Name : Devasy Patel

Roll No : 20BCE057

Date : 27/03/2023

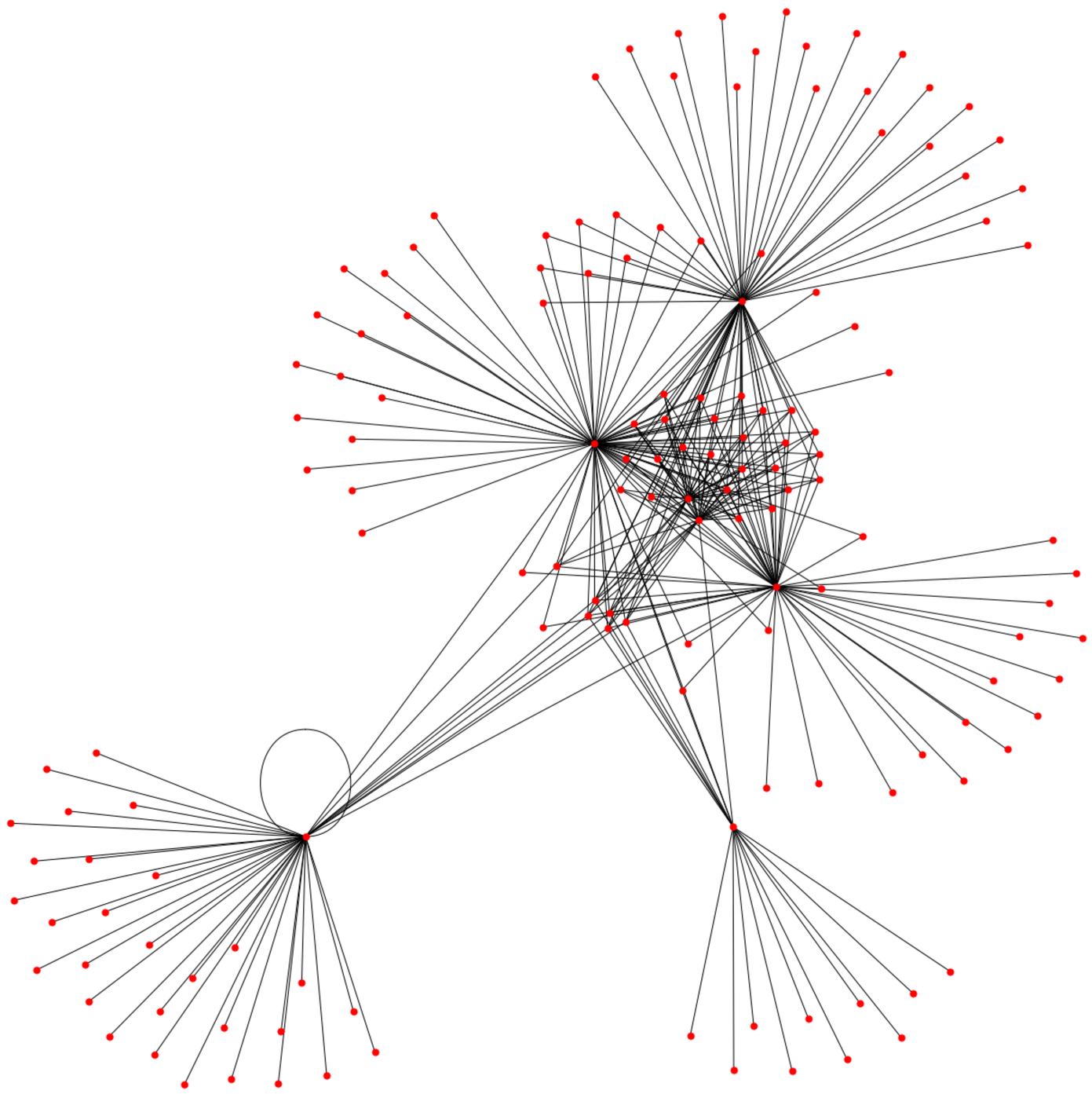
Practical 7 : Implement a Web crawler using simple queue data structure.

