2017 B.Tech CSE Batch

Literature Review Document

Approved Problem Statement: To automate the process of Hardware Recon

Zeroth review minutes of the meeting:

We have discussed the problem statement and our proposal with our mentor and finalised the problem statement.

Group Members

Name	Roll No.
Aathira Dineshan	AM.EN.U4CSE17101
Kalpesh Gupta	AM.EN.U4CSE17135
Amrita Nair	AM.EN.U4CSE17504
Jishnu Ganesh	AM.EN.U4CSE17133

Amrita Nair

Field -> Area-> Sub-area->topic: Text extraction from images -> OCR and image manipulation techniques

Literature Review:

Guide approval of selected papers:

S#	Name	Roll Number	Paper Title	URL/Reference Link of the paper selected	IEEE/ACM Publication	Year	Scopus Journal with impact factor/ Transaction/ conference	How the selected paper is relevant for your project?
1	Amrita Nair	AM.EN.U4CSE 17504	An Overview of the Tesseract OCR Engine	https://ieeexplore.ieee.org/document/4376991/	IEEE	2007	ICDAR 2007	This paper walks us through the working of the tesseract and tells us how tesseracts works on a given image, giving us an idea on what how to pre processing the data
2	Amrita Nair	AM.EN.U4CSE 17504	Deep Statistical Analysis of OCR Errors For Effective Post-OCR Processing	https://ieeexplore.ieee.org/document/8791206	IEEE	2019	ACM/IEEE 2019 JCDL	This paper helps in understanding all the possible errors post OCR and gives us an idea on how to overcome such errors
3	Amrita Nair	AM.EN.U4CSE 17504	Adaptive thresholding: A comparative study	https://ieeexplore.ieee.org/document/6993140	IEEE	2014	ICCICCT 2014	This paper helps analyzing how adaptive thresholding works better compared to other threshold techniques like Watershed and Otsu.

Key Points:

1. **An Overview of the Tesseract OCR Engine**
    - The paper describes in brief how the Tesseract OCR engine recognises words and characters from an image.
    - The engine usually follows four major steps - Processing, Recognition, Classification and the Final phase.
    - Processing
        - This involves finding the outline of the components, followed by doing a page layout analysis, to figure out the height which assists in differentiating the punctuations from letters. (using a height filter)
        - The base lines of the text are and these lines are further help in finding if it's a fixed pitch text. On identifying this, we can easily "chop" into characters.
        - In case, of non proportional pitched text, Tesseract recognises the gaps and chops the characters of the word
    - Recognition
        - This is a 2 pass process - first pass involves recognising the words, and if satisfactory, use an adaptive classifier to recognise the words.
        - In the second pass, the words which were not properly identified, are passed through associators which figure out the gaps and help in identifying the words.
    - Classification
        - This is a 2 step process - a class pruner is used to identify characters which might match. The characters counts are taken and are matched with the lookup table. Each character is associated with 2 parameters: confidence and rating.
        - Uses an adaptive classifier, which normalises based on isotropic baseline/x-height normalization, as opposed to an static classifier
    - Final phase
        - This step looks for fuzzy spaces and uses alternative hypotheses to look for small-cap text.

2. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing

- This paper gives us an idea of all viable errors post OCR process. This helps us recognise what can be done to improve quality of the text traction and reduce the possible errors.
- According to the paper, the most common errors post OCR are misspellings, word length differences, etc.
- The human generated errors/misspellings are compared to those generated by the OCR and it is understood that most of them are first position errors and single letter errors. This is mostly caused due to the noise around the starting and the ending of the word.
- The approaches to look for errors where based on two approaches: Dictionary based and context based
- Dictionary based - doesn't look up the word context or the semantics, just performs a lookup
- Context based - looks up the grammar and semantics along with the context of the word nearby
- Standard mapping helps us identify the similar patterns in the misspellings

3. Adaptive Thresholding - a comparative study

- This paper compares different types of thresholding techniques to get a adaptive threshold
- Otsu's thresholding - Most widely used and fastest method
    - it selects the threshold by maximizing the within class variance between two groups of pixels.
- Rosin's method - takes the highest peak from the peak, and a straight line to the last non empty bin
    - the maximum deviation between the line and histogram curve gives the adaptive threshold
- Kapur's method - Uses's Shannon's rule of entropy
    - uses two different probability distributions : for foreground and background, their individual entropy are combined and maximized later.

