

# The Faculty of Information and Communication Technology Mahidol University

# ITCS414 Information Storage and Retrieval Semester 1, 2023

# **Project 2:**

# **Prototype IR System**

**Submission Date:** September 24, 2023

**Section:** 1

**Student ID / Name:** 6488179 Ponnapassorn Iamborisut

6488181 Thadeeya Duangkaew

6488210 Ravikarn Jarungjitvittawas

**Instructors:** Assoc. Prof. Dr. Suppawong Tuarob

(suppawong.tua@mahidol.ac.th)

# **Executive Summary**

This report is a part of the ITCS414 Information Storage and Retrieval project focused on developing an advanced search engine to retrieve articles on astrology. Leveraging Elasticsearch and Flask, the system employs sophisticated indexing and ranking algorithms, allowing users to search for astrology-related content with enhanced precision. The implementation includes features like one-word, multi-word queries, partial matching, and a dynamic ranking system based on relevance and creation date. Additionally, the project integrates user-friendly elements such as pagination and keyword highlighting for an improved user experience. While successfully addressing the core requirements, the team recognizes opportunities for future enhancements, including the implementation of advanced search features, optimization of system performance, and ongoing user feedback mechanisms to refine and evolve the search engine for greater usability and effectiveness.

# **Table of Contents**

Executive Summary	a
Introduction	1
Problem(s) that you are trying to solve	1
Existing relevant systems	2
Implementation	5
I. Data collection, example documents, and data statistics	5
II. Tools and software	7
III. System diagram	8
IV. Snapshots of the system	9
V. Example step-by-step search sessions that highlight the following fund	ctionality 11
1. One word query	11
2. Multiple word query	11
3. Partial match	12
4. Ranking	12
Discussion	14
Conclusion	17
Reference	18

### AstroSage 101: A Search Engine for Astrology

(a)

#### Introduction

- Name of the Search System: AstroSage 101
- Type of Information: AstroSage 101 is an innovative astrology-focused search engine designed to provide a comprehensive and intuitive platform for astrology enthusiasts, practitioners, and curious individuals seeking astrological insights, guidance, and knowledge. Our motivation behind creating AstroSage 101 stems from the recognition of the increasing interest and curiosity about astrology and its potential to empower and enrich lives. AstroSage 101 is a dedicated astrology search engine crafted to simplify astrology, nurture curiosity, provide accurate information, and offer a sense of community to individuals interested in exploring the cosmic influences that shape our lives. We believe in the transformative power of astrology and aim to make it accessible and beneficial for everyone.

(b)

#### **Problem(s)** that you are trying to solve.

AstroSage 101 is designed to address several key problems related to astrology and its accessibility:

- Complexity and Confusion in Astrology: Astrology can be intricate and overwhelming for those new to the field. Understanding astrological terms, charts, and interpretations can be challenging. AstroSage 101 aims to simplify astrology and present it in an understandable and digestible format for both beginners and those seeking to deepen their knowledge.
- Lack of Reliable and Centralized Information: There is a vast amount of astrological
  information available online, but it's dispersed across various sources and platforms, often
  leading to confusion and misinformation. AstroSage 101 seeks to centralize reliable and
  accurate astrological information, becoming a trusted hub for individuals seeking genuine
  astrological insights.

- Inaccessibility to Quality Astrological Guidance: Finding accurate and personalized astrological guidance can be difficult, especially for those who don't have easy access to astrologers. AstroSage 101 aims to bridge this gap by offering accessible and reliable astrological information that can guide individuals in understanding their birth charts, zodiac signs, and planetary influences.
- Empowerment Through Self-Awareness: Many individuals are interested in using astrology as a tool for self-discovery and personal growth. AstroSage 101 addresses this by providing resources that enable users to delve into astrology, understand their unique traits and potential life paths, and make informed decisions aligned with their astrological profile.
- Promoting Community and Learning: Astrology enthusiasts often desire a sense of community and a platform to discuss, share, and learn from one another. AstroSage 101 endeavors to foster an active community where users can engage with each other, share experiences, and exchange astrological insights, creating a supportive space for learning and growth.

By addressing these problems, AstroSage101 strives to make astrology more approachable, trustworthy, and empowering, enhancing the astrological journey for individuals seeking knowledge and guidance in this fascinating field.

