Project 2: Prototype IR System Report

By

Thanachot Onlamoon 6588062
Punnut Sawetwannakul 6588142
Bhurinat Kanchanasuwan 6588150

This project is partial fulfillment of the requirements for
ITCS414 Information storage and retrieval
Academic year 2024/ 1st semester
Faculty of ICT, Mahidol university

Table of content

Introduction

In this report, we present the development of a prototype information retrieval (IR) system focused on enhancing search functionality for information related specifically to Doraemon gadgets. Targeting both Doraemon enthusiasts and toy designers, this system aims to provide a streamlined solution for accessing detailed data on specific gadgets, which is currently challenging due to limited specialized search options. Using web scraping and indexing techniques with tools like GoColly, Elasticsearch, and Kibana, our team built an interface to support intuitive and efficient queries, offering a customized experience for users interested in this unique niche. This report outlines our process, from identifying the problem and analyzing existing systems to implementation and discussion of technical challenges, lessons learned, and system limitations.

Problems that we are trying to solve

Problem that we are trying to solve is:
There are:

- Many doraemon fans and doraemon merchandise collectors who want to know more information that are related to some doraemon gadgets.
- Many toy designers and toy creators who also want to know more information that are related to some doraemon gadgets to design and create these gadgets in real life.

But they don't have a decent search system which can search for information that are related to doraemon gadgets only.

Existing relevant systems

There are some existing relevant systems which can search for information that are related to doraemon but there are 0 existing relevant systems which can search for information that are related to doraemon gadgets only.

These are the main existing relevant systems that are mentioned above.

- Doraemon Wiki
- Doraemon.com

Implementation

Data collection

Our group uses “web scraping” as our data collection method. Web scraping is a type of data collection method that uses softwares to collect data such as texts, pictures and much more from a certain website.
Our group uses Doraemon Wiki as our source website.

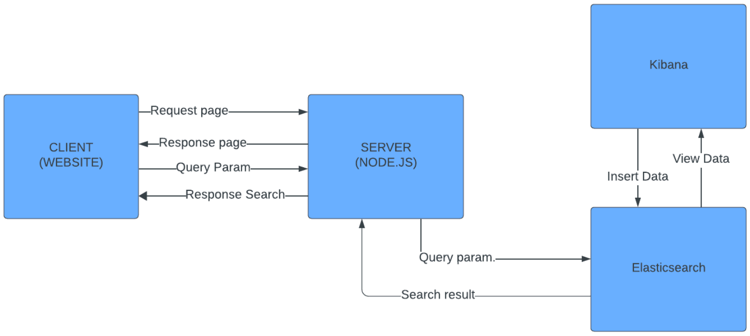
Software and Tool

These are softwares and tools that our group uses:

- Data collection: our group uses “Go-Colly” to collect data.
  Go-Colly is a type of web scraping tool which is written in Go programming language and can automate a process of collecting data such as texts, pictures and much more from a website.
- Data indexing: our group uses “Elasticsearch (Okapi BM25)” and “Kibana” to index data.
  Elasticsearch (Okapi BM25) is a type of search engine which is a part of an elastic stack and can search a certain data from a large database/ store a certain data into a large database.
  Kibana is a type of data visualization tool that is also a part of an elastic stack and can view a large database which is stored in Elasticsearch through a graph or a dashboard.

- Search system interface designing: our group uses “Node.js”, “Express.js”, “HTML”, “CSS”, and “Javascript” to design our search system interface.

