Software Requirement Specification Document for SEO Pro

Farah Khaled, Farah Essam, Hanien Samy, Hashim Said
Supervised by: Dr. Fatma Helmy, Eng. Nour El Huda Ashraf

May 16, 2023

Table 1: Document version history

Version	Date	Reason for Change
1.0	24-Nov-2022	SRS First version's specifications are defined.
2.0	4-Dec-2022	SRS updated specifications are defined.
3.0	2-May-2023	New features are added and updated specifications are defined.

GitHub: `https://github.com/farahkhaled1/seopro`

Contents

Abstract

Since the pandemic, the importance of a digital presence for all businesses worldwide had been exponentially growing. Hence comes the fact that the electronic market is overwhelmed by many websites. The challenge is to ensure that your target audience sees the content you create in their search query results. Many companies in Egypt manually provide search engine optimization (SEO) services, but none offer automation for this process. Arabic websites also grabbed our attention since very few Arabic SEO tools are introduced to the middle eastern market. Our project aims to help business owners by facilitating the SEO process through keyword extraction in both Arabic and English, as well as website content substitution with better synonymous keywords that rank. Furthermore, the process will be eased by analyzing each keyword's CPC and search volume, as well as generating ads and blog content for the website itself. We apply state-of-the-art natural language processing techniques to offer a solid academic background for our project.

1 Introduction

1.1 Purpose of this document

The objective of this document is to present and highlight the distinct requirements needed to develop SEOPro's desired web application. This document illustrates the main elements required in the system's development process, including the functional and non-functional requirements. The intended audience for this document includes all stakeholders, from students to professors, as well as any developer who works on this project. We also provide future project users with a complete description of each processing stage, inputs, and outputs.

1.2 Scope of this document

SEOPro provides an editor for both those who already have a website and those who are still trying to develop theirs, either in English or Arabic. **The scope of this document** is to show the foundation of our project so that any developer who is a part of this project or might take part in the future development or maintenance of this web application could understand the solution we are providing. It describes the application's requirements and functionalities to provide a clear outline of the system.

1.3 Business Context

From Professional Services, to Home Repairs, Medical Practices, Online Businesses or even Freelancers, all these businesses and more need Search Engine Optimization in order to increase their Return on Investment. Here comes the role of SEO or Search Engine Optimization, it is a variety of techniques and strategies used to make the website's owner's content as much as relevant to his potential customer which leads to more organic exposure on the Search Engine Results Page. SEO automation had been out-of-focus in the Arabic world neglecting a whole new market of Arabic website owners who would majorly need such an Arabic based service. Our goal is to make an all inclusive SEO tool that analyzes a website rank loading time aiming for website optimization while generating keywords in both Arabic and English. We have introduced new features to optimize website performance. The Domain Analysis tool offers insights into domain authority, backlinks, organic search traffic and other crucial metrics to measure ranking analysis. The Website Speed Checker measures a website's loading time and recommends efficient solutions for website optimization. These features help users enhance online visibility and drive profitability.

2 Similar Systems

2.1 Academic

In this paper [1], A completely generalized unsupervised machine learning algorithm is used to obtain content that is high in quality as the content is typically displayed by the current search engines depending on a variety of factors, not just the content's quality. The output from the top-ranked URLs will be shown to the user and compared to the results from conventional search engines in terms of quality and relevance of information. The ranking algorithm used consists of the following four phases:

1. TF-IDF algorithm used to calculate the numeric value of each keyword.
2. Singular value decomposition (SVD) algorithm used to decompose the TF-IDF matrix.
3. K-means clustering algorithm used to cluster the documents and choose the largest cluster to apply the rating algorithm to.
4. Sort the documents by how far apart they are from the cluster representative.

With the help of a webpage content extractor, the rating algorithm's input data is gathered. Data extraction is simplified by the usage of Python. The rating algorithm is made up of sub-processes that cluster and classify the textual material into clusters before ranking each search result according to its content.

In this paper [2], four methods for extracting keyphrases from Arabic documents are compared. The KP-Miner keyphrase extraction system is used in the first technique. The second method filters some patterns that can be weighted using the token frequency-inverse document frequency (TF-IDF) algorithm using Arabic natural language processing tools (stemmer and part of the speech tagger). The third approach measures the similarity between the candidate pattern and the document title to determine the weighting of the resultant patterns using Google's Word2Vec package. The weighting results from the second and third methods are combined in the fourth approach. To create a list of candidate keyphrase patterns that are weighted based on the specified methods, the system combined the POS tagger and the stemmer. The system experimented on a corpus of Wikipedia articles to gauge how well the methods utilized were working.

The aim of this paper [3] is to determine the keyword combinations that will improve search engine optimization using the technologies of Term Frequency-Inverse Document Frequency (TF-IDF), K-means clustering and indexing quality inspection. The study showed that it can successfully improve a website's growth in search engine ranking, boost website exposure level, and increase click through rate. TF-IDF is a technique that makes use of word frequency statistics. It is a weighting approach that is frequently applied to text mining and data mining to disclose the level of a word in a certain article. while the goal of K-means classification is to group a large number of items in two or more dimensions according to an attribute of those elements in order to determine the medium in which each element's natural distribution occurs. Users are more accustomed to using the data needed by active search engines, so keyword advertising really links the information customers are now looking for to the advertisement and displays it to them. figure 1 shows how it appeared when it was uploaded on the search engine results page (SERP).In a nutshell, Utilizing the research's suggested technique, it is possible to efficiently retrieve key word combinations that will benefit the website.

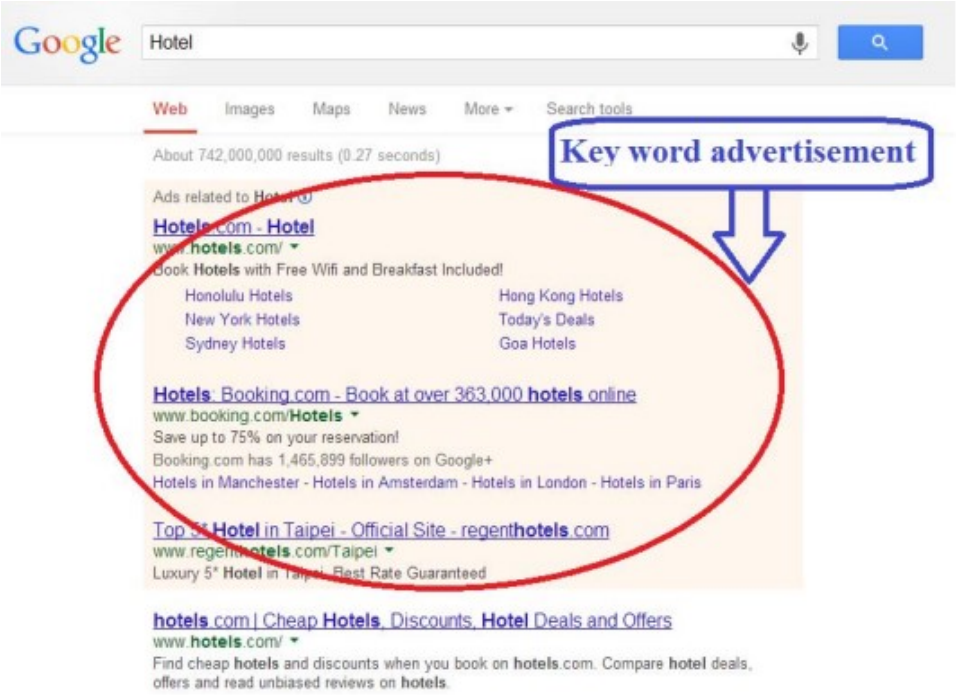


Figure 1: SERP

2.2 Business Applications

1. Wordtracker[4] : "All you need to do is start searching" .As shown in figure 2 wordtracker helps businesses grow, expand and reach a bigger audience by getting keyword results one can rely on and Find lucrative new market areas.

