CS242: Information Retrieval and Web Search

Project Report

1. Collaboration Details:
2. Abhishek Sharma: Lucene Indexing implementation
3. Gaurav Ratnakar: Lucene Indexing implementation
4. Neil Castellino: Web crawler implementation
5. Rajat M. Jain: Web crawler implementation
6. Overview of the crawling system:

Our crawler crawls website of [The New York Times](https://www.nytimes.com/) for news articles and stores them to our data set.

1. Architecture:
2. main.py:

This file is the top-level code or the entry point which acts as a driver and runs our crawler program. It imports our crawler module and initializes some important variables like:

first\_article\_link: The homepage of The New York Times

url\_per\_file: Which stores the URLs collected from the page

url\_limit = The stopping condition for our crawler

1. crawler.py:

The implementation begins with calling the crawl() method which very firstly initializes a queue. This queue is for the program to keep track of the URLs we visit and begin back from the last URL in cases where the program crashes or runs into errors resulting the termination of the process. The program then checks for the start\_url which when none, triggers to begin the program with the homepage or if start\_url is not none we get the list of the href for that starting URL.

1. The Crawling Strategy:
2. Overview of the Lucene indexing strategy