Building A Search Engine

By Sahil Mutneja
under the mentorship of Dr. Dileep A.D.

Problem Statement

- Search Engine that provides the user with a list of experts when queried against a specialization
  - The experts will be from the premium Institutes(IIT's, NIT's) spanning across India
- The list will be such that the most relevant of the result will be towards the top.

Final Outcome

- The result will comprise of the links of web pages of the specialist along with the top-notch work done on the queried specialization.
- After analysing the crawled content, we will be building a web graph/network wherein nodes will be the specialization and the edges will be the links of the webpages or some relevant content

Description of the project

Decomposing the Problem Statement gives us three subparts:

1. Web Crawling
2. Indexing
3. Searching

Plan as per the last review meeting

Work to be Completed	Tentative Date	Current Status
Build a Crawler	11th October 2014	1. Built a crawler that crawls the relevant content of IIT Mandi
Formatting via Parser	7th November 2014	1. Added a parser that checks for keywords in the url, content of the page & filters out the best 2. Formed a Indexer wherein (content, url) pair is the basis 3. Devised a lookup mechanism in which when queried with a keyword, output is the list of web pages of relevant specialist.

Web Crawling and Parser

```
def crawl_web(seed):
    #procedure to crawl the whole of web
    tocrawl = [seed]#starts with the page entered by the user
    crawled = []#will list all the pages that are crawled
    index = []#will contain the word to url mapping
    while tocrawl:
        link = tocrawl.pop()
        #page stores the link of the last popped out item from the tocrawl list
        if link not in crawled:
            #to avoid repetition checking if the page is already explored
            #it avoids cycles
            content = get_page(link)#stores the content of the page
            indexing.add_page_to_index(index, link, content)#add it to indexer
            union(tocrawl, get_all_links(content, link))
            crawled.append(link)
    #will list all the pages that will be crawled starting from the seed page
    return index
```

Seed Page from which crawling begins

http://www.iitmandi.ac.in/institute/faculty.html

Faculty

- School of Computing and Electrical Engineering
- School of Basic Sciences
- School of Engineering
- School of Humanities and Social Sciences

Indexer

- For the initial testing purpose, rather than crawling the entire nation institutes, we are only focussing on IIT Mandi.
- Once we start dealing with a distributed system, we will start crawling a wide number of institutes that we have initially targeted.
- A list data structure is formed where first half is the content from the page and the second half is the url it is present in.
- When queried against a specialization, list lookup will start (currently **O(n)**) in order to scan all the pages that contains the searched keyword.
- For IIT Mandi, list contains 18,377 entries in the Indexer.

Indexing & Keyword Lookup

```
def add_page_to_index(index, url, content):

    lines = content.split('\n')
    #content is divided based on new line

    for line in lines:
        #every line is checked and then added to indexer DS
        add_to_index(index, line, url)
```

```
def add_to_index(index, line, url):

    for entry in index:
        if(line in entry[0] and url not in entry[1]):
            #will append the url if the line already exists
            entry[1].append(url)
            return
    #will add a new item in the list along with the URL
    index.append([line, [url]])
```

```
def lookup(index, keyword):

    links = []#will store the links where the keyword is found

    for entry in index:#will check the first half of evey value in index
        if( keyword.lower() in entry[0].lower() and entry[1] not in links):
            links.append(entry[1])#will add the link in the links list

    #after traversing the entire indexer will return the links list
    return links
```

Keyword Lookup on the formed Indexer

```
Indexer is ready, Enter your string :: Image Processing


Links with the keyword Image Processing
['http://faculty.iitmandi.ac.in/~rajendra/']
['http://faculty.iitmandi.ac.in/~sarita/']
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~padman/pubs.html']
['http://faculty.iitmandi.ac.in/~addileep']
['http://faculty.iitmandi.ac.in/~arnav']
['http://faculty.iitmandi.ac.in/~arnav', 'http://faculty.iitmandi.ac.in/~anil/']
['http://faculty.iitmandi.ac.in/~anil/']
```

```
Indexer is ready, Enter your string :: Machine Learning


Links with the keyword Machine Learning
['http://www.iitmandi.ac.in/institute/faculty_bs.html']
['http://faculty.iitmandi.ac.in/~manoj/']
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~addileep']
['http://faculty.iitmandi.ac.in/~arnav']
```

```
Indexer is ready, Enter your string :: Pattern Recognition


Links with the keyword Pattern Recognition
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~padman/']
['http://faculty.iitmandi.ac.in/~addileep']
['http://faculty.iitmandi.ac.in/~arnav']
['http://faculty.iitmandi.ac.in/~anil/']
```

```
Indexer is ready, Enter your string :: Computer Networks


Links with the keyword Computer Networks
['http://faculty.iitmandi.ac.in/~arti/']
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~anand/']
['http://faculty.iitmandi.ac.in/~tag/']
['http://faculty.iitmandi.ac.in/~samar/']
```

```
Indexer is ready, Enter your string :: Kernel Methods


Links with the keyword Kernel Methods
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~addileep']
```

```
Indexer is ready, Enter your string :: Communication


Links with the keyword Communication
['http://faculty.iitmandi.ac.in/~varun/', 'http://faculty.iitmandi.ac.in/~varun']
['http://faculty.iitmandi.ac.in/~sudhir/']
['http://faculty.iitmandi.ac.in/~shekhar/']
['http://faculty.iitmandi.ac.in/~ajay/']
['http://faculty.iitmandi.ac.in/~vkn/']
['http://faculty.iitmandi.ac.in/~arti/']
['http://faculty.iitmandi.ac.in/~prasanth/']
['http://faculty.iitmandi.ac.in/~abhimanew/']
['http://faculty.iitmandi.ac.in/~achakraborty/']
['http://faculty.iitmandi.ac.in/~abbas/']
['http://www.iitmandi.ac.in/institute/faculty_c&ee.html']
['http://faculty.iitmandi.ac.in/~anand/']
['http://faculty.iitmandi.ac.in/~bsr/']
['http://faculty.iitmandi.ac.in/~padman/pubs.html']
['http://faculty.iitmandi.ac.in/~ramesho/']
['http://faculty.iitmandi.ac.in/~tag/']
['http://faculty.iitmandi.ac.in/~addileep']
['http://faculty.iitmandi.ac.in/~arnav']
['http://faculty.iitmandi.ac.in/~samar/']
['http://faculty.iitmandi.ac.in/~anil/']
```

Things in Hand and Future Targets

- Working on building a search engine covering IIT Mandi in its domain.
- Built an Indexer for the same which supports addition and lookup operations.
- Search Optimization i.e. making lookup faster than O(n).
- Ranker that arranges the pages on the basis of their ranks.
- Indexer enhancement via looking into ML, text classification techniques and much more.
- Making the system scalable to a large number of machines for storing the processed indexer data.
- After ensuring the scalability, focus will be on the front end and the improvement of existing structure.

Course of Action with Time Frame (7th and 8th Sem included)

Work to be Completed	**Tentative Date**
Research on Text Classification, ML Algorithms for Indexer Enhancement and Algorithms for Ranker	In Winter Vacations
Search Optimization Techniques	7th March 2015
Implementation of Ranker and Searcher for IIT Mandi	15th March 2015
Extending domain of Crawler and Parser to more institutes and build Indexer	27th March 2015
Setting up of Distributed System for huge Indexer data	15th April 2015
Ranker Collaboration with large data	19th April 2015
Building Front End and Final Touches	30th April 2015