In [ ]:
```python
import requests
from bs4 import BeautifulSoup
from collections import deque

def crawl(start_url):
    visited = set()
    queue = deque([(start_url, 0)])
    tree = {}


    while queue:
        url, depth = queue.popleft()
        tree[url] = []
        if depth > 3:
            print('Max depth reached')
            return tree
        if url not in visited:
            visited.add(url)
            print(f'Visiting: {url}')
            try:
                response = requests.get(url)
                soup = BeautifulSoup(response.text, 'html.parser')
                for link in soup.find_all('a'):
                    href = link.get('href')
                    if href and href.startswith('http'):
                        tree[url].append(href)
                        queue.append((href, depth + 1))
            except KeyboardInterrupt:
                return tree
            except:
                pass
    return tree

# tree = crawl('https://www.isro.gov.in/')
tree = crawl('https://www.nirmauni.ac.in/')
```

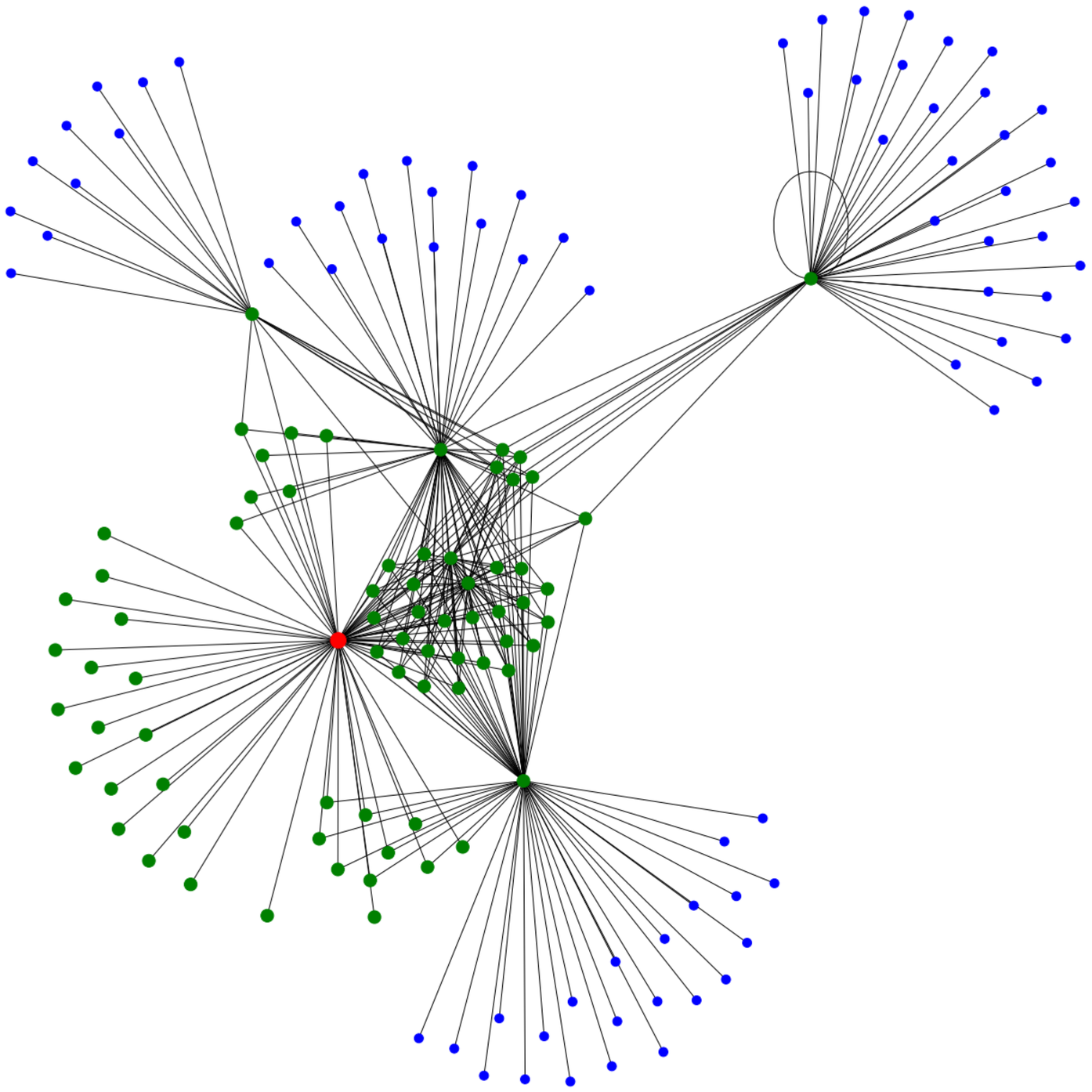
```
Visiting: https://www.nirmauni.ac.in/
Visiting: https://nirmauni.ac.in
Visiting: https://nirmauni.ac.in/admissions-aid/
Visiting: https://nirmauni.ac.in/academics/
Visiting: https://nirmauni.ac.in/campus-life/
Visiting: https://nirmauni.ac.in/research-at-nirma/
Visiting: https://nirmauni.ac.in/placement/
Visiting: https://nirmauni.ac.in/alumni/
Visiting: https://nirmauni.ac.in/about/
Visiting: https://nirmauni.ac.in/announcement/phd-admissions-2023-24/
Visiting: https://www.facebook.com/sharer/sharer.php?u=https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Visiting: https://twitter.com/home?status=https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Visiting: https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Visiting: https://nirmauni.ac.in/announcement/
Visiting: http://architecture.nirmauni.ac.in/
Visiting: http://commerce.nirmauni.ac.in/
Visiting: http://design.nirmauni.ac.in/
Visiting: http://fdsr.nirmauni.ac.in/
Visiting: http://law.nirmauni.ac.in/
Visiting: http://management.nirmauni.ac.in/
Visiting: http://pharmacy.nirmauni.ac.in/
Visiting: http://science.nirmauni.ac.in/
Visiting: http://technology.nirmauni.ac.in/
Visiting: https://nirmauni.ac.in/academics
Visiting: https://nirmauni.ac.in/research-at-nirma/directorate-of-research-innovation/bank-of-thrust-areas/
Visiting: https://nirmauni.ac.in/research_stories/
Visiting: https://nirmauni.ac.in/research-at-nirma/research-activities/events/
Visiting: https://nirmauni.ac.in/funded-projects/
Visiting: http://fdsr.nirmauni.ac.in/doctoral-students/current/
Visiting: https://nirmauni.ac.in/research-at-nirma/research-activities/undergraduate-research/
Visiting: https://nirmauni.ac.in/admissions/
Visiting: https://commerce.nirmauni.ac.in/admission-aid/postgraduate-programme/mcom-accounting-taxation/
Visiting: http://commerce.nirmauni.ac.in/admission-aid/under-graduate/b-com-hons/
Visiting: https://admissions-id.nirmauni.ac.in/student/default.aspx
Visiting: http://law.nirmauni.ac.in/admission-aid/ba-bcom-llb-hons/