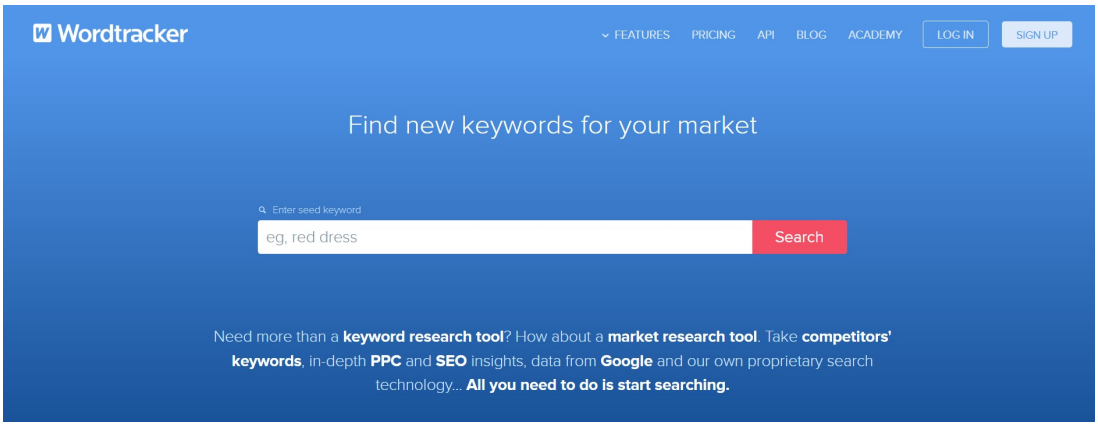


Figure 2: wordtracker

2. Umbrellum[5] : Umbrellum contains great features hence it can be used as a ranking tracker, It can be used to track SERP(Search Engine Results Page) features. one can also keep track of their competition. All in all, Wordstream is considered a great SEO(Search Engine Optimization) tool as shown in figure 3 and in figure 4.

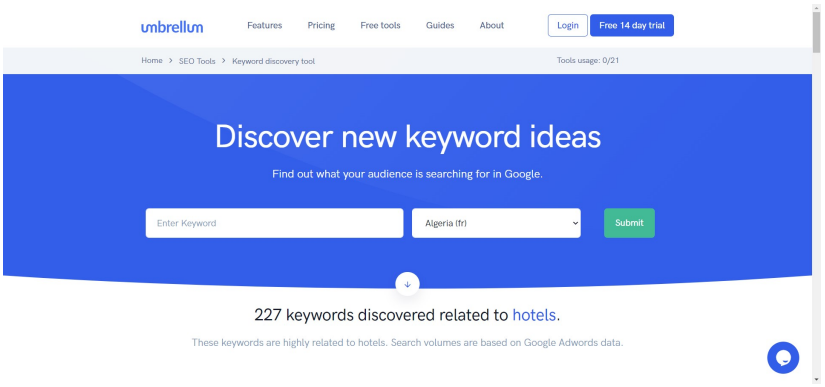


Figure 3: Umbrellum

Figure 4: Umbrellum

3 System Description

3.1 Problem Statement

SEOPro aims to ease the process of SEO for businesses by making the SEO hassle easier for them. Improving a website's performance and search engine ranking has always been a rough process. Our contribution to the problem is:

• For those who are just starting to develop their web applications, we aim to generate high-ranking content such as blogs in English in order to ease the process of website content creation for them. Additionally, our new feature, the Domain Analysis Tool, provides insights into domain authority, backlinks, and organic search traffic to identify areas for improvement.

• For those who already have existing web applications, we aim to generate synonyms for both English and Arabic web applications to help them find their weak points in their existing websites. Additionally, The Website Speed Checker measures loading times and suggests areas of improvement for better load times. Our new feature, Image Optimizer, optimizes website images for faster load times, which is important for SEO.

3.2 System Overview

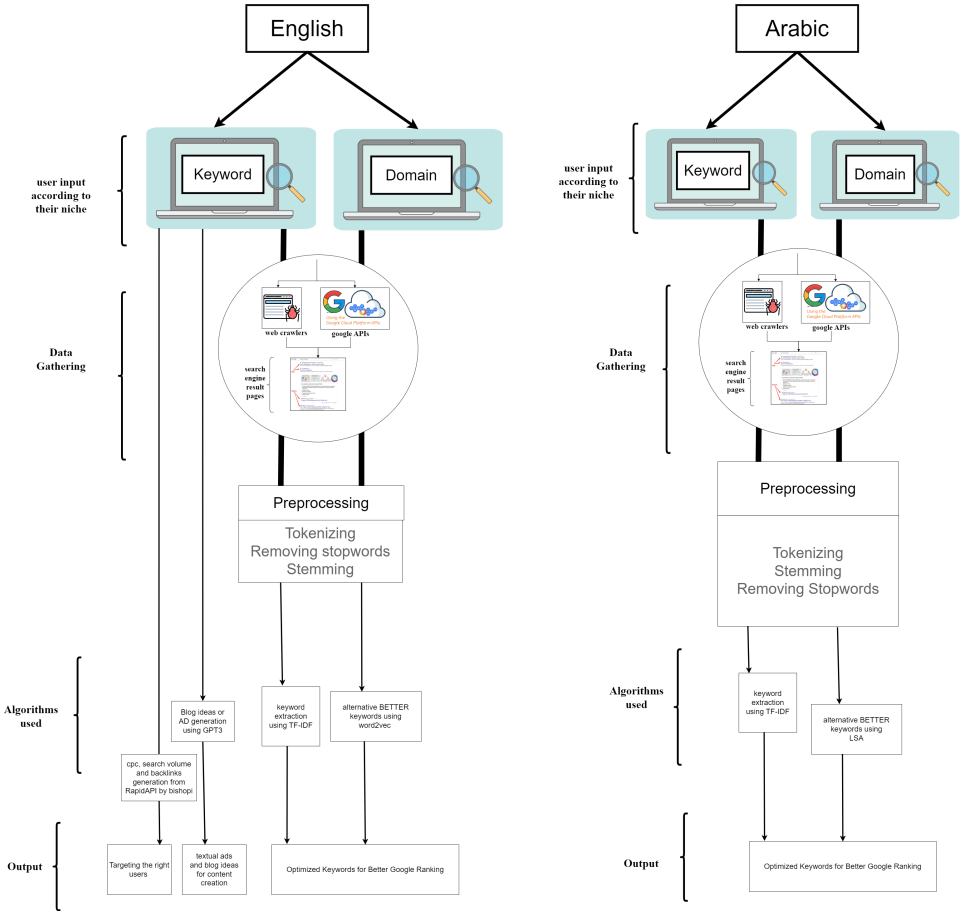


Figure 5: SEOPro System Overview

Figure 5 shows the flowchart of the proposed project. It contains the following steps:

- User input: A string input will be required from the end user in order to understand the keywords we need to optimize the search engine on.
- Data Gathering : The string input will be used as the key in order to target the data needed to be gathered. By using google APIs and web crawling method, The data will be extracted.
- Preprocessing : In order to function the extracted data, some preprocessing will be applied such as the process of tokenizing, stemming and removing 'stop words' in order to function some methods.
- Methods:
  - 'TF-IDF' algorithm will be implemented to learn the importance and relevance of the words in the generated sentences. A static measure will be applied to the data and a classified output will be claimed.
  - To generate blog ideas or ad copy, GPT-3 will be used only based on the keyword's the user provided related to his niche.

  - word2vec algorithm will be used to extract synonyms by using point-wise mutual information to detect the similarity of a pair of words in English, then replace them with higher ranking keywords from the previously generated data. However, LSA algorithm will be employed to perform the same function on the Arabic keywords.

- Output:
  - A more suitable alternative keywords will be suggested to rank higher on google search.
  - Textual advertisements and blog ideas will be generated for the user which will help him in the website content creation process.
- End target: A Detailed user friendly interface will be carried out to the end users showing the proposed keywords that can be used with statistics and other perks.

3.3 System Scope

This system is mainly designed to help our users easily manage and optimize their website's search engine optimization (SEO) efforts. This implies that users will be able to improve their website's ranking in organic search results, which may result in more traffic to their website and potential new customers. Our goal is to provide a user-friendly and efficient tool that assists users in their SEO efforts. The proposed system is capable of:

Keyword Research: Using Natural Language Processing, this feature provides users with high-ranking keywords in both Arabic and English to help them optimize their website's content and improve their SEO. The tool can be used by both those who are starting to develop their websites and those with already existing websites.

