Report 2: Prototype IR System

By

Mr. Waris Damkham 6388014

Mr. Thanawath Huayhongthong 6388016

Miss Pattanan Korkiattrakool 6388022

Submitted to

Assoc. Prof. Dr. Suppawong Tuarob

A Report Submitted in Partial Fulfillment of

ITCS414 Information Storage and Retrieval

Faculty of Information and Communication Technology

Mahidol University

2021

Table of Contents

Introduction

Nowadays, the exploration and discovery of marine life have advanced and expanded. Currently, there are many species and types that are divided into different classes. In some species, there are physical similarities. But there are some differences in observation and behavior, for example, when talking about whales, many people will be confused and guess that each species they encounter is the same species.

Sea me like I see you are a search engine that provides general information about marine life. Data from the search engine includes details of scientific names and physical features such as shape, size, and body landmarks. Our searches also describe habitats discovered by marine life and settlements. Including foraging behavior that tells the type of food, they use to live and also mentions how the way of life in each season changes according to the changes of the world, etc.

Problem(s) that you are trying to solve

Sea me like I see you is designed for people interested in research and education about marine life such as divers, students, and students, including those who research the details of various marine life. Sea me like I see you is designed to classify each animal for most people or some groups of people who may be confused about the types of animals in the sea to understand and learn to know marine life correctly and provide detailed information about animals in the sea such as physical appearance Including the behavior of hunting food in their habitat as well. To reduce the misunderstanding of encountering marine animals, such as poisonous and dangerous marine animals, be cautious and coexist with these marine animals without disturbing each other to preserve these marine animals to live on.

Existing relevant systems

1. Marinebio

   Website citation: https://www.marinebio.org/creatures/

   - Marinebio is an online database for the most common and endangered marine species, including referenced taxonomic, morphological, behavioral, dietary, habitat, reproductive, and conservation status information. It will also include high-quality photographs, video, or access to video, as well as a variety of online resources for deeper species research. Marine algae and plants, marine worms, hard and soft corals (and other cnidarians like jellyfish), plankton (phytoplankton and zooplankton), echinoderms, crustaceans, cephalopods, commercial, reef, and deep-sea fishes, sharks, marine birds, sea turtles (and other marine reptiles), and marine mammals are among the species.

2. Fishbase

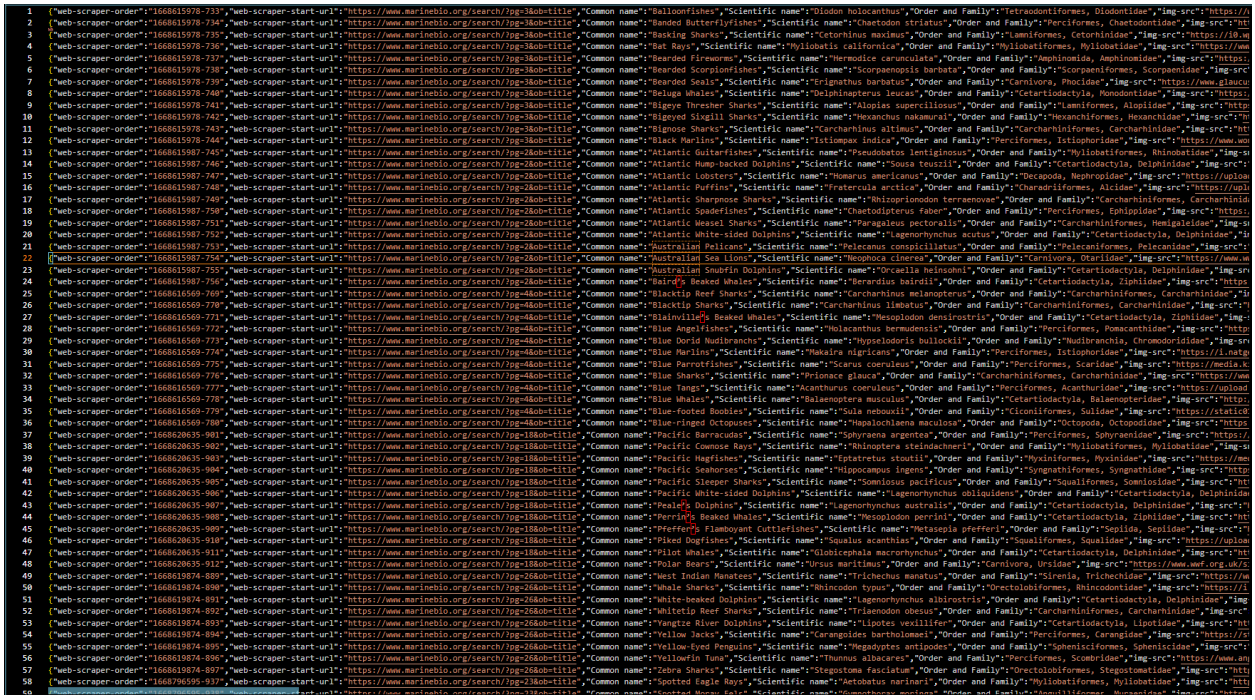
   Website citation: https://www.fishbase.se/NoRecordForCommonName.php?crit1_operator=EQUAL&CommonName=Sharks

