Skill Development Portal for Targeted Audience

Krishna Vinay 201501222
G. Sai Sriragh 201501113
Instructor :- Prof. P Krishna Reddy

Problem Statement :-

Most of the people are not able to find various skill development programmes by easy surfing. They feel it as a difficult task to know about the suitable skills for the qualification. In this project, the target audience are the people who are interested in learning skills for their employment opportunities. We develop a online portal which suggests suitable skills and the sources for learning them easily to the users.

This portal will help the users to fulfil their needs at one place. They can have access to different resources for learning a particular skill without having to search the web to get a suitable website.

These resources can be acquired by focused crawling. Once we get the list of resources for every skill we need to build the portal which suggests the skills based on the user's input.

Project Requirements

- We aim to build a skill development portal which targets various kinds of audience and will advise them on what skills they can develop or achieve.

- To do this we need to extract data from various internet sites and store them in a database and then give advice to that particular audience based on his personal data which we will store as well.

Goal of the Project

- We need to ask the user various questions and based on the answers we advice them on probable skills to be developed.

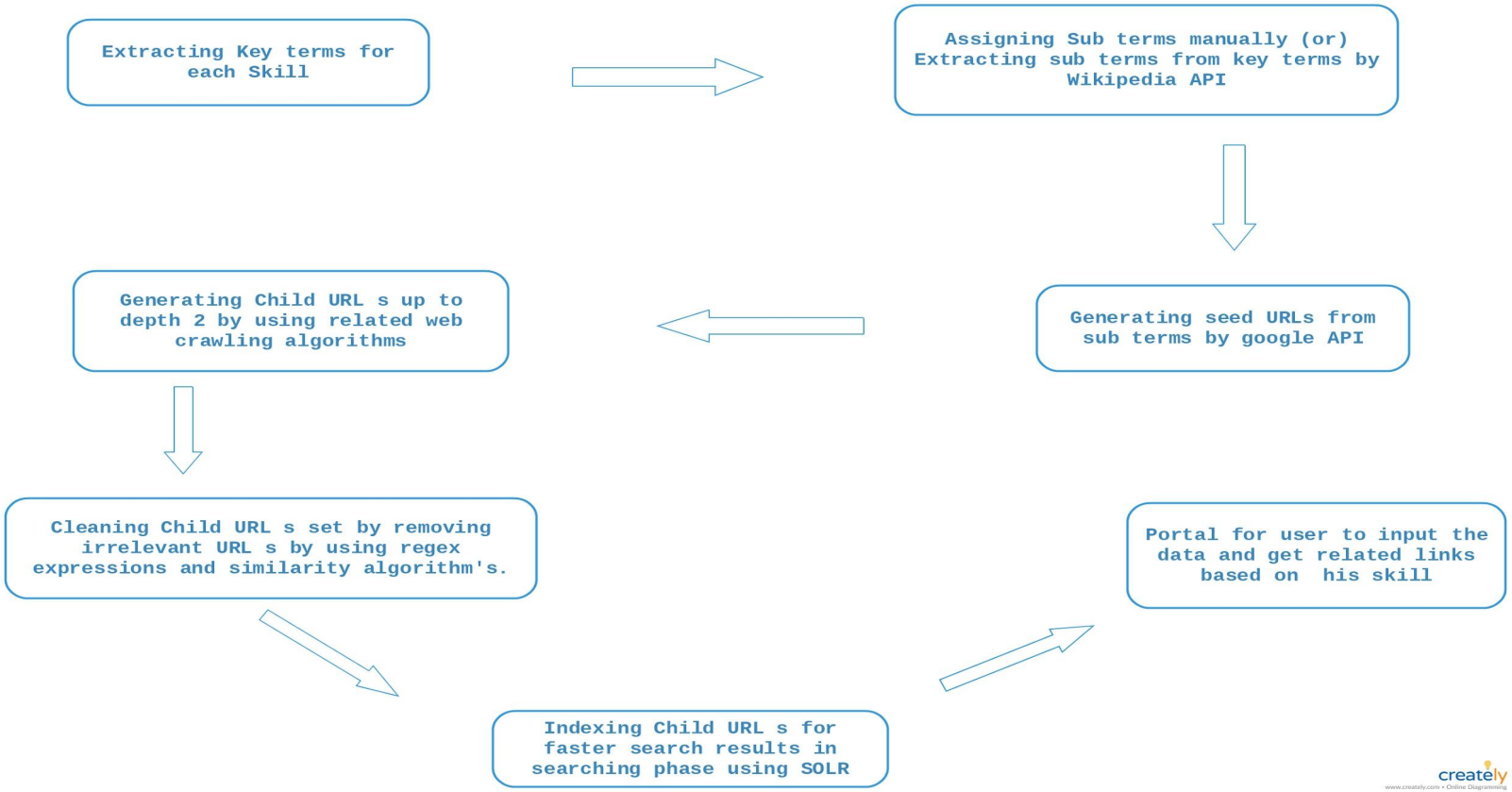
Project Plan - “Module-1”

