2', '3'],
  'Authors': ['Author A', 'Author B', 'Author C'],
  'Title': ['Title 1', 'Title 2', 'Title 3'],
  'Reference': [
    'Reference with https://doi.org/10.3390/f11030326',
    'Reference with doi:10.2134/agronj14.0597',
    'Reference with https://czasopisma.uni.lodz.pl/space/article/view/22196'],
  'Year_Pub': ['2020', '2021', '2022']})

search_studies(Studies)

with pd.option_context(
  #'display.width', None,
  #'display.max_colwidth', None,
  'display.max_rows', None,
  'display.max_columns', None):
  print(Studies)
```

```
  StudyId   Authors    Title  \
0       1  Author A  Title 1
1       2  Author B  Title 2
2       3  Author C  Title 3

                                          Reference Year_Pub  \
0   Reference with https://doi.org/10.3390/f11030326     2020
1           Reference with doi:10.2134/agronj14.0597     2021
2  Reference with https://czasopisma.uni.lodz.pl/...     2022
```

```
                          doi_link                   doi_code  \
0  https://doi.org/10.3390/f11030326                       None
1                               None  doi:10.2134/agronj14.0597
2                               None                       None

                                           URL  found_doi_link  \
0                                         None            True
1                                         None           False
2  https://czasopisma.uni.lodz.pl/space/article/v...          False

                            Authors_found_doi_link  \
0                                  Getzner, Michael
1  Error searching article with DOI link None: In...
2  Error searching article with DOI link None: In...

                              Title_found_doi_link  \
0  The Benefits of Local Forest Recreation in Aus...
1                                         None
2                                         None

          Reference_found_doi_link Year_Pub_found_doi_link  found_doi_code  \
0  https://doi.org/10.3390/f11030326                  2020           False
1                               None                  None           False
2                               None                  None           False

                            Authors_found_doi_code Title_found_doi_code  \
0                                         None                 None
1  Error searching article with DOI link https://...                 None
2                                         None                 None
```

```
  Reference_found_doi_code Year_Pub_found_doi_code  found_URL  \
0                     None                     None      False
1                     None                     None      False
2                     None                     None       True

                                  Authors_found_URL  \
0  Error searching article with URL None: Invalid...
1  Error searching article with URL None: Invalid...
2                                   Dietwald Gruehn

                                    Title_found_URL  \
0                                              None
1                                              None
2  \n\t\tEconomic Valuation of Urban Open Spaces ...

                                Reference_found_URL Year_Pub_found_URL  \
0                                              None               None
1                                              None               None
2  https://czasopisma.uni.lodz.pl/space/article/v...               None

   found_columns                                Authors_found_columns  \
0           True                                     Getzner, Michael
1          False  Error searching article with DOI link https://...
2           True                                      Dietwald Gruehn

                                Title_found_columns  \
0  The Benefits of Local Forest Recreation in Aus...
1                                              None
2  \n\t\tEconomic Valuation of Urban Open Spaces ...

                           Reference_found_columns Year_Pub_found_columns
```

```
0                    https://doi.org/10.3390/f11030326                    2020
1                                                 None                    None
2  https://czasopisma.uni.lodz.pl/space/article/v...                    None
```

**Testando com algumas linhas do ESVD:**

```
esvd = pd.read_csv(
    r"..\downloads\Esvd_Full_Data_10th-Sep-2024_08-19-31_Database_Version_APR2024V1.1.csv",
    dtype = str, na_filter=False)
print(n_esv := len(esvd), "linhas no arquivo ESVD.")
print(esvd["ValueId"].nunique(), "registros no ESVD.")

df = (esvd.groupby(
    ['StudyId', 'Authors', 'Title', 'Reference', 'Year_Pub'])
    .first()
    .reset_index())
print(len(df), "estudos fonte distintos no ESVD.")