System diagram

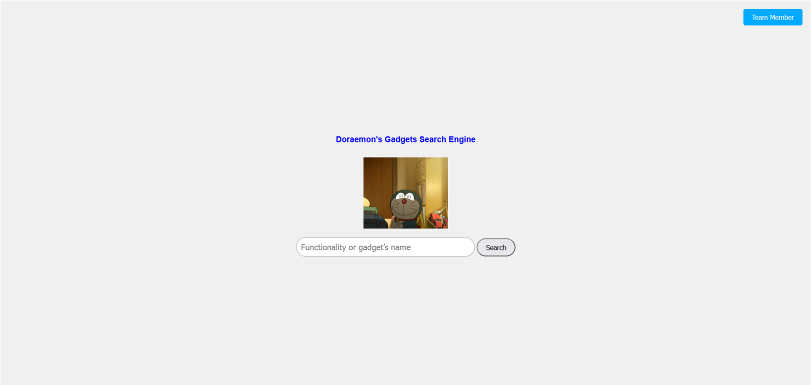


System snapshots

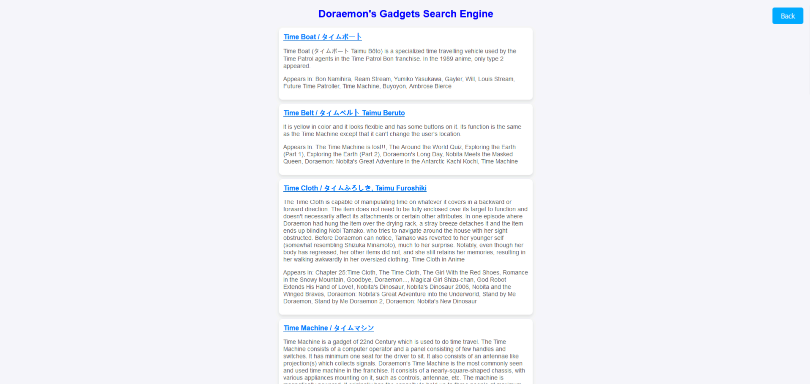
Link to the system: https://doraemon.mahumeaw.com/
Link to GitHub Repository: https://github.com/Yobubble/doraemon-gadgets-search-engine.git

These are our search system interface:

- Search page: this page is a page where users can search for a certain information by typing a query.


Doraemon's Gadgets Search Engine

- Search results page: this page is a page where users can see a certain information from their typed query.



Doraemon's Gadgets Search Engine

Time Boat / タイムボート

Time Boat (タイムボート Taimu Bōto) is a specialized time travelling vehicle used by the Time Patrol agents in the Time Patrol Bon franchise. In the 1989 anime, only type 2 appeared.

Appears In: Bon Namihira, Ream Stream, Yumiko Yasukawa, Gayler, Will, Louis Stream, Future Time Patroller, Time Machine, Buyoyon, Ambrose Bierce

Time Belt / タイムベルト Taimu Beruto

It is yellow in color and it looks flexible and has some buttons on it. Its function is the same as the Time Machine except that it can't change the user's location.

Appears In: The Time Machine is lost!!, The Around the World Quiz, Exploring the Earth (Part 1), Exploring the Earth (Part 2), Doraemon's Long Day, Nobita Meets the Masked Queen, Doraemon: Nobita's Great Adventure in the Antarctic Kachi Kochi, Time Machine

Time Cloth / タイムふろしき, Taimu Furoshiki

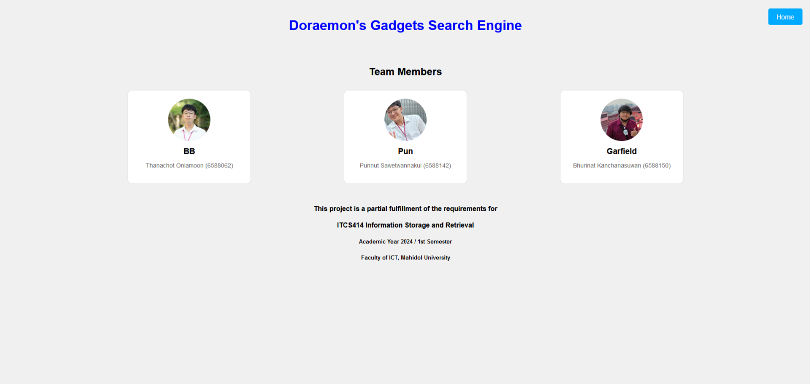
The Time Cloth is capable of manipulating time on whatever it covers in a backward or forward direction. The item does not need to be fully enclosed over its target to function and doesn't necessarily affect its attachments or certain other attributes. In one episode where Doraemon had hung the item over the drying rack, a stray breeze detaches it and the item ends up blinding Nobi Tamako. who tries to navigate around the house with her sight obstructed. Before Doraemon can notice, Tamako was reverted to her younger self (somewhat resembling Shizuka Minamoto), much to her surprise. Notably, even though her body has regressed, her other items did not, and she still retains her memories, resulting in her walking awkwardly in her oversized clothing. Time Cloth in Anime

