|  |  |
| --- | --- |
| ­­­­­­­­Academic Year | 2023-24 |
| Batch | 2022-24 |
| Trimester | 6th |
| Programme  (PGDM / PGDM-FS / PGDM-RM) | PGDM & PGDM-FS |
| Name of Course | Big Data NLP |
| Section | A |
| Name of Faculty | Dr. Purnendu Shekhar Pandey |

­­­­­­­­­­­­­­­­­­

|  |  |
| --- | --- |
| Nature of Submission  (Group Assignment / Group Project Report) | Group Assignment |
| Topic of Group Assignment / Group Project | Enhancing Keyword Search efficiency in big data: A comparative analysis of algorithms and Techniques |
| Deadline for Submission | 05-04-2024 |
| Group/ Learning Team Number | LT-07 |
| Maximum Marks Allotted |  |

**Tick on appropriate**

We hereby declare that this project/assignment does not contain any AI generated content (e.g. ChatGPT etc.)

**I** We hereby declare that this project/assignment has used AI generated content (e.g. ChatGPT etc.) as supporting resource for the completion of project/assignment.

**Contribution of Group/LT members in the Assignment/Project**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Name & enroll. No. of Student** | **Contribution** | **Signature** |
| 1 | Victor Bhattacharya (JL22PG233) | 100% | Victor |
| 2 | Adeeba Shareef (JL22PG243) | 100% | Adeeba |
| 3 | Udita Saini (JL22PG224) | 100% | Udita |
| 4 | Akanksha Shukla (JL22FS007) | 100% | Akanksha |
| 5 | Aviral Srivastava (JL22014) | 100% | Aviral |
| 6 | Abhishek Singh (JL22FS002) | 100% | Abhishek |

Enhancing Keyword Search Efficiency in Big Data: A Comparative Analysis of Use of Voice Assistant and Search Engines

Victor Bhattacharya, Udita Saini, Adeeba Sareef, Abhishek, Aviral, Akansha

***Jaipuria Institute of Management, Lucknow (226010), INDIA***

## Abstract

In this case study methodology, our primary focus revolves around delving into the intricacies of improving keyword search efficiency within the expansive realm of Voice Assistants and Search Engines. To achieve this, we've opted for a diverse research approach, combining structured questionnaires containing 33 carefully crafted inquiries with qualitative interviews. Our participant selection process was meticulous, considering various demographic factors and participants' familiarity with search technologies. We then collected data through these methods, ensuring thoroughness and reliability. Following this, we carefully examined the information, utilizing both quantitative and subjective methods to gather bits of knowledge into member ways of behaving and insights. Our point is to reveal nuanced examples and experiences that shed light on the intricacies of catchphrase search inside the immense scene of large information.

## Introduction

In the present scenario, keywords have turned into the foundation of online hunt and data recovery, filling in as the essential structure hinders that associate clients with important substances. These `keywords, whether a solitary term or a mix of expressions, go about as the doorway to getting to the huge store of data accessible on the web. Their importance lies in their capacity to smooth out the pursuit cycle, permitting clients to communicate their data needs briefly and effectively. The idea of keywords rose to unmistakable quality with the appearance of web indexes, especially Google, which altered the manner in which people explore the web. Google's inquiry calculations depend intensely on keywords to comprehend the unique circumstance and aim behind clients' questions, guaranteeing that list items are exact and custom-fitted to their necessities. This dependence on keywords prodded the advancement of site design improvement (Web optimization), a training zeroed in on enhancing site content to line up with famous hunt terms and further develop perceivability on web crawler results in pages (SERPs). Past their part in customary web look, keywords have pervaded different aspects of online connections. They assume a vital part in web-based publicizing, where promoters bid on catchphrases to show designated promotions to clients looking for related content. Watchwords additionally support semantic examination and normal language handling (NLP) advancements, empowering frameworks to decipher and answer client questions in a more human-like way. In addition, as innovation keeps on propelling, catchphrases have developed to envelop different types of searches, including voice search and picture acknowledgment. Voice colleagues like Google Partner influence watchwords to grasp spoken inquiries, extricate significant data, and give precise reactions. Similarly, Google's image search utilizes keyword metadata and visual analysis techniques to categorize and retrieve images based on descriptive keywords.

##### Statement of problem

### With the blast of information in the advanced world, finding what we want rapidly has become progressively testing. Keywords, which we use to look for data on the web, are at the core of this test. We're confronted with questions like: How would we make keywords look more effective, particularly in the huge ocean of large information? As though that weren't sufficient, the ascent of voice colleagues like Siri and Google Collaborator adds one more layer of intricacy. In this way, the issue we're handling is the way to make catchphrase look through work better, particularly in the period of large information and voice partners. It's tied in with understanding how individuals search, what they battle with, and tracking down ways of making the entire cycle smoother and more compelling. Eventually, by taking care of this issue, we can make it simpler for everybody to find the data they're searching for in the midst of the consistently developing expanse of computerized information.

##### Research Question

How Enhancing Keyword Search Efficiency can be increased through the help of Voice Assistant and Big Data?

##### Literature Review

As the web keeps on growing at a remarkable rate, finding the right data rapidly has become progressively troublesome. Customary strategies utilized via web crawlers to rank sites are battling to stay aware of this development. To handle this issue, specialists are investigating better approaches to make web crawlers more astute. One such study by(Amer, 2015 ) suggests a system that can understand web page content and organize it in a way that matches what users are looking for. By assigning weights to different topics on a page and comparing them to a user's query, this system helps search engines show more relevant results. The consequences of this study were promising, showing critical upgrades in search precision contrasted with existing techniques. Notwithstanding more intelligent hunt calculations, site proprietors are additionally utilizing procedures like website streamlining (Search engine optimization) to further develop their permeability on the web.(Gandour, 2011) studied how a website called Fragfornet used SEO techniques to attract more visitors. By focusing on specific topics and keywords within their content, Fragfornet was able to climb higher in search engine rankings, bringing in more traffic to their site.

Furthermore, there's an emerging trend of using social annotations to improve search results. (Hsu, 2011) looked into predicting tags for web pages based on what users are saying about them. By examining how individuals label pages and utilizing that data to recommend related content, web crawlers can give more exact outcomes. This approach has shown promising outcomes, making it more straightforward for clients to find what they're searching for on the web.

