



# Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

Department of Artificial Intelligence & Data Science

## Report on One Day Workshop

**Title:** One Day Workshop on Web Retrieval and Crawling

**Date:** 27<sup>th</sup> September 2024

**Time:** 9.30 AM to 4.30 PM

**Venue:** Main Building, Date Centre Lab - 105.

**Target Audience:** B.E AIDS Students

**Number of Participants:** 68

### **Objectives of the Programme:**

1. To learn web crawling and its role in collecting data from the web
2. To extract text, image, video, ppt from web and save it in working directory
3. To learn Importance of respecting the robots.txt file
4. To learn strategies for dealing with various data formats.
5. To learn page ranking algorithm.

**Speakers:** Rushikesh Tanksale (Professional Freelancer, Piyush IT Services Pune)

**Inauguration:** The workshop started with an inauguration session by Rohini Naik (Asst Prof, AI&DS, VPKBIET), she welcomed the guest and all the participants. She introduced speaker to the audience and also discussed the rationale of the workshop programme and its key components. She motivated to take the benefit of skills through this workshop. Mrs. Rohini Naik addressed audience with her words of wisdom and expressed that all members have the ability and responsibility to change themselves to changing technology. She added that she is fully confident that teachers are always ready to and would successfully cope up with new challenges for the benefit of the students and institution.

The programme was one day Workshop on Web Retrieval and Crawling. It included two short sessions. Both sessions were delivered by the session speaker. It was a hands-on session.

**Session I:** Session I, of the programme focussed on “Basics of Web Retrieval and Crawling” and was delivered by Speaker Rushikesh Tanksale, Professional Freelancer Piyush IT Services Pune.

He explain what web retrieval is and why it's essential. Discuss the key concepts, including web resources, URLs, and HTTP requests. Getting Started with Python. He briefly introduced Python as the primary programming language for web crawling and guide participants through setting up their Python environment, making HTTP requests etc...Also, explained use of Python libraries like HTTP requests to web servers. Demonstrated how to retrieve web content (HTML) from a webpage and parsing HTML document which helped to extract meaningful information. Introduced BeautifulSoup as a popular library for parsing HTML. Guided participants through extracting data from a sample webpage, a hands-on example.

**Session II:** Demonstrated how to crawl the website with hands on IMDB website. Extracting the movie names and rating from the IMDB website. Cleaning the data and converting it to pandas dataframe for further analysis using Machine Learning technique.

**Session III:** In session III Speaker focussed on “Hands-On Web Crawling”

Speaker highlighted some main points on web crawling. He defined web crawling and its role in collecting data from the web. Explained with use cases for web crawling in various industries, Robots.txt and Ethical Crawling. A hands-on through creating a simple web crawler using Python. Setting up initial URLs, crawl through links, and store data. Handling Different Data formats. Strategies for dealing with various data formats, such as JSON and XML. Demonstrate parse and extract data from different file types. Dealing with Pagination and Infinite Scrolling. Also, explained with hands-on the page ranking algorithm. Encouraged participants to explore these areas further after the workshop.

After the session, area was open for Q&A and troubleshooting

The session ended with a vote of thanks by Dr. Chaitanya Kulkarni, HoD AIDS, VPKBIET. He thanked participants for their valuable time for the session. He also thanked all the participants for attending the sessions and believed that they will use the tools and technical skills in their day-to-day activities so as to make the process dynamic. Dr Chaitanya Kulkarni thanked the workshop co-ordinator for organising the workshop and smooth conduction of the same.

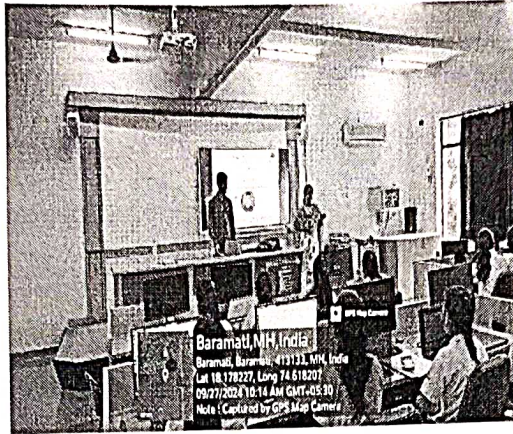
Speaker summarized the key takeaways from the workshop. Shared additional resources and references for participants to continue their learning journey.