(c)

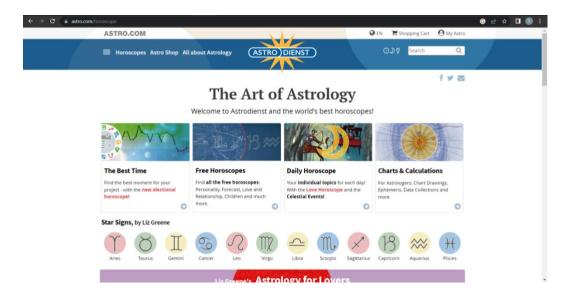
## **Existing relevant systems**

(i.e., Are there any other similar search engines as yours?)

There are several popular astrology-related websites, platforms, and apps that offer search functionalities and comprehensive astrological information. Some of these platforms provide search capabilities within their platforms or websites:

# • Astro.com (Astrodienst):

Astro.com is a widely recognized astrology website that offers a variety of astrological services, including birth chart interpretations, horoscopes, and astrological articles. Users can search for specific astrological information and generate detailed birth charts [1].



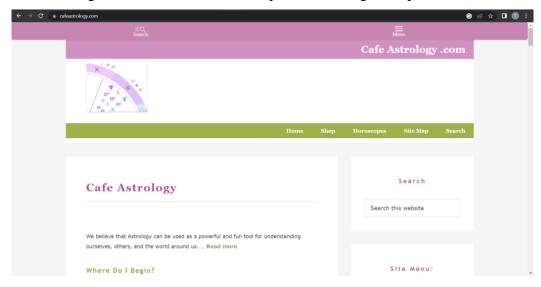
# • Astrology.com:

Astrology.com is a popular astrology website that provides daily and weekly horoscopes, tarot readings, astrology-related articles, and the ability to search for various astrological content [2].



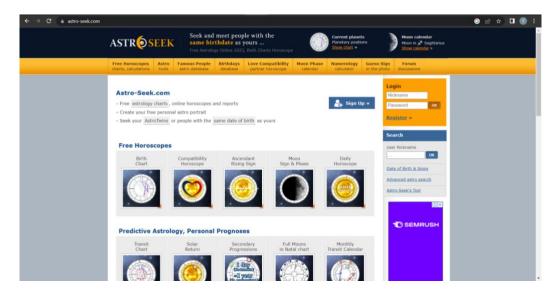
## • Cafe Astrology:

Cafe Astrology is known for its extensive collection of astrology-related content, including birth chart interpretations, compatibility analyses, horoscopes, and educational articles. Users can navigate the site and search for specific astrological topics [3].



## • Astrology Weekly:

Astrology Weekly is a platform that offers a range of astrological services, including horoscopes, articles, and educational resources. Users can search for astrological articles and content [4].



#### Astro-Seek:

Astro Seek is a website that provides astrology-related tools and resources, including birth chart interpretations and various astrological calculators. Users can search for specific astrological information and access their astrological charts [5].



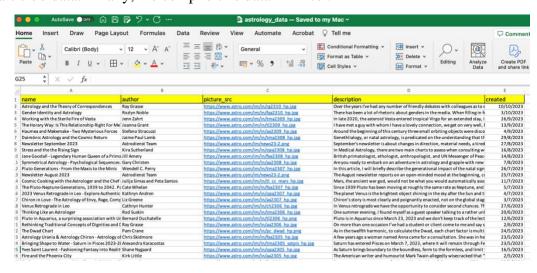
(d)

## **Implementation**

## I. Data collection, example documents, and data statistics

#### • Data Collection:

We begin by collecting data such as name, description, picture source (picture\_src), author, and creation or publication date from a website, which is Astro.com [1]. After collecting the data, we modify the format of the article's name to include the author's name. We separate the article's name from the author's name and then verify the accuracy of the source article's data. Finally, we compile this data in Excel.



After collecting the data in Excel, we performed a JSON transformation and condensed it into a one line to import the data into Elasticsearch and Kibana.

```
() articles-volution : X
() articles-volution for the study is Children's horoscope of Meronian", "Bayand", "picture_src": "https://www.astrg.com/inv/in/rg238_bp.jp c "hame": "Case Study is Children's horoscope of Meronian", "author": "Gristi Gelamon", "picture_src": "https://www.astrg.com/inv/in/rg238_bp.jp c "hame": "Case Study is Children's horoscope of Meronian", "author": "Gristi Gelamon", "picture_src": "https://www.astrg.com/inv/in/vap238_bp.jp c "hame": "To Sture Return and Quarter Square", "picture_src": "https://www.astrg.com/inv/in/vap238_bp.jp courtes d"hame": "Basin Return and Quarter Square", "picture_src": "https://www.astrg.com/inv/in/vap238_bp.jp courtes d"hame": "Astrology and the Theory of Gerrepondences", "author": "Ang Grasse", "picture_src": "https://www.astrg.com/inv/in/vap238_bp.jp c"hame": "Basin William ("hame": "Basin Meronian ("hame": "Sture San Meronian ("hame": "Sture San
```