##### Methodology

###### Data Summary

Our research focuses on improving keyword search efficiency in the vast domain of big data by comparing different algorithms and techniques. To gather information, we reached out to people through online platforms like social media and messaging apps. We asked them questions about how they use voice assistants, their preferences when searching for information, and any concerns they have about privacy and security. The survey lasted for 4 days, during which we collected responses from 167 individuals, mainly middle-aged people from Northern India. We wanted to ensure we had a diverse group of participants to get a better understanding of the topic. After collecting the data, we analyzed it to identify any trends or patterns that could help us enhance keyword search efficiency in big data.

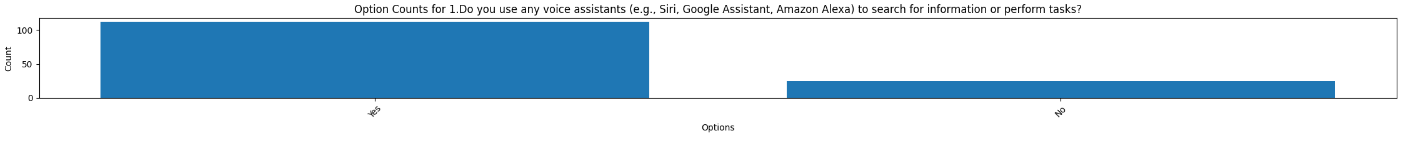
##### Test Done

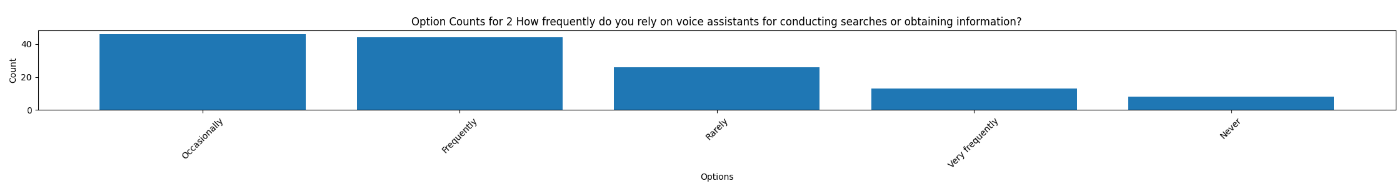
1. **Text Classification with Naive Bayes:** Text classification is akin to teaching a computer to understand and categorize human language. Initially, we divide our dataset into two segments: one for training the computer to recognize textual patterns and another for evaluating its learning. However, computers don't inherently grasp words as humans do. Therefore, we convert the text into numerical representations using a method called Count Vectorizer. This transformation essentially assigns a numerical value to each word, allowing the computer to process and analyze the text. Next, we employ a specialized algorithm known as Multinomial Naive Bayes for training. This algorithm operates by assessing the occurrence and frequency of words in different categories, calculating probabilities to determine which category a given piece of text most likely belongs to. Once the computer is trained, we assess its predictive accuracy by testing it on unseen data. The characterization report furnishes us with itemized experiences into the model's presentation, including measurements like accuracy, review, and F1-score.
2. **Top Features Analysis:** The pith of top highlights examination lies in uncovering the most effective words or elements that impact the PC's dynamic cycle. To accomplish this, we scrutinize the text data and evaluate the likelihood of each word belonging to various categories. Words with the highest likelihood for a particular category are deemed the "top features" for that category. These top features offer valuable insights into the factors driving the computer's predictions. For example, in opinion examination, words like "cheerful," "upbeat," or "positive" may arise as top elements for the "positive feeling" classification, revealing insight into the phonetic prompts adding to opinion grouping.
3. **Correlation Analysis:** Connection examination fills in as a way to comprehend the complex connections between various elements in our dataset and the objective variable we look to foresee. For numerical features such as age or income, we employ the Point Biserial Correlation method to gauge the strength of their association with the target variable. This statistical measure elucidates whether variables like age or income significantly impact the likelihood of using voice assistants, for instance. Conversely, for categorical features like gender or occupation, we leverage the Chi-squared statistic to ascertain if there exists a meaningful association between the feature and the target variable. By discerning these correlations, we gain valuable insights into the predictive power of various features and their significance in driving accurate predictions.
4. **Option Counts Analysis:** The option counts analysis entails a meticulous tally of the frequency of each response option for every question in our survey. This approach grants us unparalleled insight into the distribution of responses, unveiling prevailing trends and preferences within our dataset. For instance, inquiring about preferred voice assistant options might reveal whether Siri, Google Assistant, or Amazon Alexa garners the most Favor among respondents. Visualizing this data using bar plots provides a comprehensive overview of respondent preferences, facilitating a nuanced understanding of user behavior and preferences across different demographic segments.