Appears In: Chapter 25:Time Cloth, The Time Cloth, The Girl With the Red Shoes, Romance in the Snowy Mountain, Goodbye, Doraemon..., Magical Girl Shizu-chan, God Robot Extends His Hand of Love!, Nobita's Dinosaur, Nobita's Dinosaur 2006, Nobita and the Winged Braves, Doraemon: Nobita's Great Adventure into the Underworld, Stand by Me Doraemon, Stand by Me Doraemon 2, Doraemon: Nobita's New Dinosaur

Time Machine / タイムマシン

Time Machine is a gadget of 22nd Century which is used to do time travel. The Time Machine consists of a computer operator and a panel consisting of few handles and switches. It has minimum one seat for the driver to sit. It also consists of an antennae like projection(s) which collects signals. Doraemon's Time Machine is the most commonly seen and used time machine in the franchise. It consists of a nearly-square-shaped chassis, with various appliances mounting on it, such as controls, antennae, etc. The machine is

- Our group member page: this page is a page where users can see our group member information.



Doraemon's Gadgets Search Engine

Team Members

BB

Thanachot Onlamoon (6588062)

Pun

Punnut Sawetwannakul (6588142)

Garfield

Bhurinat Kanchanasuwan (6588150)

This project is a partial fulfillment of the requirements for

ITCS414 Information Storage and Retrieval

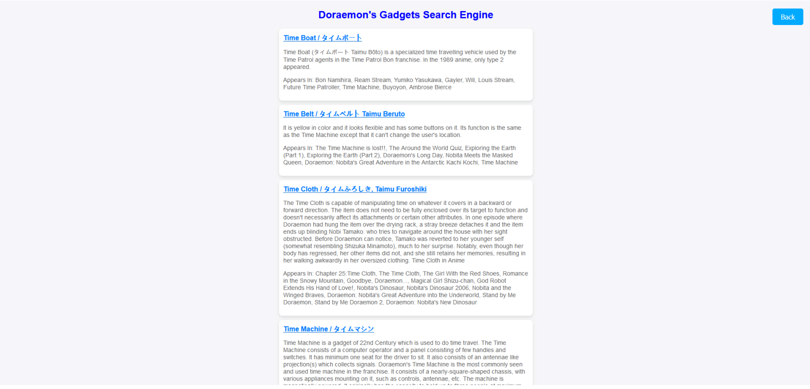
Academic Year 2024 / 1st Semester

Faculty of ICT, Mahidol University

System snapshots: query and rank

This is our search results page when:

- Users type "time" (an example of 1 word query) in our search page.



Doraemon's Gadgets Search Engine

Time Boat / タイムボート

Time Boat (タイムボート Taimu Bōto) is a specialized time travelling vehicle used by the Time Patrol agents in the Time Patrol Bon franchise. In the 1989 anime, only type 2 appeared.

Appears In: Bon Namihira, Ream Stream, Yumiko Yasukawa, Gayler, Will, Louis Stream, Future Time Patroller, Time Machine, Buyoyon, Ambrose Bierce

Time Belt / タイムベルト Taimu Beruto

It is yellow in color and it looks flexible and has some buttons on it. Its function is the same as the Time Machine except that it can't change the user's location.