- Analyze the skill development domain and collect all related URL s for web crawling
- Develop a system and method to crawl publicly available web-based resources for harvesting occupation-specific skill development related content.

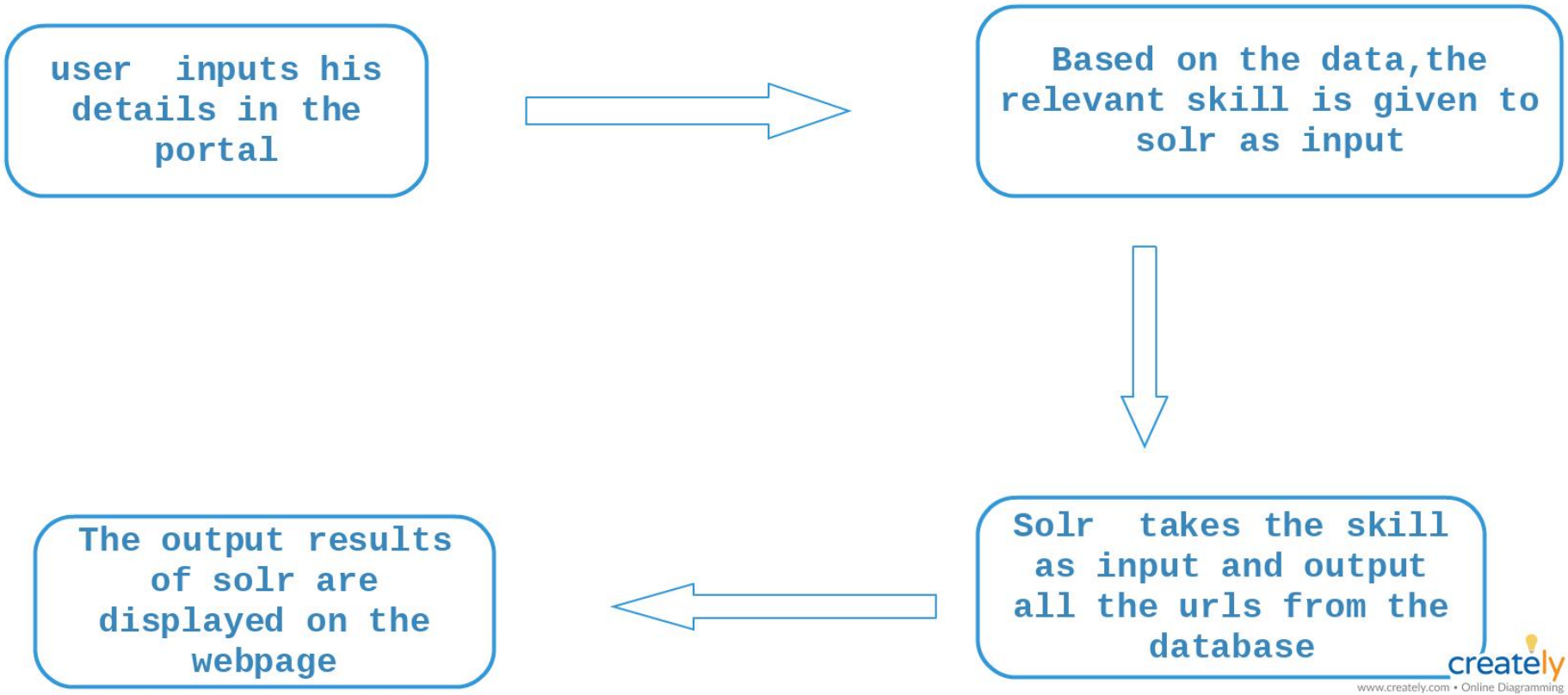
Project Plan - “Module-2”