## • Example Documents:

Name: Astrology and the Theory of Correspondences

Author: Ray Grasse

o **Picture\_src:** https://www.astro.com/im/in/rg2310\_hp.jpg

Obescription: Over the years I've had any number of friendly debates with colleagues as to exactly how astrology works. During those exchanges I've come across multiple theories which try to answer that question, drawing on such wideranging concepts as geo-magnetism, quantum non-locality, fractals, gravity, string theory, sacred geometry, Jung's theory of "synchronicity," or perhaps some subtle-but-as-yet-unknown energy of nature.

o Created: 10/10/2023

#### • Data Statistics:

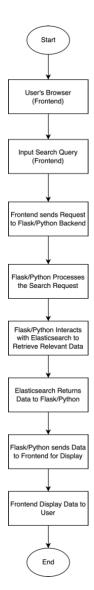
- Total number of documents.
- Word count distribution.
- o Common phrases or keywords.

## II. Tools and software

### • Tools:

- Elasticsearch and Kibana: Our team utilized Elasticsearch and Kibana for indexing and searching astrology-related documents efficiently, both of which are open-source search and analytics engines designed for various data types.
  - Elasticsearch, an open-source enterprise search engine, is widely distributed and scalable based on Lucene. It is released under the Apache License and stands as the most widely used enterprise search engine. Apache Solr, which also employs Lucene, follows closely behind Elasticsearch in popularity. Elasticsearch's ability to rapidly respond to searches is attributed to its approach of searching indexes rather than direct messages. This scalability allows Elasticsearch to expand to thousands of servers and handle petabytes of data.
  - **Kibana** serves as a free and open user interface, allowing us to visualize our Elasticsearch data and navigate the Elastic Stack. It facilitates various tasks, ranging from monitoring query loads to comprehending the flow of requests through our applications.
- Flask Application/ Python: Flask is a lightweight web application framework written in Python. We leveraged Flask to develop both the frontend and backend components of our search engine. Python, as the programming language, provides a versatile and efficient foundation for building the logic and functionality of the application. Flask, being a micro-framework, allowed us to structure and organize our code seamlessly, ensuring a scalable and maintainable solution for the search engine.
  - Flask Application: Flask served as the backbone of our application, handling requests, managing routes, and interfacing with the backend data processing.
- o **HTML**, **CSS**, **JavaScript**: These web technologies played a crucial role in crafting the user interface (UI) for our search engine.

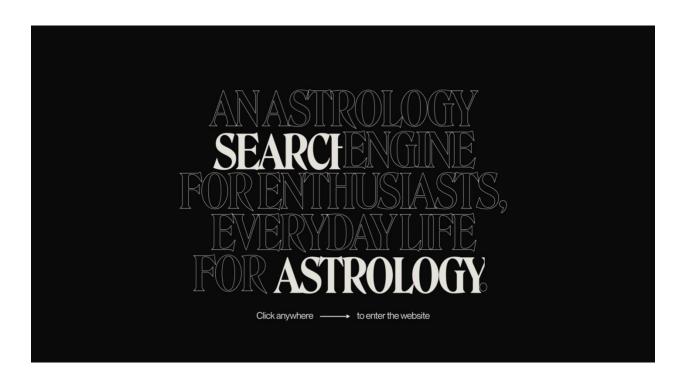
# III. System diagram



- 1. **User's Browser**: The starting point where the user interacts with the web application.
- 2. **User Inputs Search Query**: The user enters a search query into the frontend.
- 3. **Frontend Sends Request to Flask/Python Backend**: The frontend sends a request to the Flask/Python backend to process the search query.
- 4. **Flask/Python Processes the Search Request**: The backend processes the search request, preparing it for interaction with Elasticsearch.
- 5. **Flask/Python Interacts with Elasticsearch**: The backend communicates with Elasticsearch to retrieve relevant data based on the search query.
- 6. **Elasticsearch Returns Data to Flask/Python**: The retrieved data is sent back to the Flask/Python backend.
- 7. **Flask/Python Sends Data to Frontend for Display**: The backend sends the data to the frontend for display to the user.
- 8. **Frontend Displays Data to the User**: The user sees the search results on the frontend.