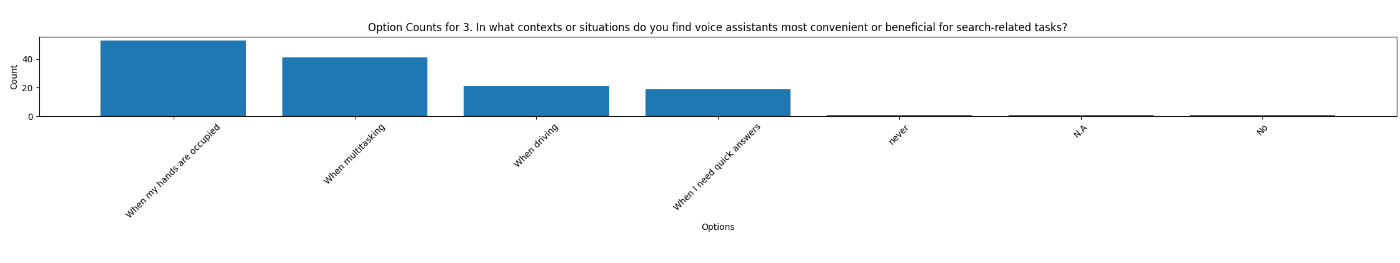
###### Data Preparation

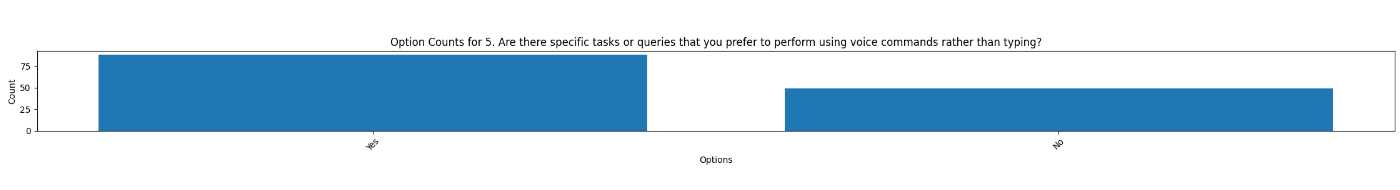
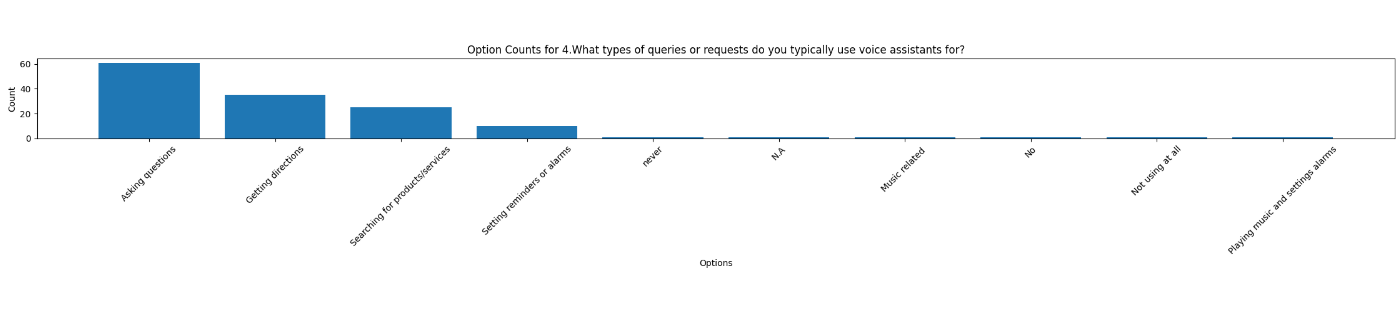
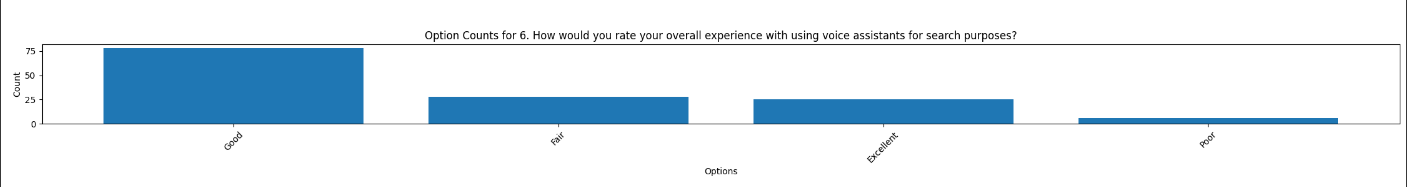
The data collected from the respondents are then pre-processed for further analysis, it includes the part of the analysis where no null values should be intertwined with our results and no abstract character should be written in the open-ended questions.