- Indexing the URL s which we get from web crawling and develop a search engine to query a large database efficiently
- Deploy a Portal where user can submit his/her own data and get suitable URL s based on his skill

BLOCK DIAGRAM FROM BACK END POINT VIEW



FLOW DIAGRAM FOR USER's PERSPECTIVE



Explanation of the Block Diagram Step by Step

Extracting Term from Each Skill

We divided the Skill domains into 6 important subdomains.

Like, mostly unemployed come from Agricultural background and even though they are unemployed they may excel in some skills like repairing works, designing, Computer data entering etc../.

Based on these we divided the level1 skills. In the next level we extend these words and get more terms called as level2 terms.

Extracting subterms using level1 terms by wikipedia API

We use wikipedia python API to get some relevant terms linked to main terms. We used them for more specific seed URL s.

```
→   k =  wikipedia.page(main_term).links

    unicodedata.normalize('NFKD', k).encode('ascii','ignore')
```

Generating seed URL s by using google API s

Using many subterms which are generated by the level1 terms we now need to get the seed URL s means the important parent URLs for crawling the data with utmost relevance

```
# need to get google API and google search APIs

→  for j in search(sub_terms, tld="co.in", num=10, stop=1, pause=2):
            file_for_seed_urls.write(j+"\n")

WHY only google API?
```

Generating child URL s upto depth 2 by efficient web crawling techniques

Web Crawling Algorithm:-

- Remove a URL from the URL list. (from the file which contaning seed URL s)
- Download the corresponding page.
- Check the Relevancy of the page.
- Extract any links contained in it.
- Add these links back to the URL list.
- After getting all the URL s we need to apply the relevance algorithm to remove all the irrelevant links which are not useful to us.

- We use requests API and beautifulsoup library for downloading any html file and for later indexing it with href’s and get URL s which are as http but not general path showing url s.

```
→   html_ page  = requests.get(url)

    soup = BeautifulSoup(html_page.content)

 for link in soup.findAll('a',attrs={'href': re.compile ("^https://")}):
         links.append(link.get('href'))
```

Cleaning child URL s by removing irrelevant ones

- Generally, in a page there are so many urls which are not connected to our topic like ad’s and social media links and also like google forms and login, signin signup pages.
- To remove these types of links, we use url parsing library to get domain of the link and the path. Based on the path and we can filter the links which are relevant or not. And also we need to maintain visited set to get rid of duplicates.

