Revised Proposal Note

This document answers the issues pointed by evaluators, interviewers and mentions the improvements done to proposal itself.

General Q&A

Why this project was chosen?

Proposed solution addresses problems that I already had faced when using web like poor quality search results, unable to find already bookmarked resource and I already know the issues in existing solutions out there.

Typically, most students will pick a regular software development project that do CRUD operations or use some API or library, I just wanted to do something different. I decided to go with ML domain since I am interested in the field myself even though I am not a veteran in the field. Other than that lecturer who conducted the lectures regarding the final project encouraged to try data science, ML fields since majority go with software engineering.

Addressing Evaluator Grading Feedback

Project description to proposed solution :

I have already followed the instructions given in the proposal guidelines document when I submitted the document, wish it had more good and bad examples since the majority of student are not good English writers. Anyway, I think the problem may be my way of expressing myself in written English in a way that others can understand without the context that I have.

Anyway according to evaluators grading feedback project proposal was improved by

- Amending the existing text to improve the clarity
- Providing more explanations, diagrams and tables.

Even though the feedback about proposal is vague, wish the feedback was more clear on what is the issue and how it could be fixed, hope I did my best to improve the mentioned issues.

Likelihood of success if questionable?

I have kept the project scope at the right size in order to make it achievable during given timeframe, instead of increasing the scope with lot of innovations and features.

Not clear project is creative or innovative?

Since I have already studied the existing solutions out there, both commercial and non-commercial, I am confident that proposed solution is definitely innovation due to following

- Proposed solution use
    - ML to auto tag bookmarks, while existing solution use manual approach.
    - On-device ML model which improve user privacy.

Potential for learning is less clear or likely?

I consider potential for learning in this project as big, since this is not a regular software development project that do CRUD operations or use some API or library.

The student has a vague idea of the domain and the project?

I personally have the faced the problems addressed in this project like poor quality search results, unable to find already bookmarked resource. I also have studied the existing solutions out there and their issues. So I think I have a better idea of the problem domain and the project.

Addressing Evaluators Comments

Comment : *Need to find a suitable dataset?*

Reply : While there are existing data sets out there like delicious dataset, datahub delicious dataset, rdrr delicious dataset transforming and cleaning those datasets are quite annoying due to numerous reasons. So we will use following two methods to generate a dataset

- Labeling unlabeled bookmarked resources using topic modeling.
- Web scraping already existing bookmarks and their tags from websites like Pinboard, Diigo

And it is important to know when doing data science you always do not get ready-made datasets, so one have to learn to extract and preprocess data themselves which usually consumes the largest amount of time in data science process.

Comment : *Student does not have enough knowledge?*

Reply : While I myself are not a machine learning expert, but I consider myself as a beginner with good understanding of ML concepts. I do also have a high interest in the ML domain. And this problem that I am tackling in my project is a problem faced by myself over the years, so suppose that I understand the problem well and have already identified good solution to the problem.

Comment : *Problem not worth solving?*

Reply : There are plenty of bookmark managers out there outside the browser built-in ones which are both commercial (Pocket, Pinboard, Diigo) and free (Shaarli, LinkAce). This is a good indicator that there exist issues that need to be solved by third party browser managers. In my project, I have proposed a solution that address issues in existing commercial and free bookmark managers. If this project implemented successfully could sell as SaaS product, this project has commercial potential since it do automatic bookmark tagging and do bookmark tagging using on-device ML which is privacy-friendly.

Addressing Interviewers Comments

Project is not enough?

While at glance the project seems simple, it is not, due to the reasons given below

- Data pre-processing step involved in this project is complex and time-consuming
  - Scraping data from the web and creating dataset is complex and time-consuming since

        - Need to handle a lot of edge cases when scraping data. Ex: non HTML web resources like PDFs, Word documents etc.
        - Have to clean and de-duplicate resources, tags.
    - Tagging the bookmarks is also challenging since due to
        - The nature of the web, i.e.: heterogeneity and versatility
        - Tags can have complex hierarchies' ex: Tags like *NeuralNetworks*, *LinearRegression* can be taken under *MachineLearning* tag.
        - In some cases need to implement custom logic to add certain tags to certain bookmarked resources ex: Adding video tag for YouTube resources
- Building an ML model that meets the following criteria is challenging
    - Reaching certain accuracy level (over 70%) is challenging due to heterogeneity and high versatility of the web.
    - The model needs to predict tags in minimum time and have to use lesser system resources when doing so.
- Developing cross-browser extension is a hassle since different browser vendors support different features and APIs.
- In addition to developing proposed solution, project include industrial best practices like writing tests, using CI/CD which also consumes time.

If still the project scope is too narrow, I could add features like

- Allow users to publicly share their catalog of tagged bookmarks and add shared bookmarks to user's existing catalog (which is known as Social bookmarking)
- Iteratively improving the ML model according to user's custom tagging.
- Using federated learning in addition to on-device ML.

ITE 3962 – Final Year Project

Project Proposal

Smart Bookmark Manager

Submitted by:

E2142007
A.G.G. Malshan

Bachelor of Information Technology (External Degree)
Faculty of Information Technology
University of Moratuwa

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	***Description***
SEO	*Search Engine Optimization*
ML	*Machine Learning*
NLP	*Natural Language Processing*
CSV	*Comma Separated Value*
JSON	*JavaScript Object Notation*
BoW	*Bag of words*
LDA	*Latent Dirichlet allocation*

API	*Application Programming Interface*
HTML	*Hyper Text Markup Language*
CRUD	*Create Read Update Delete*
UI	*User Interface*
UX	*User Experience*
VS Code	*Visual Studio Code*
CSS	*Cascading Style Sheets*

LIST OF APPENDICES

1. Introduction

Today most of the people use search engines to find new information in the vast sea of information expecting to get high quality accurate information, but in recent years quality of search engine results were degrading due to numerous reasons like

- Web pages solely created for higher search engine ranking with lower focus on content quality, which is known as SEO spam.
- Appearance of duplicated machine generated web pages of authoritative websites.
- Conflicting interest of search engine companies, i.e.: Advertising revenue vs high quality search results.

And the drawbacks of strictly depending on search engines to find quality information is also coming to the light like

- Search engines invade user privacy by collecting information about user behavior on the internet to serve ads.
- Regular change in search engine algorithm resulting in different search results over the time for the same query.
- Promoting biased search results over others for business gain and strengthening certain views.

In order to overcome these drawbacks of search engines is to maintain your own collection of high quality resources, bookmark managers help to achieve this. But existing bookmark managers out there have limited flexibility like

- Bookmark search is limited to page title text and URL text.
- Most browsers use folders to categorize bookmarks, so cannot categorize a single resource under multiple categories.

And involve manual work for categorizing like

- Have to categorize bookmarks by hand using folders or tags.

In order to overcome these drawbacks, we suggest a smart bookmark manager application which address previously mentioned issues as follows

- Use tags to categorize bookmarks instead of folders, which allow categorizing particular bookmarked resources under multiple topics.
- Generate those tags using machine learning considering the content of the bookmarked resource, which increases the chance of finding the resource for later use.