```

In [ ]:
```python
# function to find the height of the tree of urls
def find_height(tree, start_url):
    height = 0
    queue = deque([(start_url, 0)])
    visited = set()
    while queue:
        url, h = queue.popleft()
        if url not in visited:
            visited.add(url)
            height = max(height, h)
            for link in tree[url]:
                queue.append((link, h+1))
    return height
# find_height(tree, 'https://www.isro.gov.in/')
```

In [ ]:
```python
# plto the tree
import networkx as nx
import matplotlib.pyplot as plt

plt.figure(figsize=(20,20))
G = nx.Graph()
for parent, children in tree.items():
    for child in children:
        G.add_edge(parent, child)


nx.draw(G, node_size=50, node_color='red', font_size=8)
plt.show()
# show the tree in larger scale
```



In [ ]:
```python
import networkx as nx
import matplotlib.pyplot as plt

def plot_tree(tree, start_url):
    G = nx.Graph()
    colors = []
    sizes = []
    for parent, children in tree.items():
        for child in children:
            G.add_edge(parent, child)
    for node in G.nodes:
        if node == start_url:
            colors.append('red')
            sizes.append(300)
        else:
            try:
                depth = nx.shortest_path_length(G, start_url, node)
                if depth == 1:
                    colors.append('green')
                    sizes.append(200)
                elif depth == 2:
                    colors.append('blue')
                    sizes.append(100)
                elif depth == 3:
                    colors.append('yellow')
                    sizes.append(50)
            except nx.NetworkXNoPath:
                # Assign a default color and size to nodes that are not reachable from the start_url
                colors.append('gray')
                sizes.append(25)
    plt.figure(figsize=(20,20))
    nx.draw(G, node_color=colors, node_size=sizes, font_size=8)
    plt.show()

plot_tree(tree, 'https://www.nirmauni.ac.in/')
```



In [ ]:
```
import plotly.graph_objs as go
import networkx as nx

def plot_tree_with_plotly(tree):
    G = nx.Graph()
    for parent, children in tree.items():
        for child in children:
            G.add_edge(parent, child)
    pos = nx.spring_layout(G)
    edge_x = []
    edge_y = []
    for edge in G.edges():
        x0, y0 = pos[edge[0]]
        x1, y1 = pos[edge[1]]
        edge_x.extend([x0, x1, None])
        edge_y.extend([y0, y1, None])
    edge_trace = go.Scatter(
        x=edge_x, y=edge_y,
        line=dict(width=0.5, color='#888'),
        hoverinfo='none',
        mode='lines')
    node_x = []
    node_y = []
    for node in G.nodes():
        x, y = pos[node]
        node_x.append(x)
        node_y.append(y)
    node_trace = go.Scatter(
        x=node_x, y=node_y,
        mode='markers',
        hoverinfo='text',
        marker=dict(
            showscale=False,
            color='red',
            size=10,
            line_width=2))
    fig = go.Figure(data=[edge_trace, node_trace],
                 layout=go.Layout(