Domain Analysis: This feature allows users to analyze their website's domain performance, including metrics such as domain authority, backlinks, organic search traffic and much more. It can help users identify areas for improvement as well as track their competitors' domains.

Website Image Optimization: Images are often a significant contributor to a website's load time, which can negatively impact SEO. This feature allows users to scrape all images within a given URL and optimize them. Users can download all their optimized images with one button press.

Website Speed Checker: This feature measures how long a website takes to load and analyzes its performance. It provides users with an analysis of the website's load time and performance. The Website Speed Checker is important for SEO because search engines prioritize websites that load quickly, and it helps users optimize their website for faster load times.

Blog Ideas and Content Generator: This feature generates new blog ideas and content to remove writer's block and ease the process of content creation. It provides users with a list of potential topics and ideas based on their website's niche and industry.

By providing these features, we aim to make SEO easier for our users, allowing them to focus on their business and website's growth.

3.4 System Context

As shown in figure 6, The application will start by requesting the user to identify themselves, If they are new to building a website or that they already have one. Then They will enter the input required.The owners will enter their domain and have suggested better keywords. On the other hand, The new builder will be shown the keywords he can insert to rank higher.

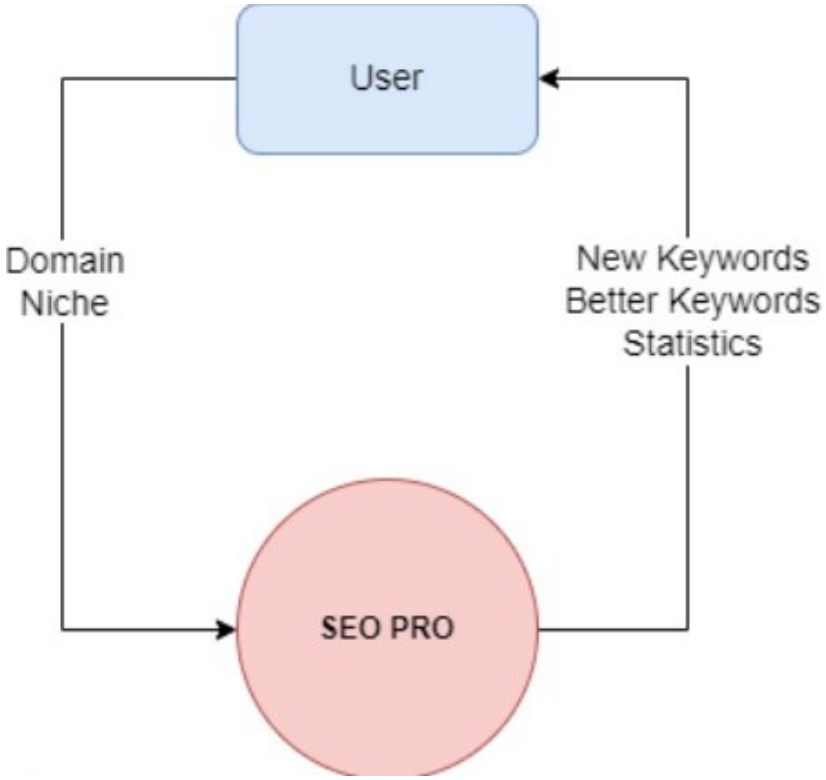


Figure 6: System context

3.5 Objectives

Our system aims to provide website owners and marketers with a hassle-free and efficient SEO experience by:

- Providing the best and most relevant keywords in Arabic and English to help improve website ranking on SERP using the TF-IDF algorithm. This saves an average of 10 days in keyword research time.
- Offering AI-powered tools such as Natural Language Processing to analyze and interpret text and spoken language like humans.
- Replacing website content with better synonymous keywords in English (using word2vec algorithm) or Arabic (using LSA algorithm) to improve website ranking.
- Offering a domain analysis tool that allows website owners to track and analyze their domain's performance, including metrics such as domain authority, backlinks, and organic search traffic. Users can also track their competitors' domains.
- Optimizing website images for faster load times using the Website Image Optimizer feature, which converts images to the best format for websites (webp) and downloads them to the user's device.

- Providing a Website Speed Checker that measures a website's loading time and suggests efficient solutions to improve load times in case it is below average.
- Generating new blog ideas and content using the GPT-3 API to help website owners avoid the hassle of finding new ideas for blog and website content creation, as well as Ad copywriting for paid Google Ads.

3.6 User Characteristics

- The user is looking for a solution to help them improve their website's SEO and save time on SEO tasks.
- The user shall understand the importance of ranking their website.
- The user can be of any age and must have internet access.
- The user must have a website or plan to create one.
- The user should have a basic knowledge of search engines and how SEO works.

4 Functional Requirements

4.1 System Functions

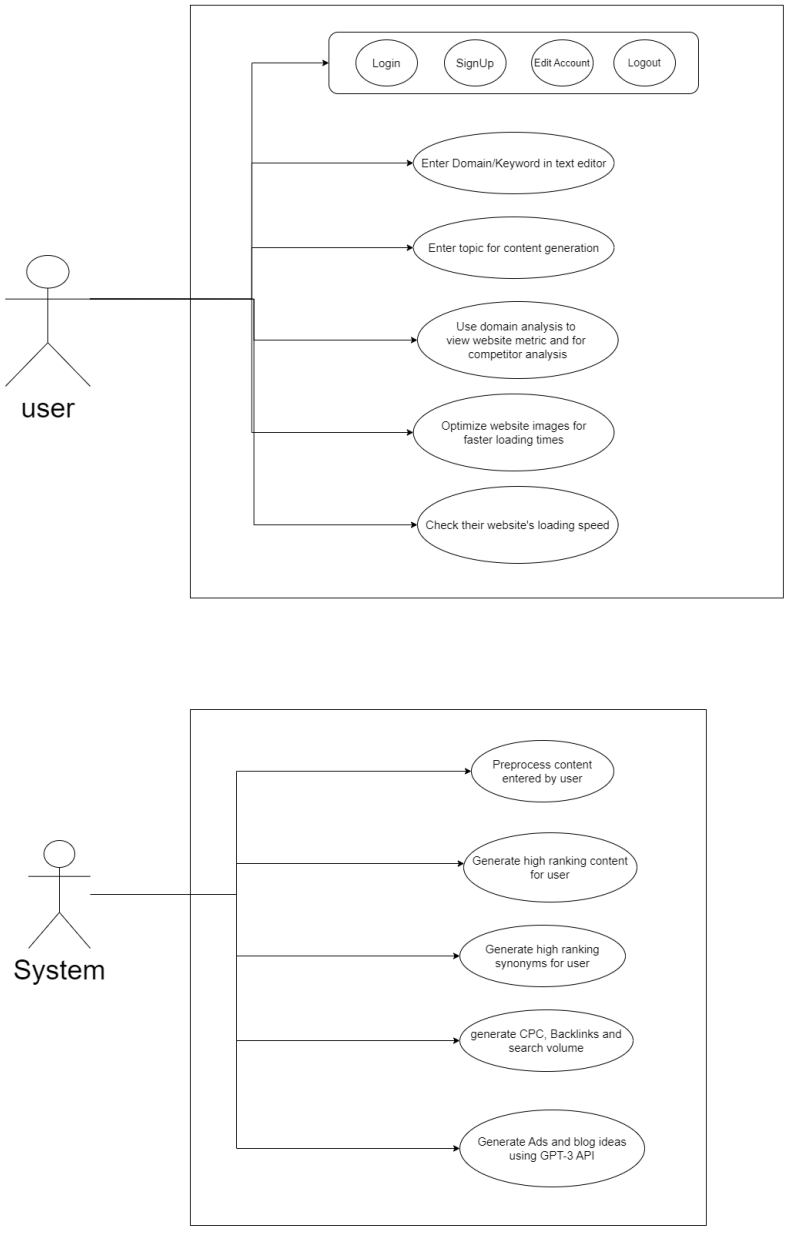


Figure 7: Use Case Diagram

- **FR01** The user shall be able to sign up to create a new account.
- **FR02** The user shall be able to login to their account.
- **FR03** The user shall be able to edit their account.
- **FR04** The user shall be able to log out of their account.