2. Background & Motivation

2.1 Increased importance of bookmarking

Typically, users use search engines to find new information, and they expect those search engines to have high quality results. But in recent years search engines are failing to meet user needs due to numerous reasons which are described in upcoming sections, these issues in search engines encourage users to maintain their own set of bookmarked resources for later reference.

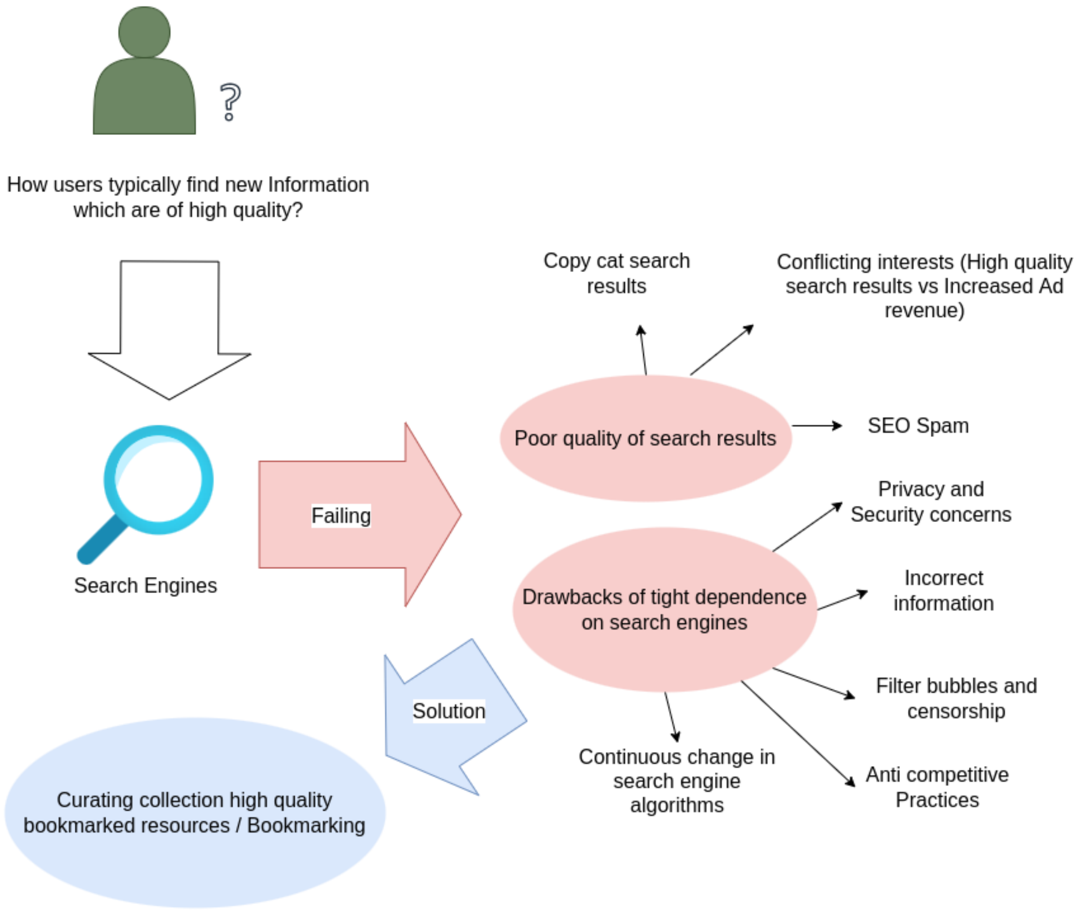

How users typically find new Information which are of high quality?
Search Engines
Failing
Poor quality of search results
Copy cat search results
Conflicting interests (High quality search results vs Increased Ad revenue)
SEO Spam
Drawbacks of tight dependence on search engines
Privacy and Security concerns
Incorrect information
Filter bubbles and censorship
Anti competitive Practices
Continuous change in search engine algorithms
Solution
Curating collection high quality bookmarked resources / Bookmarking

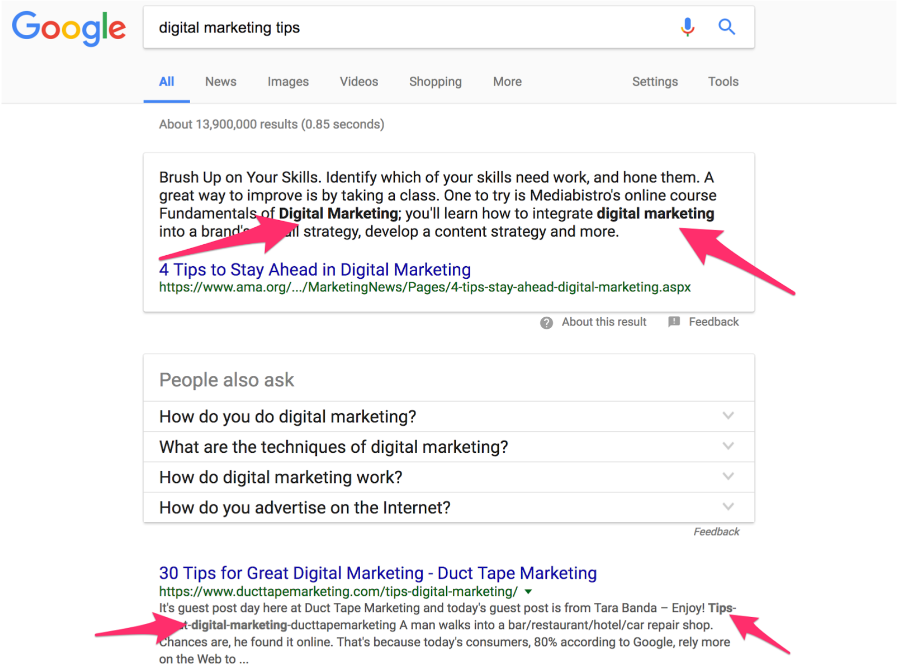
Figure 1 : *Drawbacks of search engines and bookmarking as a solution*

2.1.1 Worsening Quality of Search Engine Results

People use search engines to find new information, but in recent years the quality of search engine results are degrading rapidly due to numerous reasons which are discussed in upcoming sections.

2.1.1.1 SEO Spam

In the digital world, there is intense competition among website / business owners to rank their websites by using SEO hacks. While search engines can detect these kinds of SEO trickery, they are not perfect. So in order to gain advantage of these weaknesses in search engines, website owners build carefully crafted webpages solely to get search engine ranking and clicks, but they are in poor quality as resources.



digital marketing tips

About 13,900,000 results (0.85 seconds)

Brush Up on Your Skills. Identify which of your skills need work, and hone them. A great way to improve is by taking a class. One to try is Mediabistro's online course Fundamentals of **Digital Marketing**; you'll learn how to integrate **digital marketing** into a brand's overall strategy, develop a content strategy and more.