   - Fishbase is a global online database of information about marine life. It aims to provide key information on the taxonomy, distribution, and ecology of all marine species in the world apart from finfish. Fishbase works with the WorldFish Center in Malaysia and the University of British Columbia's UBC Institute for the Oceans and Fisheries. (Daniel Pauly is the principal investigator, and it is coordinated by Maria Lourdes D. Palomares. As of October 2016, it had descriptions of 74,000 species, 47,700 common names, 12,400 pictures, and references to 31,700 works in the scientific literature.

Sea Me Like I See You Search System Implementation

Data collection

The data we have used is from source data which are the 'Fishbase' and 'Marinebio websites. We use 'web scrapper' to retrieve data and then store it in an NDJSON file. For images, we use the URL of each image from its image source and then stored it in an NDJSON file as well as other data.



Example of data document (Dataset)

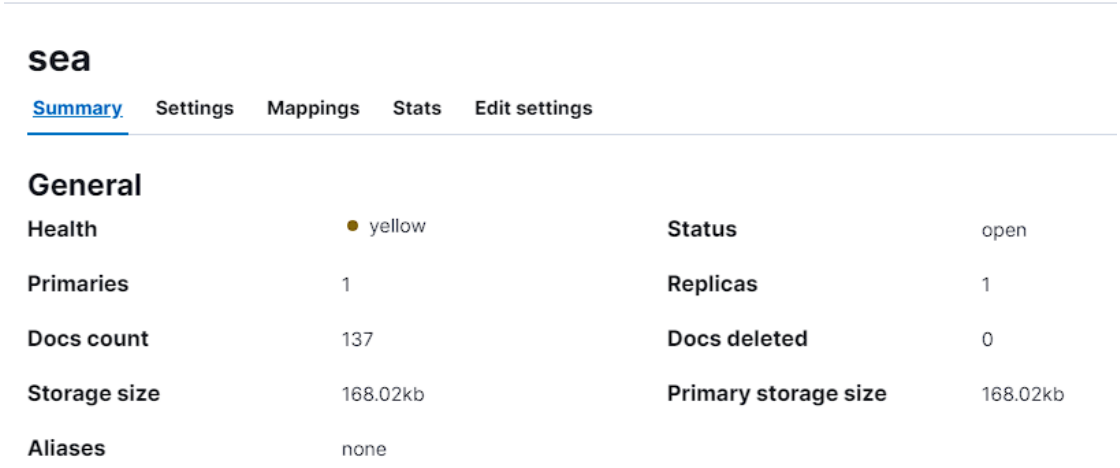
Example documents

The information we collect is the basic information that most users would like to know about the aquatic species, namely the common name, scientific name, order and family, image, and description which includes its common and distinctive features and behavior also habitat and eating habits.

Data statistics

The JSON data is loaded as the sea index into the Elasticsearch search engine. The JSON data file contains 137 records, which are analyzed and translated into 137 documents for the marine index.