                    showlegend=False,
                    hovermode='closest',
                    margin=dict(b=20,l=5,r=5,t=40),
                    annotations=[ dict(
                        text="",
                        showarrow=False,
                        xref="paper", yref="paper",
                        x=0.005, y=-0.002 ) ],
                    xaxis=dict(showgrid=False, zeroline=False, showticklabels=False),
                    yaxis=dict(showgrid=False, zeroline=False, showticklabels=False)))
    fig.show()

plot_tree_with_plotly(tree)
```

In [ ]:
```
import requests
from bs4 import BeautifulSoup
from collections import deque
from urllib.parse import urlparse, urljoin
from urllib.robotparser import RobotFileParser

def polite_crawl(start_url):
    visited = set()
    queue = deque([start_url])
    tree = {}
    rp = RobotFileParser()

    while queue:
        url = queue.popleft()
        tree[url] = []
        if url not in visited:
            visited.add(url)
            print(f'Visiting: {url}')
            try:
                # Check robots.txt for politeness
                parsed_url = urlparse(url)
                base_url = f"{parsed_url.scheme}://{parsed_url.netloc}"
                rp.set_url(urljoin(base_url, '/robots.txt'))
                rp.read()
                if not rp.can_fetch('*', url):
                    print(f"Not allowed to crawl {url} due to robots.txt restrictions")
                    continue

                response = requests.get(url)
                soup = BeautifulSoup(response.text, 'html.parser')
                for link in soup.find_all('a'):
                    href = link.get('href')
                    if href and href.startswith('http'):
                        tree[url].append(href)
                        queue.append(href)
            except:
                pass
    return tree

tree = polite_crawl('https://www.nirmauni.ac.in/')
```

```
Visiting: https://www.nirmauni.ac.in/
Visiting: https://nirmauni.ac.in
Visiting: https://nirmauni.ac.in/admissions-aid/
Visiting: https://nirmauni.ac.in/academics/
Visiting: https://nirmauni.ac.in/campus-life/
Visiting: https://nirmauni.ac.in/research-at-nirma/
Visiting: https://nirmauni.ac.in/placement/
Visiting: https://nirmauni.ac.in/alumni/
Visiting: https://nirmauni.ac.in/about/
Visiting: https://nirmauni.ac.in/announcement/phd-admissions-2023-24/
Visiting: https://www.facebook.com/sharer/sharer.php?u=https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Not allowed to crawl https://www.facebook.com/sharer/sharer.php?u=https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/ due to robots.txt restrictions
Visiting: https://twitter.com/home?status=https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Visiting: https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/
Not allowed to crawl https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/phd-admission-page/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/announcement/
Not allowed to crawl https://nirmauni.ac.in/announcement/ due to robots.txt restrictions
