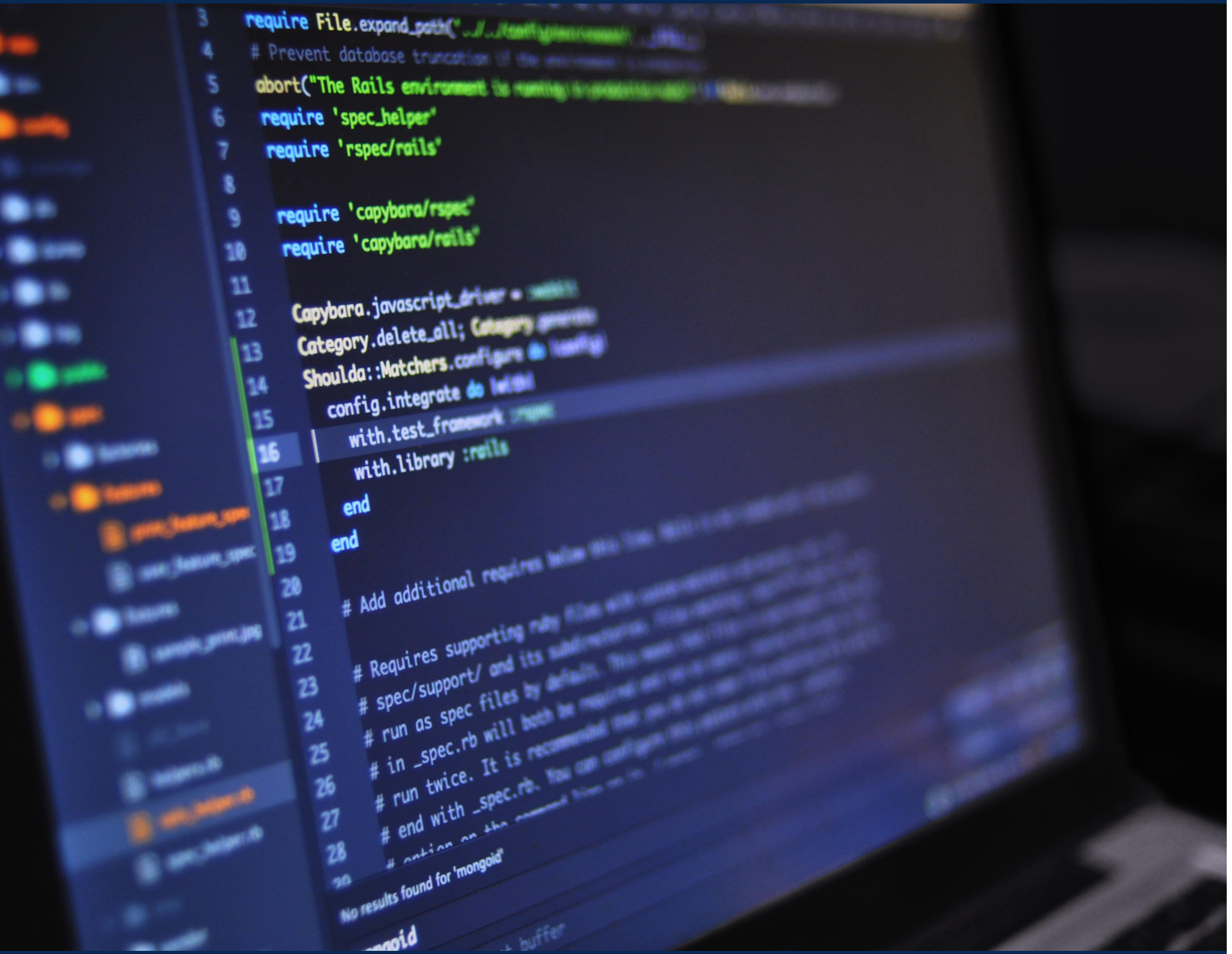
EXAMEN FINAL

INTEGRANTES:

- CASTRO FERNANDEZ, JHORDY
- HERRERA MENDOZA, ELVIS
- ROCA MATÍAS, BETSABÉ



```
require File.expand_path('../../config/environment', [illegible]
# Prevent database truncation [illegible]
abort("The Rails environment [illegible]
require 'spec_helper'
require 'rspec/rails'

require 'capybara/rspec'
require 'capybara/rails'

Capybara.javascript_driver = [illegible]
Category.delete_all; Category [illegible]
Shoulda::Matchers.configure do [illegible]
  config.integrate do [illegible]
    with.test_framework :rspec
    with.library :rails
  end
end

# Add additional requires below this line. [illegible]

# Requires supporting ruby files [illegible]
# spec/support/ and its subdirectories. [illegible]
# run as spec files by default. [illegible]
# in _spec.rb will both be required [illegible]
# run twice. It is recommended [illegible]
# end with _spec.rb. You can configure [illegible]
```

No results found for 'mongoid'