4 Tips to Stay Ahead in Digital Marketing
https://www.ama.org/.../MarketingNews/Pages/4-tips-stay-ahead-digital-marketing.aspx

People also ask

How do you do digital marketing?

What are the techniques of digital marketing?

How do digital marketing work?

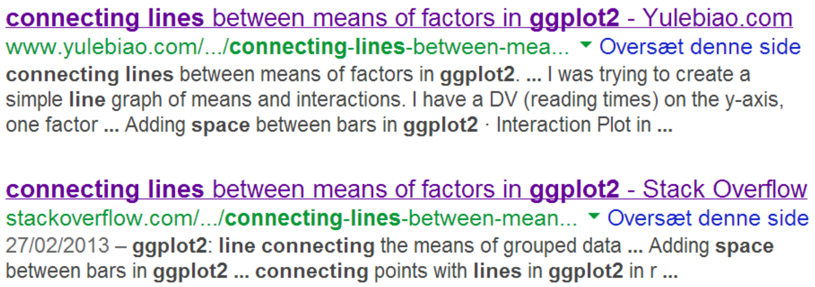
How do you advertise on the Internet?

30 Tips for Great Digital Marketing - Duct Tape Marketing
https://www.ducttapemarketing.com/tips-digital-marketing/
It's guest post day here at Duct Tape Marketing and today's guest post is from Tara Banda – Enjoy! **Tips-great-digital-marketing**-ducttapemarketing A man walks into a bar/restaurant/hotel/car repair shop. Chances are, he found it online. That's because today's consumers, 80% according to Google, rely more on the Web to ...

Figure 2 : Google Search result showing SEO spamming.
Image Credit : https://neilpatel.com

Due to this so-called SEO spam, most users have to add additional keywords to the search query to find good search results. Ex: Appending “reddit" to the search query where you can find quality resources recommended by communities focused on that specific area or have to go to a specific website and use the website search functionality or have to use a browser extension to filter out spam results automatically. Among those solutions, referring to a bookmarked resource is the most reliable solution, since search results for the same query can change over time and browser extension filter lists can be outdated.

2.1.1.2 Copycat search results



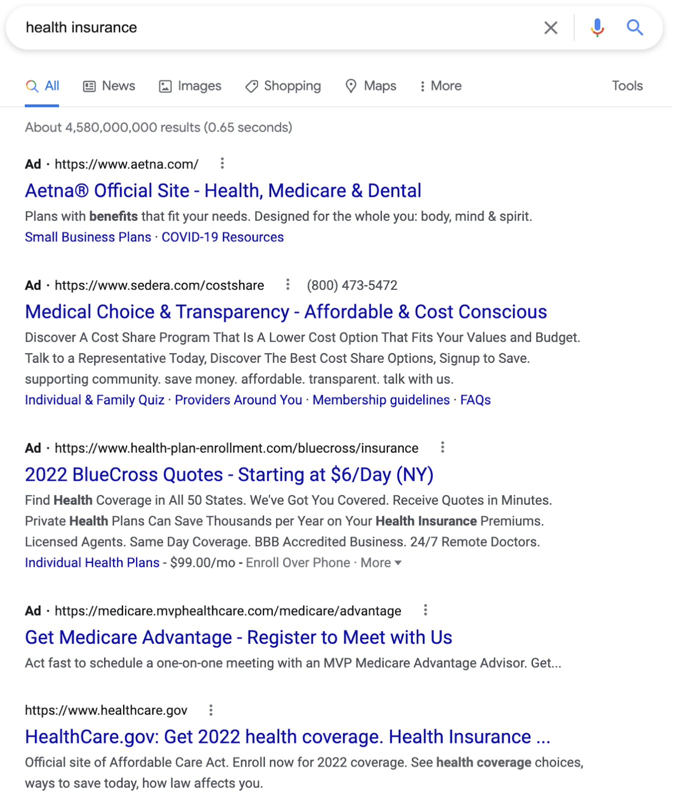
connecting lines between means of factors in **ggplot2** - Yulebiao.com
www.yulebiao.com/.../**connecting**-**lines**-between-mea... ▾ Oversæt denne side
connecting lines between means of factors in **ggplot2**. **...** I was trying to create a simple **line** graph of means and interactions. I have a DV (reading times) on the y-axis, one factor **...** Adding **space** between bars in **ggplot2** · Interaction Plot in **...**

connecting lines between means of factors in **ggplot2** - Stack Overflow
stackoverflow.com/.../**connecting**-**lines**-between-mean... ▾ Oversæt denne side
27/02/2013 – **ggplot2**: **line connecting** the means of grouped data **...** Adding **space** between bars in **ggplot2** **... connecting** points with **lines** in **ggplot2** in r **...**

Figure 3 : Stack Overflow copycat website ranking higher for a search query
Image Credit : https://i.stack.imgur.com

In recent years web search results have been populated by copycat websites of well-known authoritarian websites ex: Clones of GitHub, StackOverflow, NPM, Wikipedia generated by bots. In order to get rid of these copycat results either you have to manually avoid them or use a browser extension[1] to filter them out, a much better hassle-free solution is to bookmark frequently accessed resources for easier access, this also comes with added security benefits like avoiding phishing websites.[2]

2.1.1.3 Advertising vs Search Result Quality



health insurance

About 4,580,000,000 results (0.65 seconds)

Ad · https://www.aetna.com/
Aetna® Official Site - Health, Medicare & Dental
Plans with **benefits** that fit your needs. Designed for the whole you: body, mind & spirit.
Small Business Plans · COVID-19 Resources

Ad · https://www.sedera.com/costshare (800) 473-5472
Medical Choice & Transparency - Affordable & Cost Conscious
Discover A Cost Share Program That Is A Lower Cost Option That Fits Your Values and Budget. Talk to a Representative Today, Discover The Best Cost Share Options, Signup to Save. supporting community. save money. affordable. transparent. talk with us.
Individual & Family Quiz · Providers Around You · Membership guidelines · FAQs

Ad · https://www.health-plan-enrollment.com/bluecross/insurance
2022 BlueCross Quotes - Starting at $6/Day (NY)
Find **Health** Coverage in All 50 States. We've Got You Covered. Receive Quotes in Minutes. Private **Health** Plans Can Save Thousands per Year on Your **Health Insurance** Premiums. Licensed Agents. Same Day Coverage. BBB Accredited Business. 24/7 Remote Doctors.
Individual Health Plans - $99.00/mo - Enroll Over Phone · More ▾

Ad · https://medicare.mvphealthcare.com/medicare/advantage
Get Medicare Advantage - Register to Meet with Us
Act fast to schedule a one-on-one meeting with an MVP Medicare Advantage Advisor. Get...

https://www.healthcare.gov
HealthCare.gov: Get 2022 health coverage. Health Insurance ...
Official site of Affordable Care Act. Enroll now for 2022 coverage. See **health coverage** choices, ways to save today, how law affects you.