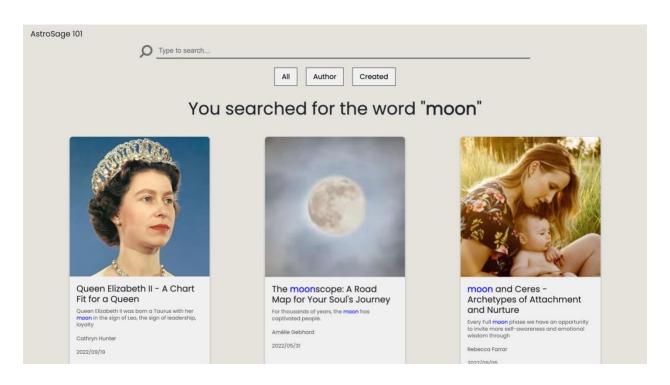
# IV. Snapshots of the system

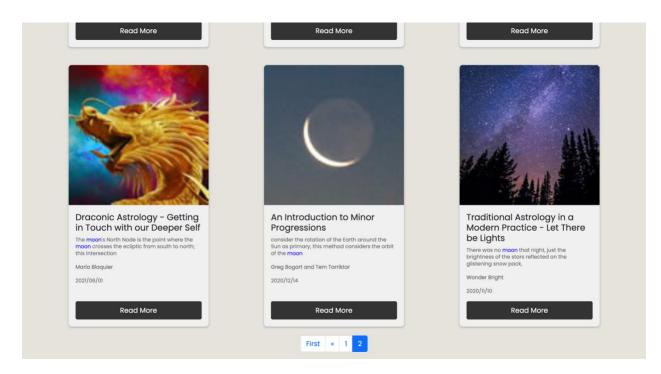
Index.html (For search keyword)





# Search.html (Display Data)

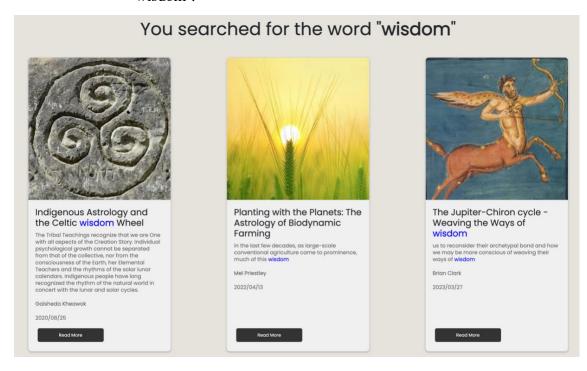




# V. Example step-by-step search sessions that highlight the following functionality:

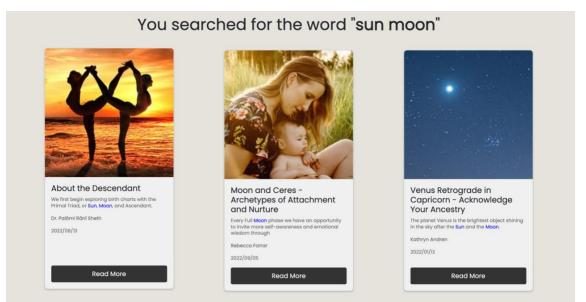
## 1. One word query

- Query: "wisdom"
- Expected Result: Information about the articles that related the word 'wisdom'.



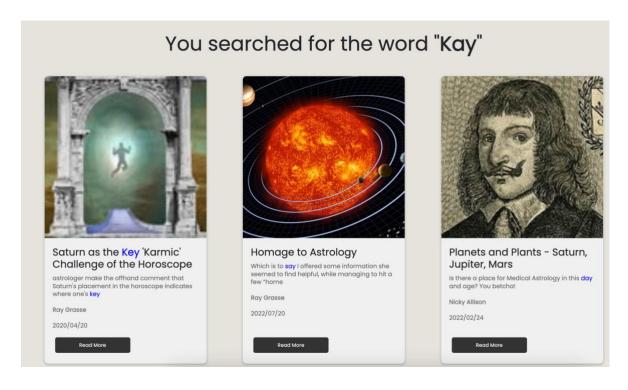
# 2. Multiple word query

- Query: "sun moon"
- Expected Result: Sun and moon related information in one article.