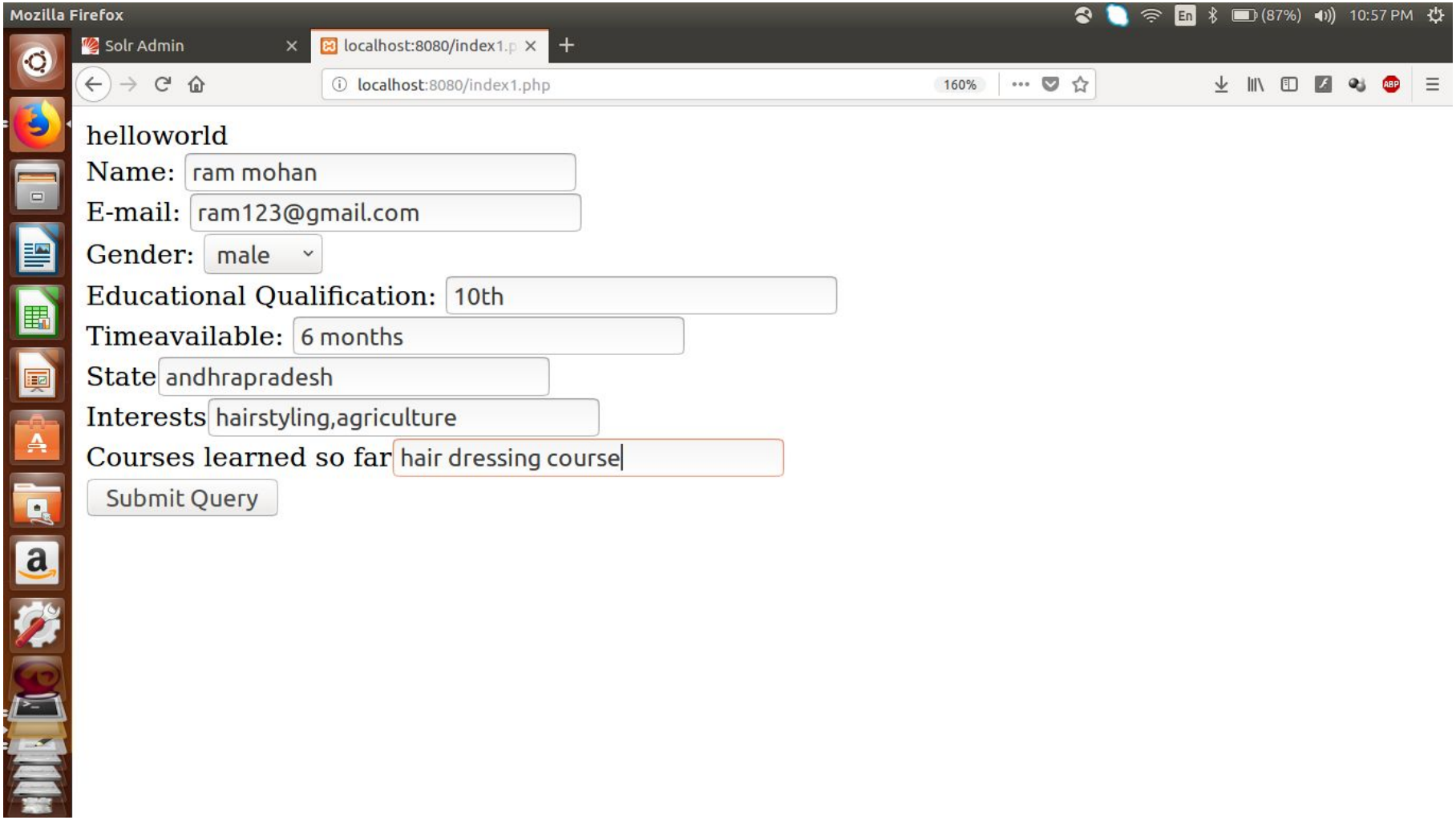
Indexing child urls for faster search results while querying