The data in each document is composed of mainly five fields: the marine animal's common name ("Common name"), the image of the marine animal in jpg and png format ("img-src"), the description of the marine animal ("Description"), the marine animal's scientific name ("Scientific name"), and the order and family of the marine animal ("Order and Famaily")



Tools and software

- **Elasticsearch**

  We have utilized Elasticsearch as our search engine. The following are the primary settings that we have implemented.

  1. For ranking, we have used the ascending (smallest value first) sorting module, which is the Standard module. Therefore, no configuration was needed.

```
"sort": [
        { "Common name": "asc" }
    ]
```

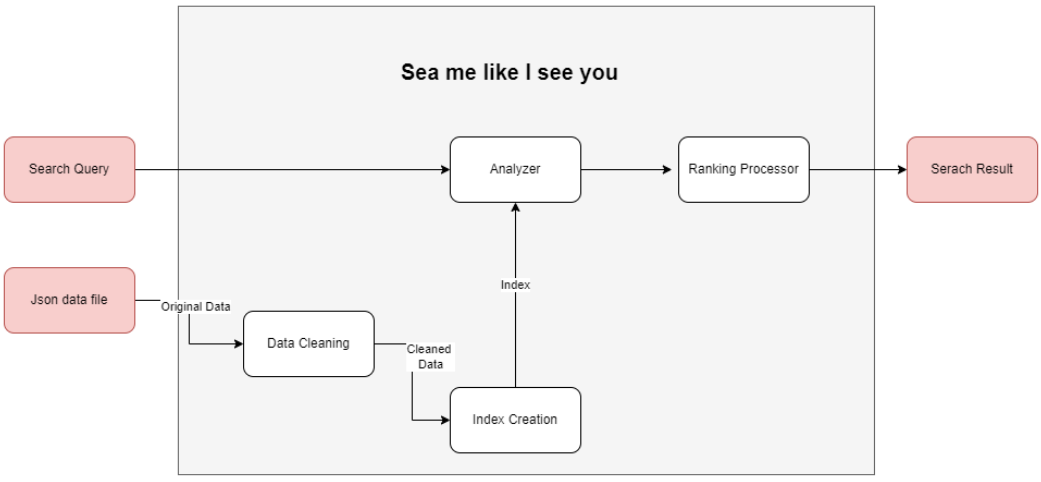
2. For the data analyzer, we have used a custom Elasticsearch analyzer, which is the implementation analyzer. We have configured this in Elasticsearch's development tool. The configuration is as follows.

```
"analysis": {
  "filter": {
    "my_stemmer": {
      "type": "stemmer",
      "language": "English"
    },
    "my_custom_stop_words_filter": {
      "ignore_case": "true",
      "type": "stop",
      "stopwords": [
        "and",
        "is",
        "the",
        "a",
        "an"
      ]
    }
  },
  "analyzer": {
    "my_analyzer": {
      "filter": [
        "lowercase",
        "my_stemmer",
        "my_custom_stop_words_filter"
      ],
      "tokenizer": "standard"
    }
  }
},
```

- **Flask**

Flask was chosen as our web service to facilitate the use of Elasticsearch as our search engine.

System diagram



Snapshots of the system

Our search system's websites are made of two pages: the inde2.html page and the search.html page. The page screenshots are displayed below.

