

LABORATORY REPORT
Application Development Lab
(CS33002)

B.Tech Program in CSE

Submitted By

Name:- Shreeya Das

Roll No: 2205158



Kalinga Institute of Industrial Technology
(Deemed to be University)
Bhubaneswar, India

Spring 2024-2025

Table of Content

Exp No.	Title	Date of Experiment	Date of Submission	Remarks
1.	Build a Resume using HTML/CSS	16/01/2025	23/01/2025	
2.	Machine Learning for Cat and Dog Classification	23/01/2025	30/01/2025	
3.	Pneumonia Detection using CNN			
4.	Regression Analysis for Stock Prediction	30/01/2025	06/02/2025	
5.	Conversational Chatbot with Any Files	06/02/2025	20/02/2025	
6.	Web Scraper using LLMs	13/03/2025	20/03/2025	
7.				
8.				
9.	Open Ended 1			
10.	Open Ended 2			

Lab Number	8
Experiment Number	6
Experiment Title	Web Scraper using LLMs
Date of Experiment	13/03/2025
Date of Submission	20/03/2025

1. Objective:-

To create a web scraper application integrated with LLMs for processing scraped data.

2. Procedure:- (Steps Followed)

1. Use Python libraries like BeautifulSoup and Requests to scrape web data.

2. You can also use LlamaIndex for Web Scraping and Ollama for open ended LLMs.

3. Integrate LLMs to process and summarize the scraped information.

4. Develop a Flask backend for handling scraping tasks and queries.

5. Create an HTML/CSS frontend to initiate scraping (like the web page to scrape) and display results.

6. You can also take a topic and search the web for a web page and then scrape it.

3. Code:-

```

1  from flask import Flask, request, jsonify
2  import re
3  import requests
4  from bs4 import BeautifulSoup
5  from youtube_transcript_api import YouTubeTranscriptApi
6  from duckduckgo_search import DDGS # Corrected import
7  import google.generativeai as genai # Using Gemini API for LLM-based text processing
8  from dotenv import load_dotenv
9  import os
10 from flask_cors import CORS
11
12 # Initialize Flask app
13 app = Flask(__name__)
14 CORS(app) # Allow all origins
15 load_dotenv() # Load environment variables from .env
16
17 # Configure Gemini API
18 genai.configure(api_key=os.getenv("GEMINI_API_KEY"))
19
20 # Function to classify user input
21 def classify_input(user_text):
22     url_pattern = re.compile(r'https?://\S+')
23     youtube_pattern = re.compile(r'(https://)?(www\.)?(youtube\.com|youtu\.?be)/.+')
24
25     urls = url_pattern.findall(user_text)
26     youtube_links = [url for url in urls if youtube_pattern.match(url)]
27
28     if youtube_links:
29         return {"type": "youtube", "links": youtube_links}
30     elif urls:
31         return {"type": "web", "links": urls}
32     else:
33         return {"type": "search", "query": user_text}
34

```

```

5 # Web Scraper
6 def scrape_website(url):
7     try:
8         response = requests.get(url, timeout=5)
9         if response.status_code != 200:
10             return "Failed to retrieve page"
11
12         soup = BeautifulSoup(response.text, 'html.parser')
13         paragraphs = soup.find_all('p')
14         text = '\n'.join([p.get_text() for p in paragraphs])
15         return text[:5000] # Limit text size
16     except Exception as e:
17         return f"Scraping error: {str(e)}"
18
19 # YouTube Transcript Extractor
20 def get_youtube_transcript(url):
21     video_id = re.search(r"(?:v=|V/)([0-9A-Za-z_-]{11})", url)
22     if not video_id:
23         return "Invalid YouTube URL"
24     try:
25         transcript = YouTubeTranscriptApi.get_transcript(video_id.group(1))
26         return '\n'.join([t['text'] for t in transcript])
27     except:
28         return "Transcript not available"
29
30 # Web Search & Scrape (Improved)
31 def search_web(query):
32     with DDGS() as ddgs:
33         results = list(ddgs.text(query, max_results=3)) # Get up to 3 results
34
35     valid_content = []
36     for res in results:
37         if 'href' in res:
38             content = scrape_website(res['href'])
39             if content and "Failed" not in content:
40                 valid_content.append(content)
41
42     if valid_content:
43         return "\n".join(valid_content[:2]) # Combine max 2 sources