- **FR05** The user shall be able to enter domain/keyword of their web application in the provided text editor.
- **FR06** The user shall be able to input text for Content generation.
- **FR07** The user shall be able to use the domain analysis tool for competitor analysis
- **FR08** The user shall be able to optimize website images for faster loading times using the website image optimizer.
- **FR09** The user shall be able to check their website's loading speed using the website speed checker.
- **FR10** The system shall be able to perform pre-processing techniques on content entered by user.
- **FR11** The system shall be able to generate high ranking content for the user to use in their web application to help them rank better on SERP.
- **FR12** The system shall be able to generate high ranking synonyms for users with an already existing website to use in their web application to help them rank better on SERP.
- **FR13** The system shall be able to generate Ads and Blog ideas using GPT3 API.
- **FR14** The system shall be able to generate the domain's global rank, search volume, traffic, domain authority, SEO difficulty, indexed pages, backlinks, CPC and many more met

4.2 Detailed Functional Specification

Table 2: User Input

Name	User input
Code	ID05
Priority	Extreme
Critical	10/10
Description	The user shall be able to enter domain/keyword of their web application in the provided text editor
Input	The user must enter a keyword/domain
Output	high ranking content or synonyms depending on if the user already has a website or just starting one.
Pre-condition	client must login/signup to SEOpro
Post-condition	The generated content is stored in database
Dependency	Client must input a keyword/domain and have an account on SEOPro.
Risk	None

Table 3: Ad generation input

Name	Ad generation input
Code	ID06
Priority	Extreme
Critical	10/10
Description	The user shall be able to input text for content generation
Input	The user must enter text.
Output	Ads will be generated for the user to use.
Pre-condition	client must login/signup to SEOpro.
Post-condition	The generated Ads is stored in database
Dependency	Client must input a keyword/domain and have an account on SEOPro.
Risk	None

Table 4: View statistics

Name	Statistics
Code	ID07
Priority	high
Critical	7/10
Description	The user shall be able to use domain analysis tool for competitor analysis.
Input	Domain Name (URL)
Output	Global rank, search volume, traffic, domain authority, SEO difficulty, indexed pages, backlinks, CPC , organic keywords for each client's web application
Pre-condition	Client must have an existing website and an account on SEOPro.
Post-condition	Client's statistics is stored in database
Dependency	Client must have an existing website.
Risk	This function may fail to show each client's statistics.

Table 5: Website image optimizer

Name	Website image optimizer
Code	ID08
Priority	Mid
Critical	6/10
Description	The user shall be able to optimize website images for faster loading times using the website image optimizer
Input	The user must enter their website URL
Output	Optimized images in Webp format with reduced file size
Pre-condition	The user must be logged in and have an existing account.
Post-condition	Optimized images are downloaded to user's device instantly.
Dependency	None.
Risk	Image quality may be compromised due to optimization.

Table 6: Website speed checker

Name	Website speed checker
Code	ID09
Priority	High
Critical	8/10
Description	The user shall be able to check their website's loading speed using the website speed checker
Input	The user must enter their website's URL
Output	Website's loading speed
Pre-condition	The user must be logged in and have an existing account.
Post-condition	In case loading time is below average, useful suggestions are displayed to work on that issue.
Dependency	None.
Risk	The website's loading speed may vary due to network connectivity.

Table 7: Content generation

Name	High ranking Content generation
Code	ID11
Priority	Extreme
Critical	10/10
Description	The system shall be able to generate high ranking content for the user to use in their web application to help them rank better on SERP.
Input	User must login and then he shall enter a keyword/topic related to their niche.
Output	High ranking keywords shall be generated for the client to help them rank higher on SERP.
Pre-condition	Client must input a keyword.
Post-condition	The generated content is stored in database
Dependency	Client must input a keyword.
Risk	None.

Table 8: Synonyms generation

Name	High ranking synonyms
Code	ID12
Priority	Extreme
Critical	10/10
Description	The system will generate high ranking synonyms for the user to help their web application to rank better on SERP
Input	The user must login and enter a string in order to understand the keywords we need to optimize the search engine on
Output	High ranking synonyms will be generated
Pre-condition	Client must input a keyword or have an existing website
Post-condition	The generated content is stored in database
Dependency	Client must input a keyword or have an existing website
Risk	None

Table 9: Ads generation

Name	GPT3 API
Code	ID13
Priority	Extreme
Critical	10/10
Description	The system will generate Ads and Blog ideas using GPT3 API
Input	The user must login and input a keyword
Output	Textual Ads and Blog ideas will be generated for content creation
Pre-condition	Client must input a keyword
Post-condition	The generated content is stored in database
Dependency	Client must input a keyword
Risk	None

5 Non-functional Requirements

- Usability : Having a user-friendly view is essential to make our user comfortable enough to use the app. Users shall find it easy to use the interface; the functionalities will not be hard which will take the users no time to understand it .
- Maintainability : the system should be sustainable , which means it can evolve to meet the changing needs .
- Extensibility : The system should cater to future changes through flexible architecture
- Availability : The system will be a web application that can work on any browser or device. An internet connection is the only thing required for the users to reach any tools within the system.
- Interoperability: The system should be able to integrate with other systems and software, allowing for data exchange and communication between different platforms.

6 Database Design

6.1 Data Description

Our dataset consists of the data scraped from google's results page regarding every and each topic or keyword which is done using Google API and python libraries and code to be able to get Google's content from the SERP (google's search engine).We run pre-processing on the scraped data that have been crawled from inside all the links in the SERP. these results count as competitor analyzing for our user in order to ease the process of keyword research for them as we already do all the competitor analyzing for them.