Appears In: The Time Machine is lost!!, The Around the World Quiz, Exploring the Earth (Part 1), Exploring the Earth (Part 2), Doraemon's Long Day, Nobita Meets the Masked Queen, Doraemon: Nobita's Great Adventure in the Antarctic Kachi Kochi, Time Machine

Time Cloth / タイムふろしき, Taimu Furoshiki

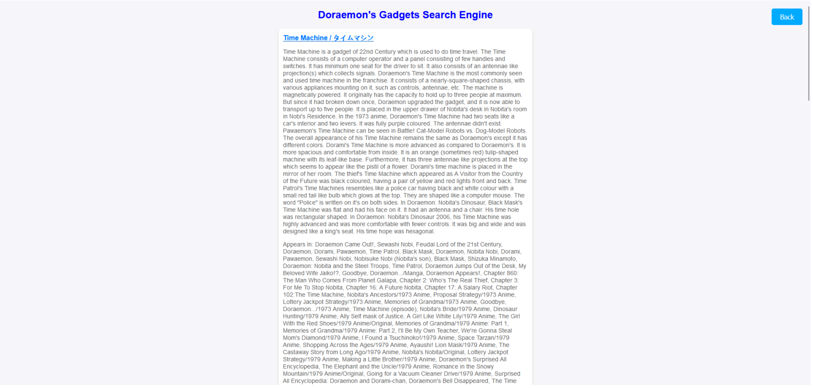
The Time Cloth is capable of manipulating time on whatever it covers in a backward or forward direction. The item does not need to be fully enclosed over its target to function and doesn't necessarily affect its attachments or certain other attributes. In one episode where Doraemon had hung the item over the drying rack, a stray breeze detaches it and the item ends up blinding Nobi Tamako. who tries to navigate around the house with her sight obstructed. Before Doraemon can notice, Tamako was reverted to her younger self (somewhat resembling Shizuka Minamoto), much to her surprise. Notably, even though her body has regressed, her other items did not, and she still retains her memories, resulting in her walking awkwardly in her oversized clothing. Time Cloth in Anime

Appears In: Chapter 25:Time Cloth, The Time Cloth, The Girl With the Red Shoes, Romance in the Snowy Mountain, Goodbye, Doraemon..., Magical Girl Shizu-chan, God Robot Extends His Hand of Love!, Nobita's Dinosaur, Nobita's Dinosaur 2006, Nobita and the Winged Braves, Doraemon: Nobita's Great Adventure into the Underworld, Stand by Me Doraemon, Stand by Me Doraemon 2, Doraemon: Nobita's New Dinosaur

Time Machine / タイムマシン

Time Machine is a gadget of 22nd Century which is used to do time travel. The Time Machine consists of a computer operator and a panel consisting of few handles and switches. It has minimum one seat for the driver to sit. It also consists of an antennae like projection(s) which collects signals. Doraemon's Time Machine is the most commonly seen and used time machine in the franchise. It consists of a nearly-square-shaped chassis, with various appliances mounting on it, such as controls, antennae, etc. The machine is

- Users type “time machine” (an example of multiple word query) in our search page.