```

```

index.html U X app.py 1, M
templates > index.html > html > head > style
2 <html lang="en">
57 <body>
58     <div class="container">
59         <h1>webscrapper using LLMs</h1>
60         <textarea
61             id="userInput"
62             placeholder="Enter text, URL, or query..."
63         ></textarea>
64         <button onclick="processInput()">Process</button>
65         <div id="result"></div>
66     </div>
67     <script>
68         function processInput() {
69             const userInput = document
70                 .getElementById("userInput")
71                 .value.trim();
72             const resultDiv = document.getElementById("result");
73
74             if (!userInput) {
75                 resultDiv.innerHTML =
76                     "<p style='color: red;'>Please enter some text.</p>";
77                 return;
78             }
79
80             resultDiv.innerHTML = "<p>Processing...</p>";
81
82             fetch("http://localhost:5000/process", {
83                 method: "POST",
84                 headers: {
85                     "Content-Type": "application/json",
86                 },
87                 body: JSON.stringify({ text: userInput }),
88             })
89                 .then((response) => response.json())
90                 .then((data) => {
91                     resultDiv.innerHTML = `<p><strong>Summary:</strong> ${data.summary}</p>`;
92                 })
93                 .catch((error) => {
94                     resultDiv.innerHTML =
95                         "<p style='color: red;'>Error processing request.</p>";
96                     console.error("Error:", error);
97                 });

```

4. Results/Output:-

WebScraper using LLMs

India News

Process

Summary: Okay, here's a summary of the news snippets you provided, categorized by topic: **"Politics & Governance"** **"Farmers' Protest"** Jagjit Singh Dalewal, a cancer-stricken farmer, is on a hunger strike advocating for a legally guaranteed MSP. He met with the Union Agriculture Minister for talks. **"Jai and Jai Sram"** I alu Yadav appeared before the Enforcement Directorate in connection with the land for job scam. Tejashvi Yadav claims the cases are politically motivated. **"Bihar Politics"** Congress leaders in Bihar are saying that Kumar's appointment was done keeping in mind the 20 per cent Dalit population in the state. **"Election Commission"** Voters declining to share Aadhaar number will have to explain why. **"Manipur Crisis"** The Finance Minister announced a contingency fund of ₹500 crore for Manipur and stated that PM Modi addressed the issue in the House. NALSA justices are visiting relief camps. **"Pakistan & Terrorism"** India accuses Pakistan of actively promoting and sponsoring cross-border terrorism, calling it a roadblock to regional peace and security. **"Law & Order/Crime"** **"Airport Waste Dumping"** The New Delhi district magistrate flagged illegal waste dumping from the Indira Gandhi International Airport, citing potential safety and health risks. **"Churachandpur Clashes"** Clashes occurred in Churachandpur town in Manipur shortly after a peace settlement between communities. **"Militant Attacks in Jammu"** Militant attacks have been reported in the Jammu region over the past two years. **"Land Trespassing"** A complaint alleges contractors of KISS-KIT trespassed and dumped materials on private land. **"2021 Valpoi Protest FIR"** An FIR was lodged against individuals in connection with a 2021 protest, including charges of rioting and obstructing a public servant. **"Dehuli Village Massacre"** On November 18, 1981, 24 Dalits were killed in Dehuli village. **"Development & Infrastructure"** **"Adhokra Development"** Adhokra, Bihar, remains largely cut off from development compared to other districts. **"Indian Railways"** The Indian Railways has not increased fares since 2020. **"International Affairs"** **"Ukraine Conflict"** The EU has a clear commitment to continue supporting the war in Ukraine. **"Bangladesh Border Security"** Assam Police and BSF are to prevent illegal entry from Bangladesh. **"Science & Technology"** **"NASA Astronauts Return"** NASA astronauts Sunita Williams and Butch Wilmore returned to Earth after nine months in the space station. PM Modi praised their resilience. **"Gaganyaan Mission"** Group Captain Prashanth Balakrishnan Nair, Angad Prathap, Ajit Krishnan and Shubanshu Shukla have been selected for the Gaganyaan mission. **"Social Issues"** **"Assistive Devices Distribution"** Under a scheme, assistive devices worth ₹59.25 lakh were distributed in Salunbar. **"Environment"** **"Air Quality in Metro Cities"** The CSE's Urban Labs assessed trends in PM 2.5 concentrations in six metro cities, finding annual and seasonal changes between 2021-22 and 2024-25. **"Arts, Culture and Media"** **"Journalism Awards"** Journalists faced many challenges in pursuit of their stories that are nominated for the awards spanning a range of subjects. **"New Media Policy"** The New Media Policy of 2020 was met with opposition from both the media as well as political parties. **"Celebrity/Entertainment"** **"Kareena & Priyanka Friendship"** Kareena Kapoor Khan addressed her changing relationship with Priyanka Chopra, noting how friendships can shift in high-pressure environments.

5. Remarks:-

Signature of the Student

(Name of the Student)

Signature of the Lab Coordinator

(Name of the Coordinator)