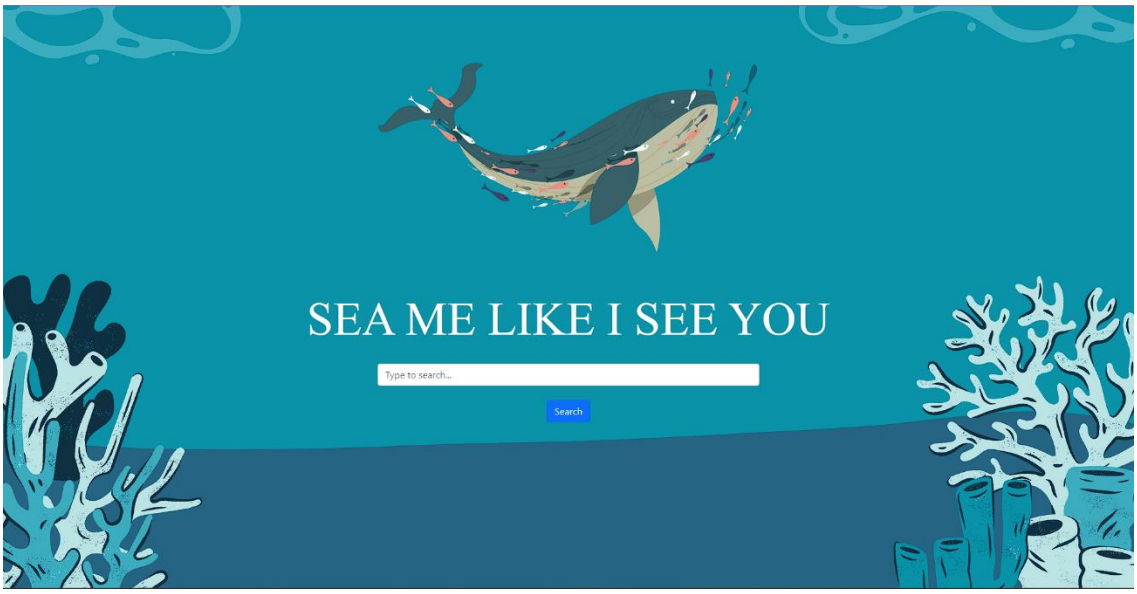


Figure 5 shows the search page of Sea Me Like I See You search system (inde2.html page)