6.2 Database design description

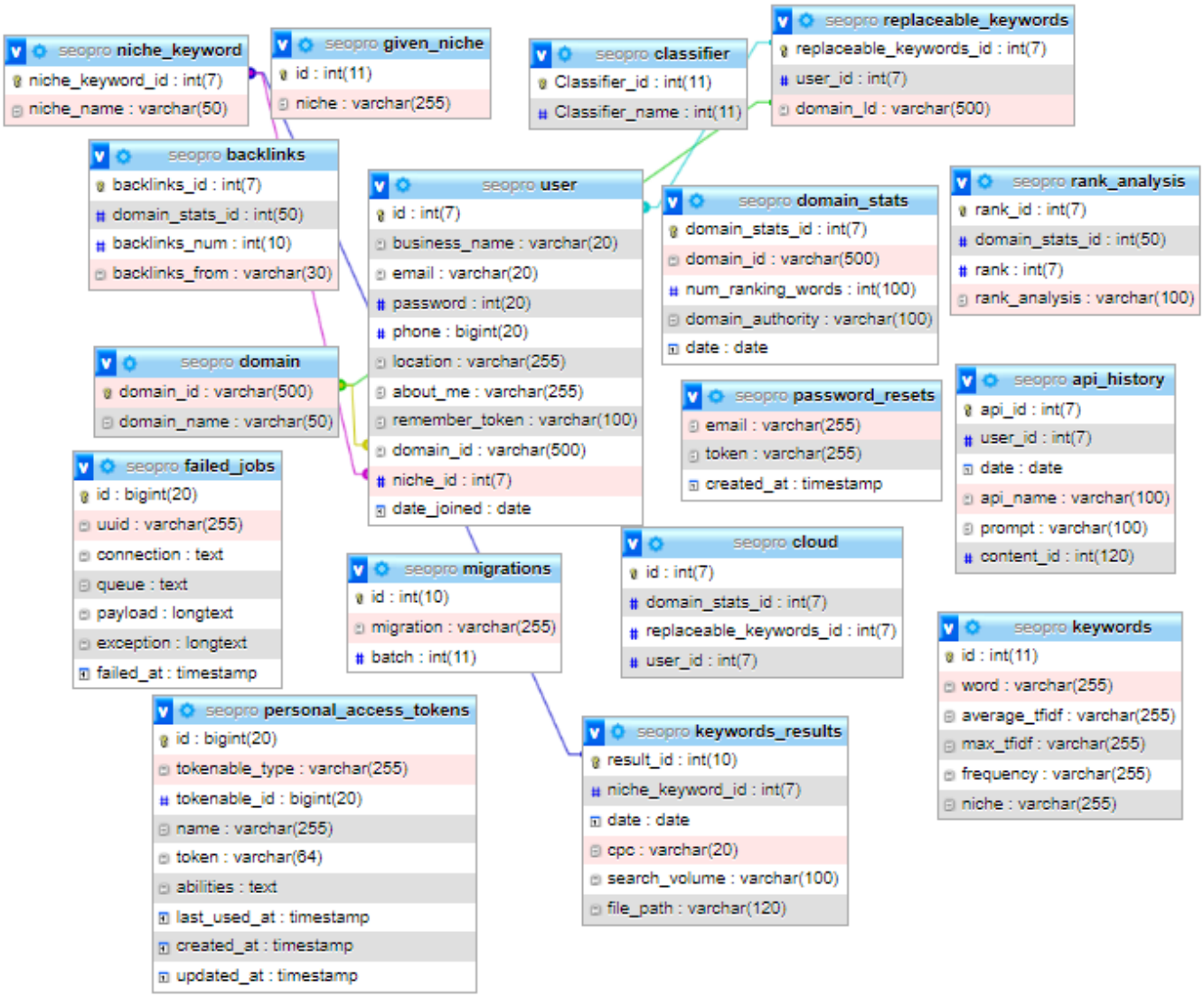


Figure 8: Database

7 Preliminary Object-Oriented Domain Analysis

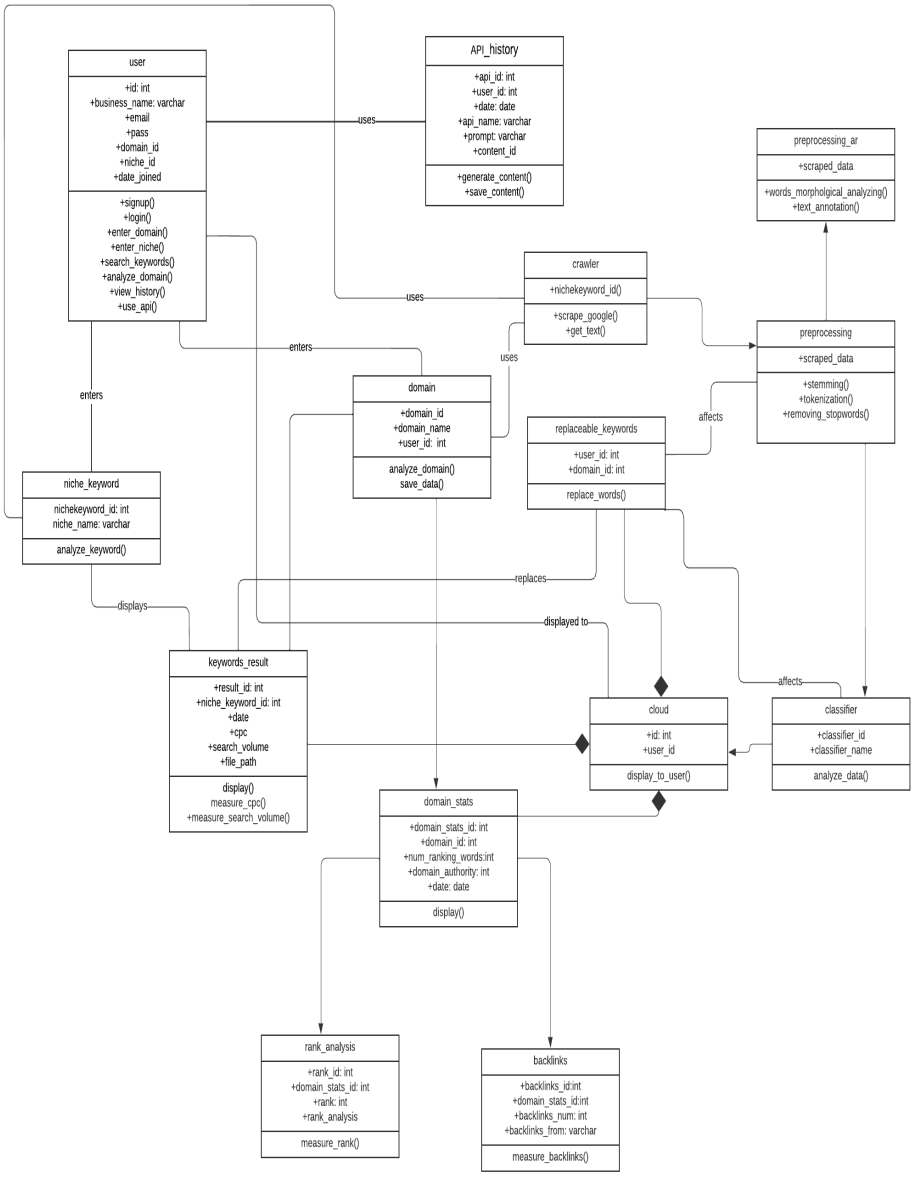
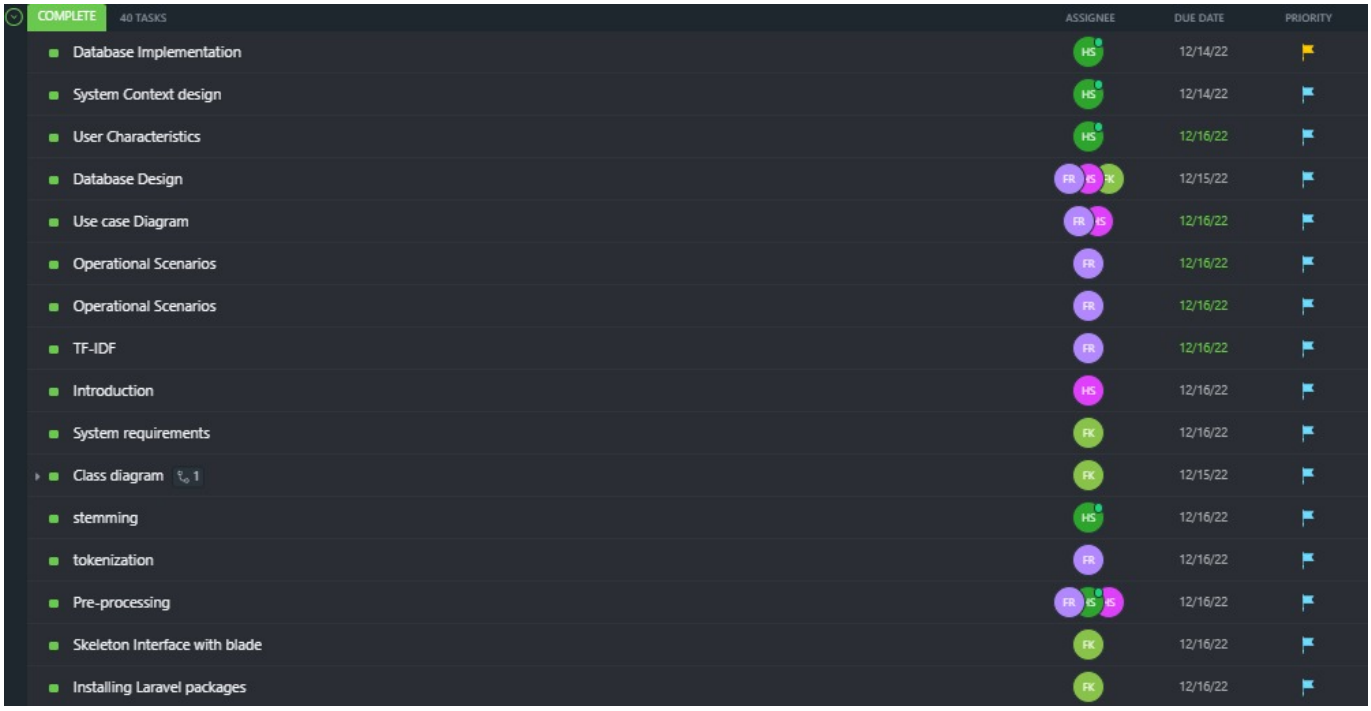