Figure 4 : Google Search result dominated by ads for the query health insurance
Image credit : https://cdn.substack.com/

Google is the most (92.01%) used search engine in the world that most people use and comes as the default in most browsers today, but they are an advertising driven company, which resulted in $43.3 billion search ad revenue for the fourth quarter of 2021. Advertising and quality search results do not go hand in hand in most cases, according to the research done by Google Co-founders at Stanford University.[3]–[5]

In order to declutter search experience from ads, you can use browser extensions like uBlock or bookmark a set of community recommended resources for easy access.

2.1.2 Other pitfalls of depending on search engine results

There are other issues users should keep in mind when tightly depending on search engines, which are discussed in upcoming sections in detail. We can easily overcome these issues by using a bookmark manager.

2.1.2.1 Security concerns

In some cases it is known that search engines like Google index search results of web pages containing malware which comes with software security risks like phishing, data theft, etc. Using bookmarks eliminates these kinds of security risks[6].

2.1.2.2 Privacy concerns

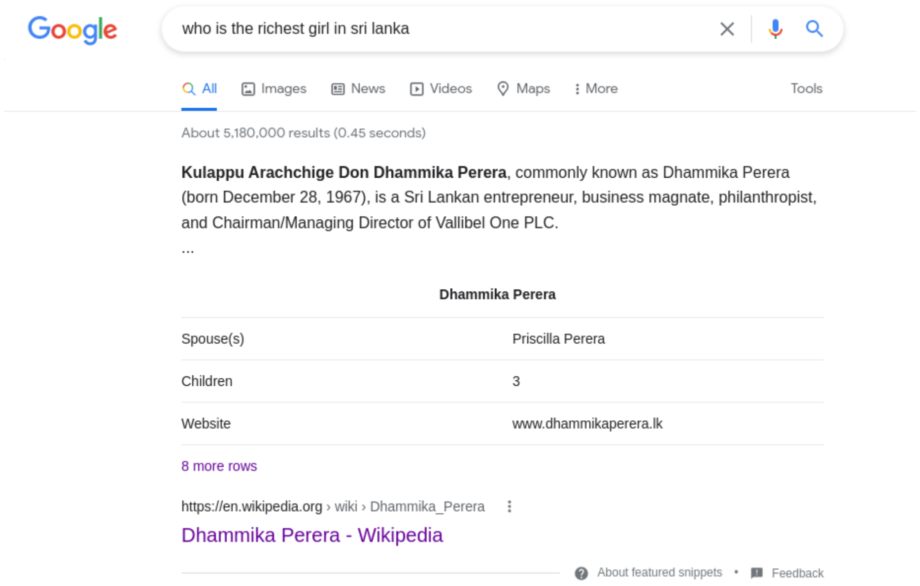
As you read earlier, the search engine market is dominated by Google, which also tracks and collects data about your behavior to show targeted ads. This in itself raises lots of privacy issues. Bookmarking a good set of resources bypasses the need to use a search engine, which results in better user privacy[7].

2.1.2.3 Filter bubble and censorship

Some search engines like Google provide users with personalized search results based on past searches and tracker info, this idea is called filter bubble which brings some dangers like reinforcing certain viewpoints of users without providing them with unbiased info. On other side Google also censor showing certain results based on location or tracker info which is considered as unethical in nature.[8], [9]

In order to know unbiased, objective information, it is much better to bookmark websites like Hacker News and other trusted sets of websites for easy access.

2.1.2.4 Erroneous instant search results



who is the richest girl in sri lanka

About 5,180,000 results (0.45 seconds)

Kulappu Arachchige Don Dhammika Perera, commonly known as Dhammika Perera (born December 28, 1967), is a Sri Lankan entrepreneur, business magnate, philanthropist, and Chairman/Managing Director of Vallibel One PLC.
...

Dhammika Perera	
Spouse(s)	Priscilla Perera
Children	3
Website	www.dhammikaperera.lk

8 more rows

https://en.wikipedia.org › wiki › Dhammika_Perera

Dhammika Perera - Wikipedia

Figure 5 : Instant search result with incorrect information

Most users find instant search results quick and instantly gratifying, but these are not correct all the time, since these knowledge graphs can contain incorrect information. Best example is if you search for "richest girl in Sri Lanka" in Google, you will get a top instant result containing Mr. Dhammika Perera's name who is male. Another thing to note is that some information like ranking changes frequently, which also result in incorrect information in instant search results.

For information like this, it is best to bookmark a set of authoritative website URLs for easy access and to gain more accurate information.

2.1.2.5 Anti-Competitive Practices

Problem with search engines like Google is that they own a plethora of different services and products like Google Cloud, Gmail, Google Play Music, Google Pixel phones which compete with other popular companies. Google is known to use their dominance in search engine space to drive out competitors by promoting their products and services over competitors using search results[10]. As consumers of these search results, we have a tendency to cultivate bias towards certain products, and we lose the discoverability of new products and services due to this anticompetitive practices of search engine companies.

It is best to bookmark a product review website or visit a forum like Reddit or Hacker News to know about a particular product.

2.1.2.6 Continuous Change in Search Engine Algorithms

An inherent issue with search engines is that search results for the same query change over time, while this has its benefits it comes with issues like not being able to find previously found resources by using the same query.

So it is good to not to strictly depend on search engines to find the info you need and use bookmarking as a tool to curate a good set of resources.