Limitations in the paper:

- **An Overview of the Tesseract OCR Engine**
  - Tesseract uses polygonal approximation instead of directly using the raw outlines obtained after recognition.
  - Polygonal approximations of curves do not perform accurately during systematic distortion i.e noise, scaling etc.
  - Polygonal approximations do not minimize the distance of the points from the polygon while retaining the number of polygon edges as small as possible.

- **Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing**
  - Doesn't take into consideration errors caused by edit distances, grouping of real/non-real words.

Summarized survey of all the papers:

S#	Paper Title	Student Reader name	Main problem addressed in the paper	Methodology used in the paper to solve the problem	Contributing Results achieved in the paper	What is the limitation of this paper in solving the problem that you address
1	An Overview of the Tesseract OCR Engine	Amrita Nair	To address the steps involved in the extraction of the text	Processing, recognition and classification of the image is down to extract the text character by character	If the pre-processing done right, Tesseract gives accurate results	Doesn't mention the pre processing requirements for the image before OCR
2	Deep Statistical Analysis of OCR Errors For Effective Post-OCR Processing	Amrita Nair	Possible errors caused by an OCR	Analyse the errors using data from 4 different datasets	Tells possible errors that can be caused by the OCR	Fails to include other possible errors caused by real/non-real words and edit distances

3	Adaptive thresholding: A comparative study	Amrita Nair	To compare how different algorithm which gives us a adaptive threshold of an image	compares the threshold values obtained from all the algorithms using correlation and SSIm calculations	Algorithms which rely on Shannon's entropy rule give better results	The paper compares the different algorithms but doesn't come up with an optimal algorithm. Doesn't include segmentation methods which can be used

Aathira Dineshan

Approved Problem Statement: Automated Hardware Recon

Field -> Area-> Sub-area->topic (from general to specific): Web -> Web scraping

Literature Review:

Guide approval of selected papers:

S#	Name	Roll Number	Paper Title	URL/Reference Link of the paper selected	IEEE/ACM Publication	Year	Scopus Journal with impact factor/ Transaction/ conference	How the selected paper is relevant for your project?
1	Aathira Dineshan	AM.EN.U 4CSE17101	A Survey on Python Libraries Used for Social Media Content Scraping	https://ieeexplore.ieee.org/document/9215357	IEEE	2020	2020 International Conference on Smart Electronics and Communication (ICOSEC)	This paper compares the performance of the web scraper used in our project – BeautifulSoup with two other popular ones.
2	Aathira Dineshan	AM.EN.U 4CSE17101	Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef	https://ieeexplore.ieee.org/document/9198450	IEEE	2020	2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)	The application constructed in the paper is very similar to the initial part of our project. Here using the tool Scrapy the recipe name, ingredients and url are stored in a database.
3	Aathira Dineshan	AM.EN.U 4CSE17101	An Approach of Web Scraping on News Website based on Regular Expression	https://ieeexplore.ieee.org/document/8878550	IEEE	2018	2018 2nd East Indonesia Conference on	The paper talks about web scraping the site using regular

							Computer and Information Technology (EIConCIT)	expression, which can be used in our project to get the information about the vulnerabilities and components easily.

Key Points:

A Survey on Python Libraries Used for Social Media Content Scraping

- This paper compares the performance of the web scraping tools BeautifulSoup, LXml and RegEx.
- The best, average and worst-case response time of each one is compared. BeautifulSoup did not perform well compared to the other two. RegEx bagged first for the best case response time, and LXml for the others.
- When the prediction accuracy is compared, Beautiful soup comes second after RegEx.
- RegEx has limited rule extraction when it comes to the web page with more inner tags, therefore RegEx is capable of performing only on moderately complex contexts.
- To recapitulate, discarding RegEx due to its limitation, BeautifulSoup can be chosen in terms of accuracy, and LXml for response time.

Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef

- The application stores the recipe name, ingredients, and the URL of the recipe in database, collected through web scraping beforehand.
- Web scraper used: Scrapy, database used: Mongo dB
- Dictionary is used to store the recipe name and ingredients, lists are used to store dictionary keys, to extract ingredients and to return as output.
- When ingredient and recipe name are typed, the URL and recipes are appended to an empty list and printed.

An Approach of Web Scraping on News Website based on Regular Expression

- In this paper, recall, precision, and F-measure are calculated to check whether usage of regular expression help in web scraping.
- Firstly, the news elements, and non-news elements are identified. The non-news elements consit of text advertisement, video advertisement, link, image, and script with different pattern for every website.
- There are two patterns created - content pattern (for extracting original text article of news) and filter pattern (for eliminating non-news elements).
- Based on the evaluations, on the three news websites: Detik and Tribunnews had recall = 1, precion = 1 and F-Measure = 100% while for Liputan6 0.95, 0.95, and 95% respectively. This means that manual copy-paste can be avoided by using Regex.

Limitations in the paper:

A Survey on Python Libraries Used for Social Media Content Scraping

The prediction accuracy was determined by using True-Positive and True-Negative mode. More accuracy metrics like dynamism in the content, originality of the content must be used to make the ultimate judgment of which web scraper provides the accurate results.

Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef

The application shows no result when there an entry is not present in the database is given. Neither does it do web scraping to provide results, nor does it store that entry in another database to run a separate web scraping to update the database. According to our project, the commonly found components are manually stored in the database. When the OCR does not give an accurate result, database can be looked up to find a similar pattern and go for web scraping, if needed.

An Approach of Web Scraping on News Website based on Regular Expression

A particular regex can be applied to that website, i.e., for different websites, the regex changes. Generalization can be a problem when dealing with different sites in our project. But this can fasten the search in selected sites, which we plan to do in our project.

Summarized survey of all the papers:

S#	Paper Title	Student Reader name	Main problem addressed in the paper	Methodology used in the paper to solve the problem	Contributing Results achieved in the paper	What is the limitation of this paper in solving the problem that you address
1	A Survey on Python Libraries Used for Social Media Content Scraping	Aathira Dineshan	To compare the response time and accuracy of three web scraping tools.	The best, average and worst-case time, and prediction accuracy are measured.	BeautifulSoup can be used in terms of better accuracy.	More prediction metrices should be used to decide on ultimate accurate web scraper.
2	Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef	Aathira Dineshan	A semi-autonomous application that can print the recipes based on recipe name and ingredient names given.	The recipe name, ingredients and recipe URL are stored in a database. The input provided is looked up in the database.	A database provides faster access.	The paper does not provide a way to get the information in case if its not present in the database.
3	An Approach of Web Scraping on News Website based on Regular Expression	Aathira Dineshan	How to accurately and easily web scrape of news website	Regex are used to make it easier to extract information. It is also very accurate.	Regex can be used to extract a component details or vulnerabilities in a website	Regular expressions vary with websites, making it hard to generalize.

Kalpesh Gupta

Field -> Area-> Sub-area->topic (from general to specific): Web -> Web scraping->getting the details of hardware (ex - previously known hardware exploits of the device, device specs)

Literature Review: (From the next page)
Guide approval of selected papers:

#	Name	Roll Number	Paper Title	URL/Reference link of the paper selected	IEEE/ ACM publication	Year	Scopus journal with impact factor/Transaction /Conference	How is the selected paper relevant for your project?
1	Kalpesh Gupta	AM.EN.U4CSE17135	Schema Inference and Data Extraction from Templatized Web Pages	https://ieeexplore.ieee.org/document/7087084	IEEE	2015	International Conference on Pervasive Computing (ICPC)	This paper talks about how to extract data from templatized web pages from the internet and thus will be helpful for us to extract the information from the internet about the device under study.
2	Kalpesh Gupta	AM.EN.U4CSE17135	Web data extraction using textual anchors	https://ieeexplore.ieee.org/document/7436204	IEEE	2015	2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)	This paper they are creating a visual tool to create a web scraper to extract the data records from the web pages, which is also relevant to data extraction information about the device under study.
3	Kalpesh Gupta	AM.EN.U4CSE17135	Web Crawling-based Search Engine using Python	https://ieeexplore.ieee.org/document/8821866	IEEE	2019	3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)	This paper suggests a web mining powered search engine for the education sector, helping users to search the information they are looking for during the times of admission. Since, we also need to extract the information about the different components under study from the web; the work will be useful for our study.