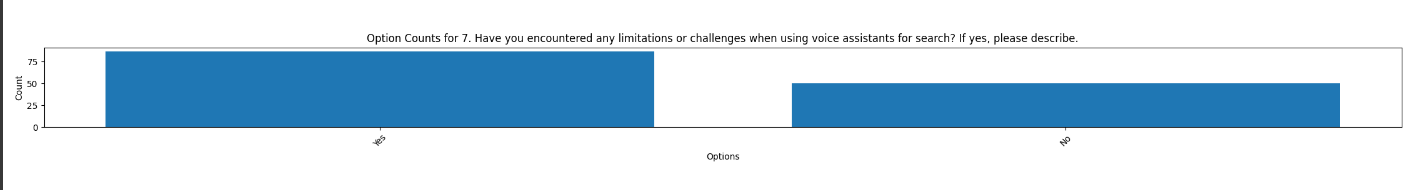
##### Empirical Insights

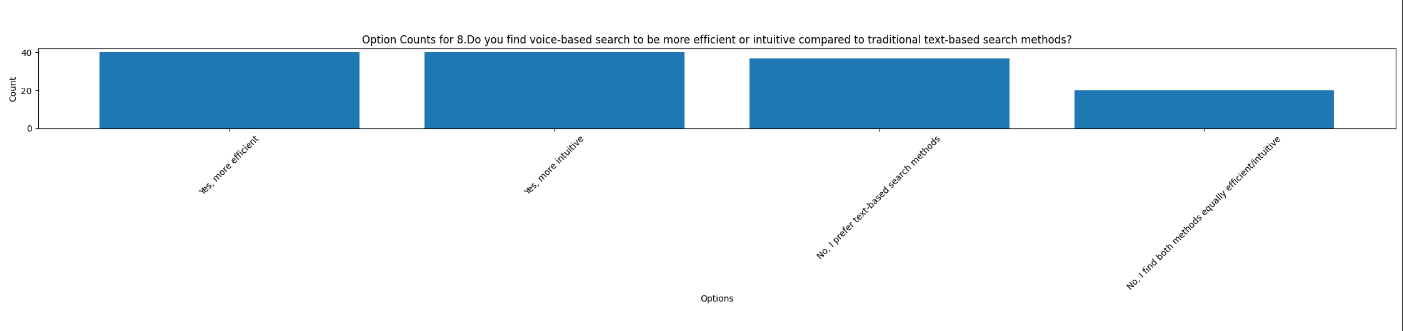


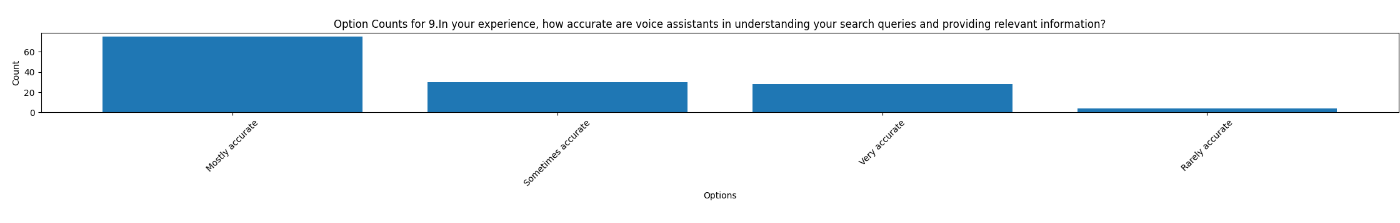


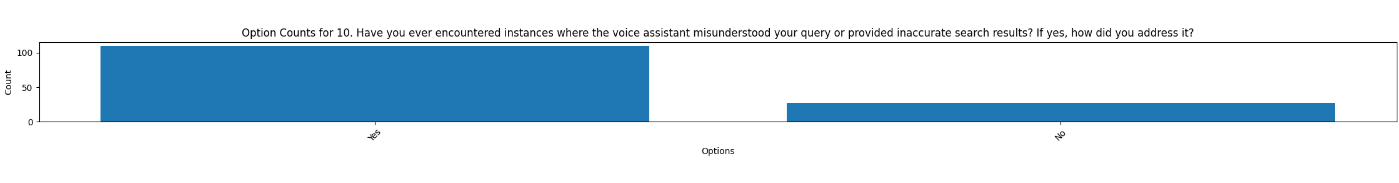


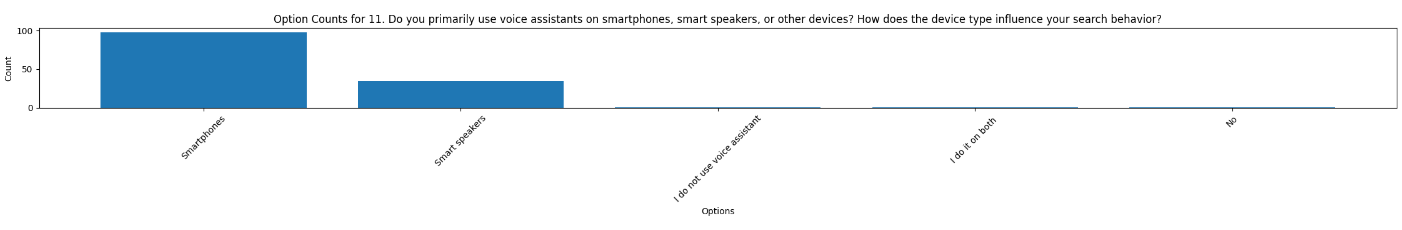
 

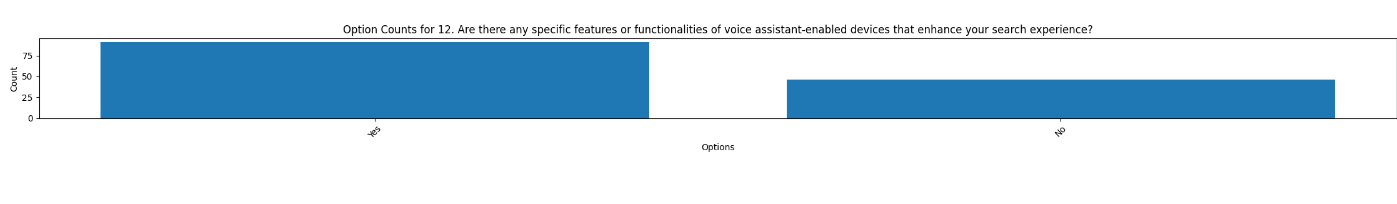


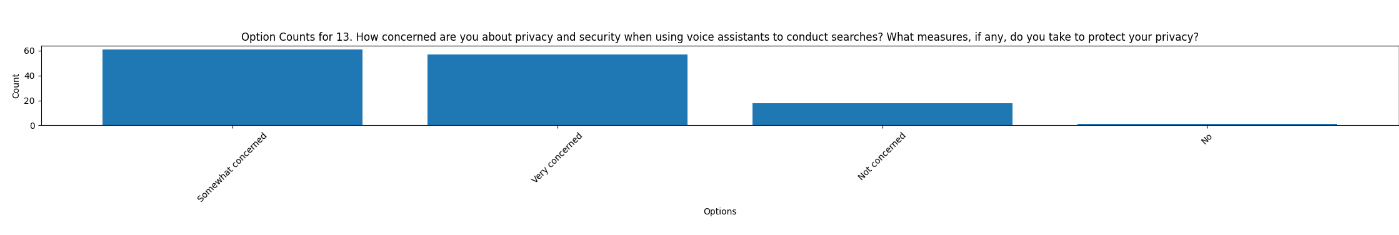


