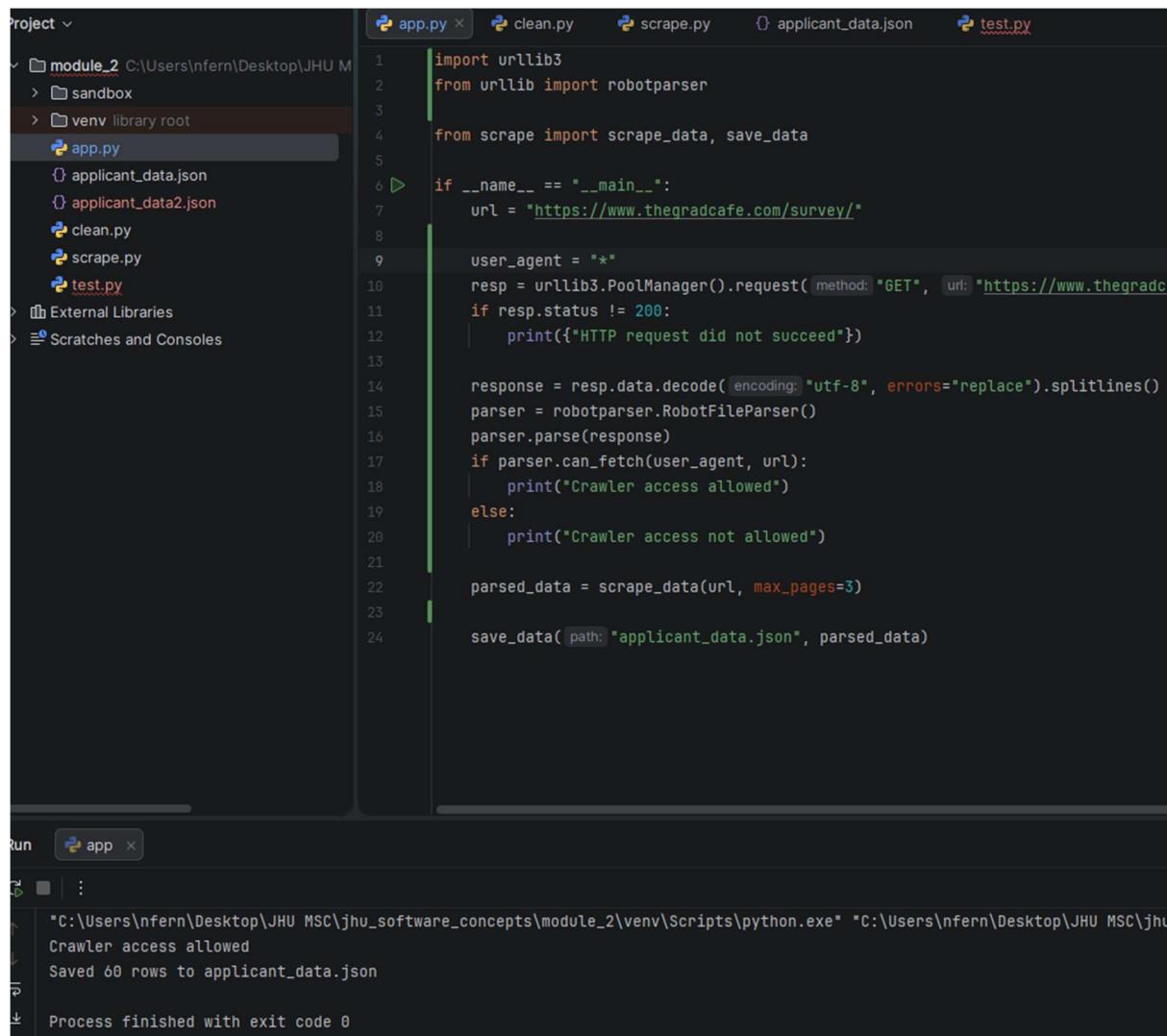


## Screenshot of output from checking robots.txt – with allowed agent

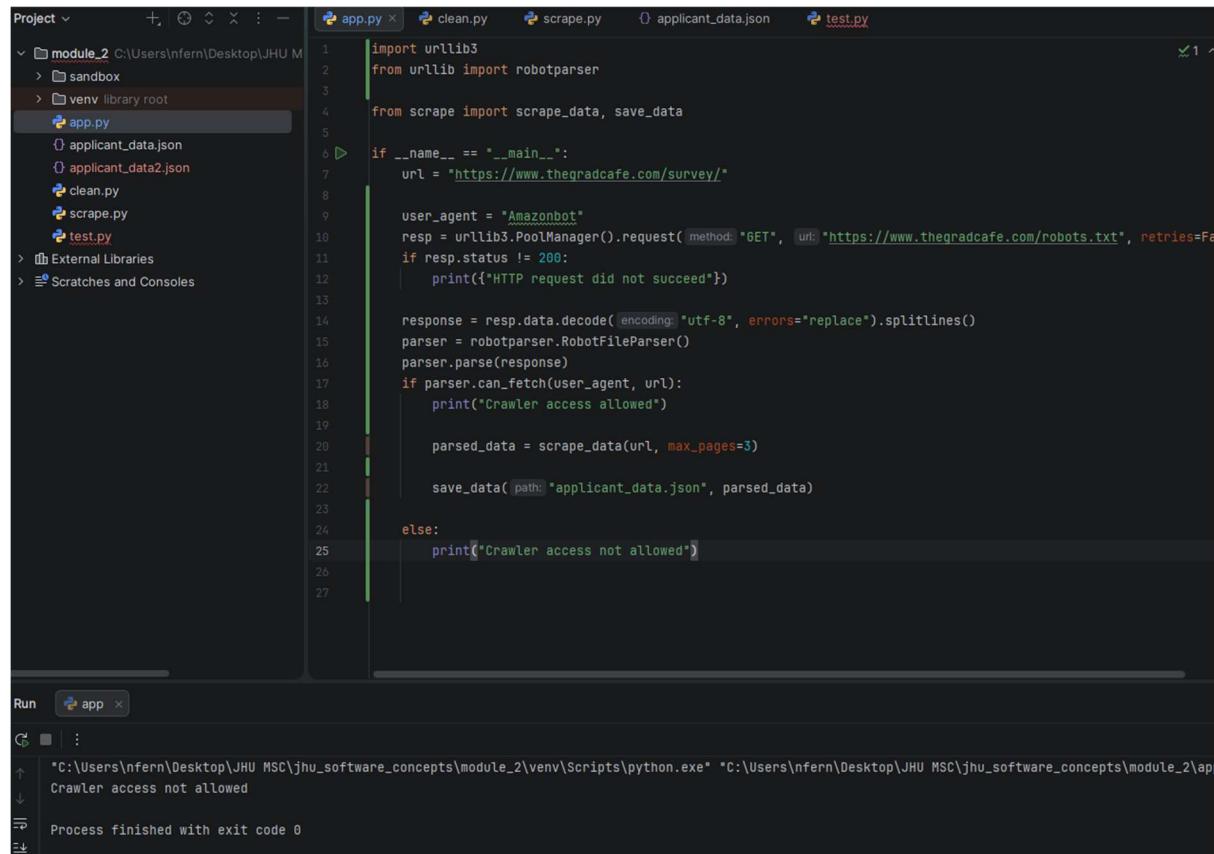


```
Project ▾
  module_2 C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\venv\Scripts\python.exe
    app.py
    clean.py
    scrape.py
    applicant_data.json
    applicant_data2.json
    test.py
  External Libraries
  Scratches and Consoles

app.py
1 import urllib3
2 from urllib import robotparser
3
4 from scrape import scrape_data, save_data
5
6 if __name__ == "__main__":
7     url = "https://www.thegradcafe.com/survey/"
8
9     user_agent = "*"
10    resp = urllib3.PoolManager().request(method="GET", url="https://www.thegradcafe.com/survey/")
11    if resp.status != 200:
12        print("HTTP request did not succeed")
13
14    response = resp.data.decode(encoding="utf-8", errors="replace").splitlines()
15    parser = robotparser.RobotFileParser()
16    parser.parse(response)
17    if parser.can_fetch(user_agent, url):
18        print("Crawler access allowed")
19    else:
20        print("Crawler access not allowed")
21
22    parsed_data = scrape_data(url, max_pages=3)
23
24    save_data(path="applicant_data.json", parsed_data)

Run app
  "C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\venv\Scripts\python.exe" "C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\app.py"
  Crawler access allowed
  Saved 60 rows to applicant_data.json