df = df.head(100)
search_studies(df)

with pd.option_context(
  #'display.width', None,
  #'display.max_colwidth', None,
  'display.max_rows', None,
  'display.max_columns', None):

  print(df[
  df['found_doi_link'] |
  df['found_doi_code'] |
```

```
  df['found_URL'] |
  df['found_columns']][[
  'StudyId', 'Authors', 'Title', 'Reference', 'Year_Pub',
  'found_doi_link', 'Authors_found_doi_link', 'Title_found_doi_link',
  'Reference_found_doi_link', 'Year_Pub_found_doi_link',
  'found_doi_code', 'Authors_found_doi_code', 'Title_found_doi_code',
  'Reference_found_doi_code', 'Year_Pub_found_doi_code',
  'found_URL', 'Authors_found_URL', 'Title_found_URL',
  'Reference_found_URL', 'Year_Pub_found_URL',
  'found_columns', 'Authors_found_columns', 'Title_found_columns',
  'Reference_found_columns', 'Year_Pub_found_columns']])
```

```
10874 linhas no arquivo ESVD.
10874 registros no ESVD.
1354 estudos fonte distintos no ESVD.
   StudyId              Authors  \
3     1001  Bowker and Didychuk
6     1004         Schaub et al
22    1019         Collof et al
28    1024        Buckley et al

                                            Title  \
3   Estimation of the nonmarket benefits of agricu...
6   Plant diversity effects on forage quality, yie...
22  Natural pest control in citrus as an ecosystem...
28  Recreational demand for farm commonage in Irel...

                                          Reference Year_Pub  \
3   Bowker, J., & Didychuk, D. (1994). Estimation ...     1994
6   Schaub, S., Finger, R., Leiber, F. et al. Plan...     2020
22  Colloff, M., Lindsay, E., & Cook, D. (2013). N...     2013
```

```
28  Buckley, C., Van Rensburg, T., & Hynes, S. (20...     2009

    found_doi_link                              Authors_found_doi_link  \
3            False  Error searching article with DOI link None: In...
6             True                                        Schaub, Sergei
22           False  Error searching article with DOI link None: In...
28           False  Error searching article with DOI link None: In...

                                Title_found_doi_link  \
3                                               None
6   Plant diversity effects on forage quality, yie...
22                                              None
28                                              None

                     Reference_found_doi_link Year_Pub_found_doi_link  \
3                                        None                     None
6   https://doi.org/10.1038/s41467-020-14541-4                     None
22                                       None                     None
28                                       None                     None

    found_doi_code Authors_found_doi_code  \
3             True             J.M. Bowker
6            False                    None
22            True                    None
28            True                    None

                                Title_found_doi_code  \
3   Estimation of the Nonmarket Benefits of Agricu...
6                                               None
22                                       Redirecting
28                                       Redirecting
```

```
                          Reference_found_doi_code Year_Pub_found_doi_code  \
3         https://doi.org/10.1017/S1068280500002331                    1994
6                                             None                    None
22  https://doi.org/10.1016/j.biocontrol.2013.07.017                    None
28  https://doi.org/10.1016/j.landusepol.2008.10.013                    None

    found_URL                                 Authors_found_URL  \
3       False  Error searching article with URL None: Invalid...
6       False  Error searching article with URL None: Invalid...
22      False  Error searching article with URL None: Invalid...
28      False  Error searching article with URL None: Invalid...

   Title_found_URL Reference_found_URL Year_Pub_found_URL  found_columns  \
3             None                None               None           True
6             None                None               None           True
22            None                None               None           True
28            None                None               None           True

   Authors_found_columns                                Title_found_columns  \
3            J.M. Bowker  Estimation of the Nonmarket Benefits of Agricu...
6         Schaub, Sergei  Plant diversity effects on forage quality, yie...
22                  None                                        Redirecting
28                  None                                        Redirecting

                          Reference_found_columns Year_Pub_found_columns
3         https://doi.org/10.1017/S1068280500002331                  1994
6        https://doi.org/10.1038/s41467-020-14541-4                  None
22  https://doi.org/10.1016/j.biocontrol.2013.07.017                  None
28  https://doi.org/10.1016/j.landusepol.2008.10.013                  None
```