Key Points from the papers:

Schema Inference and Data Extraction from Templatized Web Pages

- The authors have presented a page-level data extraction system that extracts schema from template generated pages automatically.
- Their system takes as input a collection of web pages.
- For each webpage, a VB tree is constructed by applying a Vision based Page Segmentation algorithm.
- Noise blocks are removed from the DOM trees.
- For fixed template pages applying FiVaTEch tree merge algorithm, a fixed/variant pattern is constructed
- Schema is inferred which identifies basic type, set type, optional type and tuple type.
- Data is extracted matching pattern tree <-> HTML tree at each level
- System was able to find schema of a website in XML like structure.

Web data extraction using textual anchors

- They have tried to stimulate the way people typically look at the web page i.e., searching for visual patterns while traversing top to down or bottom to up.
- Anchors contribute key elements within region patterns and a pattern can be created using one or multiple anchors.
- The main steps involved in their proposed system:
  - A user creates a wrapper for a representative page which serves a file of reference for similar pages
    - For this, it loads a web page to identify the anchors in the embedded browser
    - User can also specify the anchor using semantic node attributes like id or class name
  - User must specify how target data regions are identified using these anchors(like Self, Parent, Following, preceding, Common Ancestor, Between)
  - These patterns are used to extract the data.
  - Then a data extraction algorithm of polynomial time complexity(i.e. of N^3 in the worst case) is applied to extract the data in XML format.
  - The proposed system relies on page contents rather than HTML alone.

Web Crawling-based Search Engine using Python

- This paper suggests a web mining powered search engine for the education sector helping users to search the information they are looking for during the times of admission
- PHP based web interface
- Two python modules are used for developing the web crawlers- solicitations and BeautifulSoap4
- After web crawling, the data is made available for the scrappers to selectively work on the provided links.
- This engine automates the manual task of finding the information about different institutes.

Limitations of the papers:

Schema Inference and Data Extraction from Templatized Web Pages

Here they were working with template-based pages taking two web pages of the same w websites as input, but there are many websites, which are dynamic in nature in today's world. Therefore, we may need to extract the information for these kinds of web pages also. Also, we need to select the proper pages manually for each website to be representational of that website.

Web data extraction using textual anchors

This tool can create a wrapper of a particular website. We want to get the data from different sources on the internet i.e. always there is no particular site from which we may need to extract the data. Also, for a particular type of website, first we need to manually create a wrapper using anchors for using this system and need to modify the wrapper for getting some specific detail from the web page.

Web Crawling-based Search Engine using Python

This system is designed for the educational domain, there can be user feedback mechanisms also included in the system to provide better and more relevant suggestions to the users and thus can reduce the irrelevant results from the obtained set of suggestions. This similar approach we may follow to avoid any wrong suggestion for the information being displayed.

Summarized survey of all the papers:

#	Paper Title	Student reader name	Main problem addressed in the paper	Methodology used in the paper to solve the problem	Contributing results achieved in the paper	What is the limitation of the paper in solving the problem that you address?
1	Schema Inference and Data Extraction from Templatized Web Pages	Kalpesh Gupta	This paper attempts to extract the data from the web pages. The authors have presented a page-level data extraction system that extracts schema from template-generated pages automatically.	Their system takes as input a collection of web pages. For each webpage, a VB tree is constructed by applying a VBPS algorithm. Then based on visual features, blocks in VB trees are compared to determine similar or variant template pages. Removal of noise blocks is done from DOM trees. Then for fixed template pages, by using the tree merging algorithm, a fixed/variant pattern tree is constructed for each Web site and the schema is identified. After construction of pattern tree, Data extraction is done by matching pattern tree and HTML tree at each level.	The proposed system was able to find the schema of a website in XML-like structure and extract the data from the web page. Also, it was more efficient than the existing available system in terms of time consumption.	Here they were working with template-based pages taking two web pages of the same w websites as input, but there are many websites which are dynamic in nature in today's world. So, we may need to extract the information for these kinds of web pages also. Also, we need to select the proper pages manually for each website to be representational of that website.