Visiting: http://architecture.nirmauni.ac.in/
Visiting: http://commerce.nirmauni.ac.in/
Visiting: http://design.nirmauni.ac.in/
Visiting: http://fdsr.nirmauni.ac.in/
Visiting: http://law.nirmauni.ac.in/
Visiting: http://management.nirmauni.ac.in/
Visiting: http://pharmacy.nirmauni.ac.in/
Visiting: http://science.nirmauni.ac.in/
Visiting: http://technology.nirmauni.ac.in/
Visiting: https://nirmauni.ac.in/academics
Not allowed to crawl https://nirmauni.ac.in/academics due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/research-at-nirma/directorate-of-research-innovation/bank-of-thrust-areas/
Not allowed to crawl https://nirmauni.ac.in/research-at-nirma/directorate-of-research-innovation/bank-of-thrust-areas/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/research_stories/
Not allowed to crawl https://nirmauni.ac.in/research_stories/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/research-at-nirma/research-activities/events/
Not allowed to crawl https://nirmauni.ac.in/research-at-nirma/research-activities/events/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/funded-projects/
Not allowed to crawl https://nirmauni.ac.in/funded-projects/ due to robots.txt restrictions
Visiting: http://fdsr.nirmauni.ac.in/doctoral-students/current/
Visiting: https://nirmauni.ac.in/research-at-nirma/research-activities/undergraduate-research/
Not allowed to crawl https://nirmauni.ac.in/research-at-nirma/research-activities/undergraduate-research/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/admissions/
Not allowed to crawl https://nirmauni.ac.in/admissions/ due to robots.txt restrictions
Visiting: https://commerce.nirmauni.ac.in/admission-aid/postgraduate-programme/mcom-accounting-taxation/
Not allowed to crawl https://commerce.nirmauni.ac.in/admission-aid/postgraduate-programme/mcom-accounting-taxation/ due to robots.txt restrictions
Visiting: http://commerce.nirmauni.ac.in/admission-aid/under-graduate/b-com-hons/
Visiting: https://admissions-id.nirmauni.ac.in/student/default.aspx
Visiting: http://law.nirmauni.ac.in/admission-aid/ba-bcom-llb-hons/
Visiting: http://law.nirmauni.ac.in/admission-aid/post-graduate/ll-m/
Visiting: https://management.nirmauni.ac.in/admission-aid/postgraduate/mba-and-mba-hrm/
Visiting: http://management.nirmauni.ac.in/admission-aid/undergraduate/bba-mba-five-year-integrated/
Visiting: http://management.nirmauni.ac.in/admission-aid/postgraduate/mba-family-business-entrepreneurship/
Visiting: https://technology.nirmauni.ac.in/admission-aid/under-graduate/bs-cse/
Visiting: https://fdsr.nirmauni.ac.in/admission-aid/doctor-of-philosophy/full-time-phd-program/
Visiting: http://internationalrelations.nirmauni.ac.in/admissions/
Visiting: https://commerce.nirmauni.ac.in/upi-a-game-changer-for-financial-inclusion-in-india/
Visiting: https://technology.nirmauni.ac.in/deputation-to-florida-atlantic-university-florida-usa-as-adjunct-professor-and-visiting-research-scholar/
Visiting: https://technology.nirmauni.ac.in/visit-to-dublin-city-university-a-lifelong-and-indelible-research-experience/