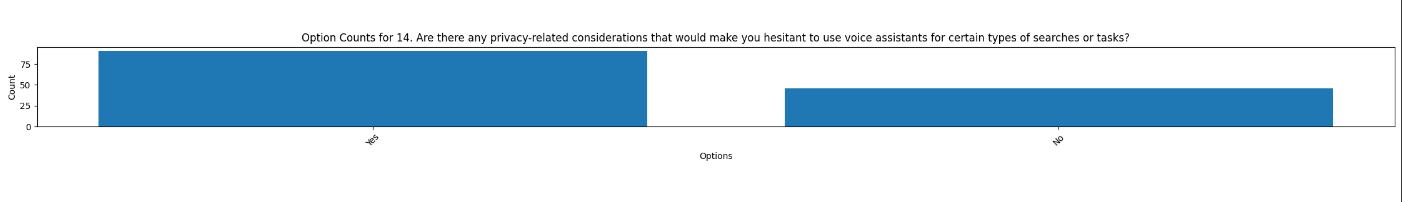


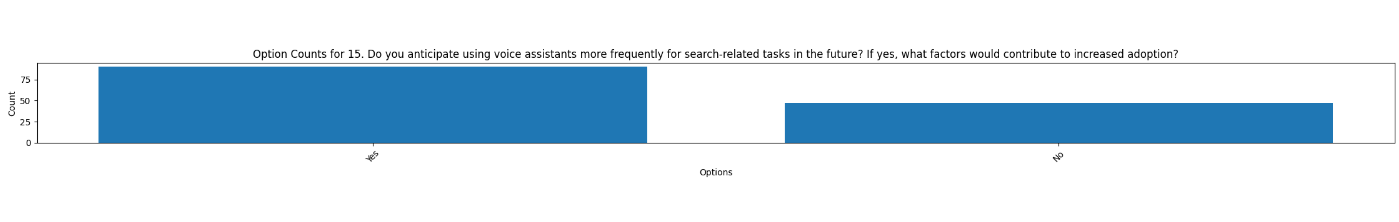


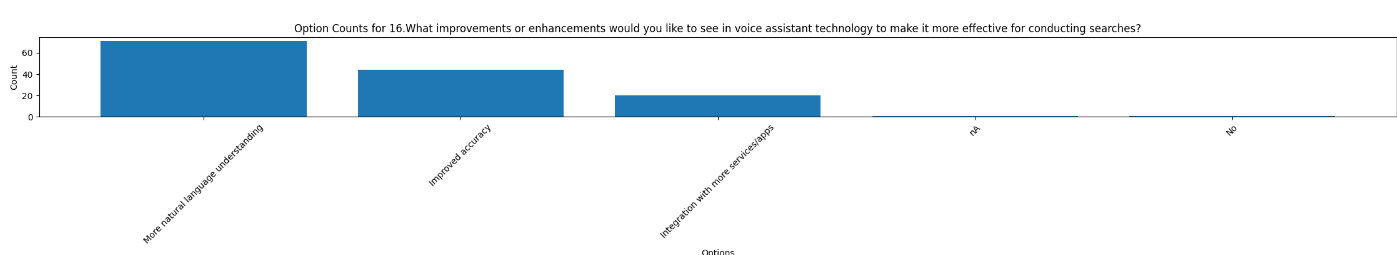


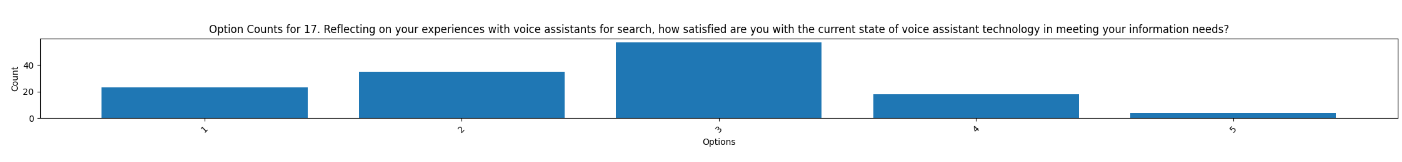


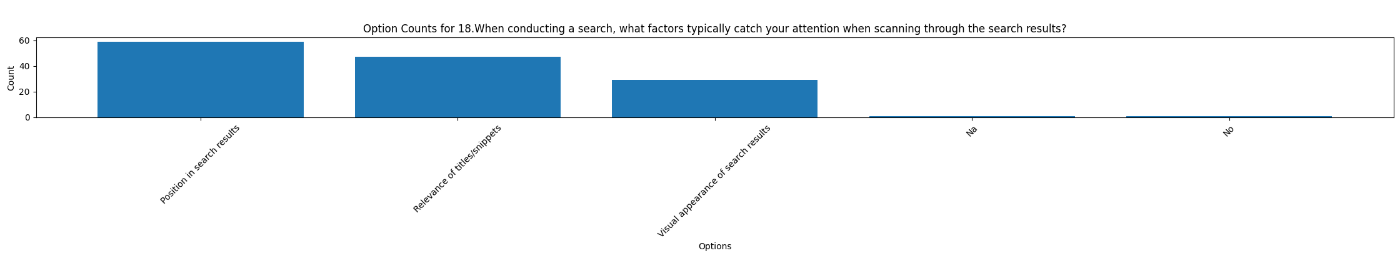


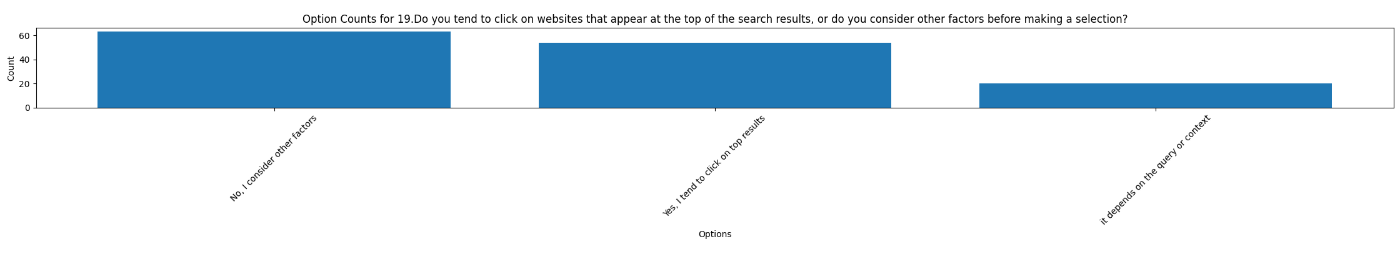
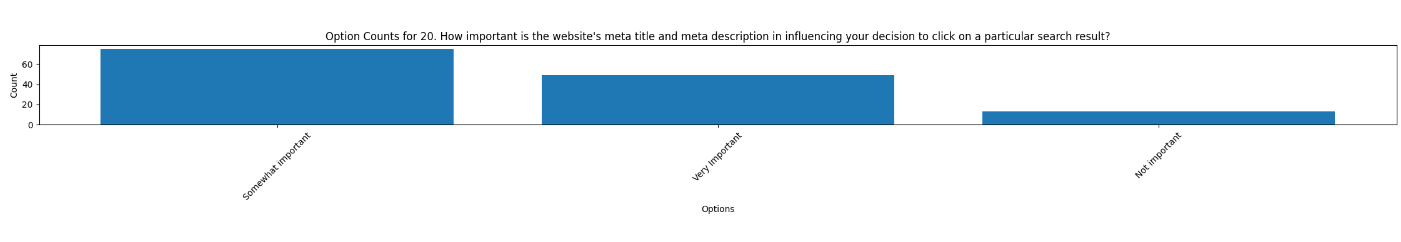


