**Proposal Presentation**

**What is your search system for?**

Our team wanted to develop a search engine where marine life-related information could be located. Why? We are interested in studying marine life and wished to develop a straightforward approach for marine life-curious readers. Therefore, with our system, users will be able to label marine life by category and habitat for any type of organism. This will generate a list of marine life that corresponds to the user's request.

**What are your target user groups?**

Our main target audience will consist of 3 groups which are sea ​​creatures enthusiasts, scuba enthusiasts, and divers. These people are more interested in the sea and underwater life than others, and when they get to the sea they enjoy it a lot more because they can immediately know what this species is. Not only are 3 main target groups but there are also students or teachers who can be our target audience as they may utilize our search engines to obtain educational content.

**What is your data source?**

For our group, we will use marinebio and sealifebase as our primary data sources.

<https://www.marinebio.org/creatures/>

<https://www.sealifebase.ca/>

https://www.fishbase.se/NoRecordForCommonName.php?crit1\_operator=EQUAL&CommonName=Sharks

<https://fishbaseapi.readme.io/reference/base-url>

<https://mycourses.ict.mahidol.ac.th/pluginfile.php/23069/mod_resource/content/1/1_2022-ITCS414-Project2.v1.pdf>

<https://www.canva.com/design/DAFNKk5Zolo/aBOiA3cDMZ-Ou06hxXNxUA/edit>

**What is a Marinebio?**

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.

**Why Marinebio?**

The reason why our group chose Marinebio to be our primary data source is

1. Marinebio is collaborative open source data, so anyone may use it without cost.
2. Contain relevant content that is suitable for our group topic.

**What is a Fishbase?**

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.

**Why Fishbase?**

The reason why our group chose sealifebase to be our primary data source is

1. Contain relevant content that is suitable for our group topic.
2. Provide APIs for us to fetch data and crawl through their website contents.

**How to preprocess your data?**

First, we will collect data from data sources to create and store data in Dataset in external file such as excel and json. In dataset, We will keep the order in alphabetical order, where each alphabet will store approximately 4-10 types of data.

Using 3 techniques to prepare data from data source in preprocess, which are

1. Stop Word
2. Stemming
3. Tokenization

* Stop Word
* Stop Word is the generally filtered out before processing a language.
* To remove stop word, In this algorithm is to cut out unnecessary words which are articles, prepositions, conjunctions, etc. and focus on important pieces of information. This reduces the size and time of searching for input queries.
* For example: “The great white shark”, so when we remove stop word, it will be “great white shark”.
* Stemming
* Stemming commonly referred simply "chop off the ends of words" as is to convert words back to its original roots of words.
* For example: If a user input “Great whity shark”, stemming algorithm will convert the word ‘whity’ back into an original form, so the word will be reverted as "white".
* Tokenization
* Tokenization is the process of splitting input textual data into separate meaningful pieces which called 'token' that can be further understood and processed by machines.
* For example: “Great White Shark” will be separated into 3 tokens which are |Great|, |White|, |Shark|.

**How to index the documents? จัดลำดับ**

We will use K-gram to indexing our document. K-gram is the division of a word proportion of 1 word 3 letters called a trigram. In this way it is suitable for our search engine because our data is very large. Therefore, by dividing that word, the search time will be faster and the closest other results can be found.

Example of K-gram in Trigram: word ‘animal’ will ‘$an ani nim ima mal al$’. We actually use the “$” symbol to denote the beginning and end of the word

**How to rank the documents? การจัดอันดับ**

First, gather all reputable sources. All sources with the same criterion or keyword should be compared to reduce inaccuracies. Then, collect all the validated data, sort it by order, and utilize the sorted data as a dataset.

Then, we divide the criteria by worksheet and sort them according to the request criteria, ranking the search by keyword. For example, put "shark" as the keyword. It sorts alphabetically and may also be sorted by nation. So, we can categorize them and give them the top retrieved documents.