Figure 9: SEOPro Class diagram

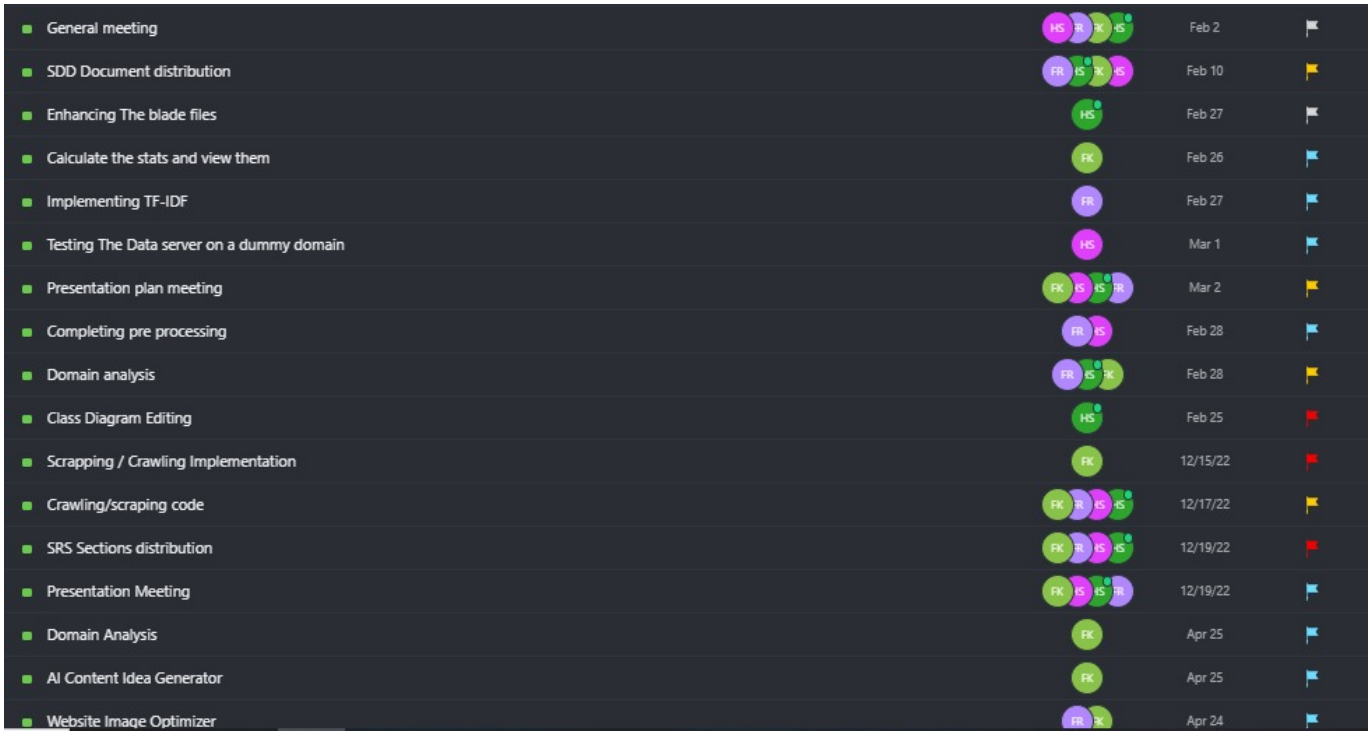
8 Operational Scenarios

- Scenario 1: User without a website
  - User inputs the topic of their website according to their niche.
  - The system generates high-ranking content ideas for English or Arabic websites to help the user rank higher on search engine results pages (SERP).
  - The system provides high ranking keywords depending on the user's niche to help him craft website content that can be crawled, indexed and ranked well by Google crawlers.
- Scenario 2: User with an existing website
  - User inputs their website's domain and better ranking keywords and synonyms are generated for English or Arabic websites to improve their website's SERP ranking.
  - The system estimates CPC, search volumes, and backlinks for the user's website to help them observe his domain analysis and work on his points of weakness in their URL..
  - The system provides a domain analysis feature that enables the user to track their website's performance metrics, such as domain's global rank, search volume, traffic, domain authority, SEO difficulty, indexed pages, backlinks, CPC and many more metrics, to identify areas for SEO improvement.
  - The system provides a website speed checker that measures the loading time of the given URL, suggests using the image optimizer to improve load time, and provides an analysis of the website's performance. This feature is important for SEO because search engines prioritize websites that load quickly, and the Website Load Time Analyzer helps users optimize their website for faster load times.
  - The system provides a website image optimizer that scrapes all pictures within the given URL, optimizes them, and converts them from jpg/png to webp. The optimized pictures can be downloaded to the user's device by pressing a button, reducing the website's load time and improving its SEO.
- Scenario 3:User with or without a website who wants to analyze competitors
  - User inputs their competitors' domain, and the system generates a competitor analysis report that includes domain authority, backlinks, organic search traffic, and number of organic keywords. This report can help the user identify areas where they can improve their own website's SEO and outrank their competitors.

9 Project Plan



COMPLETE 40 TASKS	ASSIGNEE	DUE DATE	PRIORITY
Database Implementation	HS	12/14/22	
System Context design	HS	12/14/22	
User Characteristics	HS	12/16/22	
Database Design	FR IS FK	12/15/22	
Use case Diagram	FR IS	12/16/22	
Operational Scenarios	FR	12/16/22	
Operational Scenarios	FR	12/16/22	
TF-IDF	FR	12/16/22	
Introduction	HS	12/16/22	
System requirements	FK	12/16/22	
Class diagram 1	FK	12/15/22	
stemming	HS	12/16/22	
tokenization	FR	12/16/22	
Pre-processing	FR IS HS	12/16/22	
Skeleton Interface with blade	FK	12/16/22	
Installing Laravel packages	FK	12/16/22	



General meeting	HS FR FK IS	Feb 2	
SDD Document distribution	FR IS FK IS	Feb 10	
Enhancing The blade files	HS	Feb 27	
Calculate the stats and view them	FK	Feb 26	
Implementing TF-IDF	FR	Feb 27	
Testing The Data server on a dummy domain	HS	Mar 1	
Presentation plan meeting	FK IS HS FR	Mar 2	
Completing pre processing	FR IS	Feb 28	
Domain analysis	FR IS FK	Feb 28	
Class Diagram Editing	HS	Feb 25	
Scrapping / Crawling Implementation	FK	12/15/22	
Crawling/scraping code	FK FR IS IS	12/17/22	
SRS Sections distribution	FK FR IS IS	12/19/22	
Presentation Meeting	FK IS HS FR	12/19/22	
Domain Analysis	FK	Apr 25	
AI Content Idea Generator	FK	Apr 25	
Website Image Optimizer	FR FK	Apr 24	