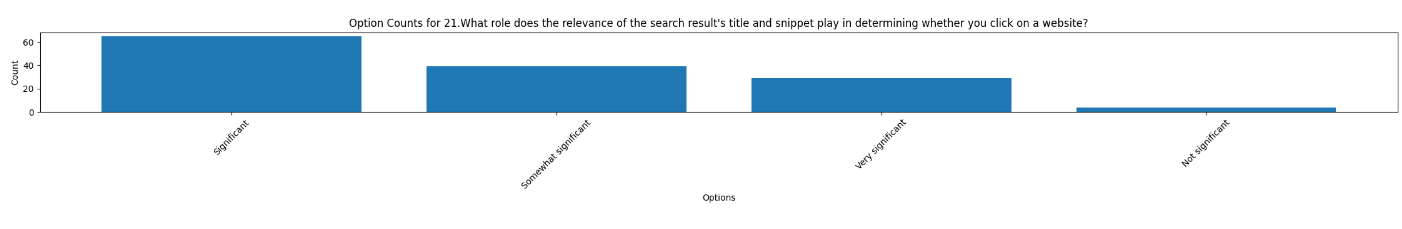


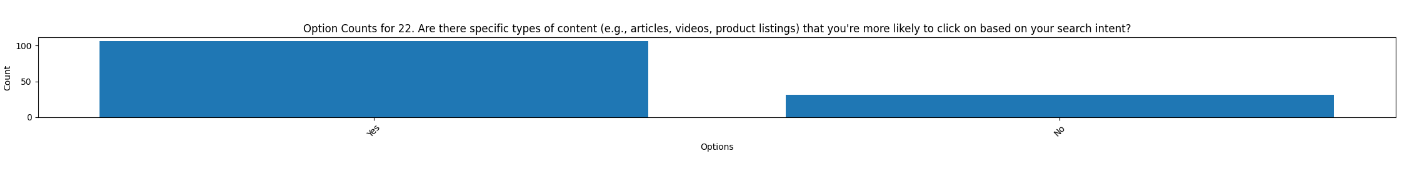


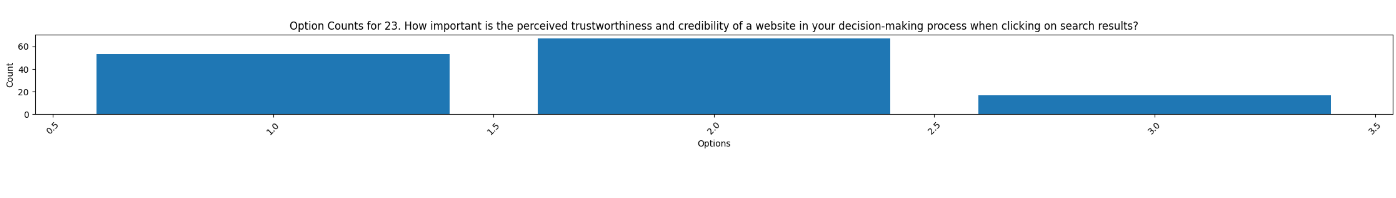


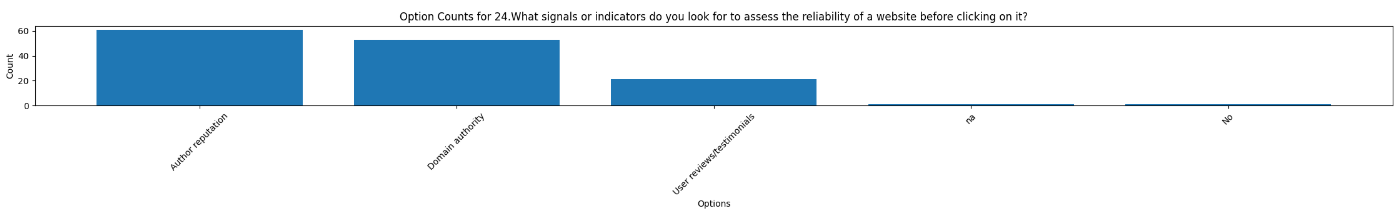


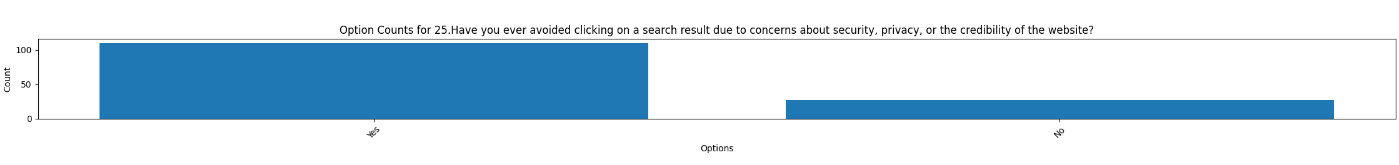
 

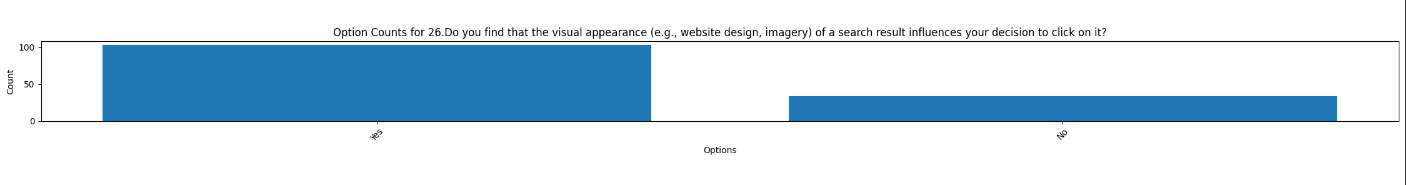


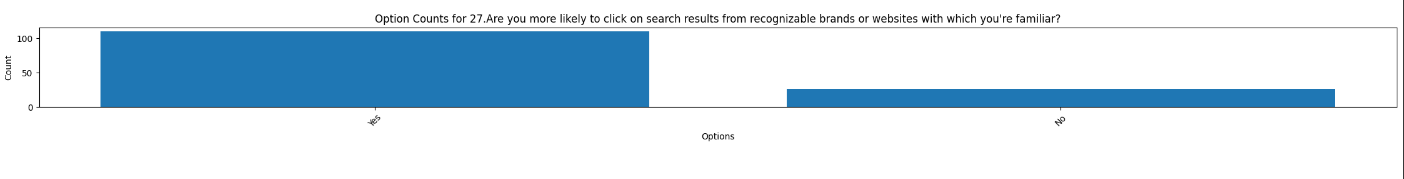


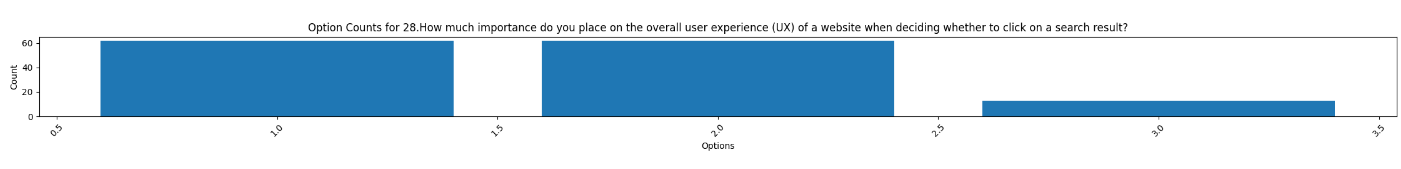


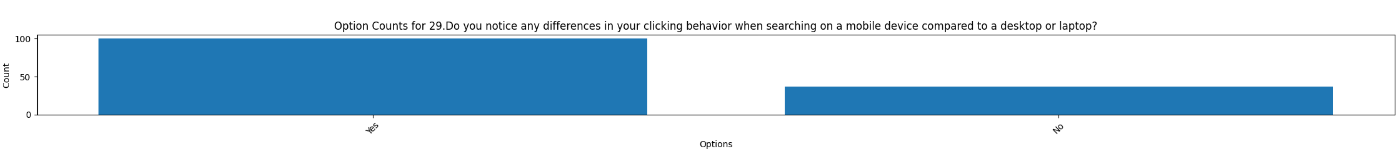


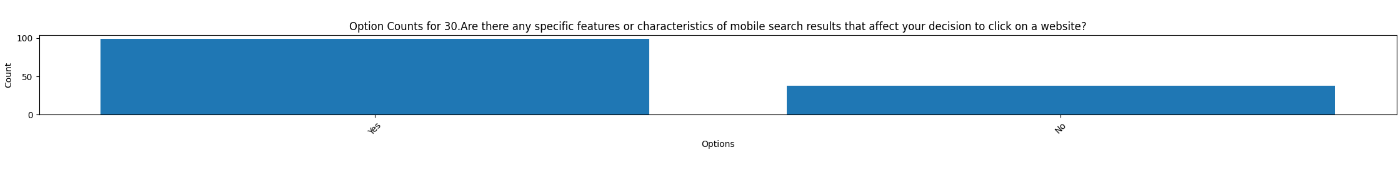












##### Findings

**Individuals Using Voice Assistants:**

* **People Utilizing Voice Associates:** Relevant Inclinations: Clients express an unmistakable inclination for voice colleagues in unambiguous settings, like driving or performing various tasks, where sans-hands communication is urgent. This shows that the comfort of voice colleagues lies in their capacity to oblige clients' exercises without requiring manual info.
* **Different Use Examples:** The study uncovers a great many undertakings performed utilizing voice collaborators, including seeking clarification on some pressing issues, getting headings, looking for items/administrations, and setting updates or cautions. This variety features the flexibility of voice partners in tending to different client needs. (Hsu, 2011).
* **Inclination for Voice Orders:** Numerous clients demonstrate an inclination for voice orders over composing for specific undertakings, underlining the effectiveness and convenience presented by voice connections. This proposes that voice colleagues are seen as a more normal and natural method for communicating with innovation.
* **Positive Client Experience:** Generally speaking, clients rate their involvement in voice aides decidedly, demonstrating fulfilment with the innovation's viability for search-related assignments. This positive opinion recommends that voice collaborators have effectively measured up to clients' assumptions and offered some benefit in their regular routines.
* **Difficulties and Worries:** While clients for the most part report a positive encounter, some recognize experiencing restrictions or difficulties while utilizing voice colleagues. Regardless of these difficulties, clients' general positive evaluations propose that any issues confronted are offset by the advantages of utilizing voice associates.
* **Proficiency and Instinct:** Clients see voice-based search as more effective and natural contrasted with conventional text-based strategies. This inclination for normal language associations mirrors the craving for consistent and easy-to-understand search encounters.
* **Precision and Unwavering quality:** Voice partners are seen as exact in understanding pursuit questions and giving pertinent data, improving client trust and fulfillment. This dependability adds to clients' trust in involving voice associates for their pursuit needs. (Amer, 2015 )
* **Gadget Inclination and Openness:** Clients use voice colleagues on cell phones, showing an inclination for convenient and promptly open gadgets. This features the significance of gadget similarity and openness in moulding clients' hunt conduct.
* **Security and Security Contemplations:** While certain clients express worries about security and security while utilizing voice collaborators, many execute measures to safeguard their protection, demonstrating a degree of confidence in the innovation's capacity to defend individual data.