2.2 Limited Bookmark Managing Functionality in Browsers

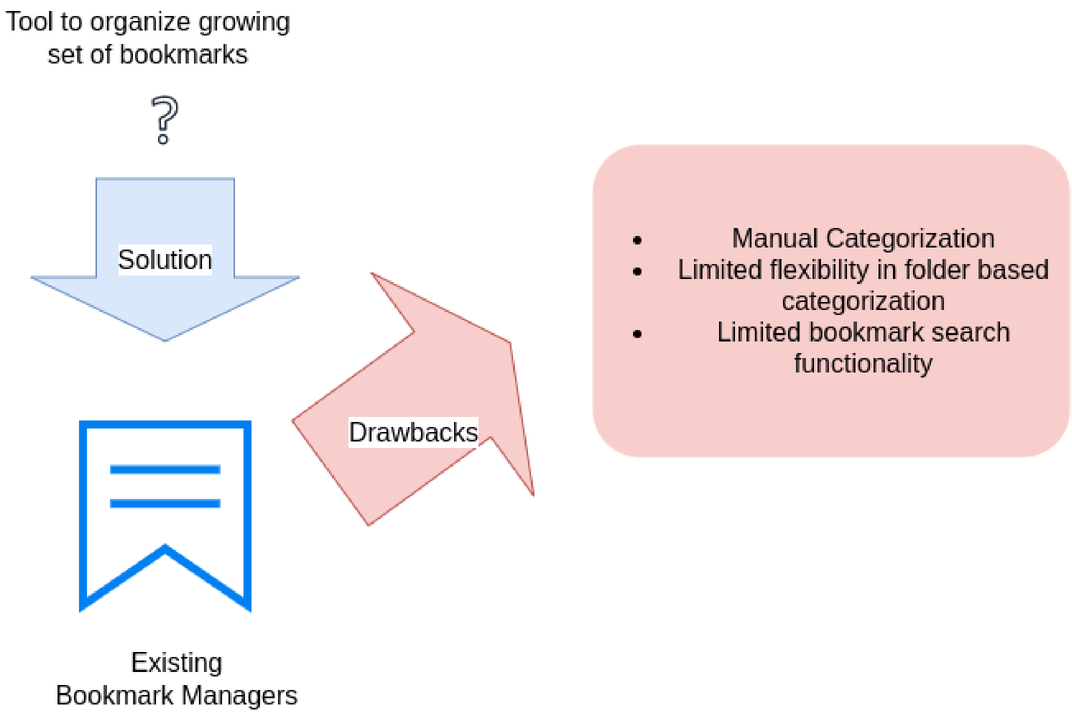

Tool to organize growing set of bookmarks
?
Solution
Existing Bookmark Managers
Drawbacks
• Manual Categorization
• Limited flexibility in folder based categorization
• Limited bookmark search functionality

Figure 6 : *Drawbacks of existing bookmarking managers and solution*

Due to these prevailing issues in search engines and their search results, one has to keep track when they found a quality resource, that is done by bookmarking which allows them to easily access that resource next time. Over time, these number of bookmarks will grow over several thousands, creating your mini version of the web with a set of high quality resources. In such cases, we need a tool to organize and search those bookmarks efficiently for later reference. But when we look at existing bookmark managers out there, they come with many drawbacks, which are described in next sections in detail.

2.2.1 Limited flexibility in Folder based Categorization

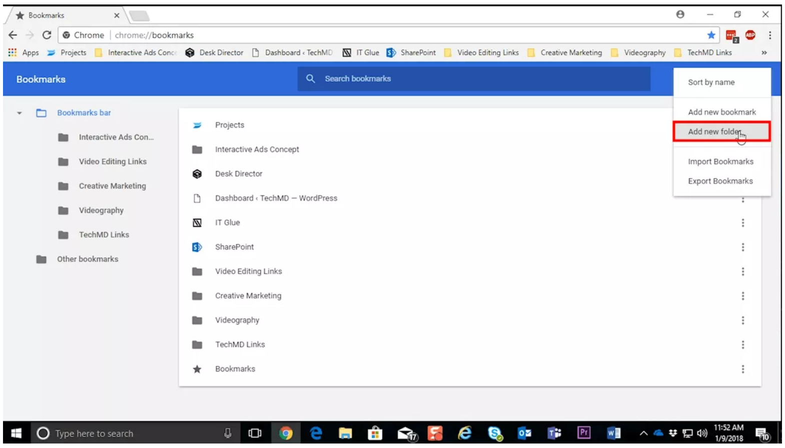


Figure 7: Folder-based categorization in Google Chrome
Image Credit : https://cdn-cpfcd.nitrocdn.com/

Most modern browsers allow categorizing bookmarks based on the folders while you can nest folders inside one another, this folder-based categorization is inflexible. If you take a web resource, it may belong to multiple distinct categories, which cannot be categorized based on the folder hierarchy. There is tag based bookmark categorization in Firefox, but it is limited to Firefox itself.

2.2.2 Manual Categorization

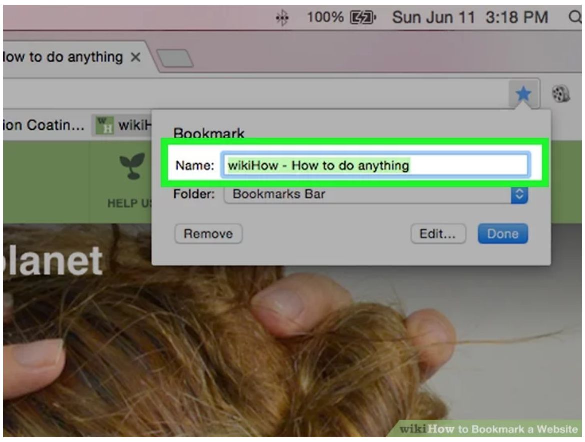
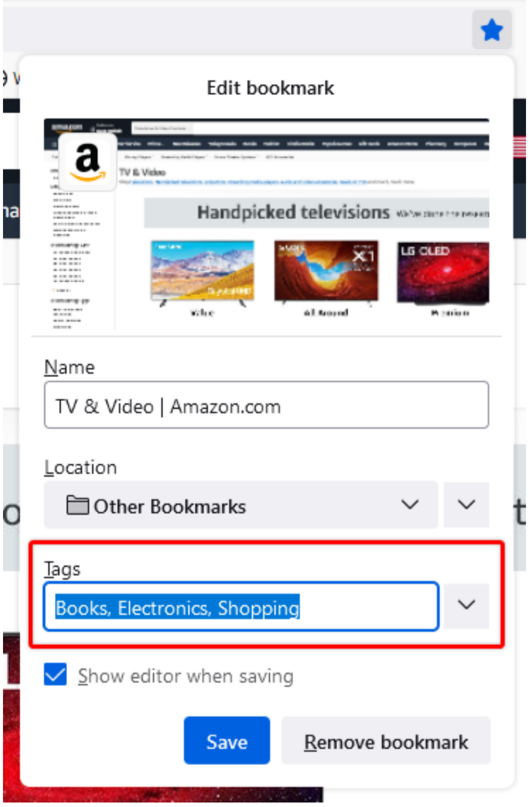

Bookmark
Name: wikiHow - How to do anything
Folder: Bookmarks Bar

Figure 8 : Folder based bookmarking on Google Chrome browser
Image Credit : https://www.wikihow.com/

Browser built-in bookmarking gives the functionality to categorize or tag bookmarks by manually creating directories / tags, which is a time-consuming tedious task. There are several bookmark managing extensions/applications out there, but they also use manual categorization.



Edit bookmark

Name

TV & Video | Amazon.com

Location

Other Bookmarks

Tags

Books, Electronics, Shopping

☑ Show editor when saving

Figure 9: Bookmark Tagging on Firefox browser
Image Credit : https://user-media-prod-cdn.itsre-sumo.mozilla.net/

Mozilla Firefox has built-in bookmark tagging functionality, but they also use manual tagging, which needs human intervention to select / create related tags. Tagging a web resource to a new category needs to create that tag manually.

2.2.3 Limited Bookmark Search Functionality

Common web browser bookmark search functionality only uses page heading and URL text to find saved bookmarks, this omits all the key text from the web page. This becomes a problem when you have thousands of bookmarks stored on your browser. You may have faced the scenario where you know that you have a bookmarked page but was not able to find it back since bookmark search query text does not match with page heading text nor URL text, forcing you to track down this page again by searching through the web which wastes a lot of time.

2.3 Personal Experience & Capability

Personally I have gathered thousands of bookmarks over the years which I had categorized using existing folder based categorization which made me realize how inflexible and time-consuming is the bookmark categorization functionality in existing bookmark managers. More than that, the most frustrating issue was finding back already bookmarked resources since search functionality in the existing bookmark manager was pretty limited.

