International Conference - Innovation-2021-Innovation-2021
International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307 (www.ijsrcseit.com)
Volume 8, Issue 3, May-June-2021

Search Engine Optimization and Report Generator

Anshuman Vats, Pranav Gholap, Pragati Tamboli, Kshitij Motke, Jayshree Chaudhari
Department of Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon, Pune, Maharashtra, India

ABSTRACT

In this project, we will be working on creating a fully automated SEO report generator based on the guidelines given by the search engines (Google/Bing/ Gropher) and create an indexing chart by going through the source code of the given website and ranking it in aspects of performance, SEO, best practices and availability. The secondary objective of the project is to recommend keywords based on the given website description (meta description tag from HTML file). To create the report we will need to rank the result that comes up after searching someone's name or their website and categorize them into three categories Good, Bad, and Critical. These can be flagged to the administrator team for content removal. For categorization, we will be building upon the Compromise NLP engine based on NODE JS environment.

Keywords : SEO, Automation, Selenium, Pagespeed Insight, Text classification

I. INTRODUCTION

A. Search Engine Optimisation

Search Engine Optimisation SEO stands for "search engine optimization." In simple terms, it means the process of improving your site to increase its visibility for relevant searches. The better visibility your pages have in search results, the more likely you are to garner attention and attract prospective and existing customers to your business.

B. How does SEO work?

Search engines such as Google and Bing use bots to crawl pages on the web, going from site to site, collecting information about those pages, and putting them in an index. Next, algorithms analyze pages in the index, taking into account hundreds of ranking factors or signals, to determine the order pages should appear in the search results for a given query.

Search ranking factors can be considered proxies for aspects of the user experience. Our Periodic table of SEO factors organizes the factors into six main categories and weights each based on its overall importance to SEO. For example, content quality and keyword research are key factors of content optimization, crawling ability and mobile-friendliness are important site architecture factors.

The search algorithms are designed to surface relevant, authoritative pages and provide users with an efficient search experience. Optimizing your site and content with these factors in mind can help your pages rank higher in the search results.

C. Web Automation

Website automation is a way to automate common web actions-like filling out forms, clicking on buttons, downloading files, and hands them over to helpful software bots. While the internet makes doing business faster and easier in countless ways, these actions can be time-consuming and prone to errors

II. PROBLEM STATEMENT

The current process of SEO optimization and report generation is a manual process. Where we search for their online reputation, search results get categorized based on textual context and the effect on one's reputation. If they have a personal website or organization's website we go to that website and based on the search engine guidelines (Google/ Bing/ Gropher) we rank the website in various aspects. After collecting all the data. A report is generated that is then delivered to the client.

Pain points:

- Manual process.
- Report generation is a repetitive task.
- Data collection from various sources is time-consuming

III. OBJECTIVE

Our objective is to Analyzing website against SEO guidelines to create the SEO ranking for a given website further Creating a text classifier to categorize the fetched result from search engines against someone's name/organization thereafter Automating the Google docs using Docs API to generate the required report based on the extracted data from previous steps and deliver it in a custom-designed google doc.

Once all the data has been collected and sorted we will use the Google Docs automation process using Python to create reports that then can be delivered in pdf format.

1. Data collection for SEO listing
2. Categorizing search engine result using NLP
3. Report automation

This is a complete solution delivery project in the scope of Full stack development based on technologies SEO, Automation, and Natural Language Processing.

The final product is going to follow the API first architecture.

IV. DISCUSSION

A. Project Scope

Creating a fully automated system for SEO report generation and optimization. Consisting of - Text classifier (NLP), Google docs Automator and web scraper (Selenium/Scrappy/Puppeteer).

- Web application adhering to the PWA standards.
- Codebase: Microservice Architecture
- Custom NLP model
- Deployed on AWS

B. Functional Requirements

The system should provide an interface where the user can enter the name of the website for which he/she wants to get the search engine ranking and page speed insights.