OBJETIVO

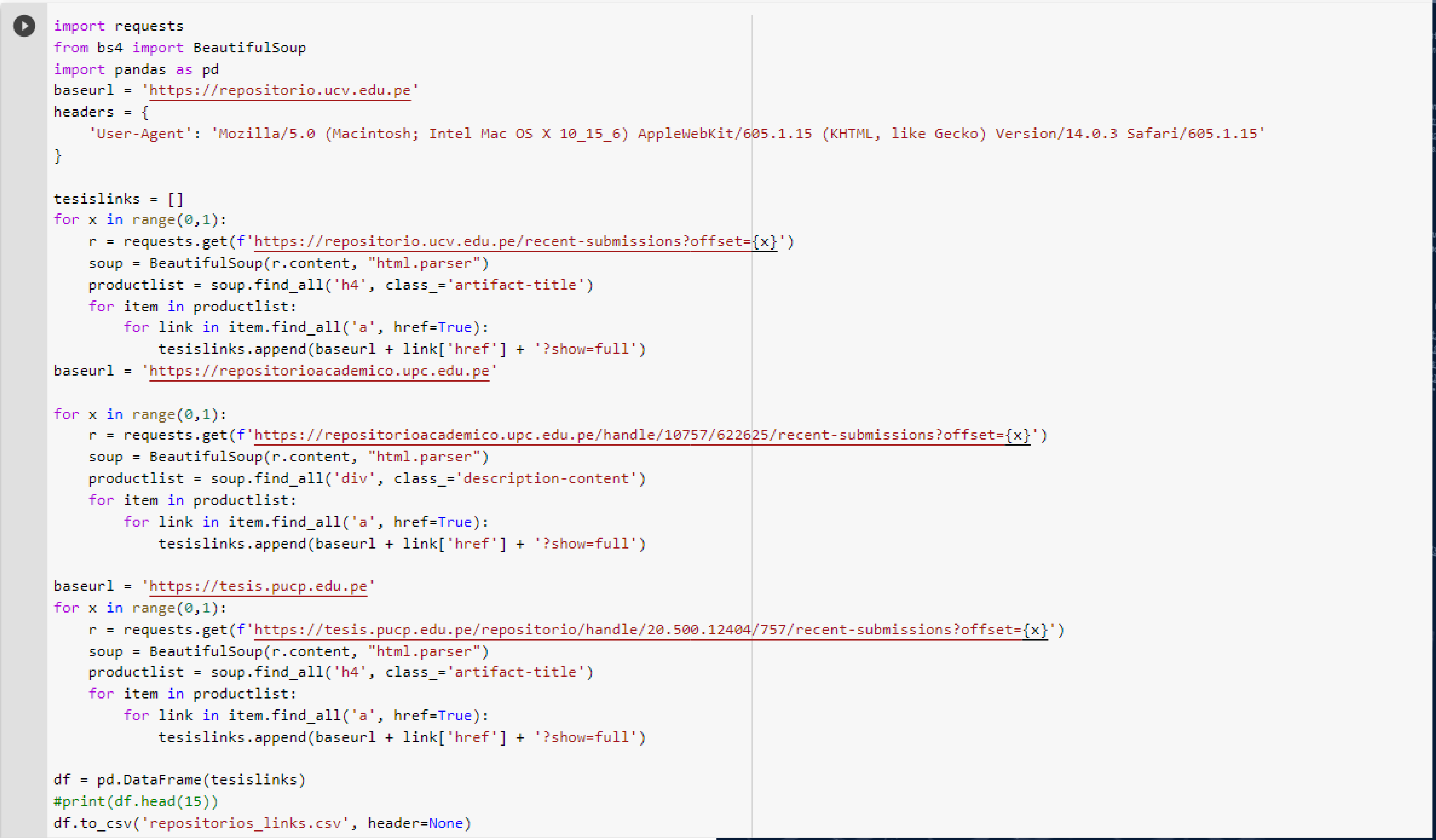
Proyecto web scraping con python para repositorios de tesis

Recopilamos información de los repositorios de las 3 universidades escogidas para facilitarnos la búsqueda de información.

