Intermediate Presentation

<Team 4>

12146323 Jeong Sang-Heon

14146320 Lee Jun-Ha

14146326 Hong Jun-Ki

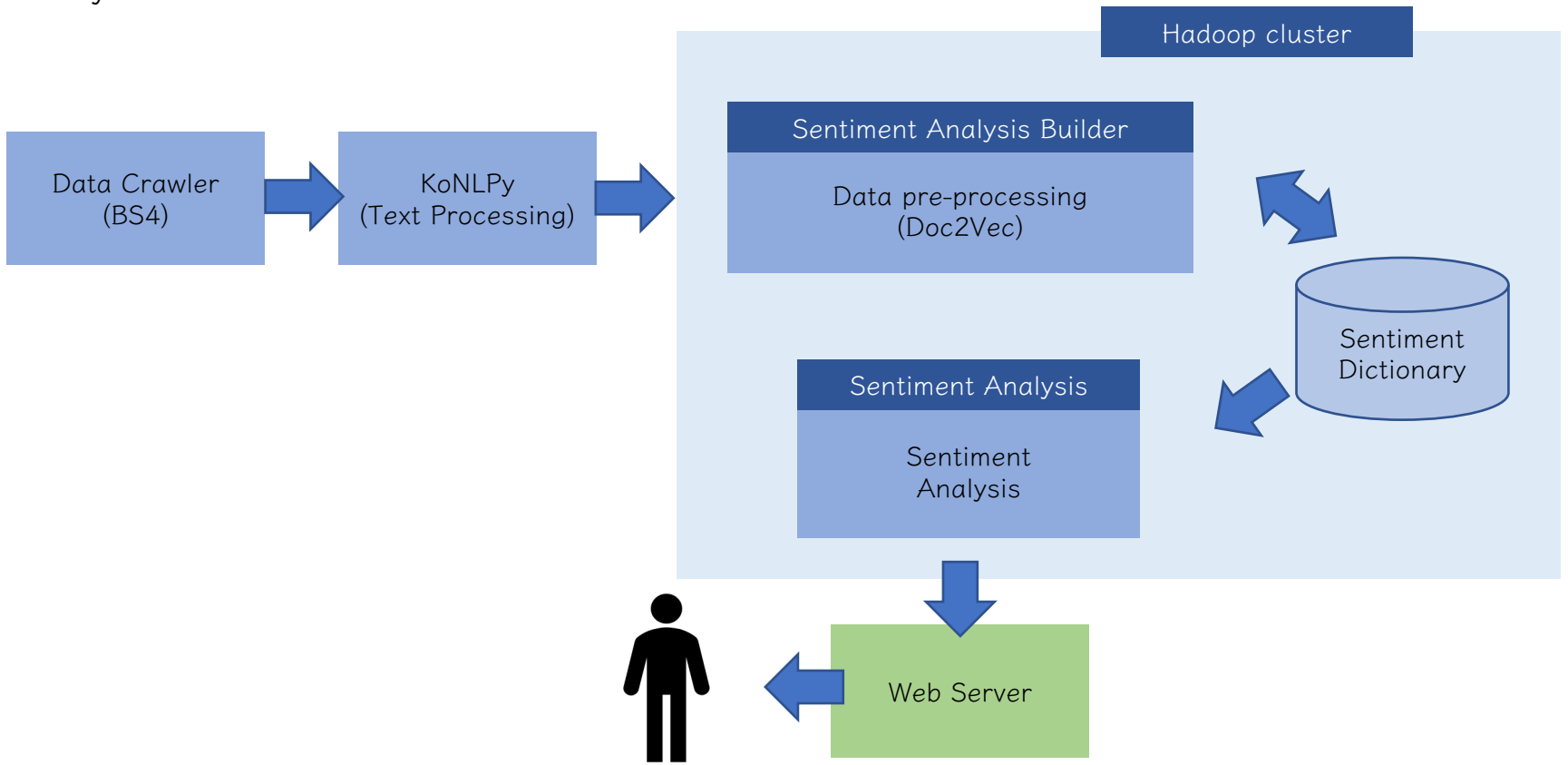
16102278 Sung Chang-Keu

Contents

1. Designing the Projects
2. Identifying important Issues
3. Result of Prototyping
4. Updated part of the SRS report and Planning

1. Designing the projects

1.1 System architecture



BS4

Python library used for parsing HTML and XML

KoNLPy

Korean morphological analysis package

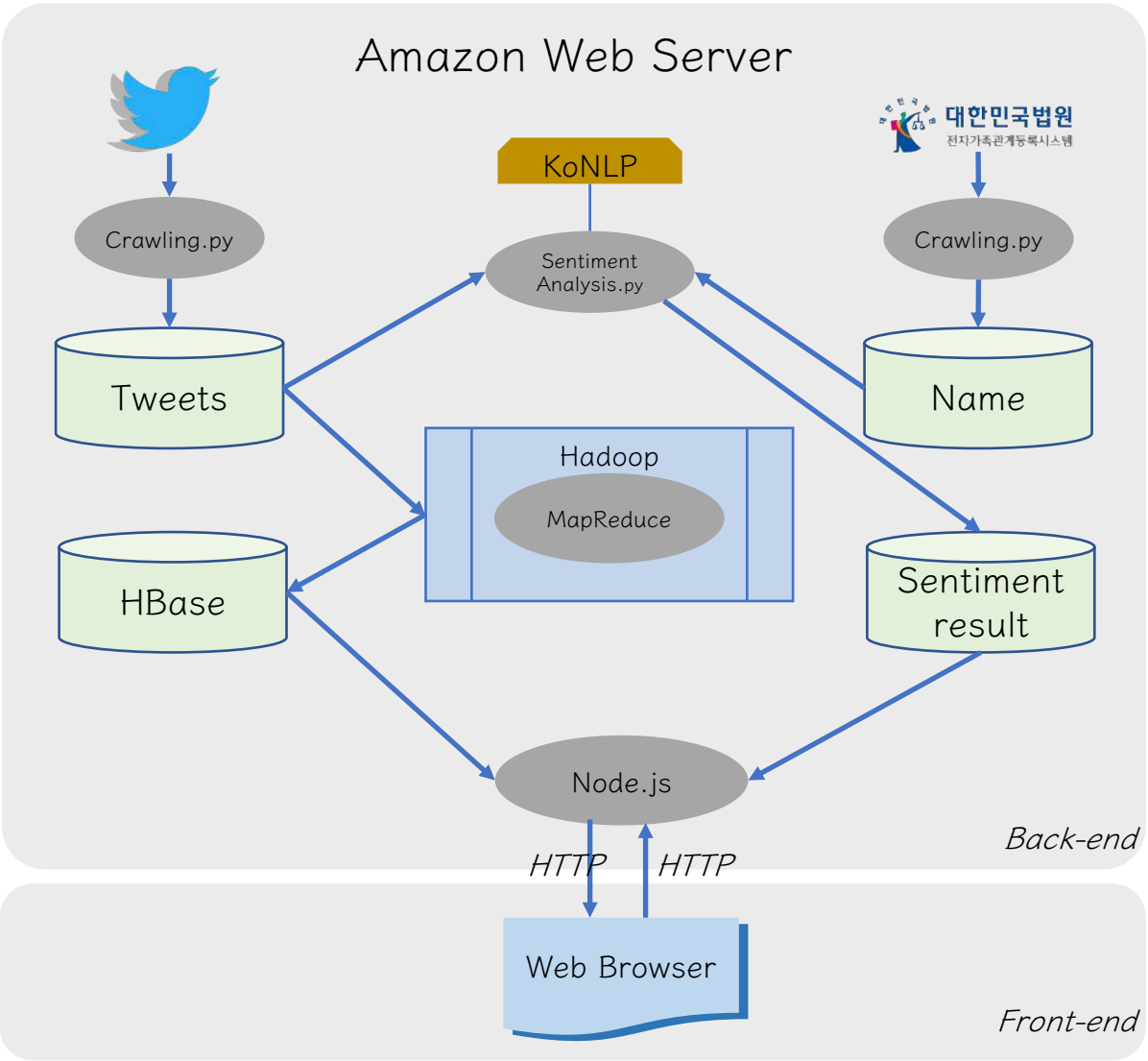
1. Data collection – collecting data using the python package (Data Crawler - BS4)