  Process finished with exit code 0
```

## Screenshot of output from checking robots.txt – with disallowed agent



```
Project ▾
  module_2 C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\venv\Scripts\python.exe
    app.py
    clean.py
    scrape.py
    applicant_data.json
    applicant_data2.json
    test.py
  External Libraries
  Scratches and Consoles

app.py
1 import urllib3
2 from urllib import robotparser
3
4 from scrape import scrape_data, save_data
5
6 if __name__ == "__main__":
7     url = "https://www.thegradcafe.com/survey/"
8
9     user_agent = "Amazonbot"
10    resp = urllib3.PoolManager().request(method="GET", url="https://www.thegradcafe.com/robots.txt", retries=False)
11    if resp.status != 200:
12        print("HTTP request did not succeed")
13
14    response = resp.data.decode(encoding="utf-8", errors="replace").splitlines()
15    parser = robotparser.RobotFileParser()
16    parser.parse(response)
17    if parser.can_fetch(user_agent, url):
18        print("Crawler access allowed")
19
20    parsed_data = scrape_data(url, max_pages=3)
21
22    save_data(path="applicant_data.json", parsed_data)
23
24 else:
25     print("Crawler access not allowed")

Run app
  "C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\venv\Scripts\python.exe" "C:\Users\nfern\Desktop\JHU MSC\jhu_software_concepts\module_2\app.py"
  Crawler access not allowed
  Process finished with exit code 0
```