* **Expected Future Utilization:** Clients expect expanded dependence on voice partners for search-related undertakings later on, recommending a developing acknowledgment and coordination of this innovation into their everyday schedules. This pattern highlights the potential for proceeding with headways and developments in voice right-hand innovation. (Gandour, 2011)

**Individuals Not Using Voice Assistants:**

* **Usage Patterns and Preferences:** Non-users may not find voice assistants beneficial in certain contexts, preferring traditional search methods where manual input is preferred. This suggests a lack of perceived value or relevance of voice assistants in their current routines.
* **Alternative Search Methods:** In the absence of voice assistants, non-users rely on alternative methods for tasks such as asking questions, indicating a preference for traditional text-based interactions. This highlights the importance of familiarity and comfort with existing search habits.
* **Comfort with Typing:** Non-users may express a preference for typing over voice commands, suggesting a comfort level with traditional search approaches. This preference may stem from familiarity with typing and a perceived lack of need for voice-based interactions.
* **Satisfaction with Current Methods:** Although not utilizing voice assistants, non-users may find satisfaction with traditional search methods, indicating adequacy with their current approach. This suggests that non-users may not perceive a need for alternative search methods.
* **Avoidance of Potential Challenges:** Non-users may avoid challenges associated with voice assistant usage, such as accuracy or privacy concerns, by sticking to traditional search methods. This indicates a reluctance to adopt new technologies without clear benefits or assurances.
* **Equal Perceptions of Efficiency:** Text-based search methods are perceived as equally efficient or preferable to voice-based methods, reflecting a preference for traditional interaction modes among non-users. This suggests a lack of perceived advantage or incentive to switch to voice-based interactions.
* **Privacy and Security Concerns:** Non-users may express heightened concerns about privacy and security associated with voice assistants, potentially influencing their decision to abstain from usage. This underscores the importance of trust and confidence in the security measures of voice assistant technology.
* **Future Adoption Considerations:** Unlike users, non-users may not anticipate using voice assistants more frequently in the future, suggesting a preference for maintaining current search habits. This indicates a potential barrier to future adoption, stemming from a lack of perceived need or relevance of voice assistant technology.

**Inferences from feature weightage analysis**

**Usage Patterns of Non-Users:**

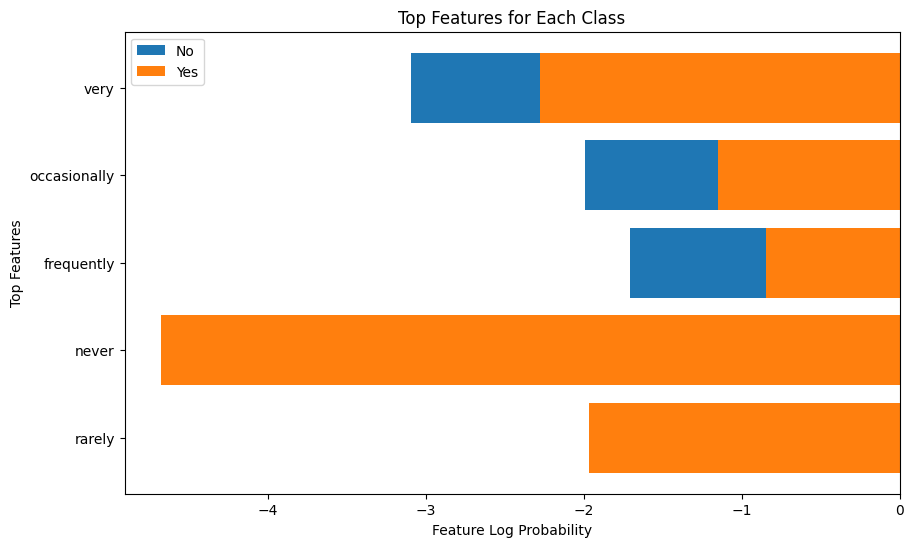
* **Rare Occurrence:** Among individuals who do not use voice assistants, the term "rarely" stands out as a top feature. This suggests that for those who refrain from using voice assistants, their reliance on them for search-related tasks is infrequent.
* **Complete Avoidance:** Additionally, the term "never" is prominent among non-users, indicating a significant portion of individuals who completely avoid using voice assistants for search purposes.
* **Intermittent Usage:** While terms like "frequently" and "occasionally" also appear, they carry lower probabilities compared to users. This implies that although some non-users might occasionally turn to voice assistants, it's not as prevalent as among users.