Since I personally know about this problem well and have researched other commercial, non-commercial bookmark managing solutions out there, I am confident that I could implement a better solution than existing ones with the help of my front-end and machine learning knowledge.

3. Major issue - Bookmark Categorization and Search

Bookmark categorization is typically done by using folder hierarchies in Chromium based browsers, which have limited flexibility for categorizing since hierarchical structure does not facilitate categorizing a single resource to multiple distinct categories at once.

Another issue with the bookmark categorization process is it is fully manual and time-consuming, even though there is tagging functionality provided by Firefox browser and third party Chrome extensions they all involve manual categorization.

Finally, the bookmark search in all browsers and third party extensions are limited to pattern matching of resource page title or URL text, which make it harder to find already bookmarked resources since page title or URL text may not match with search query text in many cases.

So we will be addressing the above-mentioned current limitations of bookmark categorization and search in our proposed solution.

4. Aim & Objectives

4.1 Aim

To develop a cross-browser extension to tag bookmarks automatically using machine learning by using bookmark resource content as the input, and to support better bookmark search using tags**.**

4.2 Objectives

- To conduct a critical review of existing bookmark manager solutions.
- To do in-depth study of ML algorithms, techniques used to label unstructured web content and to do the labeling (i.e.: bookmark tagging) in a resource efficient and privacy-friendly manner.
- To build an ML model with higher accuracy that can be used for bookmark tagging.

- To develop a cross-browser extension that uses a developed ML model locally to auto tag bookmarks efficiently and in a privacy-friendly manner.
- To conduct the formal evaluation of the browser extension with the accompanied ML model.

5. Smart Bookmark Manager Using Machine Learning

We propose improved bookmark manager extension that address previously described drawbacks of existing solutions, as described below

Drawbacks	Solutions
Limited flexibility in folder based bookmark categorization and bookmark tagging, only being limited to Firefox.	Cross-browser bookmark manager extension which allows categorizing bookmarks using tags which allow categorizing a bookmarked resource under multiple distinct topics simultaneously using tags.
Manual categorization of bookmarks.	Using ML to automatically tag bookmarks based on the content of the bookmark resource (ex: web page content).
Limited bookmark search functionality.	Since ML generates bookmark tags based on the content of the bookmarked resource, it increases the chance of finding the bookmark later on through search. As an additional benefit, it also facilitates filtering down large numbers of bookmarks using multiple tags.

Table 1 : Solutions to existing drawbacks with the smart bookmark manager

In addition to addressing above-mentioned drawbacks, this browser extension will also provide the following benefits

- Keep user bookmark data private by tagging bookmarks locally on the client using a downloaded machine learning model. (on-device machine learning) Otherwise by analyzing a collection of bookmarks of a user, we can predict that user's interests and behavior which can be considered as invasion of user privacy.
- Browser extension will utilize less client resources by using
    - Less resource intensive ML algorithms over Neural Networks.
    - On-device ML model, which does not involve network requests.

5.1 Use case Diagram

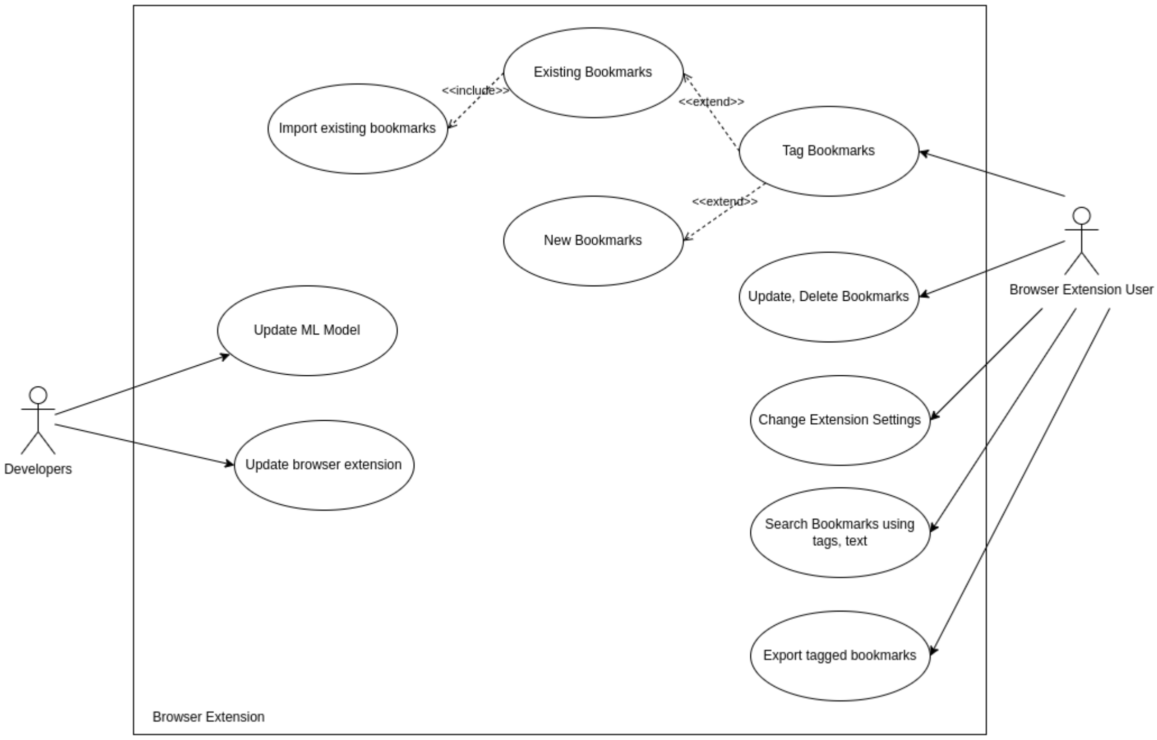


Figure 10: Use Case Diagram

Given above is a use case diagram of the system which depict major use cases and actors. According to the proposed solution, the browser extension user can do following

- Tag existing and new bookmarks.
- Update, delete existing bookmarks / tags.
- Search bookmarks using tags, text.
- Export already tagged bookmarks.
- Tweak extension settings to configure extension's default behavior.