UCV
UNIVERSIDAD
CÉSAR VALLEJO

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ

UPC
Universidad Peruana
de Ciencias Aplicadas



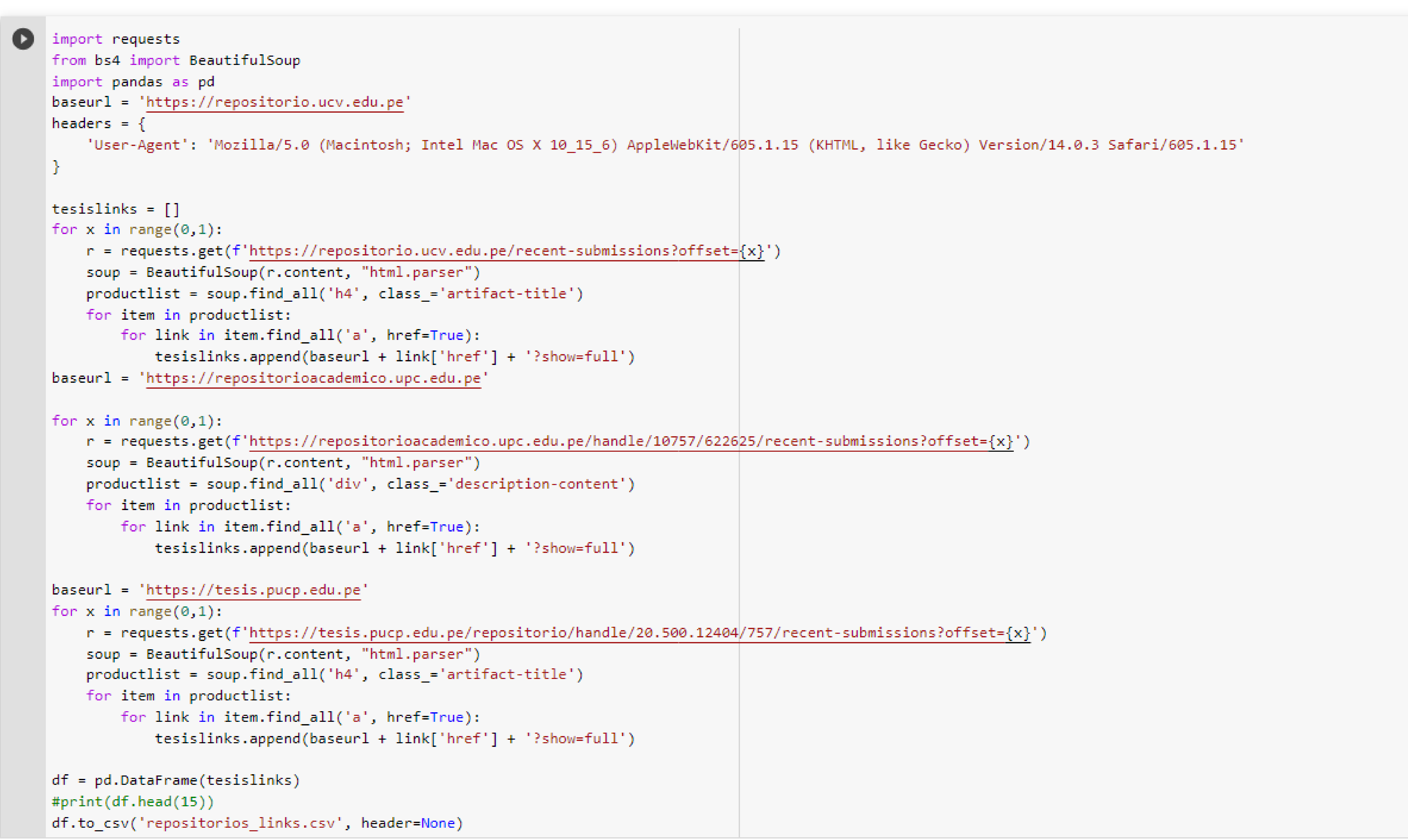
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
baseurl = 'https://repositorio.ucv.edu.pe'
headers = {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_6) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0.3 Safari/605.1.15'
}

tesislinks = []
for x in range(0,1):
    r = requests.get(f'https://repositorio.ucv.edu.pe/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('h4', class_='artifact-title')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')
baseurl = 'https://repositorioacademico.upc.edu.pe'

for x in range(0,1):
    r = requests.get(f'https://repositorioacademico.upc.edu.pe/handle/10757/622625/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('div', class_='description-content')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')

baseurl = 'https://tesis.pucp.edu.pe'
for x in range(0,1):
    r = requests.get(f'https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/757/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('h4', class_='artifact-title')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')

df = pd.DataFrame(tesislinks)
#print(df.head(15))
df.to_csv('repositorios_links.csv', header=None)
```



```python
import requests
from bs4 import BeautifulSoup
import pandas as pd
baseurl = 'https://repositorio.ucv.edu.pe'
headers = {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_6) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0.3 Safari/605.1.15'
}

tesislinks = []
for x in range(0,1):
    r = requests.get(f'https://repositorio.ucv.edu.pe/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('h4', class_='artifact-title')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')
baseurl = 'https://repositorioacademico.upc.edu.pe'

for x in range(0,1):
    r = requests.get(f'https://repositorioacademico.upc.edu.pe/handle/10757/622625/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('div', class_='description-content')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')

baseurl = 'https://tesis.pucp.edu.pe'
for x in range(0,1):
    r = requests.get(f'https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/757/recent-submissions?offset={x}')
    soup = BeautifulSoup(r.content, "html.parser")
    productlist = soup.find_all('h4', class_='artifact-title')
    for item in productlist:
        for link in item.find_all('a', href=True):
            tesislinks.append(baseurl + link['href'] + '?show=full')

df = pd.DataFrame(tesislinks)
#print(df.head(15))
df.to_csv('repositorios_links.csv', header=None)
```

¡GRACIAS!