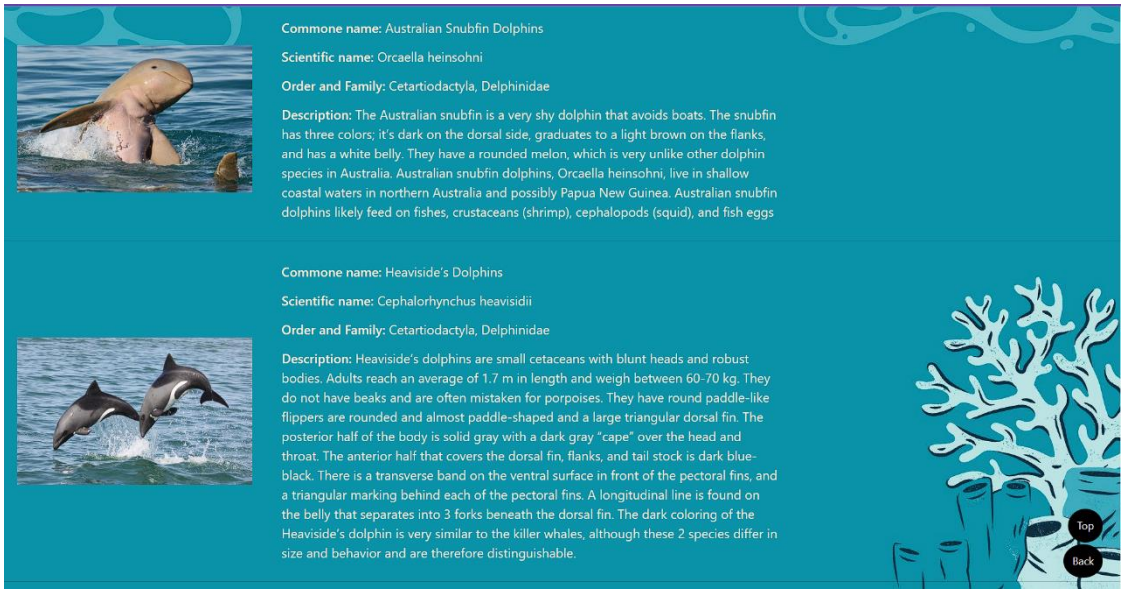


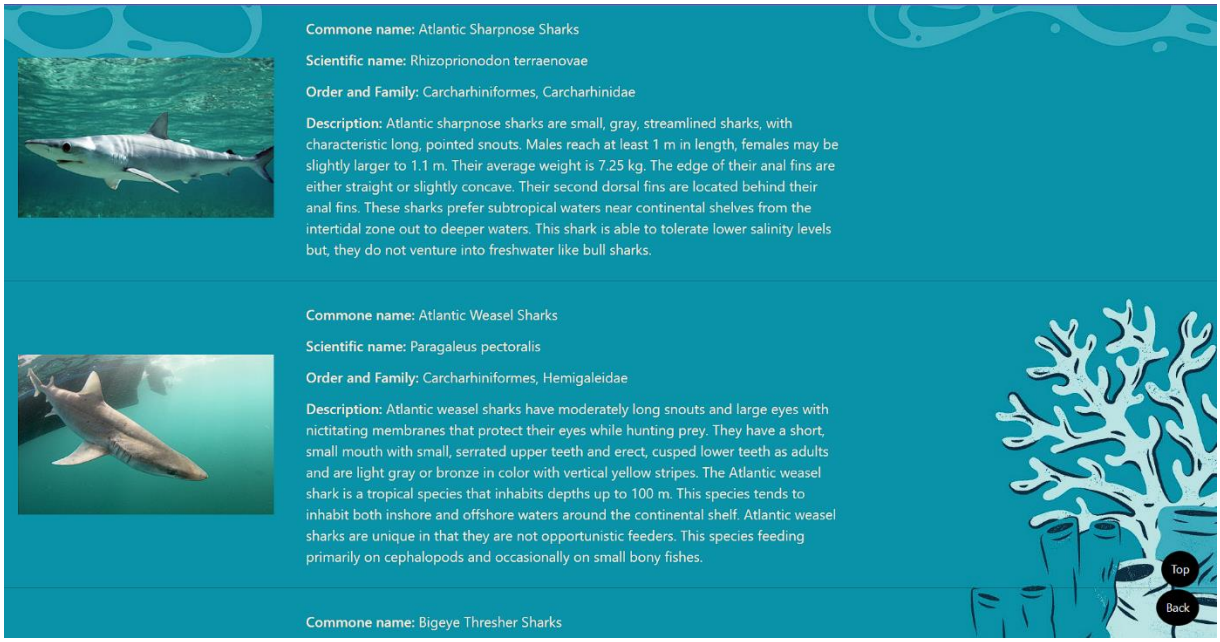
Figure 6 shows the search results page of Sea Me Like I See You search system (search.html page)

Search Sessions Examples

One word query

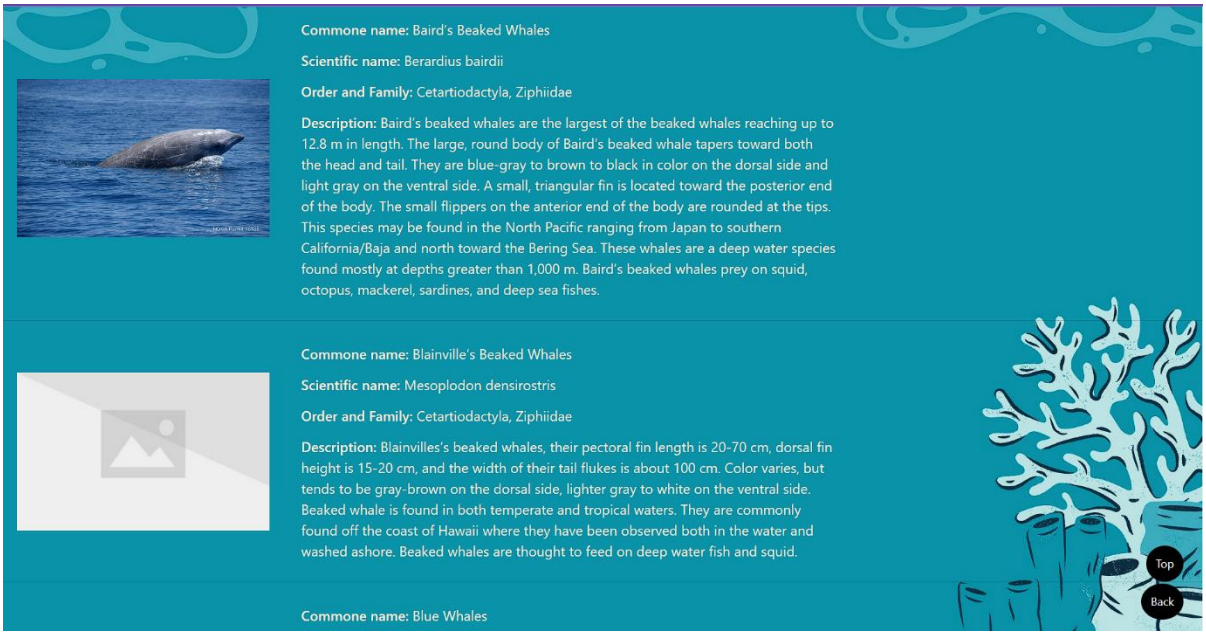
An example query is “Shark”