**Outcomes of Programme:**

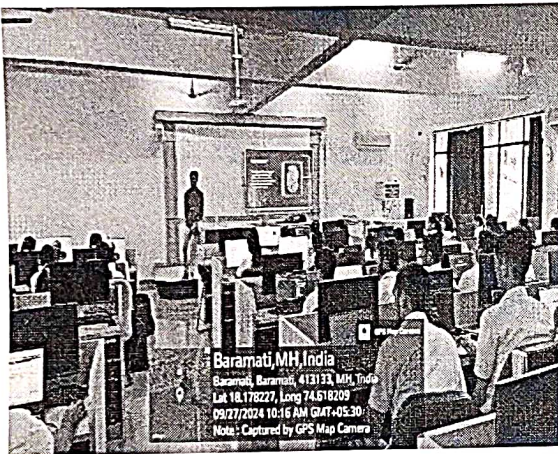
1. Learned Basics of Web Retrieval and Crawling
2. Understood techniques for handling web pages
3. Learned the importance of respecting the robots.txt file to avoid legal
4. Understood and implemented page ranking algorithm



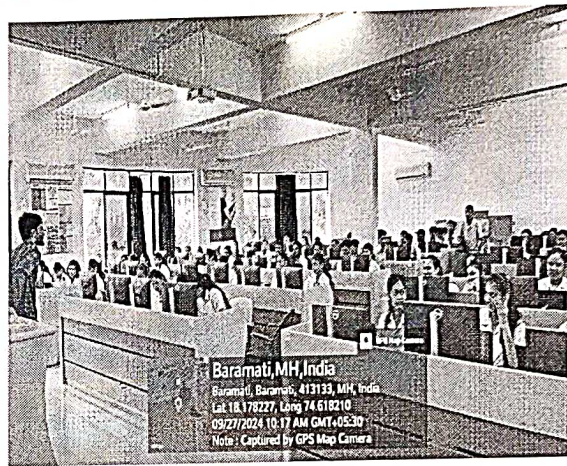
## Glimpses of the event



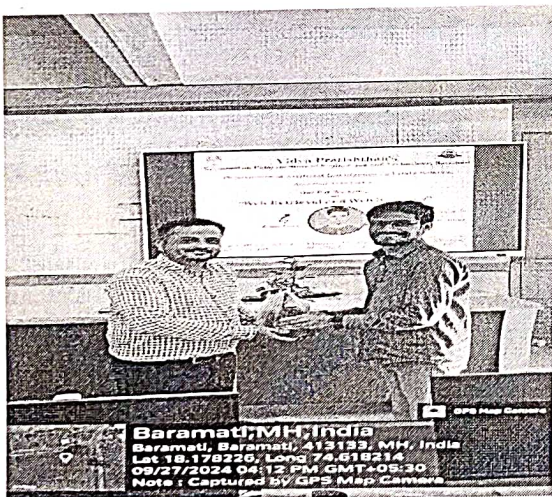
Inauguration and Introduction by Rohini Naik, Asst Prof



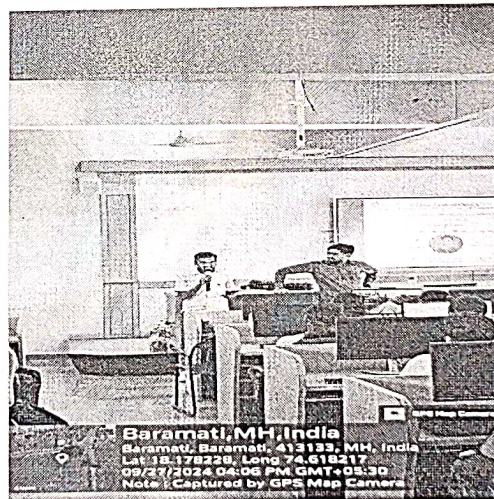
Workshop Speaker Rushikesh Tanksale



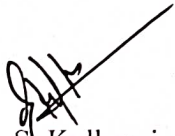
Session Participants B.E AIDS Students



Vote of thanks by Dr. Chaitanya Kulkarni, HoD



Students feedback

  
Dr. C. S. Kulkarni

HoD,  
Dept of AI&DS



Rohini Naik  
Assistant Professor  
Dept of AI&DS