2. Data analysis – text processing (morphological analysis) and sentiment analysis using 'KoNLPy' & Doc2Vec

3. Visualization – via Webserver (TCP/IP)

1. Designing the projects

1.2 System flow



Bring the text on Twitter via **python code**

- ✓ Search for text about specific words requested by USER and import id, username, text, date, retweets, favorites, mentions, hashtags, and geo information.
- ✓ What to implement: Python Crawling Code
- ✓ Useful packages: tweepy, BS4

Sentiment analysis

analyzer.js	→ make tweet's score (pos/non-pos/moderate)
morphemeServer.py	→ do the morphological analysis with json file
plot.R	→ make a graphical result (easy to show result)

Implement web server

- ✓ What to implement: server hosting(AWS), implement web page
- ✓ Web page: HTML(home.html), NodeJS(server side platform, result.js)

1. Designing the projects

1.3 System requirement

Functional Requirement	**System**	**Sentiment Database**	Select the appropriate information from the DB and print it to the user
	User	**Web Page**	Users search through web pages to get the information they want
Non-functional Requirement	**System**	**Language**	System language type
		Accessibility	The users can use service without difficulty and don't need guide to implement
		Update Cycle	Update system's modification
		Data-backup Cycle	Backup the system data
		Information Accuracy	Whether the system can provide users with validated data
		Cope with Errors	How to deal with errors
		Data Access	Correct access for the data
	User	**Search the Preferences**	Users can check the preferences for smartphone

Key function checks

- ✓ Our system must store the results from the crawler and R software in the database for 24 hours – data access
- ✓ In prototyping, we use an Apache web server to ensure that data is generated well – language, cope with errors, data access
- ✓ In addition, instead of storing results in a database, test the results by storing them in a specific directory – data access
- ✓ We should ensure that user requests are delivered in Get format and that the resulting graphs are printed on the web page – web page

2. Identifying important Issues

2.1 Crawling Issues

2.1.1 Crawling period

- How much do we need to separate crawled data

2.1.2 Duplicated tweets

- Retweeted tweets should be a fatal noise of the dataset.

2.1.3 Search Function

- Certain names are not crawled from Tweet crawler.

2.2 Sentiment Analyzing Issues

2.2.1 Result of analysis

- Whether sentiment analysis is possible using tweet data

2.2.2 Problems in module

- Difficult to set the essential module to analysis

2. Identifying important Issues

2.1 Crawling Issues

2.1.1 Crawling period : How much do we need to separate crawled data.

Our prototype deployed tweepy API to get the very recent tweets from Twitter.

http://docs.tweepy.org/en/v3.5.0/getting_started.html

Although tweepy supports fast lookup of keyword result, if we encounter the situation to investigate **historical data**, it doesn't work

"Have to clarify the user's requirement. If user wants to know the result of recent social buzzes, tweepy will be a good module."

2.1.2 Duplicated tweets : Retweeted tweets should be a fatal noise of the dataset.

```
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
RT @gemini_0613_: [1박2일 김준호×김종민×이용진 시키면 한다 OK tv 1부] 중 태형이 문자 cut
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
RT @gemini_0613_: [1박2일 김준호×김종민×이용진 시키면 한다 OK tv 1부] 중 태형이 문자 cut
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
'연애의 맛' 김종민 황미나, 100일 연애 계약 종로 앞두고 또 위기?
 https://t.co/4NmLAgkSfZ
RT @gemini_0613_: [1박2일 김준호×김종민×이용진 시키면 한다 OK tv 1부] 중 태형이 문자 cut
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
RT @gemini_0613_: [1박2일 김준호×김종민×이용진 시키면 한다 OK tv 1부] 중 태형이 문자 cut
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
RT @gemini_0613_: [1박2일 김준호×김종민×이용진 시키면 한다 OK tv 1부] 중 태형이 문자 cut
#BTS @BTS_twt #방탄소년단 https://t.co/dQn34ebpOw
황미나♥김종민 공개연애 상처줄까봐… 충격적인 사실이 드러났다.
```

```
for tweet in tweepy.Cursor(api.search, q=keyword, since='2018-1
    if (not tweet.retweeted) and ('RT @' not in tweet.text):
        StatusObject = tweet._json
        dict1 = {
                'id': StatusObject['id_str'],
```

(**However, retweeted tweets can be the weighted index factor to support the people's feelings on the target name.)

Reason to choose Twitter channel.

- Twitter
  - ✓ It is a short-sentence form for a particular subject.
  - ✓ Discussions are expanded and reproduced in the form of **retweeting by followers** in the articles of a small number of **influencers** that post their opinions.
  - ✓ If more retweets are generated in a short period of time for a particular negative issue, they are perceived to be stronger than other channels.
- Blog
  - ✓ Post-type articles on a daily topic.
  - ✓ A channel that distributes **lifestyle information** related to everyday life rather than emotional arguments.
- Community
  - ✓ Bulletins discussing daily life and concerns and articles in question and answer format
  - ✓ A channel that exchanges **specific information** in clusters.
- Media
  - ✓ Articles containing daily newspapers, economic journals, and online media.
  - ✓ A channel in which information that **relies on the agency or press released** is posted.

"Twitter is the appropriate channel to get the sentiment text database for a particular subject!"