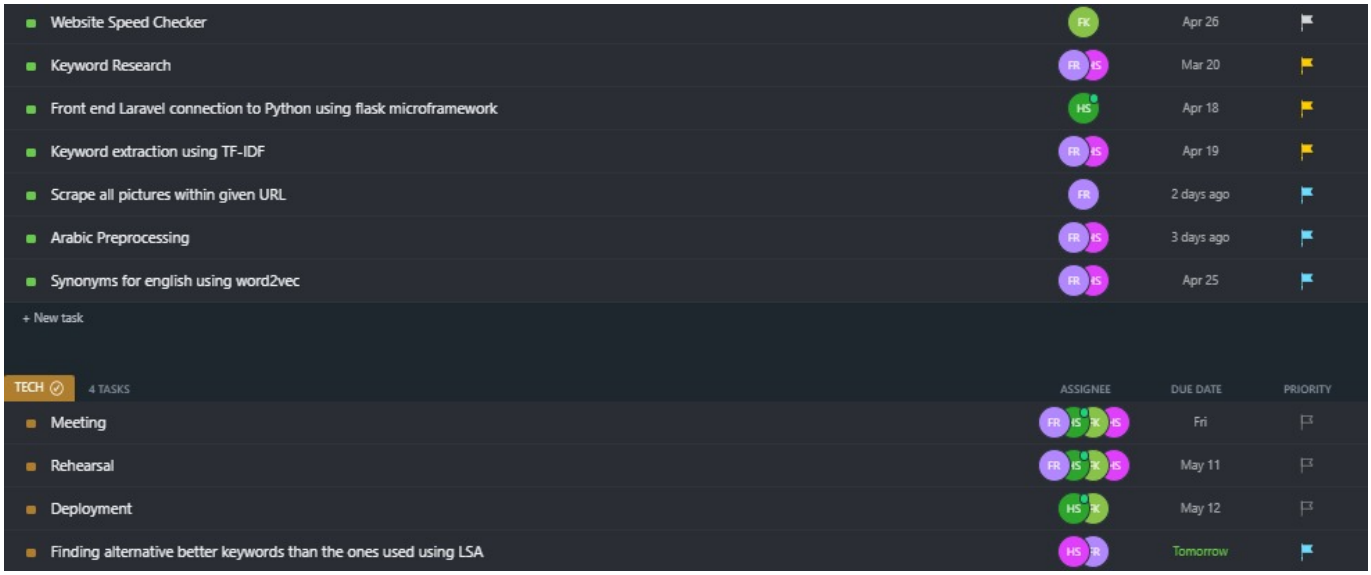


Figure 10: Timeplan

10 Appendices

10.1 Supportive Documents

- Dell competition

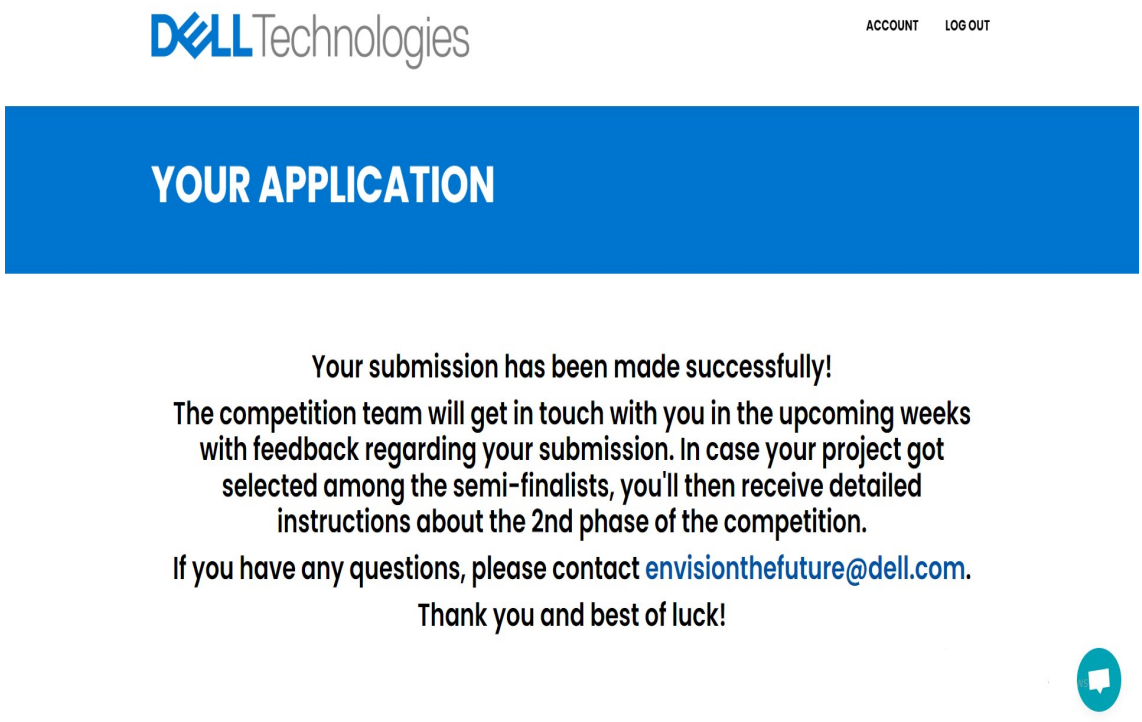


Figure 11: DELL competition

- Dataset

In the proposed system, our dataset is scraped from the Internet's search engine results page as well as users search queries. The user enters a keyword and the serp's results are scraped and pre-processed accordingly to prepare for the keyword extraction's process. It is now obvious that data scraping is crucial to a company's success, whether it is used to increase sales and expand the company's consumer base. The future of data scraping also seems busy, as more and more publicly available data will be made available online as it becomes the primary source of insight for enterprises.

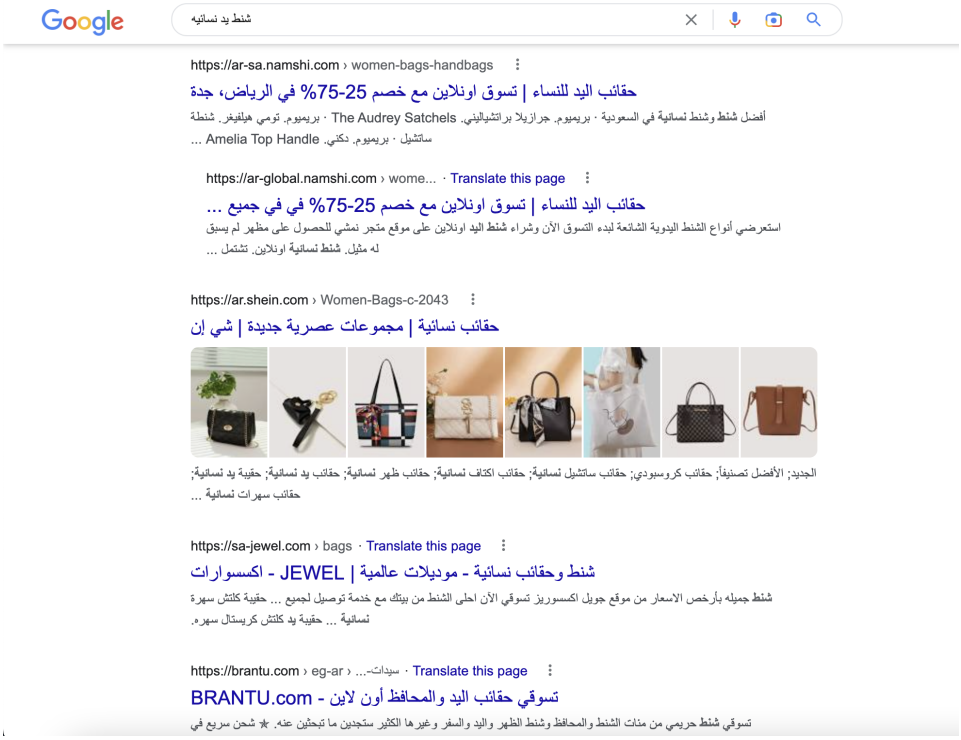


Figure 12: Google Search Engine Results Page

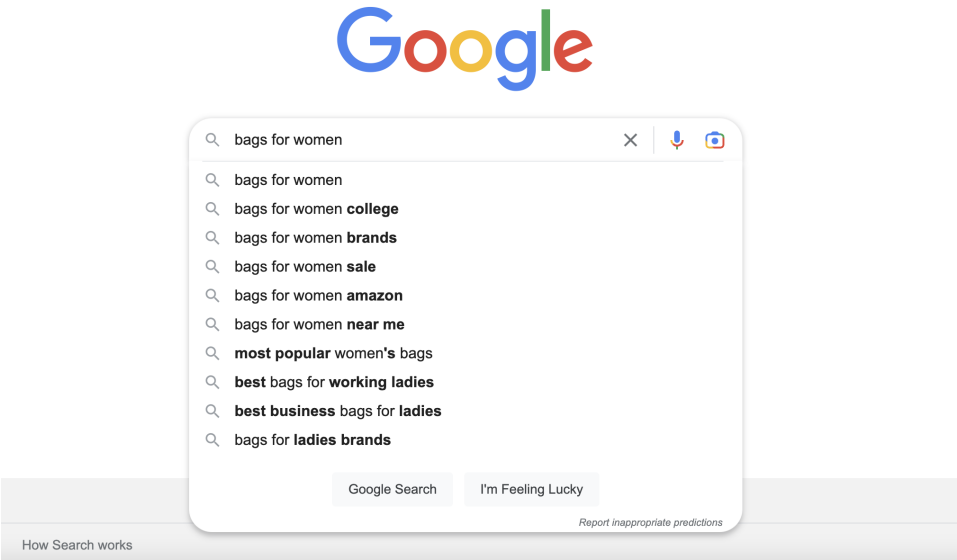
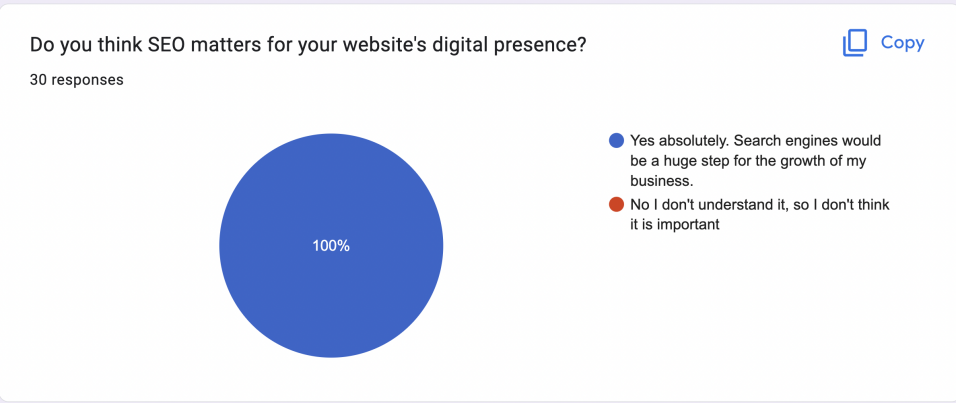