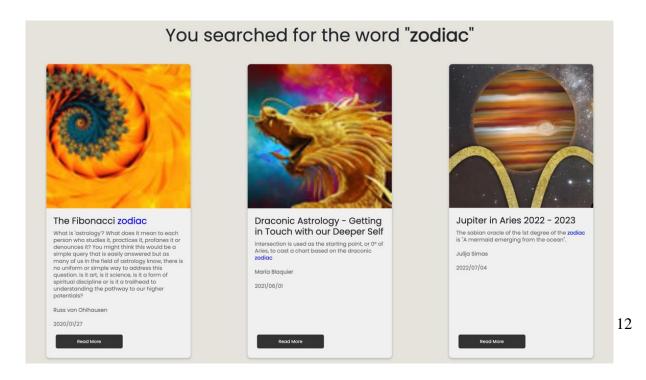
## 3. Partial match

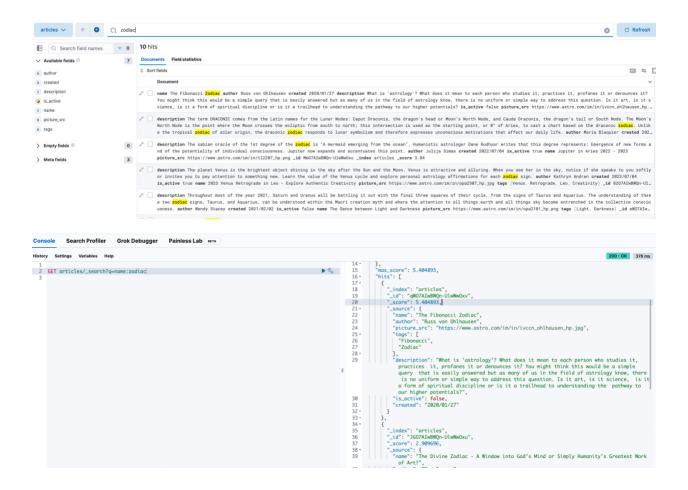
- Query: "Kay"
- Expected Result: Results including "Key," "Day," etc.



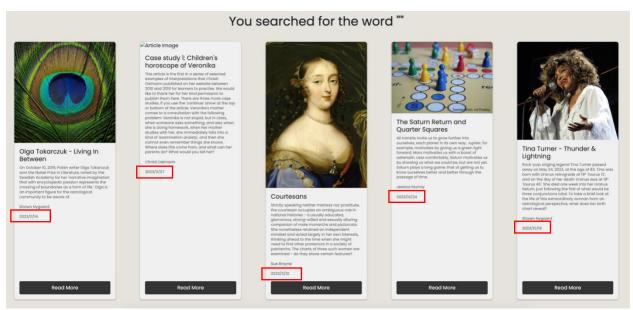
# 4. Ranking

- Query: "zodiac"
- Expected Result: Rank the results based on the score in descending order of articles related to the word 'zodiac.





- Query: ""
- Expected Result: Ranks results by date or created in descending order of all articles. (By default)



(e)

#### **Discussion**

## I. Limitations of our system

- a. **Scalability**: Depending on the volume of data and user queries, the system's performance may be impacted. ElasticSearch itself is scalable, but proper infrastructure planning is crucial.
- b. **Complexity of Queries**: While ElasticSearch provides powerful querying capabilities, complex queries may still require fine-tuning and experimentation.
- c. **Indexing Time**: The time it takes to index new documents could be a limitation, especially if dealing with a large number of documents frequently updated.
- d. Advanced Search Features: Advanced search features, such as faceted search or natural language processing (NLP), pose a limitation of our system in terms of complexity and potential resource requirements. These features often demand a deeper level of sophistication and may extend beyond the capabilities of the existing system.