Developers who develop the extension can do following

- Update extension source code ex: bug fixes, new features
- Update ML model with an improved model

5.2 Building & Evaluating the machine learning Model

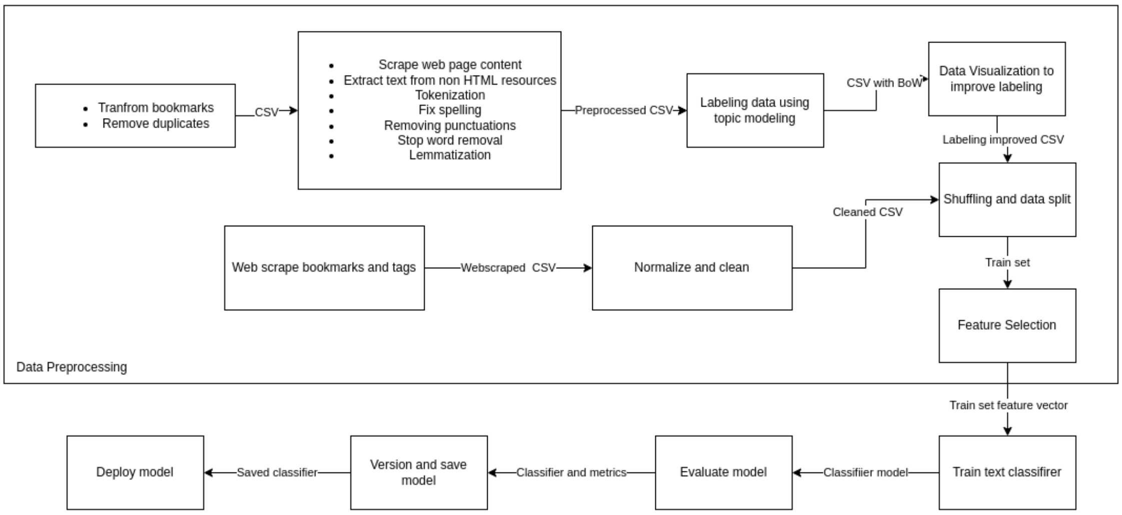


Figure 11:Building & Evaluating the machine learning model

The diagram given above provides an overview of how the ML model was built and evaluated. Main steps of this model building and evaluating process include

1. Data Pre-processing
2. Training Classifier / ML model and Versioning and saving the ML model
3. Evaluating trained ML model
4. Deploying the model

Each of these steps will be described in detail in upcoming sections.

5.2.1 Data Pre-processing

In order to train the ML model we will clean and transform data which is called data pre-processing. Data pre-processing includes following steps

- Transform browser exported bookmark HTML to ML friendly input format like .CSV or .JSON while using existing directories names, tags as tags.
- Removing duplicate bookmarks.
- Scrape the bookmark resource page content and exclude common webpage content like headers, navigation and footers. For images, extract their alt text.
- Scraping text from non HTML resources using Textract.
- Pre-process the scraped content by applying the following natural language preprocessing techniques
  - Tokenization : Breaking unstructured data into chunks or smaller subtexts. This is useful to identify meaningful keywords from the text.
  - Fix spelling mistakes programmatically.
  - Removing punctuations.

    - Stop word removal : Removing words that commonly occur across the document. This removes articles and prepositions.
    - Lemmatization : Analyzing word forms and identifying the context. Then converting the word into its correct base form. It groups different forms of the same word, which is very useful in reducing redundant words. For example, lemmatization identifies the correct base form of the words "studying, studied" as study.

In addition to pre-processing browser exported bookmarks, we will also web scrape publicly available bookmarks and their tags from websites like Diig[11], Pinboard[12], Bibsonomy[13] in order to increase the size of the dataset. Increasing the size of the data set is important since bookmarked resource can be belonged to a vast array of topics.

5.2.2 Labeling data

- After data preprocessing, that data needs to be labeled in order to feed into a classification model. Initially we will be using unsupervised topic modeling specifically LDA Algorithm from Genism to help with labeling of data since manually labeling a thousand bookmarks is a tedious and time-consuming task. The topic modeling approach uses unsupervised machine learning to analyze and create BoW for every resource, where the top ten words in BoW with the highest frequency would be considered as the tags for a particular resource.[14], [15]
- Then we will explore the labeled data through data visualization and improve the labeling where it is necessary.
- After we shuffle the data to avoid ordering bias and split the labeled data to training set and testing for later use. For this, we will be using Scikit Learn[16].
- Then we will decide on which features to use for training using manual selection and Scikit feature selection algorithm.

5.2.3 Training the model

We will be training a classifier model using a text classification algorithm from Scikit learn by using previously created train set as the input. Then trained models will be versioned and saved to the disk for evaluation.

5.2.4 Evaluating & Improving the model

We will test the accuracy of trained models using a test set. Then we will re-train and re-evaluate the models by changing

- Feature set
- Classifier model parameters

- Dataset

We will keep a record of evaluation metrics like accuracy for each classifier model and select the model with the best performance to use with the browser extension.

5.3 Designing & Developing Cross-browser Extension

After the building and evaluating the ML model, we will start designing and developing the accompanied browser extension in following main steps.

1. Designing cross-browser extension
2. Developing cross-browser extension
3. Testing
4. Deploying the extension.

Each of these steps will be described in the following sections.

5.3.1 Designing Cross-browser Extension

In this phase we will

- Identify and analyze system requirements.
- Draw UI mockups following best UX practices.
- Design JavaScript module structure for more modular, testable code.

5.3.2 Developing Cross-browser Extension

In the developing phase, we will start doing following

- Bootstrapping Cross-Browser Extension Project
- Develop User Interfaces
- Integrating browser extension with ML model
- Implementing rest of the browser extension functionality

5.3.2.1 Bootstrapping Cross-Browser Extension Project

Bootstrap cross-browser extension project using a browser extension development GitHub template, which eliminates the boilerplate code needed for cross-browser extensions.

5.3.2.2 Develop User Interfaces

Developing UIs for following interfaces.

- Importing existing bookmarks for auto tagging.
- Managing bookmarks and tags.

- Suggest relevant bookmark tags when a user makes a new bookmark, and allow users to change the suggested tags.
- Search bookmarks using tags and text.
- Settings to automatically download and use new models if a new ML model update is available.
- Show Help and About information.

5.3.2.3 Integrating browser extension with ML model

- Using ML model to tag existing bookmarks.
- Using ML model to suggest bookmarks when a new bookmark creation event is triggered.
- Check for a new version of ML model for download.

5.3.2.4 Implementing rest of the browser extension functionality

- Update, delete bookmarks and tags.
- Search for bookmarks based on tag filtering.

5.3.4 Testing

Testing whether browser extension work according to the expectations. This includes writing

- Unit tests
- Integration tests
- End to end functional tests for major use cases.

5.3.5 Deployment

Releasing installable extensions as a GitHub release.

5.4 Technologies

The table given below lists which technologies, libraries are used to develop the proposed solution with the task they are used for.

Task	**Technologies / Libraries Used**
Data Preprocessing	Pandas, NLTK[17], Autocorrect[18], Scikit Learn
Web Scraping	Beautiful Soup[19]

To extract text from different file formats	Textract[20]
Topic modeling for labeling data	Gensim[21]
Training classifier model	Scikit Learn
Browser Extension development	Chrome Extension and Firefox Addon APIs, JavaScript / Typescript
UI styling	CSS
Building UI components	React & UI components
Bookmark and tag storage	SQLite
Testing	Jest, Playwright
Source control and storage	Freely available (Git and GitHub)
Task management	Freely available (GitHub issues)
Continuous deployment	Freely available (Github Actions)
Evaluating and recording model metrics	Jupyter Notebooks

Table 2 : Technologies used for training ML model and building extension.