The search result is the document that contains the term “Sharks”.



Multiple word query

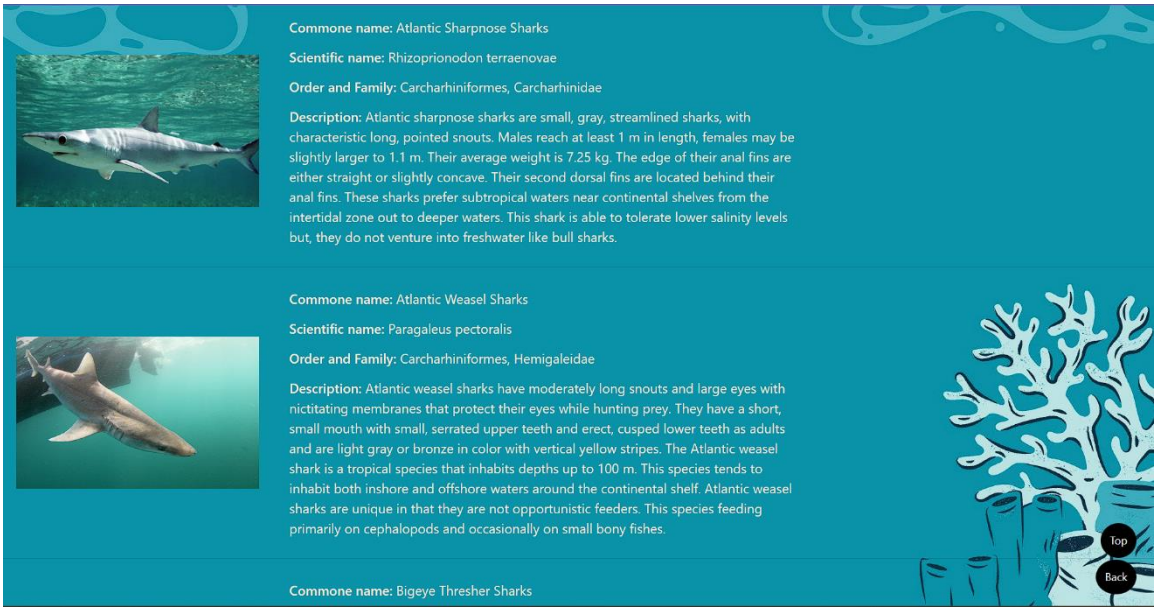
An example query is “large whales” The top two search results are the documents that contain the terms “large “and “whales”.



Partial match

An example query is “large Great shark”