## II. Technical difficulties, challenges, and lessons learned.

#### a. Technical difficulties:

i. Elasticsearch and Kibana

## `\$ ./elasticsearch`

rs [2023-11-26T17:56:02,901][INFO ][o.e.c.r.a.AllocationService] [LAPTOP-PT7SFDLG] current.health="YELLOW" message="Cluster health status changed from [RED] to [YELLOW] (reason: [shards started [[.apm-source-map][0]]])." previous.health="RED" reason="shards started [[.apm-source-map][0]]"

#### `\$ ./kibana`

[2023-11-27T07:28:15.334+07:00][ERROR][plugins.taskManager] [Task Runner] Task reports:monitor failed to release claim after failure: [object Object]
[2023-11-27T07:28:15.338+07:00][ERROR][plugins.taskManager] Failed to mark Task reports:monitor "reports:monitor" as running: Request timed out

Our Group encountering technical difficulties has been a recurring challenge in our project, primarily stemming from Elastic Search and Kibana usage. We frequently encounter errors when running Elastic Search, with a persistent yellow cluster status displayed, hindering the effective utilization of Kibana due to its close integration with Elastic Search. A noteworthy incident underscores the complexity of the issue. On a previous occasion when we successfully initiated Elastic Search and Kibana, we opted to keep my machine operational until the completion of the project. Unexpectedly, our machine remained powered on and plugged in for over 24 hours, highlighting the intricate nature of addressing and maintaining the stability of these components.

## b. Challenges:

## i. Scoring and Ranking Logic:

- For designing an effective scoring and ranking mechanism to ensure relevant search results can be complex and may require continuous refinement.
- **Approach**: Experiment with different scoring models, gather user feedback, and adjust the ranking logic based on the observed user behavior.

#### ii. Highlighting Logic:

- To implementing a robust highlighting logic that works smoothly with various search scenarios and handles HTML markup can be tricky.
- **Approach**: Test the highlighting logic thoroughly with different types of search queries and documents. Consider user experience in different scenarios.

#### c. Lessons learned:

i. Querying Elasticsearch: Elasticsearch queries, both single- and multiword queries, can be constructed logically from user input by using the keyword.

- **ii. Flask and Elasticsearch Integration:** The project efficiently makes use of Flask for web development and integrates Elasticsearch for search capabilities.
- **iii. Sorting and Ranking:** According to the date of creation of the article, the code performs a sort. For more advanced search engines, additional ranking algorithms may be necessary.
- **iv.** Comprehending Elasticsearch Index: The code demands an "articles" index, highlighting how crucial it is to comprehend the Elasticsearch index structure in order to successfully design queries.
- v. Information Retrieval Principles: Acknowledge the basics of indexing, querying, and relevance ranking in information retrieval.
- vi. Query Parsing and Analysis: Gain proficiency in handling various query types and efficiently parse and analyze user queries.
- **vii. Testing:** Implement thorough testing, including unit and integration tests, to identify and address issues as soon as possible.

## **III.** Opportunities for future improvements

- Enhanced Ranking Logic: Further refining the ranking logic based on user behavior, feedback, or additional criteria can improve the relevance of search results.
- Advanced Search Features: Introducing advanced search features such as
  faceted search, synonym handling, and natural language processing could enhance
  the user experience.
- **Real-time Indexing**: Depending on the use case, implementing real-time indexing to reduce the latency between document creation and availability in search results could be beneficial.
- **Security Improvements**: Ensuring robust security measures, especially when dealing with user inputs and sensitive data, is an ongoing process. Regular security audits and updates are essential.

(f)

## Conclusion

Elasticsearch and Kibana are open-source search tools that we use to create an "astrology" search system. This tool is designed for searching articles related to astrology. We have developed this search system for individuals who are interested in finding information about astrology. However, it is not limited to enthusiasts; it also includes anyone who wants to explore or study astrology-related information or read astrology articles. Elasticsearch is a flexible and open-source search and analytics tool for all types of data, developed in many Languages such as Python. The flexible search can display real-time results quickly. We use Kibana as our UI dashboard to visualize the search results. In addition, we use HTML, CSS, and JS to create our website pages for the astrology article search tool. Ultimately, our website search provides a satisfying experience for our target audience in finding articles that interest them.

# Reference

- [1] A. AG, "Astro," [Online]. Available: https://www.astro.com/horoscope. [Accessed 20 September 2023].
- [2] A. Team, "Astrology.com," [Online]. Available: https://www.astrology.com/us/home.aspx. [Accessed 20 September 2023].
- [3] C. team, "Cafeastrology," [Online]. Available: https://cafeastrology.com/. [Accessed 20 September 2023].
- [4] astrologyweekly Tea, "astrologyweekly," [Online]. Available: https://www.astrologyweekly.com/. [Accessed 20 September 2023].
- [5] a. s. team, "astro seek," [Online]. Available: https://www.astro-seek.com/. [Accessed 20 September 2023].