5.5 Feasibility Study

In this section, we will answer the question "Is it possible to implement the proposed solution under the current constraints?". Feasibility study will be done in following areas

Feasibility	Purpose
Technical	Check if the project is feasible resource and technical skill wise.
Economical	Check if the project can be completed without big economic expense.
Functional	Check if the project can meet expected functional criteria.
Operational	Check if the project can be developed, maintained under the given constraints. (ex: Time, Maintainability)

Table 3 : Type of feasibility study and their purpose.

5.5.1 Technical Feasibility

5.5.1.1 Resource Feasibility

Resource	Availability / Pricing
Programming device (laptop or desktop)	Already available
Code editor	Freely available (VS Code)
At least two browsers installed, one from Chromium based (ex: Google Chrome, Edge, Brave) and other from Firefox or Safari	Freely available (Google Chrome and Firefox)
Internet Connection	Already available
Collection of bookmarks with or without tags/ directories	Already available
Libraries for data pre-processing and ML	Open sourced and freely available
Resources to learn about data preprocessing and ML.	Freely available as library documentation, blog posts and books.
Model hosting	Freely available (GitHub releases)

Table 4 :Resource Feasibility

5.5.1.2 Technical Skill feasibility

Skill	Possible?	Reasoning
Understanding ML and training, evaluating ML models	Yes	● Have a high interest in ML domain. ● Course work includes a lot of material related ML and data science. ● Plenty of free resources to learn about ML, Data Science concepts.
Programming and best practices	Yes	● Have prior programming experience with the course work and professionally. ● Documentation, tutorials and forums like StackOverflow can be used to learn new unfamiliar technologies, APIs.

Table 5: Technical Skill feasibility

5.5.2 Economic Feasibility

Since most resources necessary are freely available and the project is developed by myself without any hires, the project development budget is close to 0 LKR which makes the project economically feasible.

5.5.3 Functional Feasibility

	Requirement	Possible?	Reasoning
UI & UX	Browser extension UI and its behavior should be self understandable and should be responsive to user feedback.	Yes	Using common UI patterns and UX best practices.
Accuracy	Model should predict tags correctly at least 70% of the time	Yes	We will use supervised machine learning to train the model, which will have better accuracy than unsupervised methods.
Performance	Model should suggest tags for bookmark under 2 seconds	Yes	ML model will be stored locally thus eliminating network latency and model will use less resource intensive ML algorithm instead of resource intensive Neural Networks.
Security	Browser Extension should be free of common security vulnerabilities	Yes	Will follow best practices outlined in OWASP [22]
Privacy	Application should not expose private data (ex: bookmarks) to any party.	Yes	None of the user bookmarks will be sent outside the client device, since bookmark tagging will be done using a local ML model.

Table 6: Functional feasibility

5.5.4 Operation Feasibility

5.5.4.1 Timeline Feasibility

In the appendix action plan (on the last page) we have doubled the required time as a good practice since humans have a tendency to underestimate time required, or we may run into unexpected roadblocks while developing the project. This intentional overestimation makes it easy to reach deadlines within the given timeframe. In case of unexpected events like falling ill we will drop unessential things like UI aesthetics, certain low priority features and certain unit tests to meet the deadlines.

5.5.4.2 Maintenance Feasibility

We will be using mature, industry widely used technologies for the project, which even allow other contributors to help with maintenance of the project in the future. And we are following industry used development best practices for the project, which ease the long term maintenance.

6. References and Citations

[1] "GitHub - quenhus/uBlock-Origin-dev-filter: Filters to block and remove copycat-websites from DuckDuckGo, Google and other search engines. Specific to dev websites like StackOverflow or GitHub." https://github.com/quenhus/uBlock-Origin-dev-filter (accessed Mar. 23, 2022).

[2] quenhus, uBlock-Origin-dev-filter. 2022. Accessed: Mar. 23, 2022. [Online]. Available: https://github.com/quenhus/uBlock-Origin-dev-filter

[3] "Google Q4 search ad revenue: $43.3 billion," Search Engine Land, Feb. 02, 2022. https://searchengineland.com/google-q4-2021-earnings-379735 (accessed Mar. 23, 2022).

[4] "Search Engine Market Share Worldwide," StatCounter Global Stats. https://gs.statcounter.com/search-engine-market-share (accessed Mar. 23, 2022).

[5] "google.pdf." Accessed: Mar. 23, 2022. [Online]. Available: http://infolab.stanford.edu/pub/papers/google.pdf

[6] "Tell HN: Google returning 'Untitled' results that redirect to malware/spam | Hacker News." https://news.ycombinator.com/item?id=30117388 (accessed Mar. 23, 2022).

[7] "Privacy concerns regarding Google," Wikipedia. Jan. 28, 2022. Accessed: Mar. 23, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Privacy_concerns_regarding_Google&oldid=1068485228

[8] "Filter bubble," Wikipedia. Feb. 26, 2022. Accessed: Mar. 23, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Filter_bubble&oldid=1074161056

[9] "Google bias and censorship revealed - YouTube." https://www.youtube.com/watch?v=uf6I1tjmm3o (accessed Mar. 23, 2022).

[10] "Google sued again over anti-competitive search practices - BBC News." https://www.bbc.com/news/business-55357340 (accessed Mar. 23, 2022).

[11] "Public library - bobducharme - Diigo." https://www.diigo.com/profile/bobducharme (accessed May 02, 2022).

[12] "pinboard." https://pinboard.in/recent/

[13] "BibSonomy." https://www.bibsonomy.org/ (accessed May 02, 2022).

[14] "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python | by Susan Li | Towards Data Science." https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24?gi=849fb8c27552 (accessed Mar. 23, 2022).

[15] "Latent Dirichlet allocation - Wikipedia."

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation (accessed Mar. 23, 2022).

[16] “scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/index.html (accessed Mar. 23, 2022).

[17] “NLTK :: Natural Language Toolkit.” https://www.nltk.org/ (accessed Mar. 24, 2022).

[18] “autocorrect · PyPI.” https://pypi.org/project/autocorrect/ (accessed Mar. 24, 2022).

[19] “Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation.” https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed Feb. 26, 2022).

[20] “textract — textract 1.6.1 documentation.” https://textract.readthedocs.io/en/stable/ (accessed Mar. 24, 2022).

[21] “Gensim: Topic modelling for humans.” https://radimrehurek.com/gensim/ (accessed Mar. 24, 2022).

[22] “OWASP Web Security Testing Guide | OWASP Foundation.” https://owasp.org/www-project-web-security-testing-guide/ (accessed Mar. 23, 2022).

Appendix : Action Plan

Task / Month No	1	2	3	4	5	6	7	8	9	10
Project Analysis and Design	🟨									
Data Preprocessing	80	🟥	🟥							
Data Labeling	80			🟦	🟦					
Training and evaluating ML model	160					🟪				
Developing browser extension	160						🟨	🟨		
Testing browser extension	80								🟩	
Delivery										🟦