**Usage Patterns of Users:**

* **Frequent Engagement:** In contrast, for individuals who use voice assistants, the term "frequently" emerges as a top feature. This suggests that users tend to engage with voice assistants on a regular basis for search-related tasks.
* **Occasional Utilization:** Similarly, the term "occasionally" is prominent among users, indicating that even occasional use is prevalent among this group.
* **Preference for Regularity:** The prominence of terms like "frequently" and "occasionally" underscores a preference among users for consistent or occasional utilization of voice assistants for search tasks.

**Overall Interpretation:**

* The analysis of top features unveils distinct usage patterns between users and non-users of voice assistants.
* Users tend to engage with voice assistants frequently or occasionally, while non-users demonstrate a propensity for infrequent or non-existent usage.
* These insights shed light on the differing adoption and usage behaviours surrounding voice assistant technology, providing valuable understanding for stakeholders in technology development and marketing.



**Survey Response Analysis:**

**Frequency of Voice Assistant Usage:**

* The survey indicates widespread use of voice assistants, with respondents relying on them occasionally or frequently for searches and information retrieval.

**Contexts Favourable for Voice Assistant Use:**

* Voice assistants are most commonly used in situations where hands-free operation is necessary, such as when multitasking or driving.

**Preferred Types of Queries for Voice Commands:**

* Users primarily employ voice assistants for asking questions, getting directions, and searching for products or services, showcasing their versatility in addressing various needs.

**Experience and Satisfaction Levels:**

* Overall, respondents rate their experience with voice assistants positively, although some encounter limitations or challenges during usage.

**Privacy Concerns and Future Adoption:**

* While privacy and security concerns exist, a majority of respondents anticipate increased usage of voice assistants in the future, indicating a growing trend despite lingering apprehensions.

**Suggestions for Improvement:**

* Feedback suggests a desire for improvements in natural language understanding, accuracy, and integration with other services, alongside concerns about privacy and data usage.

**Future of AI Integration:**

* Respondents foresee continued advancements in AI leading to more sophisticated search experiences, with expectations for greater accuracy and innovation shaping the future of search engines

##### Conclusion

The deductions shed light on how much individuals depend on voice colleagues for their day-to-day undertakings. It's interesting to see that when individuals are busy with their hands, similar to while driving or performing multiple tasks, they will more often than not go to voice aides for help. This recommends that in spite of the prevalence of customary hunt techniques, such as composing questions into web search tools, voice partners offer a helpful other option, particularly in circumstances where it isn't doable to utilize our hands. Fascinating that individuals aren't simply utilizing voice associates haphazardly. They're decisively integrating them into their schedules. For instance, when they're performing multiple tasks and need speedy responses without halting what they're doing, they depend on voice orders. Practically like having a supportive collaborator consistently prepared to help, or for this situation, a voice. However, it's not all going great with voice associates. Numerous clients report cases where their inquiries are misconstrued or the outcomes aren't precise. This can be disappointing in light of the fact that you're anticipating that the aide should comprehend you impeccably, yet it misses the mark once in a while. It's like attempting to have a discussion with somebody who continues mishearing you - it very well may a piece irritate you. Security is one more large worry for clients. There's this fundamental stress over whether the voice partner is continuously tuning in on our discussions. It resembles having somebody continually listening in on you, and that can feel meddlesome. Despite these worries, a few clients truly value the trend-setting innovation behind voice collaborators. The people who comprehend things like normal language handling (NLP) see the potential it needs to make cooperations smoother and more natural. Like having a savvy buddy comprehends you nearly as well as a human would. With regards to list items, clients are really fussy. They need results that are important as well as dependable. They will more often than not click on the top outcomes, (Hsu, 2011) whether they're utilizing voice search or composing physically. Everything revolves around finding data rapidly and dependably.

##### Suggestions

##### In the current fast world, voice partners have transformed into our go-to accomplices, especially while we're rearranging various tasks or our hands are bound. In any case, imagine if these associates could achieve some different option from give quick reactions. Envision a situation where they could gather information from wherever on the web and present it to us in a way that is redone to our necessities. That is the vision we're researching here: a future where voice accomplices become like individual investigation colleagues, organizing data from various sources and passing it on to us without any problem. This would mean we would have no need to contribute energy actually searching for information any longer. Taking everything into account; we could rely upon our good 'ol partner to accomplish the difficult work for us. Moreover, it's not just about solace - by re-trying how information is presented, whether it's in a summary or a point-by-point assessment, these teammates would truly become basic accomplices as we kept searching for data. This shift isn't just about additional creating development; it's connected to working on our relationship with it, making an association where individuals and PC-based insight participate perfectly to achieve our targets.

##### Appendix

Survey Questionnaire Link: -

[Google Form](https://docs.google.com/forms/d/e/1FAIpQLSeqwgTOmVwHgaPPgIWNIgdjTQ5ax6HJ2-YAAO8ekaiXP-XszQ/viewform?usp=sf_link)

##### Citations

Gandour, Aurélie & Regolini, Amanda. (2011). Web site search engine optimization: A case study of Fragfornet. Library Hi Tech News. 28. 6-13. 10.1108/074 (Amer, 2015 ) (Hsu, 2011)19051111173874. (Gandour, 2011)

Hsu, M.-H. and Chen, H.-H. (2011), Efficient and effective prediction of social tags to enhance web search. J. Am. Soc. Inf. Sci., 62: 1473-1487. <https://doi.org/10.1002/asi.21558>

E. Amer, "Enhancing Efficiency of Web Search Engines through Ontology Learning from Unstructured Information Sources," 2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 2015, pp. 542-549, doi: 10.1109/IRI.2015.87.