2	Web data extraction using textual anchors	Kalpesh Gupta	This paper focuses on creating a tool to generate web scrapers to extract the data records from the web page.	Tried to stimulate the way people typically look at the web page. They have used text features such as textual delimiters, keywords, constants or text patterns, (anchors) to create patterns for the target data regions and data records. Anchors contribute key elements within region patterns and a pattern can be created using one or multiple anchors. Each region pattern may contain some sub patterns for the regions within it. Extracted data is directly mapped onto the hierarchical XML structure as the output of the system.	The system was found to create wrappers for websites like Amazon and eBay, IMDB movies, world weather and they required only a few anchors and patterns, getting a 100% recall in these cases.	This tool can create a wrapper of a particular website. We want to get the data from different sources on the internet i.e. always there is no particular site from which we may need to extract the data. Also, for a particular type of website, first we need to manually create a wrapper using anchors for using this system and need to modify the wrapper for getting some specific detail from the web page.
3	Web Crawling-based Search Engine using Python	Kalpesh Gupta	The authors want to develop a website which should be able to crawl the useful information from the school and provide aid to the parents in Delhi NCR region.	A PHP based web application is used. They are using two different python modules for developing web crawlers- solicitations and BeautifulSoap4 and using the API services provided by few websites to scrap the web data. After crawling the web, the data is made available for the scrappers to selectively work on the provided links. Then, they store the data in a managed database.	The main contribution of this paper is a generic crowdsourcing framework for automatic and scalable semantic annotations of HTML5 videos.	This system is designed for the educational domain, there can be user feedback mechanisms also be included in the system to provide better suggestions to the users. This similar approach we may follow to avoid any wrong suggestion for the information being displayed.

Jishnu Ganesh

Field -> Area-> Sub-area->topic (from general to specific): Web->Web Scraping->Report generation

Literature Review:

Guide approval of selected papers:

S#	Name	Roll Number	Paper Title	URL/Reference Link of the paper selected	IEEE/ACM Publication	Year	Scopus Journal with impact factor/ Transaction/ conference	How the selected paper is relevant for your project?
1	Jishnu Ganesh	AM.EN.U4CSE17133	Recommending Personalized Search Terms for Assisting Exploratory Website Search	https://ieeexplore.ieee.org/document/8791155	IEEE	2019	JCDL	It tells about recommendations of search terms ,which can be used in the information retrieval.
2	Jishnu Ganesh	AM.EN.U4CSE17133	A Review on Web Scraping and its Applications	https://ieeexplore.ieee.org/document/8821809	IEEE	2019	ICCCI	It tells about web scraping and its applications and why we prefer libraries like beautiful soup.
3	Jishnu Ganesh	AM.EN.U4CSE17133	A Semi-Automatic Data–Scraping Method for the Public Transport Domain	https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8782469	IEEE	2019	ACCESS journal	It tells about a semi automatic web scraper tool.

1. Recommending Personalized Search Terms for Assisting Exploratory Website Search

Key Points:

- This paper is about a personalized search term recommender system called mySearchCLUE,which provides potential search terms within a website search engine using collaborative filtering.
- Recommender Systems help users find the items(search terms) most useful to them by providing personalized recommendations.
- Searching for information within a website can be categorized into lookup search or exploratory search.In lookup search the goals are usually well defined i.e. the users know what they are looking for while in exploratory search, the goals are not well defined.
- mySearchCLUE system recommends personalized words, in the absence of any given partial search term, to help guide both lookup and exploratory searches.
- Here search term recommendation is based on user-to-user neighborhood-based collaborative filtering(Collaborative filtering is a family of algorithms where there are multiple ways to find similar users or items).
- Uses three recommendation approaches for generating personalized search terms:

  • Approach 1: Recommendation based on the past history of search terms that users entered into the website .