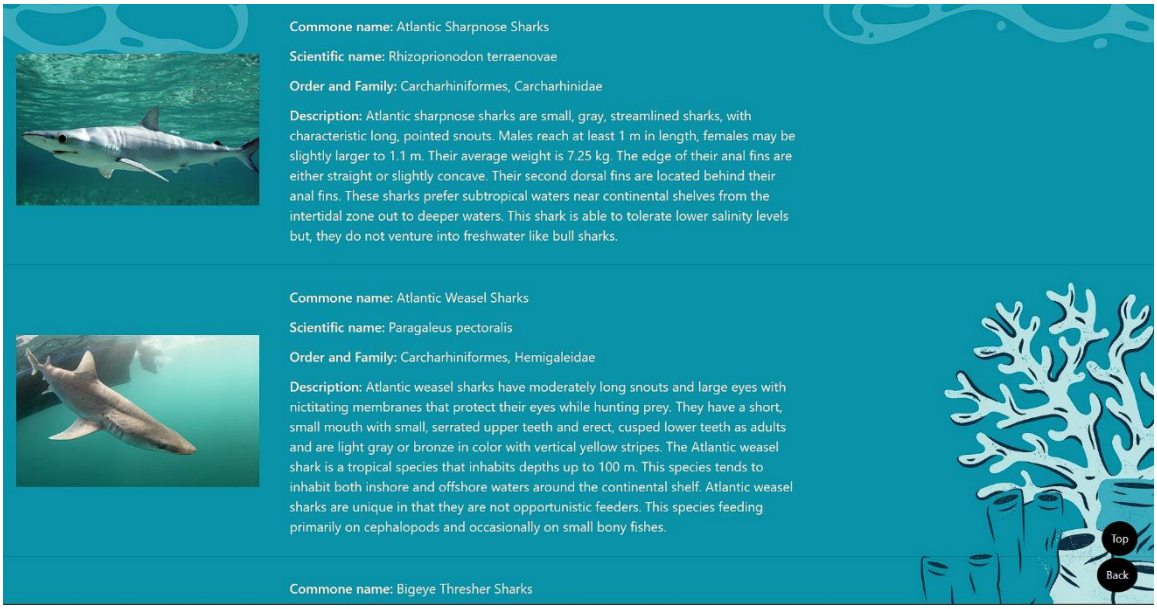
The top two search results are the documents that contain the terms “large” and “shark”, but do not contain the term “great”.



Ranking

An example query is “Sharks”

The first search result contains the word "Sharks" and is therefore very relevant to the search query. This search result is sorted by ascending order in addition to the common name.



Discussion

- **Limitations of your system**

  Our search systems are unable to find general information including habitat depth and elevation or body length, e.g. whales with swimming or diving levels between 50-100 meters. Because the numerical depth is in the form of a string which is not an integer value. Another limitation is that each recording has a different unit, for example, in recording the sea level depth, some types are recorded in meters or feet. The search system is therefore unable to group the sea surface depths where the aquatic animals reside for users to search. which in our system can display only text data.

- **Technical difficulties, challenges, and lessons learned**

  The challenge and difficulty in making search engines are that storing image data for elastic search results is quite difficult as not all image addresses are available to display images on web pages. Moreover, 'Elasticsearch' is quite easy to use, but with the setting up of the working environment cause sometimes it is difficult and complicated to use because there are some errors such as the stop working problem, so the operation is quite a waste of time and a hindrance.

  The lesson that we have learned in a search engine project is to store a large size of data into a dataset and import it into Elasticsearch by using ‘Web Scraper’ which is a tool to help in the gathering data process to make it easier. Moreover, we also have experience in using Elasticsearch as well as learning more about marine life.

- **Opportunities for future improvements**

  In terms of improving the search engine of our website in the future, we will improve our search engine to include more insights into each animal with high resolution and high relevance and will add a description search to make it more responsive to users. At the moment our search engine can only find results from the words in the description in some cases which are not covered in all cases.

Conclusion

The Sea me like I see you Search system is for scuba divers, marine scientists, marine enthusiasts also teachers, and students. It is a tool that helps these users be able to study in depth and explore relevant information about marine life. Sea me like I see you provides basic information that most users would like to know about aquatic species. The information consists of the common name, scientific name, order and family, image, and description which includes its common and distinctive features, behavior and habitat, and eating habits.

In terms of data integration and implementation for the search system, we use 'web scraper' to retrieve data from source data and then stored it into a NDJSON file. Next, is to import a NDJSON file into the 'Elasticsearch' engine, to test that the imported file can be successfully imported and implemented in the Elasticsearch, if successful, the data in the dataset will be displayed through the website that we created.

Even though our search system provides a lot of information to users when searching but even then there are some limitations in the search as well. That is, the information about some animals may not be complete or not enough for the user, especially some specifics such as the depth of the dwelling, lifespan, or animal involved. In addition, we are also trying to improve the query to be more detailed in order to meet the needs of users.