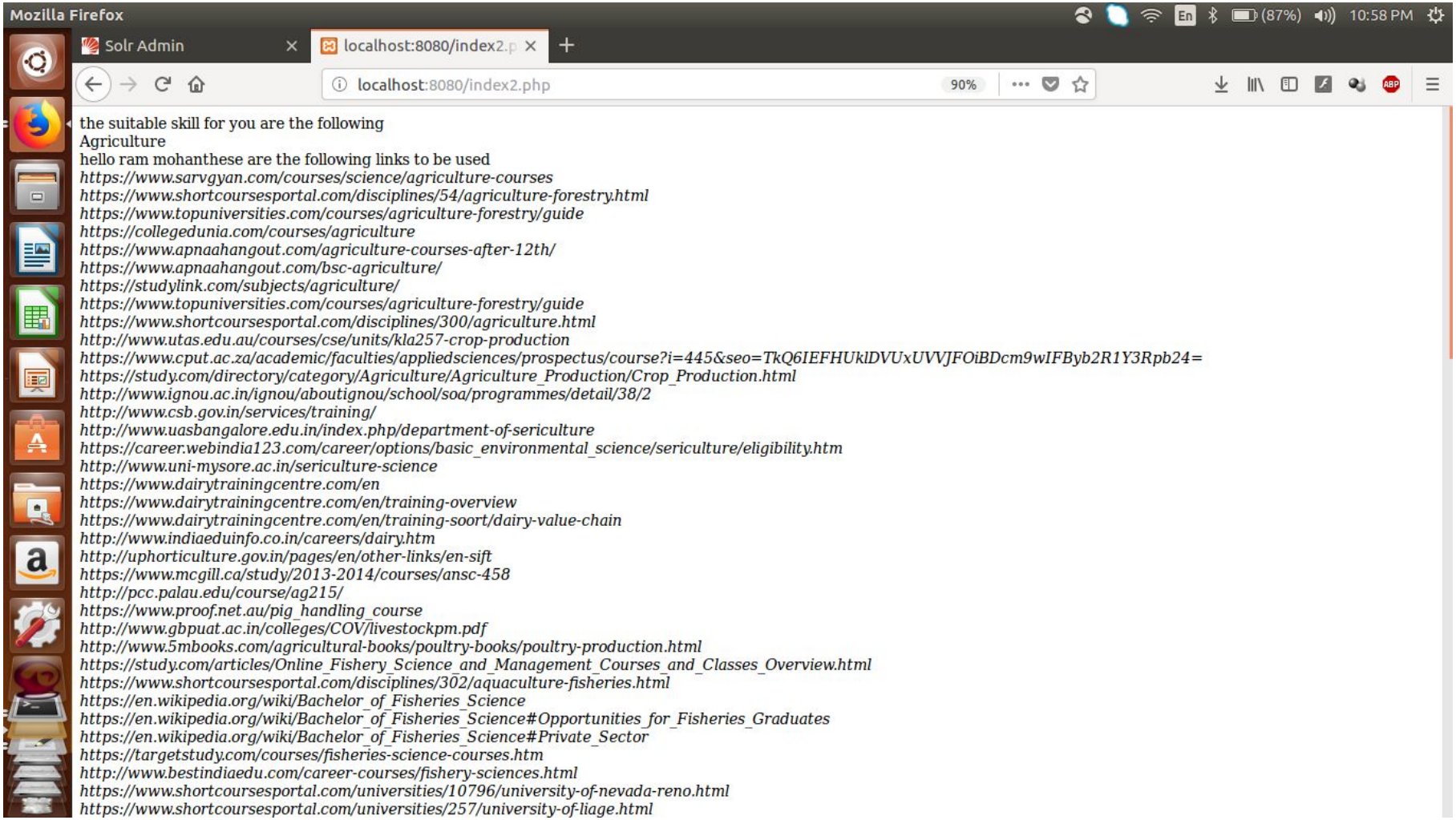
- The child urls obtained during the web crawling process are indexed according to the type of domain the url belongs to.
- For the purpose of indexing and querying we used solr. The indexing is done by solr and search can be done based on the type of domain we are interested in.

Portal for user to input the data and get relevant urls based on his skill

- Finally we designed a portal which takes user's data such as his interests,qualification,previous courses done as input and gives relevant urls based on the suitable skill he can learn.

- For this purpose we have used php for designing portal.Based on the skills he is suitable we output the related urls by querying the solr according to domain and displaying the query results of solr on to the webpage.