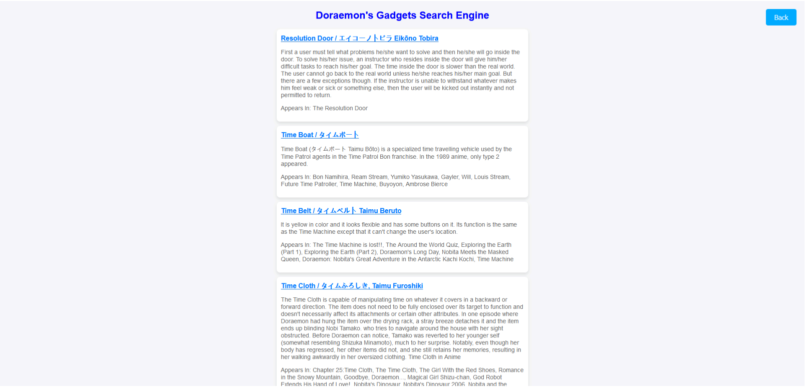
Doraemon's Gadgets Search Engine

Time Machine / タイムマシン

Time Machine is a gadget of 22nd Century which is used to do time travel. The Time Machine consists of a computer operator and a panel consisting of few handles and switches. It has minimum one seat for the driver to sit. It also consists of an antennae like projection(s) which collects signals. Doraemon's Time Machine is the most commonly seen and used time machine in the franchise. It consists of a nearly-square-shaped chassis, with various appliances mounting on it, such as controls, antennae, etc. The machine is magnetically powered. It originally has the capacity to hold up to three people at maximum. But since it had broken down once, Doraemon upgraded the gadget, and it is now able to transport up to five people. It is placed in the upper drawer of Nobita's desk in Nobita's room in Nobi's Residence. In the 1973 anime, Doraemon's Time Machine had two seats like a car's interior and two levers. It was fully purple coloured. The antennae didn't exist. Pawaemon's Time Machine can be seen in Battle! Cat-Model Robots vs. Dog-Model Robots. The overall appearance of his Time Machine remains the same as Doraemon's except it has different colors. Dorami's Time Machine is more advanced as compared to Doraemon's. It is more spacious and comfortable from inside. It is an orange (sometimes red) tulip-shaped machine with its leaf-like base. Furthermore, it has three antennae like projections at the top which seems to appear like the pistil of a flower. Dorami's time machine is placed in the mirror of her room. The thief's Time Machine which appeared as A Visitor from the Country of the Future was black coloured, having a pair of yellow and red lights front and back. Time Patrol's Time Machines resembles like a police car having black and white colour with a small red tail like bulb which glows at the top. They are shaped like a computer mouse. The word "Police" is written on it's on both sides. In Doraemon: Nobita's Dinosaur, Black Mask's Time Machine was flat and had his face on it. It had an antenna and a chair. His time hole was rectangular shaped. In Doraemon: Nobita's Dinosaur 2006, his Time Machine was highly advanced and was more comfortable with fewer controls. It was big and wide and was designed like a king's seat. His time hope was hexagonal.

Appears In: Doraemon Came Out!, Sewashi Nobi, Feudal Lord of the 21st Century, Doraemon, Dorami, Pawaemon, Time Patrol, Black Mask, Doraemon, Nobita Nobi, Dorami, Pawaemon, Sewashi Nobi, Nobisuke Nobi (Nobita's son), Black Mask, Shizuka Minamoto, Doraemon: Nobita and the Steel Troops, Time Patrol, Doraemon Jumps Out of the Desk, My Beloved Wife Jaiko!?, Goodbye, Doraemon.../Manga, Doraemon Appears!, Chapter 860: The Man Who Comes From Planet Galapa, Chapter 2: Who’s The Real Thief, Chapter 3: For Me To Stop Nobita, Chapter 16: A Future Nobita, Chapter 17: A Salary Riot, Chapter 102:The Time Machine, Nobita's Ancestors/1973 Anime, Proposal Strategy/1973 Anime, Lottery Jackpot Strategy/1973 Anime, Memories of Grandma/1973 Anime, Goodbye, Doraemon.../1973 Anime, Time Machine (episode), Nobita's Bride/1979 Anime, Dinosaur Hunting/1979 Anime, Ally Self mask of Justice, A Girl Like White Lily/1979 Anime, The Girl With the Red Shoes/1979 Anime/Original, Memories of Grandma/1979 Anime: Part 1, Memories of Grandma/1979 Anime: Part 2, I'll Be My Own Teacher, We're Gonna Steal Mom's Diamond/1979 Anime, I Found a Tsuchinoko!/1979 Anime, Space Tarzan/1979 Anime, Shopping Across the Ages/1979 Anime, Ayaushi! Lion Mask/1979 Anime, The Castaway Story from Long Ago/1979 Anime, Nobita's Nobita/Original, Lottery Jackpot Strategy/1979 Anime, Making a Little Brother/1979 Anime, Doraemon's Surprised All Encyclopedia, The Elephant and the Uncle/1979 Anime, Romance in the Snowy Mountain/1979 Anime/Original, Going for a Vacuum Cleaner Drive/1979 Anime, Surprised All Encyclopedia: Doraemon and Dorami-chan, Doraemon's Bell Disappeared, The Time

- Users type “tim” (an example of partial match query) in our search page.



Doraemon's Gadgets Search Engine

Resolution Door / エイコーノトビラ Eikōno Tobira

First a user must tell what problems he/she want to solve and then he/she will go inside the door. To solve his/her issue, an instructor who resides inside the door will give him/her difficult tasks to reach his/her goal. The time inside the door is slower than the real world. The user cannot go back to the real world unless he/she reaches his/her main goal. But there are a few exceptions though. If the instructor is unable to withstand whatever makes him feel weak or sick or something else, then the user will be kicked out instantly and not permitted to return.

Appears In: The Resolution Door

Time Boat / タイムボート

Time Boat (タイムボート Taimu Bōto) is a specialized time travelling vehicle used by the Time Patrol agents in the Time Patrol Bon franchise. In the 1989 anime, only type 2 appeared.

Appears In: Bon Namihira, Ream Stream, Yumiko Yasukawa, Gayler, Will, Louis Stream, Future Time Patroller, Time Machine, Buyoyon, Ambrose Bierce

Time Belt / タイムベルト Taimu Beruto

It is yellow in color and it looks flexible and has some buttons on it. Its function is the same as the Time Machine except that it can't change the user's location.

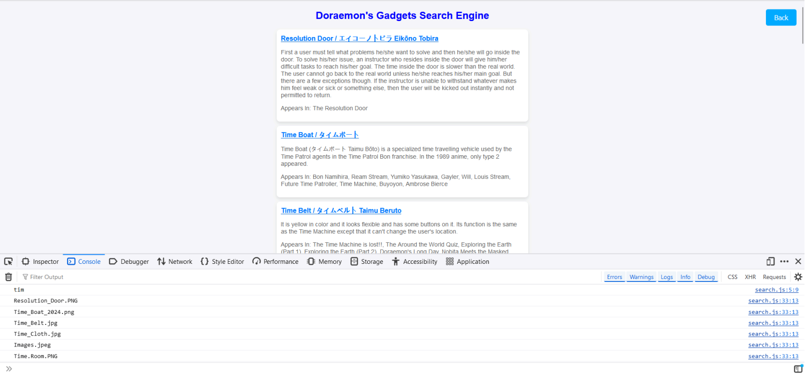
Appears In: The Time Machine is lost!!, The Around the World Quiz, Exploring the Earth (Part 1), Exploring the Earth (Part 2), Doraemon's Long Day, Nobita Meets the Masked Queen, Doraemon: Nobita's Great Adventure in the Antarctic Kachi Kochi, Time Machine

Time Cloth / タイムふろしき, Taimu Furoshiki

The Time Cloth is capable of manipulating time on whatever it covers in a backward or forward direction. The item does not need to be fully enclosed over its target to function and doesn't necessarily affect its attachments or certain other attributes. In one episode where Doraemon had hung the item over the drying rack, a stray breeze detaches it and the item ends up blinding Nobi Tamako. who tries to navigate around the house with her sight obstructed. Before Doraemon can notice, Tamako was reverted to her younger self (somewhat resembling Shizuka Minamoto), much to her surprise. Notably, even though her body has regressed, her other items did not, and she still retains her memories, resulting in her walking awkwardly in her oversized clothing. Time Cloth in Anime

Appears In: Chapter 25:Time Cloth, The Time Cloth, The Girl With the Red Shoes, Romance in the Snowy Mountain, Goodbye, Doraemon..., Magical Girl Shizu-chan, God Robot Extends His Hand of Love!, Nobita's Dinosaur, Nobita's Dinosaur 2006, Nobita and the

- Users click “inspect” in our search results page (an example of ranking).



Doraemon's Gadgets Search Engine

Resolution Door / エイコーノトビラ Eikōno Tobira

First a user must tell what problems he/she want to solve and then he/she will go inside the door. To solve his/her issue, an instructor who resides inside the door will give him/her difficult tasks to reach his/her goal. The time inside the door is slower than the real world. The user cannot go back to the real world unless he/she reaches his/her main goal. But there are a few exceptions though. If the instructor is unable to withstand whatever makes him feel weak or sick or something else, then the user will be kicked out instantly and not permitted to return.