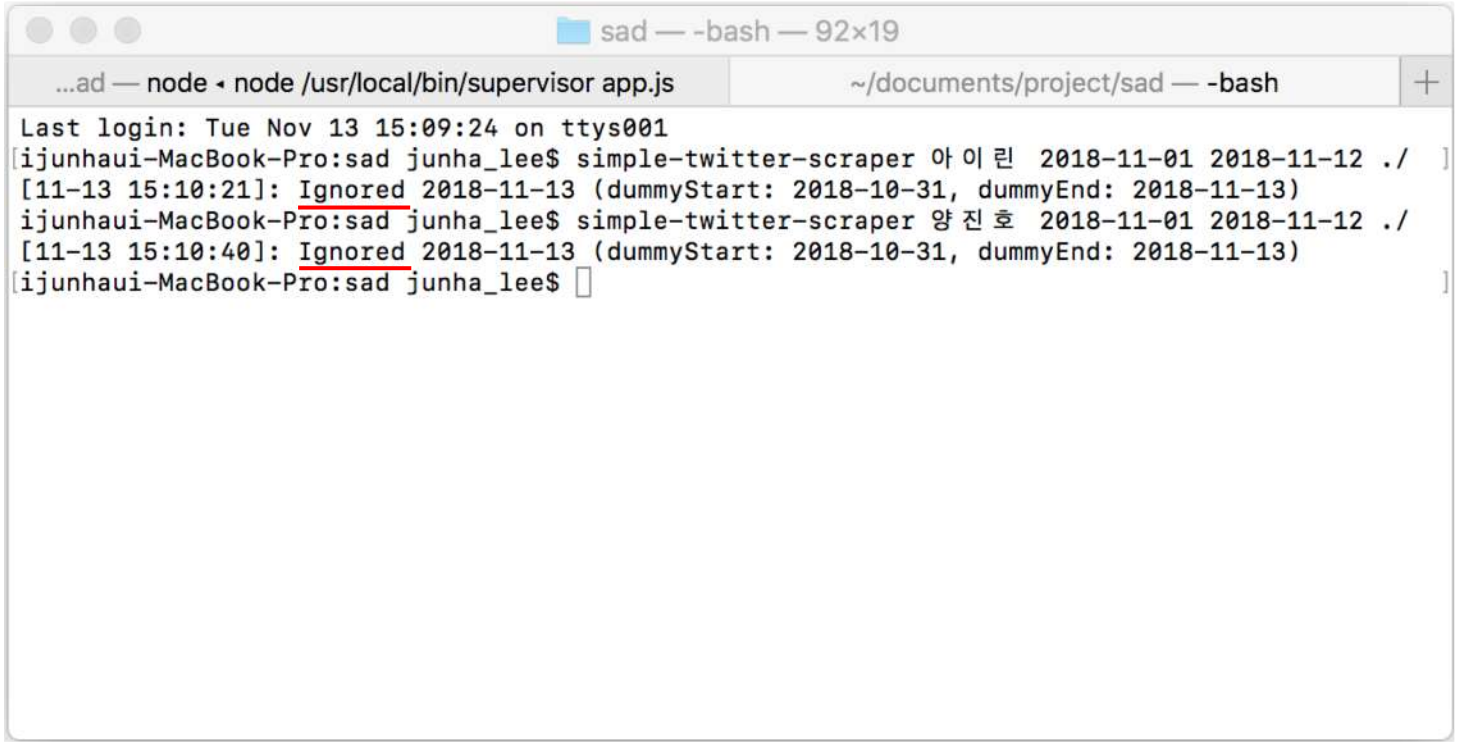
2. Identifying important Issues

2.1 Crawling Issues

2.1.3 Search Function : Certain names are not crawled from Tweet crawler.

We firstly collected name database by using Twitter Scraper package. The module we deployed did not fully worked.

When we search for specific name, it sometimes did not show up the crawled results.



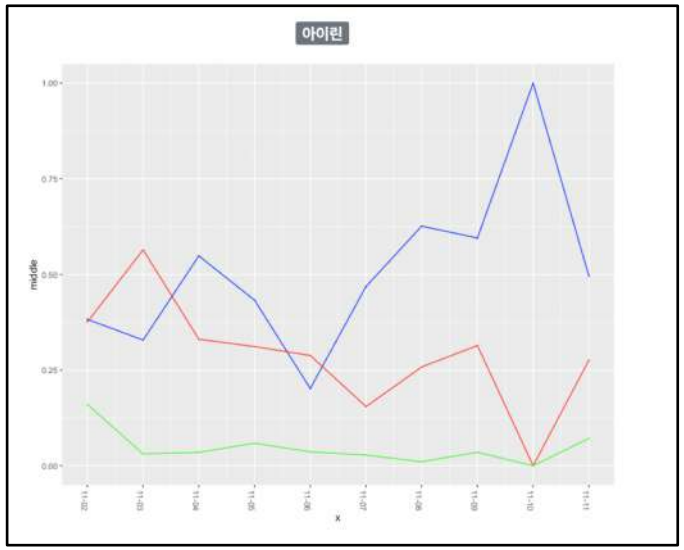
Crawler ignores the input name

2. Identifying important Issues

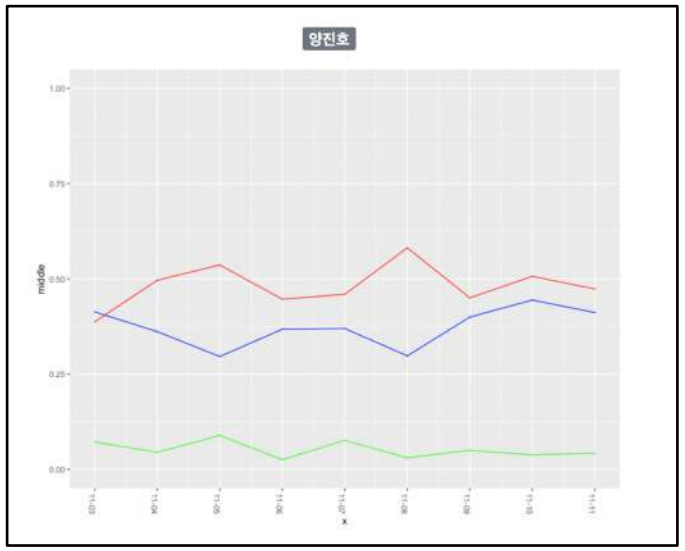
2.2 Sentiment Analyzing Issues

2.3.1 Result of analysis : Whether sentiment analysis is possible using tweet data

It was necessary to check whether the result of the sentimental analysis on the name could be evident as a positive or negative result.



Positive



Negative

2.3.2 Problems in module : Difficult to set the essential module to analysis

The Korean sentiment analyzer that we have built on one's OSX environment did not work on other's computer.

We will set up Linux based AWS(Amazon Web Service) EC2, and install the node.js and npm to successfully install KSA(Korean Sentiment Analyzer) module.

2. Identifying important Issues

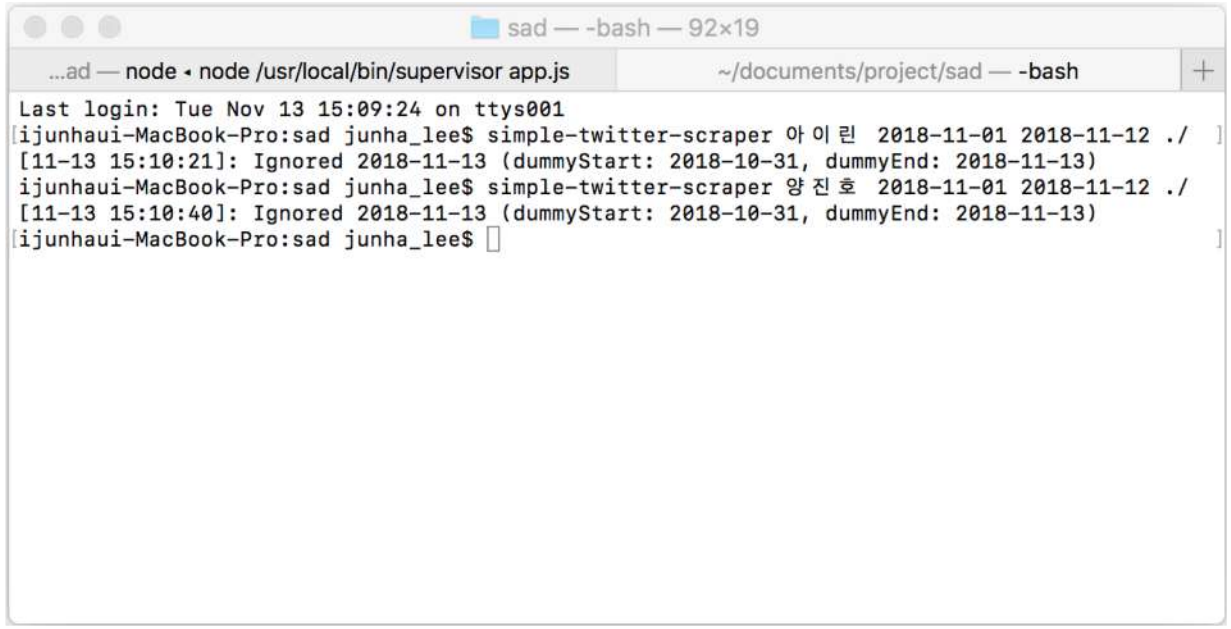
Summary.

Issues	as-is	to-be	solved
Crawling period	Tweepy package supports the very recent period of data.	Clarify user requirement.	
Duplicated tweets	Retweeted tweets make noise.	Remove “RT:” – contained tweets.	✓
Search Function	Twitter scraper does not react to the certain name.	Develop another crawler instead.	✓
Result of Analysis	Could sentimental analysis make the identifiable positive or negative result.	Testing for the analysis to use real name that people think positive and negative.	✓
Problems in module	Node.js based sentiment analyzer is not installed on every OS environment.	AWS Linux EC2 successfully handle npm.	

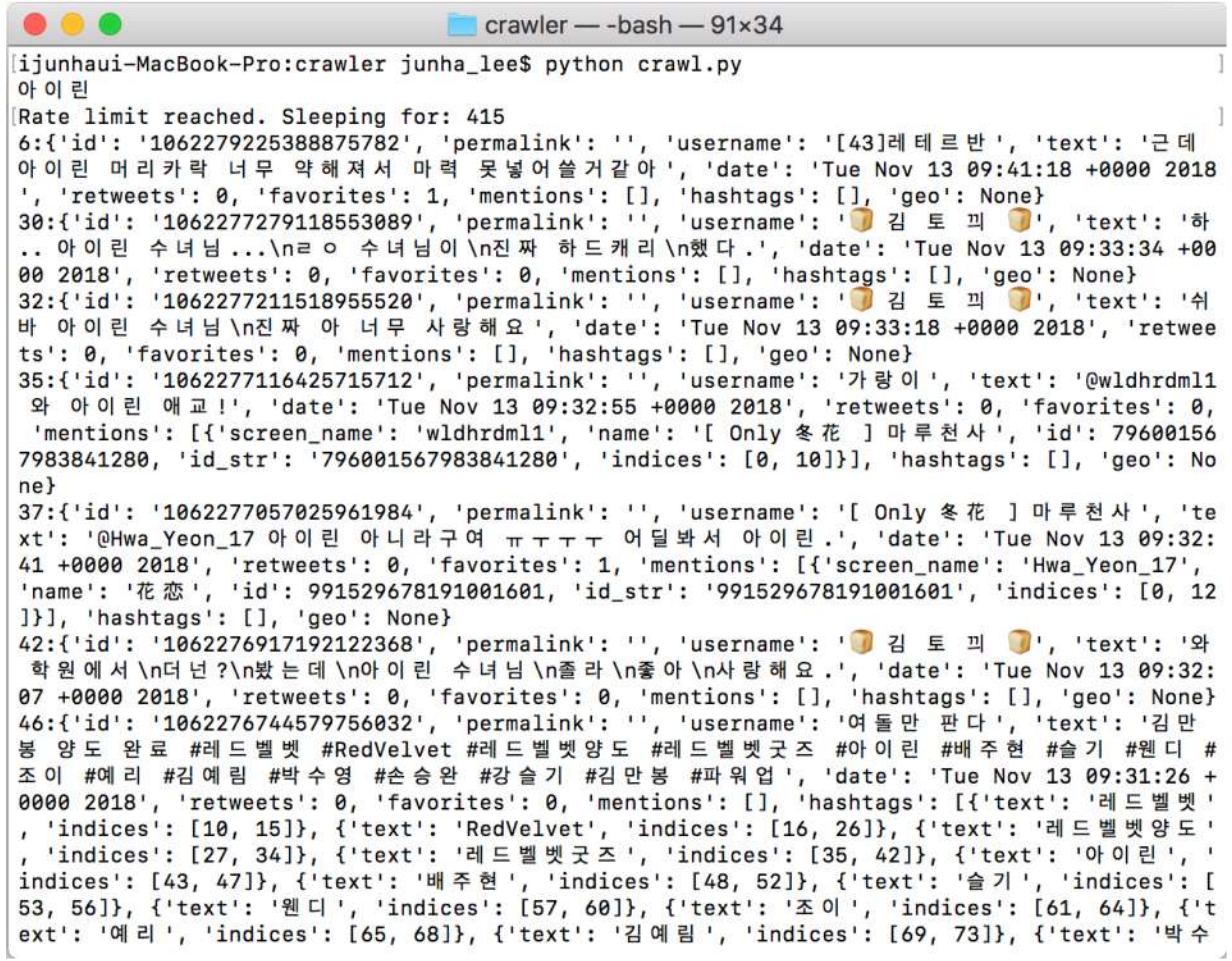
3. Result of Prototyping

3.1 Crawling

BEFORE

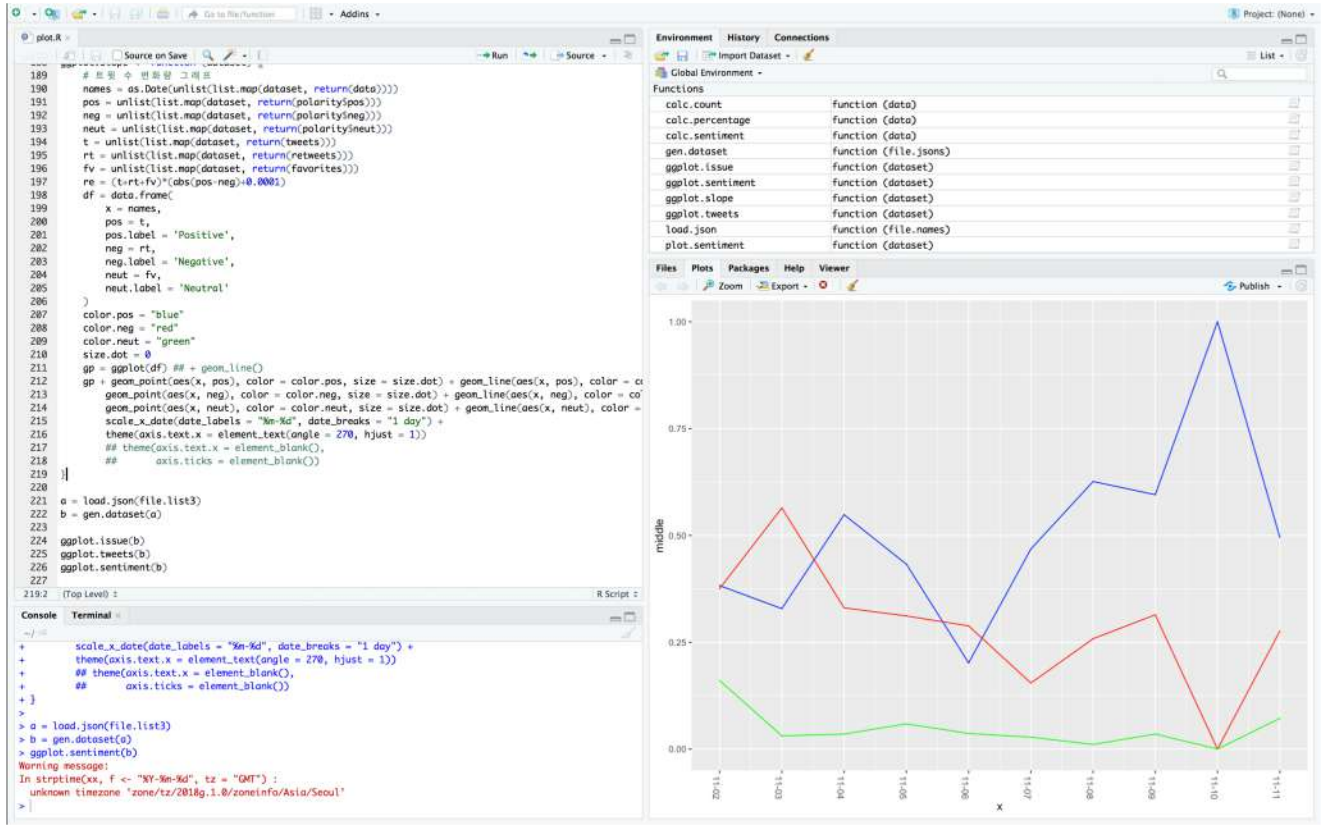


AFTER



3. Result of Prototyping

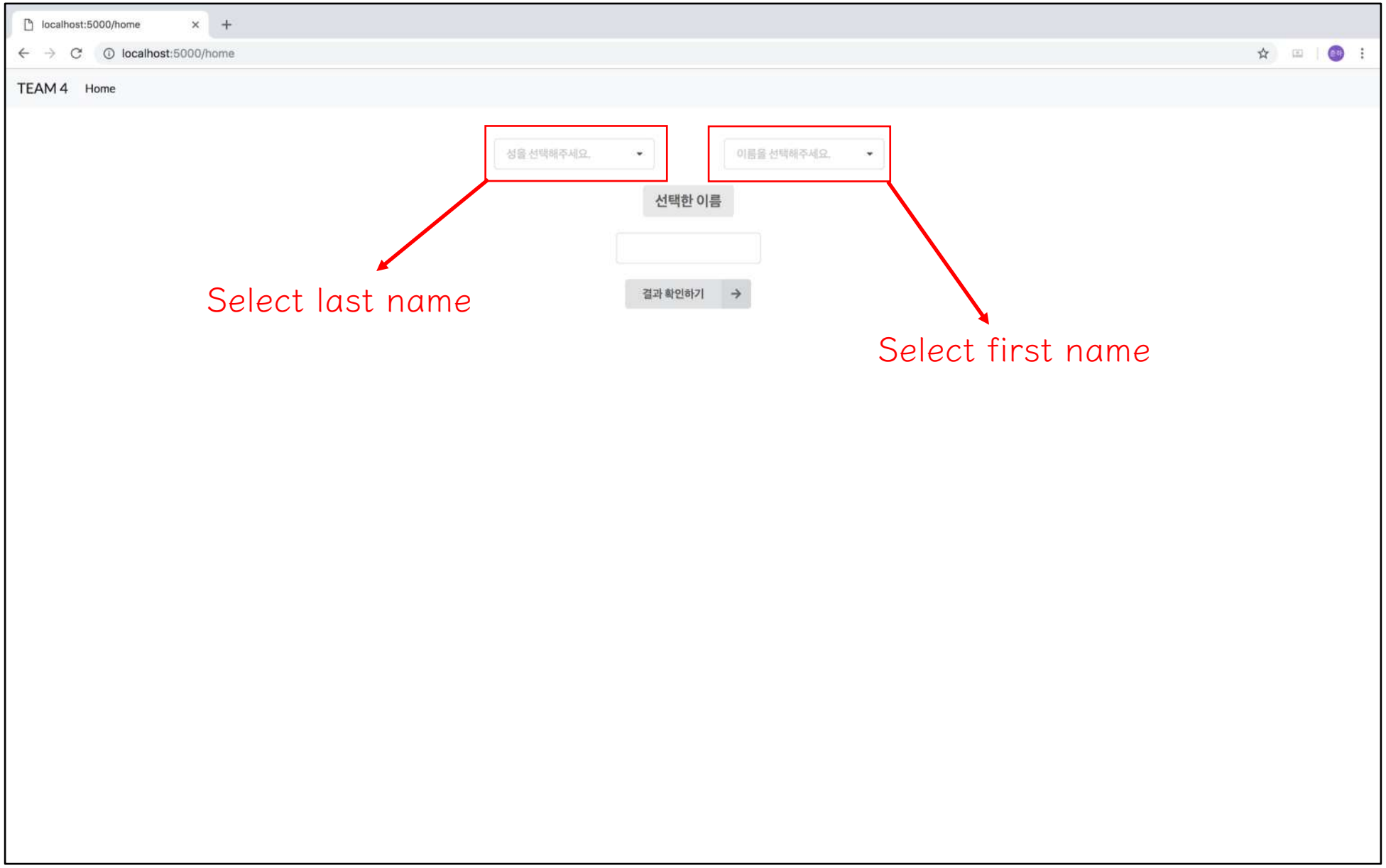
3.2 Sentiment Analysis



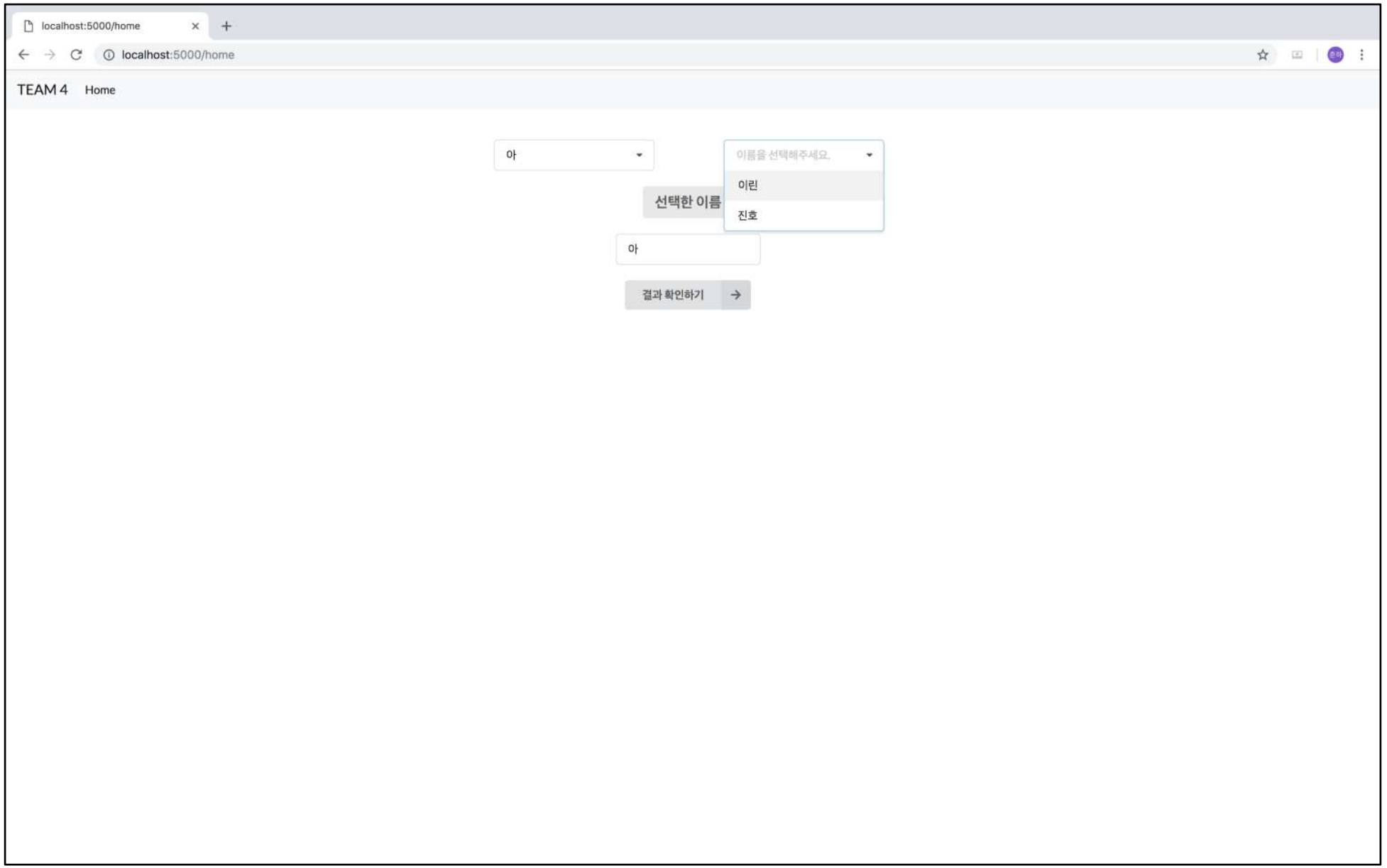
Result of 아이린

As we checked the result, so we can have confidence that this analysis is accurate.

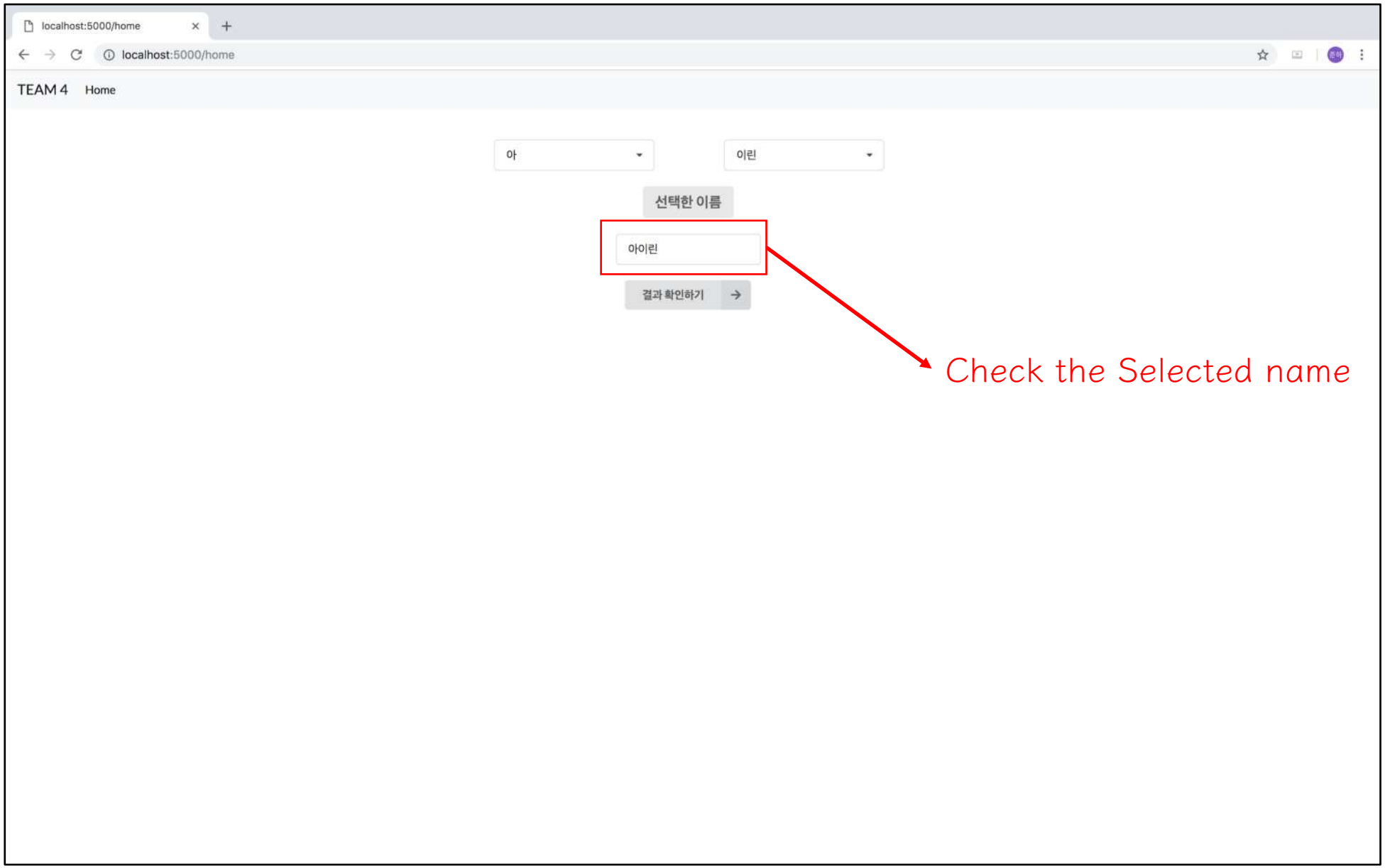
3. Result of Prototyping



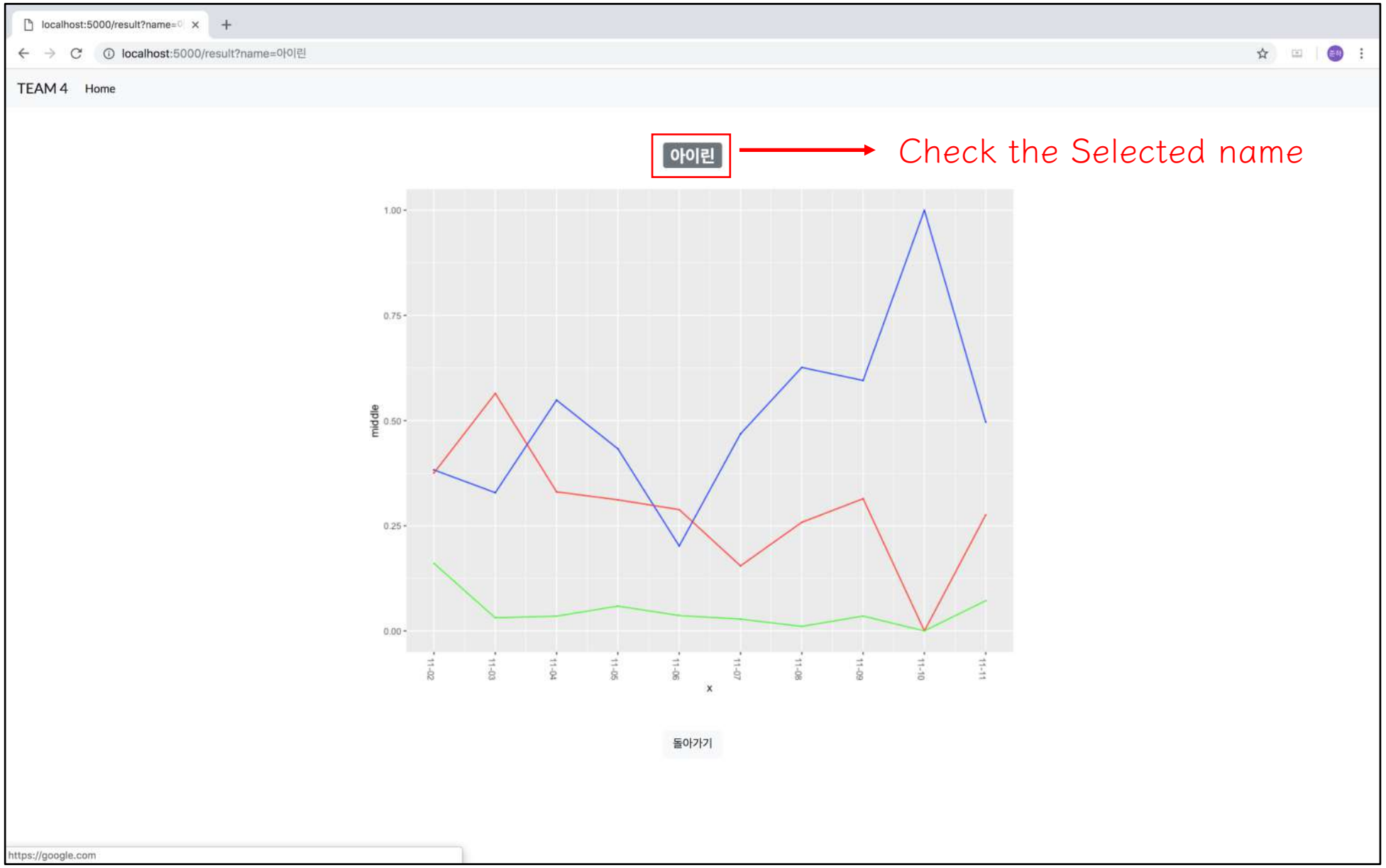
3. Result of Prototyping



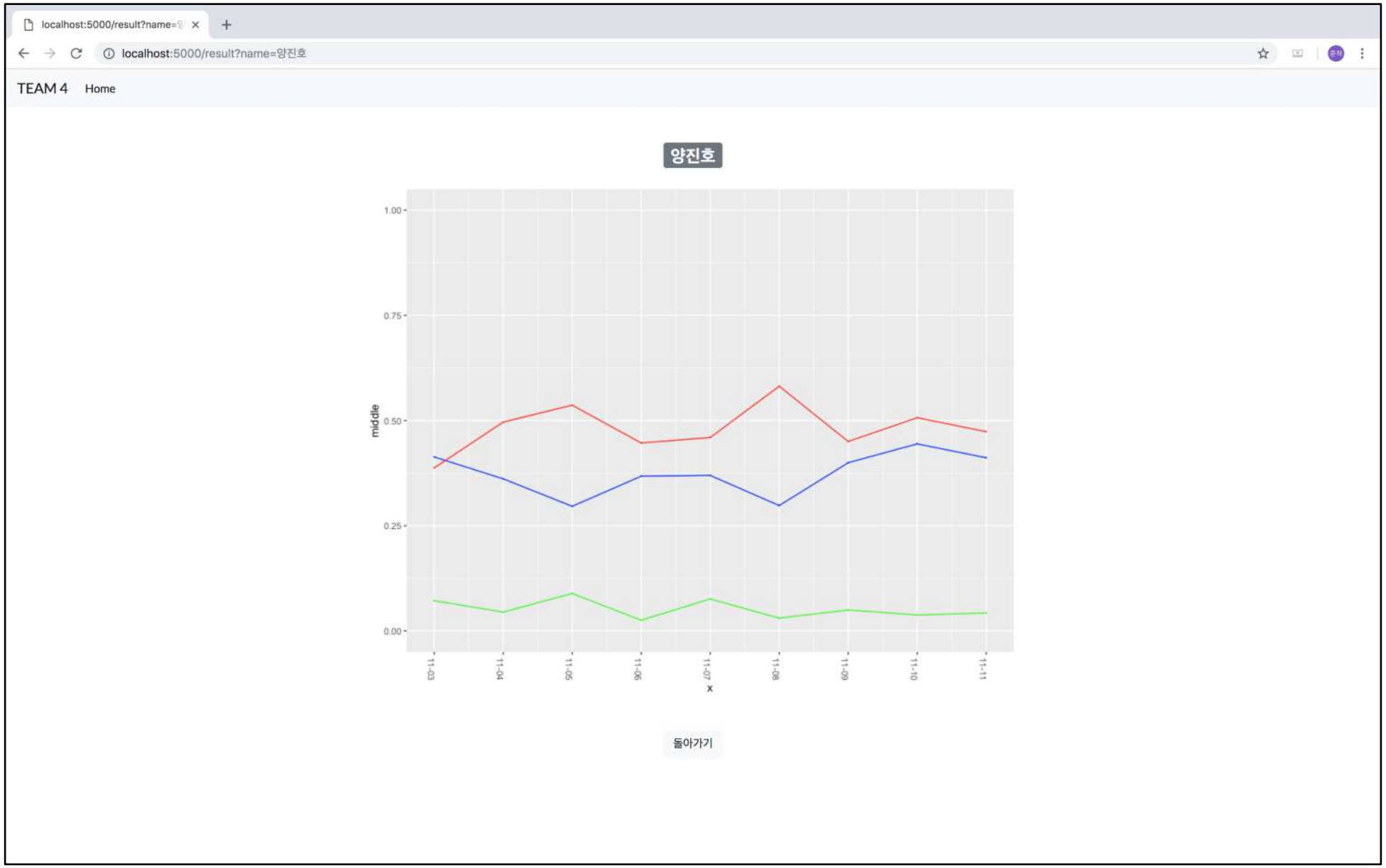
3. Result of Prototyping



3. Result of Prototyping

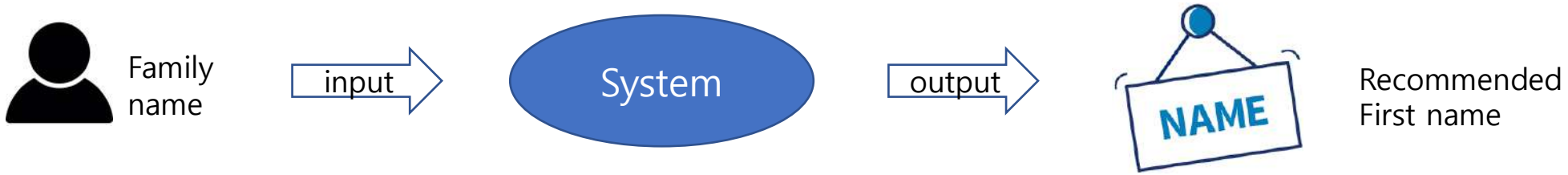


3. Result of Prototyping



4. The updated part of the SRS report and planning

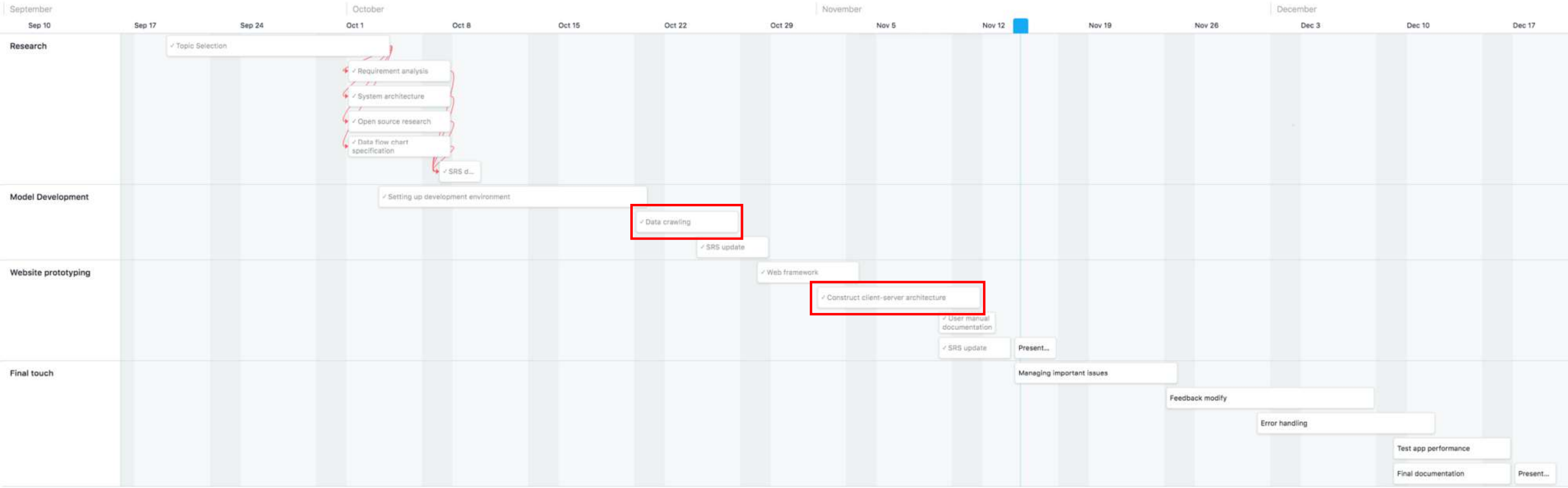
- The SRS report is updated entirely according to changes in topic
  - The overall systems and requirements are almost same as before
  - Small changes from modification of topic are reflected
  - New function for recommending First name to selected Family name will be provided at final version



- The ultimate goal of the project is to provide the preference information for most of keywords
  - Cannot sure the feasibility
  - However, the utilization of the system would be hugely improved
  - The ultimate version of this system

4. The updated part of the SRS report and planning

- The initial plan



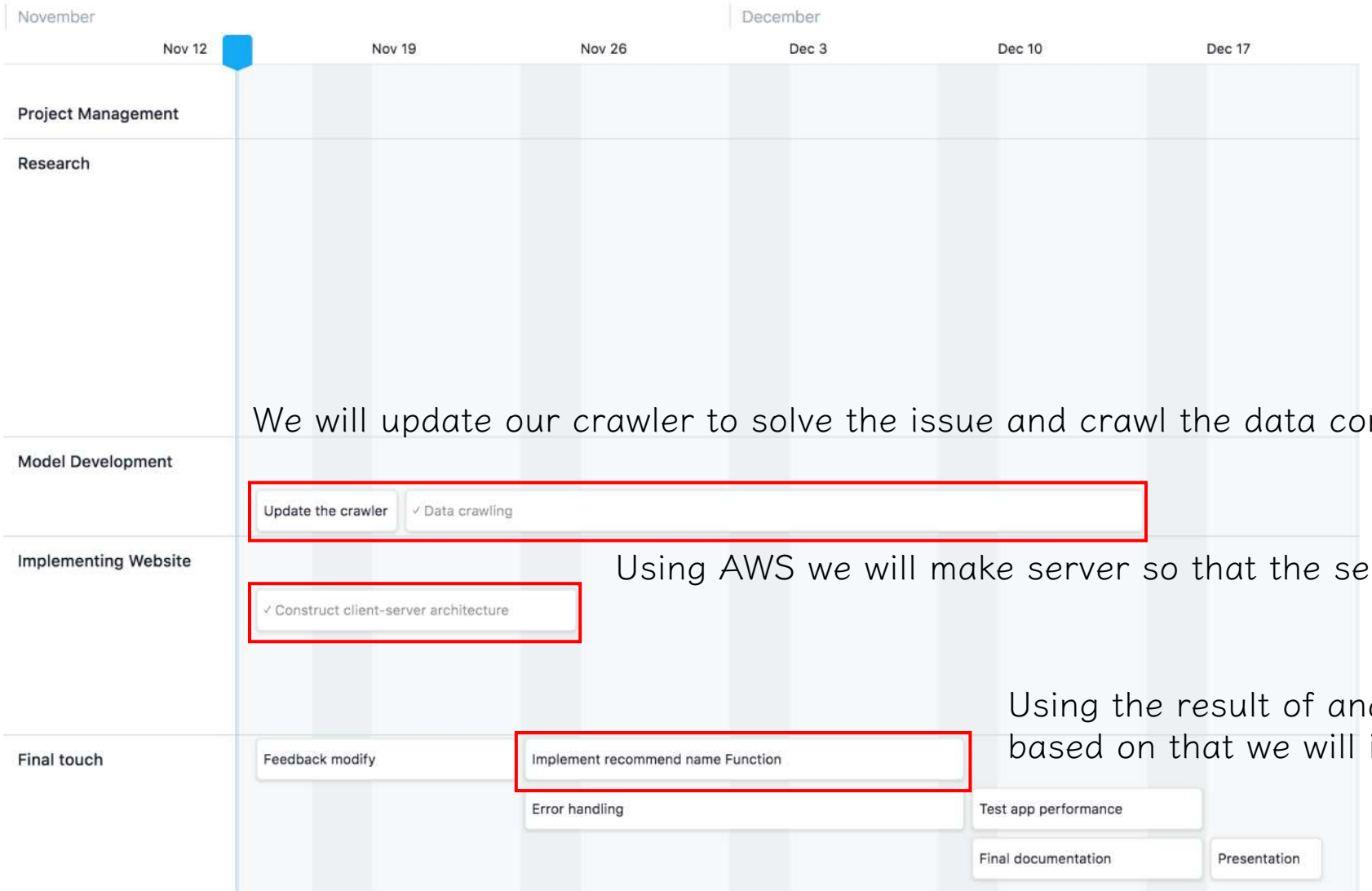
Unsolved issues : 1. Crawling Period 2. Problems in Module

New Function : Recommend name that people prefer

So we modify 'Data Crawling', 'Construct client-server architecture' tasks and add 'implement recommend'.

4. The updated part of the SRS report and planning

- The updated plan



We will update our crawler to solve the issue and crawl the data continuously.

Using AWS we will make server so that the server process the analysis task.

Using the result of analysis, we will make the ranking and based on that we will implement the recommend function.


Thank you