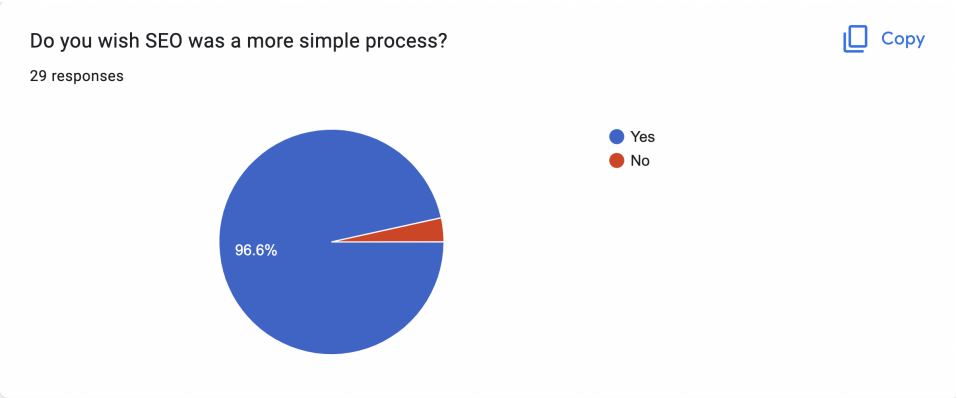
Figure 13: Google Users Search Queries

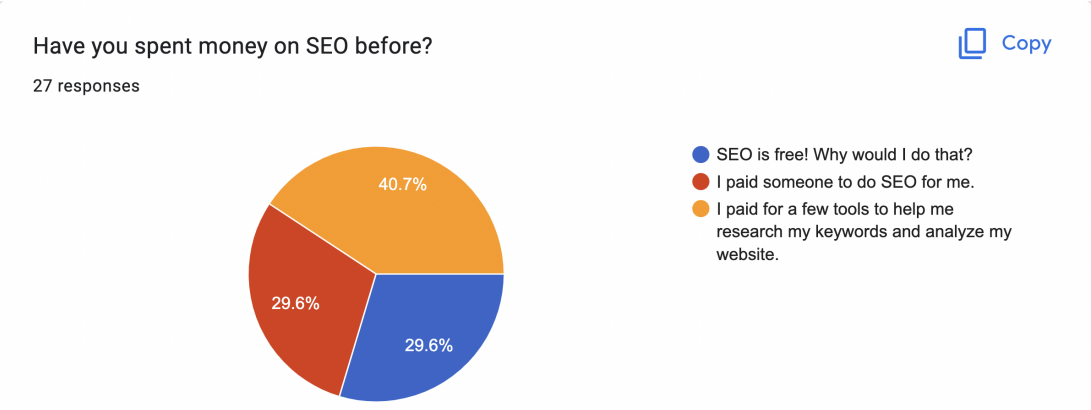
- Survey

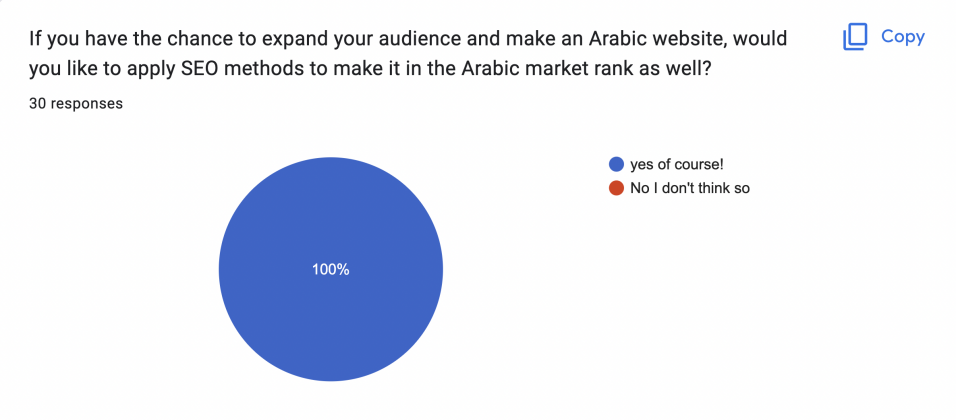
A survey had been done to get a wider scope on our market's needs and expand our market's research. 30 business and website owners have responded from various parts around the world, 26 from Egypt, 2 from USA, 1 from Germany and 1 from Palestine. The responses' niches varied from copywriting niche, to various e-commerce stores including fashion stores/blogs & clothing brands, digital marketing agencies, online tutoring, non profit organization and an entrepreneurship blog.

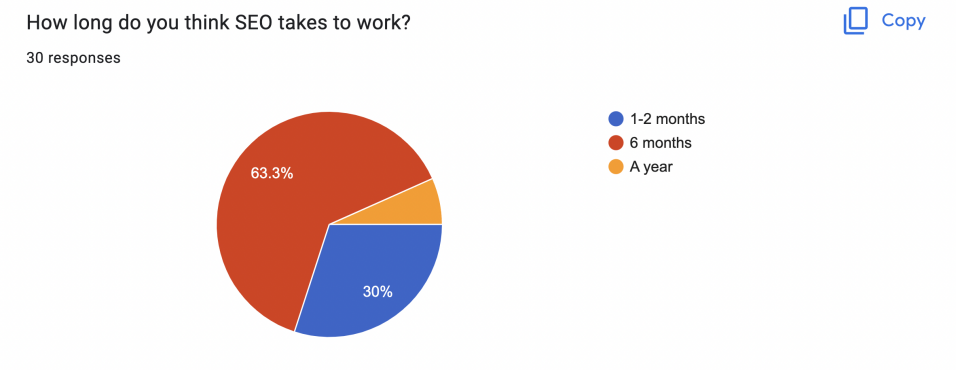
The results were indicating how much our respondents were interested in SEO and keen to apply it on their business.











References

[1] Bela Joglekar, Rohan Bhatia, Soumya Jayaprakash, et al. "Search Engine Optimization Using Unsupervised Learning". In: *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. 2019, pp. 1–5. DOI: 10.1109/ICCUBEA47591.2019.9129011.

[2] Mahmoud Nabil, Amir F. Atiya, and Mohamed Aly. "New Approaches for Extracting Arabic Keyphrases". In: *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. 2015, pp. 133–137. DOI: 10.1109/ACLing.2015.26.

[3] Tsung Fu Lin and Yan Ping Chi. "Application of Webpage Optimization for Clustering System on Search Engine V Google Study". In: *2014 International Symposium on Computer, Consumer and Control*. 2014, pp. 698–701. DOI: 10.1109/IS3C.2014.186.

[4] *wordtracker*. URL: https://www.wordtracker.com/.

[5] *Umbrellum*. URL: https://umbrellum.com/tools/keyword-discovery.