Visualization of the score based on the result returned from the servers and suggestion to get a higher score if not sufficiently based on the business requirements.

An auto report generator to provide the end-user with the ORM sheet that can later be used to pitch the client.

All this needs to be done using a headless browser instance running in the background and calculating all the scores by running the required test on the given website provided by the user.

C. Non-Functional Requirements

1) Performance Requirements

The model should efficiently work as per the given conditions and evolve accordingly

2) Safety Requirements

This model does not possess the capabilities to be used in a negative manner and as such there will be no repercussions for it.

3) Security Requirements

The model should be secure as it does not contain any components that may possess a risk factor.

D. System Requirements

1) Software Requirements

- VUE js
- Node
- Modern Browser (chromium-based preferred)

2) Hardware Requirements

- Processor i3 5th Gen & above
- Hard Drive: 10GB
- RAM Size: 4GB

V. SYSTEM DESIGN

A. Models

Initializing all the required systems in the NODE environment.

Module 1: Performance testing

We take the input from the user and perform various test to finally conclude how does the website perform in different scenarios and give insights on how to take correct measures

Module 2: SEO testing

We test the website against the set SEO parameter for which we assign a rank and give input on how to improve the SEO rank

Puppeteer JS (NPM)

```
const puppeteer = require('puppeteer');

(async () => {
  const browser = await puppeteer.launch();
  const page = await browser.newPage();
  await page.goto('https://example.com');
  await page.screenshot({path: 'example.png'});
  await browser.close();
})();
```

Module 3: User reputation (Text classification)

We search for the user's name and see if the search result contains -ve or +ve result based on our NLP engine

Natural JS

```
// Configuring Natural js for natural language processing
var Analyzer = natural.SentimentAnalyzer;
var stemmer = natural.PorterStemmer;
var analyzer = new Analyzer("English", stemmer, "afinn");
var tokenizer = new natural.WordTokenizer();
```

Module 4: Report generation

Using print.js for exporting all the JSON data that can be pitched to the clients.