  • Approach 2: Recommendation based on the past history of web pages that users visited.

  • Approach 3: Recommendation based on combination of Approach 1 and Approach 2.

Limitations in the paper:

- The above approach can be used in the information retrieval part of our problem.
- The main limitation is that this approach uses recently entered search terms,web pages visited by the users, in our problem we have to extract data on the basis of hardware devices i.e. our search term is the name of a hardware device.
- We can create a list of most common devices and websites as the input which can solve our issue.

2.A Review on Web Scraping and its Applications

Key Points:

- This paper is about various aspects of web scraping, i.e. the basic introduction and a brief discussion on various software's and tools used for web scraping.Web scraping is the method of spontaneous collection of information from the World Wide Web.Meaningful data from the HTML of websites are extracted and stored into a central local database or spreadsheet.
- The main applications of web scraping includes Data mining , Research , Marketing etc.
- rvest (the stored html is traversed using rvest) and regular expression are some of the tools used for web scraping.
- The prevalent methods in the implementation of Web data scraper can be grouped in three main categories:

  (i) Libraries for general-purpose programming languages:

  Grant access to the site with the help of the client side of the HTTP protocol and parsing is done using regular expression ,tokenization etc ,example Beautiful soup.

  (ii) Frameworks

  (iii) Desktop-based environments.

Limitations in the paper:

- The usage of libraries like beautiful soup requires API for accessing the web and parsing it.The retrieval mainly depends on the html contents.Any change in the html content to be reflected back in our database we have to again recompile the code .This can be avoided by using frameworks and desktop based environments. Both of which are harder to implement.Since our data is static we prefer libraries like beautiful soup .

3.A Semi-Automatic Data–Scraping Method for the Public Transport Domain

Key Points:

- This paper proposes a method for the semi-automatic generation of a web scraper tool for the public transport domain.
- It is first necessary to retrieve the HTML data from a data source (website), then parse the data using scraping techniques and store the target information.Multiple websites were used to scrape the data.Manually programming for these websites was a tedious and very time-consuming task owing to the unstructured nature of the website.
- The semi automatic data extraction starts with the selection of web sources(urls) and corresponding domain models i.e. the public transport domain model.

- The main concepts i.e. information like stations,lines,facilities available for the domain model were searched and mapped to the elements found in the URL.

Limitations in the paper:

- The Semi automatic Data Scraper could only be used in the Public Transport Domain as the data was extracted on the basis of lines (tracks) between stations and the number of stations.ie the web scraper was not generic.

Summarized survey of all the papers:

S#	Paper Title	Student Reader name	Main problem addressed in the paper	Methodology used in the paper to solve the problem	Contributing Results achieved in the paper	What is the limitation of this paper in solving the problem that you address
1	Recommending Personalized Search Terms for Assisting Exploratory Website Search	Jishnu Ganesh	To help users during exploratory and lookup search by recommending search terms.	The search term recommender is based on user-to-user neighborhood-based collaborative filtering.	A search term recommender that recommends search terms on the basis of recent search terms and websites visited by the user.	The recommender system requires a list of common hardware devices.
2	A Review on Web Scraping and its Applications	Jishnu Ganesh	This paper focuses on various aspects of web scraping and introduction to tools used for web scraping.	Advantages and disadvantages of tools used for web scraping.	Different tools used for web scraping.	Frameworks are preferred over libraries.
3	A Semi-Automatic Data–	Jishnu Ganesh	To provide sufficient information about the	This paper specifically focuses on the data	A web tool specifically designed	The web tool can only be applied to

	Scraping Method for the Public Transport Domain		accessibility of routes in a public transport domain for specially abled peoples.	extraction and processing of the existing information on the web concerning public transport and storing it in an open data repository .	to extract the data from a public transport domain	the transport domain.Its not generic .