Mozilla Firefox
Solr Admin
localhost:8080/index1.p
localhost:8080/index1.php
160%
helloworld
Name: ram mohan
E-mail: ram123@gmail.com
Gender: male
Educational Qualification: 10th
Timeavailable: 6 months
State andhrapradesh
Interests hairstyling,agriculture
Courses learned so far hair dressing course
Submit Query



the suitable skill for you are the following
Agriculture
hello ram mohanthese are the following links to be used
https://www.sarvgyan.com/courses/science/agriculture-courses
https://www.shortcoursesportal.com/disciplines/54/agriculture-forestry.html
https://www.topuniversities.com/courses/agriculture-forestry/guide
https://collegedunia.com/courses/agriculture
https://www.apnaahangout.com/agriculture-courses-after-12th/
https://www.apnaahangout.com/bsc-agriculture/
https://studylink.com/subjects/agriculture/
https://www.topuniversities.com/courses/agriculture-forestry/guide
https://www.shortcoursesportal.com/disciplines/300/agriculture.html
http://www.utas.edu.au/courses/cse/units/kla257-crop-production
https://www.cput.ac.za/academic/faculties/appliedsciences/prospectus/course?i=445&seo=TkQ6IEFHUklDVUxUVVJFOiBDcm9wIFByb2R1Y3Rpb24=
https://study.com/directory/category/Agriculture/Agriculture_Production/Crop_Production.html
http://www.ignou.ac.in/ignou/aboutignou/school/soa/programmes/detail/38/2
http://www.csb.gov.in/services/training/
http://www.uasbangalore.edu.in/index.php/department-of-sericulture
https://career.webindia123.com/career/options/basic_environmental_science/sericulture/eligibility.htm
http://www.uni-mysore.ac.in/sericulture-science
https://www.dairytrainingcentre.com/en
https://www.dairytrainingcentre.com/en/training-overview
https://www.dairytrainingcentre.com/en/training-soort/dairy-value-chain
http://www.indiaeduinfo.co.in/careers/dairy.htm
http://uphorticulture.gov.in/pages/en/other-links/en-sift
https://www.mcgill.ca/study/2013-2014/courses/ansc-458
http://pcc.palau.edu/course/ag215/
https://www.proof.net.au/pig_handling_course
http://www.gbpuat.ac.in/colleges/COV/livestockpm.pdf
http://www.5mbooks.com/agricultural-books/poultry-books/poultry-production.html
https://study.com/articles/Online_Fishery_Science_and_Management_Courses_and_Classes_Overview.html
https://www.shortcoursesportal.com/disciplines/302/aquaculture-fisheries.html
https://en.wikipedia.org/wiki/Bachelor_of_Fisheries_Science
https://en.wikipedia.org/wiki/Bachelor_of_Fisheries_Science#Opportunities_for_Fisheries_Graduates
https://en.wikipedia.org/wiki/Bachelor_of_Fisheries_Science#Private_Sector
https://targetstudy.com/courses/fisheries-science-courses.htm
http://www.bestindiaedu.com/career-courses/fishery-sciences.html
https://www.shortcoursesportal.com/universities/10796/university-of-nevada-reno.html
https://www.shortcoursesportal.com/universities/257/university-of-liage.html

What we have learnt

1) We have studied nearly 4 research papers for understanding different web crawling techniques and efficient identification of relevant child URL s
2) Understanding so many skill development terms and to find the best seed urls for them. And also using different API s at different levels to get required outputs

These are the main research related things we learnt from this project and there are also so many software related things which we learnt like indexing and solr searching and so many.

What are the struggles we faced?

1) First, to find seed URL s. As if we give manually only 2 to 3 URL s then whole crawling algo should will to those and get the limited links???
2) The main and important problem was removing the irrelevant URL s from the bunch which we get as output from web crawling algo.
3) Searching the large database file for getting the URL s relevant to specific domain.

What are the solution we come up with??

1) Wikipedia API s used to increase the key terms and then getting the large URL s from google search API.
2) Upon reading so many web crawling techniques like dfs,bfs,incremental and maxflow and mincut algos we choose bfs web crawling technique.
3) To query faster we used SOLR for indexing the data and querying.

References:-

1) Fine Grained Approach for Domain Specific Seed URL Extraction by Lalit Mohan S
2) Focused crawling: a new approach to topic-specific Web resource discovery by Soumen Chakrabarti
3) Efficient Identification of Web Communities by G FLAKE

THANK YOU (: