Task 1 ReadMe

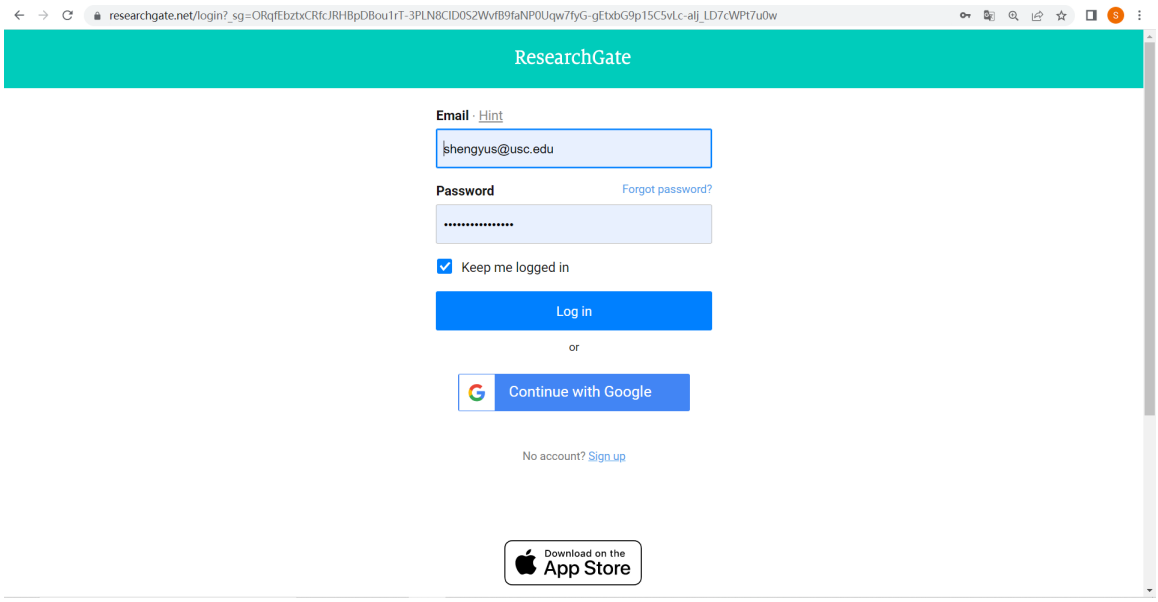
This part is used to crawl the pdf files of papers in bik dataset.

After homework1, we found that there is no single website for storing all required paper PDF files. Therefore, in this crawling task, we selected the following two websites:

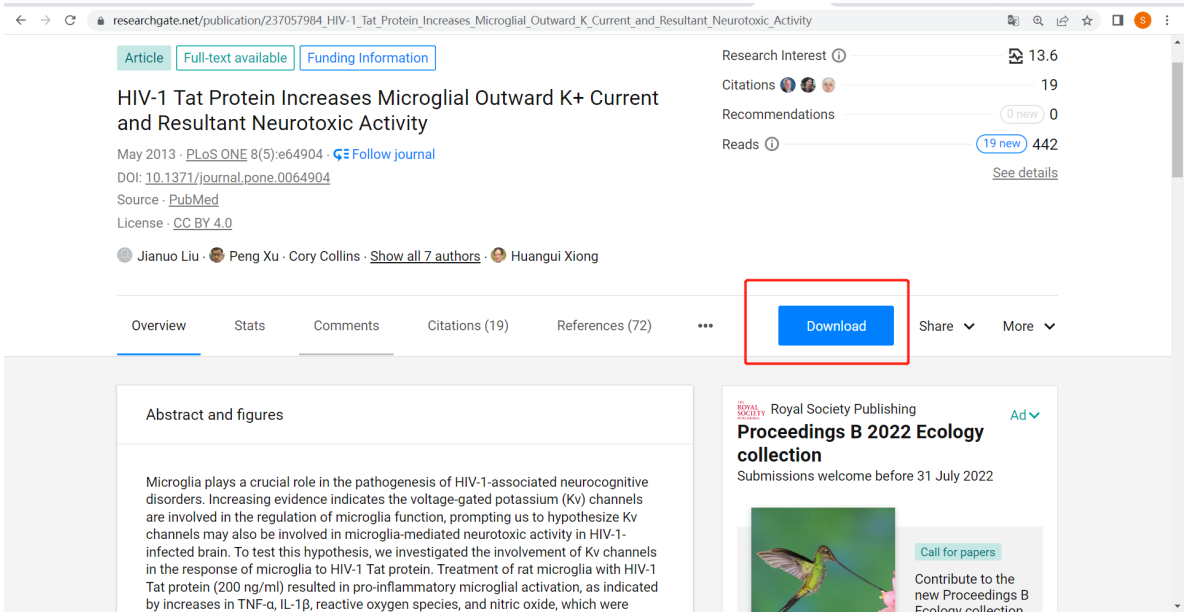
```
Research Gate: https://www.researchgate.net/
USC library: https://libraries.usc.edu/
```

Since the member who do the web crawling task is not in the United States, he need VPN to access the USC library. However, VPN will conflict with the requests library. Therefore, we chose to use selenium to simulate browser access and download for crawling.

The code below shows our first step: simulate the browser to query and crawl the Research Gate. To crawl this website, we need to login first, the user_email and user_pwd variables are the exact account information.

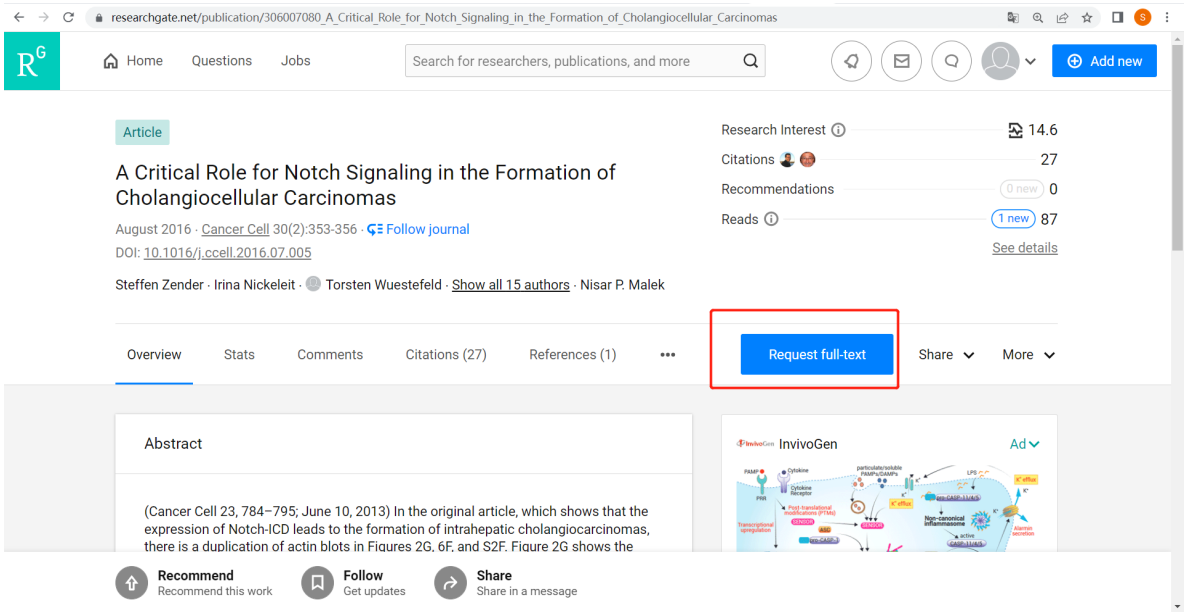


After login at the Research Gate, we can search the papers by DOI number. By simulating the browser, we can click the Download button and finish crawling.

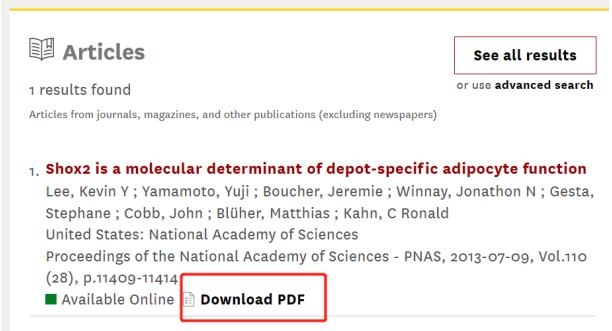



```
river=webdriver.Chrome()
driver.get("https://www.researchgate.net/login")
driver.find_element(By.ID, "input-login").send_keys(user_email)
driver.find_element(By.ID, "input-password").send_keys(user_pwd)
driver.find_element(By.CLASS_NAME, "nova-legacy-c-
button__label").find_element(By.XPATH, "./..").click()
for i in range(165,len(bik)):
    curr_page = 'https://www.researchgate.net/search/publication?q='+bik['DOI']
[i]
    driver.get(curr_page)
    bs_test=bs(driver.page_source,'html.parser')
    if bs_test.find('a',{'class':'nova-legacy-c-button nova-legacy-c-button--
align-center nova-legacy-c-button--radius-m nova-legacy-c-button--size-m nova-
legacy-c-button--color-blue nova-legacy-c-button--theme-solid nova-legacy-c-
button--width-auto'})!=None:
        driver.find_element(By.LINK_TEXT, "Download").click()
```

However, when the search result is Request full-text instead of Download, we need to find another way to crawl.



There are still 100+ papers could not find in Research Gate, so we access USC library to finish the crawling task of the remain papers. The code below shows the process of downloading pdf files from USC library by simulating Chrome browser as well.




```
else:
        curr_page = 'https://www.researchgate.net/search/publication?
q='+bik['Title'][i]
        driver.get(curr_page)
        bs_title=bs(driver.page_source,'html.parser')
        if bs_title.find('a',{'class':'nova-legacy-c-button nova-legacy-c-
button--align-center nova-legacy-c-button--radius-m nova-legacy-c-button--size-s
nova-legacy-c-button--color-blue nova-legacy-c-button--theme-ghost nova-legacy-
c-button--width-auto'})!=None:
            url=bs_title.find('a',{'class':'nova-legacy-c-button nova-legacy-c-
button--align-center nova-legacy-c-button--radius-m nova-legacy-c-button--size-s
nova-legacy-c-button--color-blue nova-legacy-c-button--theme-ghost nova-legacy-
c-button--width-auto'})['href']
            pdf_url='https://www.researchgate.net/'+url
            driver.find_element(By.LINK_TEXT, "Download").click()
    print('finish '+str(i))
```

After these two steps, we can collect 200+ papers mentioned in the bik dataset.