B. Process Diagram

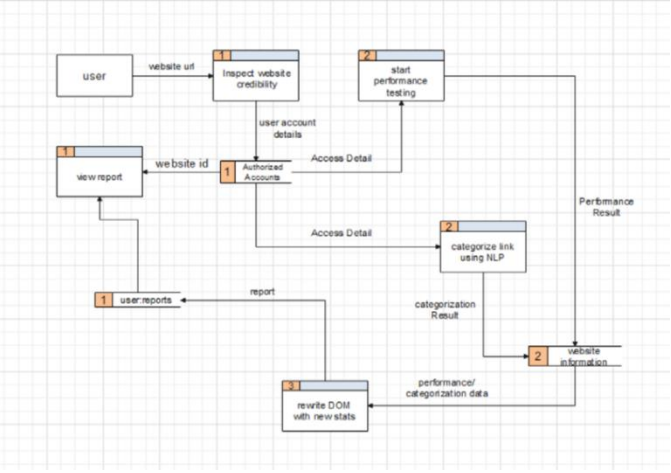

figure 2: Process Diagram

C. System Architecture

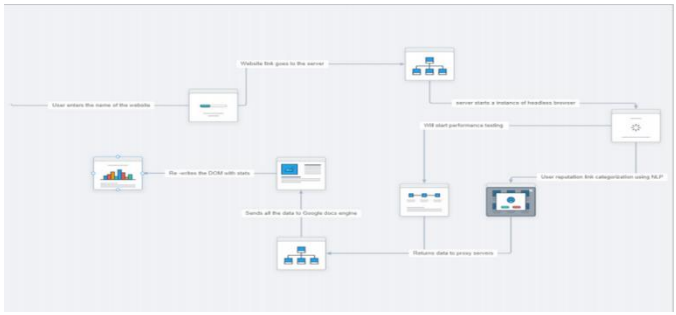

figure 3: System Architecture

D. Functionality

The user enters the name of the website for which we will be doing the analysis.
That website link is sent to the backend server and this where all the processing will be carried out.
Once the name is received the backend server will start to create a headless browser instance.
In these instances, we will do our performance and SEO testing. There will be another instance that will go on to check the search result on Google and then mark and categorize the response to Good, Bad, or Critical.
All this raw data will be sent to a proxy server which will then convert this data into a JSON object.
To access the google docs engine we need a proxy server because the remote server doesn't allow direct access to themselves hence our proxy server will feed the data to the Google engine which will, in turn, give us the required report.

VI. ADVANTAGES

Unlike traditional "outbound" advertising channels, which involve reaching out to consumers whether they want to hear from you or not, inbound methods center on making it easy for your audience to find you when they want information.
Google's organic rankings are based entirely on what its algorithm determines to be the best results for any given query. This means that once you've created a page that the search engine deems worthy of directing their users to, it can continue to attract traffic to your site for months (or even years) after you publish it. Of course, researching and writing high-quality content requires an investment. That investment will either be in the form of time if you choose to create it yourself or money if you choose to hire a digital marketing agency to create it for you.
It's difficult to say why this is, though the most logical conclusion is that users trust Google's algorithm. They know which spots advertisers are paying for, and they choose to instead visit the pages the search engine has determined to be the best.
Earning links from reputable websites is a main component of any SEO strategy. This means that one of the biggest parts of an SEO professional's job is to identify opportunities for placement or coverage on industry blogs, news publications, and other relevant sites

VII. RESULT

We tested our system on one of the most popular website "Microsoft - Official Home Page" and the co-founder of Microsoft "Bill Gates" following were the results as of 3rd June 2021 from India:

1) Performance Testing

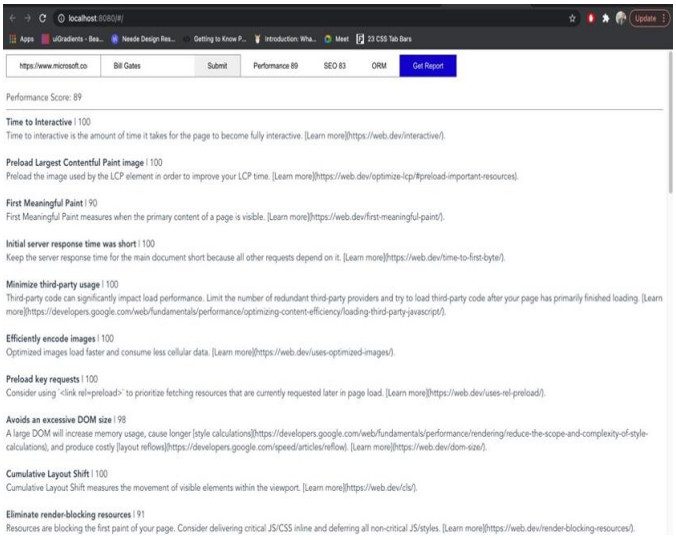

figure 4: Performance testing result

2) SEO Testing

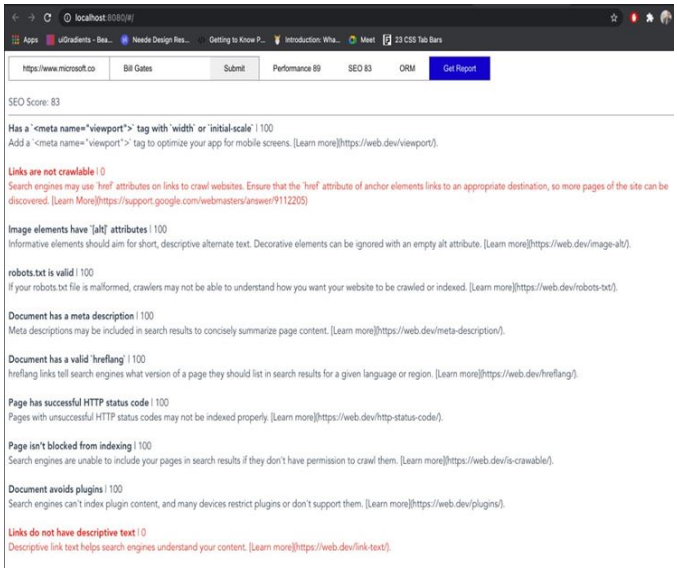

figure 5: SEO testing result

3) ORM Testing

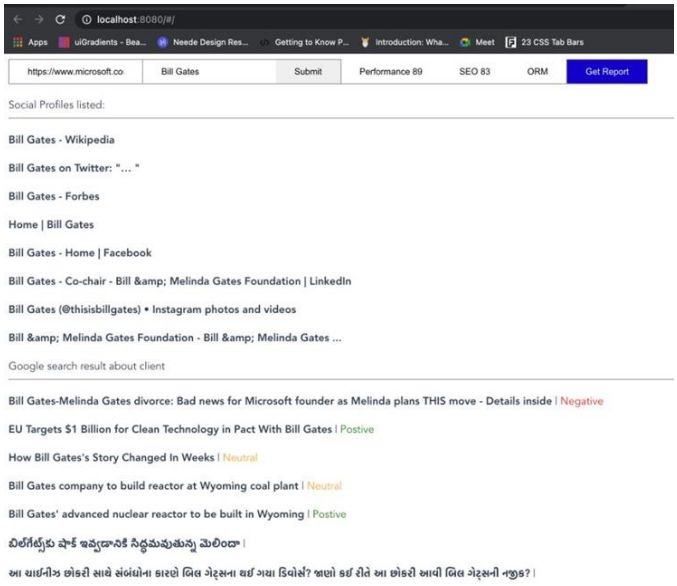

figure 6: ORM testing result

VIII. CONCLUSION & FUTURE SCOPE

In this, we have successfully built an automated system that can perform automated performance testing, SEO testing and also use an NLP engine to dissect whether a person or company's reputation is categorically good, bad, or critical.

This could be scaled up to a full-fledged business solution to provide search engine optimization solutions and online reputation management and can be set up as a microservice to create a subscription-based system.

IX. ACKNOWLEDGMENT

It gives us great pleasure in presenting the paper on "Search Engine Optimisation and Report Generator". We would like to take this opportunity to thank Dr. Pankaj Agarkar, AP, and Head of Computer Engineering Department, DYPSOE, Pune for giving us all the help and support we need during the course of the Paper writing work. We are grateful to him. Our special thanks to Dr. Ashok Kasnale, Principal DYPSOE who motivated us and created a healthy environment for us to learn in the best possible way. We also thank all the staff members of our college for their support and guidance

X. REFERENCES

[1]. D. Pratiba, Abhay M.S, Akhil Dua, Giridhar K. Shanbhag, Neel Bhandari, Utkarsh Singh, "SEO TECHNIQUES FOR VARIOUS APPLICATION-A COMPARATIVE ANALYSIS AND EVALUATION," Published in IEEE 20-22 Dec 2018

[2]. Gowtham Aashirwad Kumar, Dr. A Ravikumar, "AN ANALYTICAL STUDY OF SEARCH ENGINE OPTIMIZATION (SEO) TECHNIQUES: TO MAXIMIZE NUMBER OF TRAVELERS ON AN E-CONTENT MATERIAL WEBSITE," Published Volume 11, Issue 1, January 2020

[3]. Peng Qi*, Yuhao Zhang*, Yuhui Zhang, Jason Bolton, Christopher D. Manning, "A PYTHON LANGUAGE PROCESSING TOOLKIT FOR MANY HUMAN LANGUAGES," Published Stanford University Stanford, CA 94305 23 Apr 2020