Appears In: The Resolution Door

Time Boat / タイムボート

Time Boat (タイムボート Taimu Bōto) is a specialized time travelling vehicle used by the Time Patrol agents in the Time Patrol Bon franchise. In the 1989 anime, only type 2 appeared.

Appears In: Bon Namihira, Ream Stream, Yumiko Yasukawa, Gayler, Will, Louis Stream, Future Time Patroller, Time Machine, Buyoyon, Ambrose Bierce

Time Belt / タイムベルト Taimu Beruto

It is yellow in color and it looks flexible and has some buttons on it. Its function is the same as the Time Machine except that it can't change the user's location.

Appears In: The Time Machine is lost!!, The Around the World Quiz, Exploring the Earth (Part 1), Exploring the Earth (Part 2), Doraemon's Long Day, Nobita Meets the Masked

```
tim                                                          search.js:5:9
Resolution_Door.PNG                                          search.js:33:13
Time_Boat_2024.png                                           search.js:33:13
Time_Belt.jpg                                                search.js:33:13
Time_Cloth.jpg                                               search.js:33:13
Images.jpeg                                                  search.js:33:13
Time.Room.PNG                                                search.js:33:13
```

Discussion

Technical difficulties/ Challenges and How to handle these technical difficulties/ challenges

These are technical difficulties/ challenges that we face in this project:

- Data Scraping Issues
    - Unspecified HTML elements complicate web scraping since targeted classes or IDs are missing.
    - Data inconsistency, as Doraemon gadgets on the website have varying properties.
- Image URL Access Issues:
    - Images from the fandom site are restricted by "Content Security Policy" (CSP) and hotlinking blocks, preventing direct embedding.
- Elasticsearch Incompatibility:
    - Some team members cannot run Elasticsearch via Docker Compose for development.

How we handle these technical difficulties/ challenges that we face in this project:

- Data Scraping Issues:
    - Used advanced GO-Colly queries (e.g., ul > li > a) to target necessary elements.
    - Filtered out gadgets missing essential properties, reducing data from 700 to around 300 gadgets.
- Image URLs:
    - Switched to a different image source and re-scraped with the new links.
- Elasticsearch Mocking:
    - Created a mock API endpoint for testing, allowing frontend work without Elasticsearch.

Learned lessons

These are learned lessons which our group learns from this project:
Our group learns:

- What is “web scraping” and how to use “GoColly” to do web scraping properly.
- What are “Elasticsearch (Okapi BM25)” and “Kibana” and how to use “Elasticsearch (Okapi BM25)” and “Kibana” to index data properly.
- How to use “Node.js”, “Express.js”, “HTML”, “CSS”, and “Javascript” to create a search system interface properly.
- What are “1 word query”, “multiple word query”, “partial match query”, and “ranking”.
- How to work as a team effectively.
- How to solve problems which occur during this project effectively.

System limitations

These are our search system limitations:

- In our search results page, users can’t see some gadgets. This happens because our source website doesn’t have any information about some gadgets, so our group decided not to put these gadgets into our search system.
- In our search results page, users can’t see gadget’s data completely (all gadgets don’t have some data). This happens because our source website doesn’t have some information about some gadgets.
- In our search results page, users can’t see gadget’s pictures. This happens because our source website doesn’t allow others to collect gadget’s picture.

System limitations: future improvements

About future improvements of our search system limitations, our group will use different source websites to make sure that users will see all gadgets (each gadget will have its complete data and its picture) in our search results page.

Conclusion

In conclusion, this project successfully demonstrates the development of a prototype information retrieval system tailored to providing precise information on Doraemon gadgets. By employing web scraping with GoColly and indexing through Elasticsearch, the system offers an effective solution for Doraemon fans and toy designers seeking detailed gadget information. Despite challenges with data inconsistencies and image accessibility, our team implemented workarounds and learned valuable lessons in web scraping, database indexing, and frontend development. While current limitations remain, such as incomplete gadget data, future improvements, including alternative data sources and enhanced scraping tools, can further refine the system's capabilities and user experience.