Visiting: https://commerce.nirmauni.ac.in/sustainable-investing-practices/
Visiting: https://nirmauni.ac.in/calendar/
Visiting: https://nirmauni.ac.in/news-and-events/
Visiting: https://nirmauni.ac.in/news-events/irs-mrs-parul-srivastava-at-womens-day-celebration/
Visiting: https://nirmauni.ac.in/news-events/showcasing-doctoral-research-at-nirma-university/
Visiting: https://nirmauni.ac.in/news-events/celebration-of-the-inter-institute-cultural-festival-nuzeal/
Visiting: https://www.facebook.com/NirmaUniOfficial
Visiting: https://twitter.com/NirmaUniTweets
Visiting: https://www.instagram.com/nirmauni/
Visiting: https://www.linkedin.com/edu/school?id=156090&trk=tyah&trkInfo=clickedVertical%3Aschool%2CclickedEntityId%3A156090%2Cidx%3A2-2-6%2CtarId%3A1439363346148%2Ctas%3ANirma%20University
Visiting: https://www.youtube.com/nirmauniversityahmedabad
Visiting: https://nirmauni.ac.in/faculty-search/
Visiting: https://mis.nirmauni.ac.in/
Visiting: http://apps.nirmauni.ac.in/webdms/
Visiting: http://lms.nirmauni.ac.in/
Visiting: http://apps.nirmauni.ac.in/rpms/
Visiting: https://nirmauni.ac.in/visitors/about-ahmedabad/
Visiting: https://nirmauni.ac.in/covid-19/
Visiting: https://nirmauni.ac.in/visitors/fcra/
Visiting: https://formbuilder.ccavenue.com/live/kotak-mahindra/nirma-university
Not allowed to crawl https://formbuilder.ccavenue.com/live/kotak-mahindra/nirma-university due to robots.txt restrictions
Visiting: https://library.nirmauni.ac.in/
Not allowed to crawl https://library.nirmauni.ac.in/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/privacy-policy/
Not allowed to crawl https://nirmauni.ac.in/privacy-policy/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/disclaimer/
Not allowed to crawl https://nirmauni.ac.in/disclaimer/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/copyright/
Not allowed to crawl https://nirmauni.ac.in/copyright/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/terms-of-use/
Not allowed to crawl https://nirmauni.ac.in/terms-of-use/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/sitemap/
Not allowed to crawl https://nirmauni.ac.in/sitemap/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/disclosure/
Not allowed to crawl https://nirmauni.ac.in/disclosure/ due to robots.txt restrictions
Visiting: https://nirmauni.ac.in/mandatory-disclosure/
Not allowed to crawl https://nirmauni.ac.in/mandatory-disclosure/ due to robots.txt restrictions
Visiting: http://architecture.nirmauni.ac.in/admission-aid/under-graduate/barch/
Visiting: http://design.nirmauni.ac.in/admission-aid/under-graduate/b-des-communication-design/
Visiting: http://pharmacy.nirmauni.ac.in/admission-aid/under-graduate/bachelor-of-pharmacy/
Visiting: http://technology.nirmauni.ac.in/admission-aid/under-graduate/b-tech/
Visiting: http://management.nirmauni.ac.in/admission-aid/postgraduate/mba-and-mba-hrm/
Visiting: http://pharmacy.nirmauni.ac.in/admission-aid/post-graduate/master-of-pharmacy/
Visiting: http://science.nirmauni.ac.in/ms-c/
Visiting: http://technology.nirmauni.ac.in/admission-aid/post-graduate/mtech/
Visiting: http://cce.nirmauni.ac.in/
```