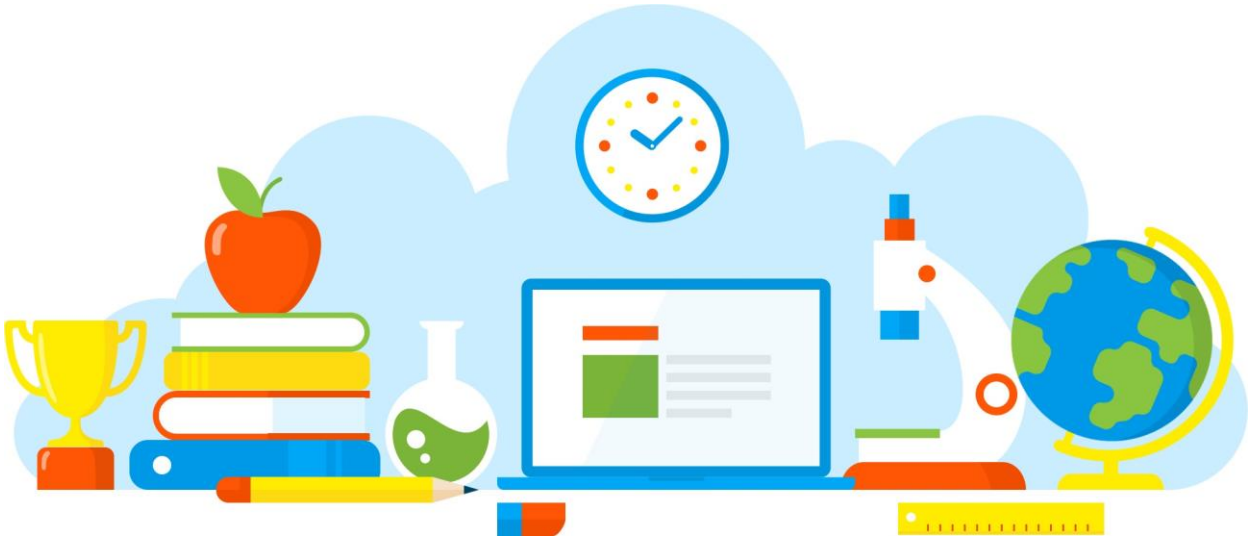


Intelligent News Express (CA1 report)

Institute of Systems Science, National University of Singapore, Singapore 119620

Team members: Sun Hang, Xu Dongbin

28/04/2020

Introduction

As information technology rapidly developed during the past 2-3 decades, the global datasphere continuously increased. It is expected the annual volume of global data will reach as high as 175ZB in 2025 as shown in Fig. 1. It's the era of information explosion. With so much data surrounded, how to get the most valuable information in a fast and effective way is a challenge. The concept of big data is proposed to deal with the mass data produced everyday.

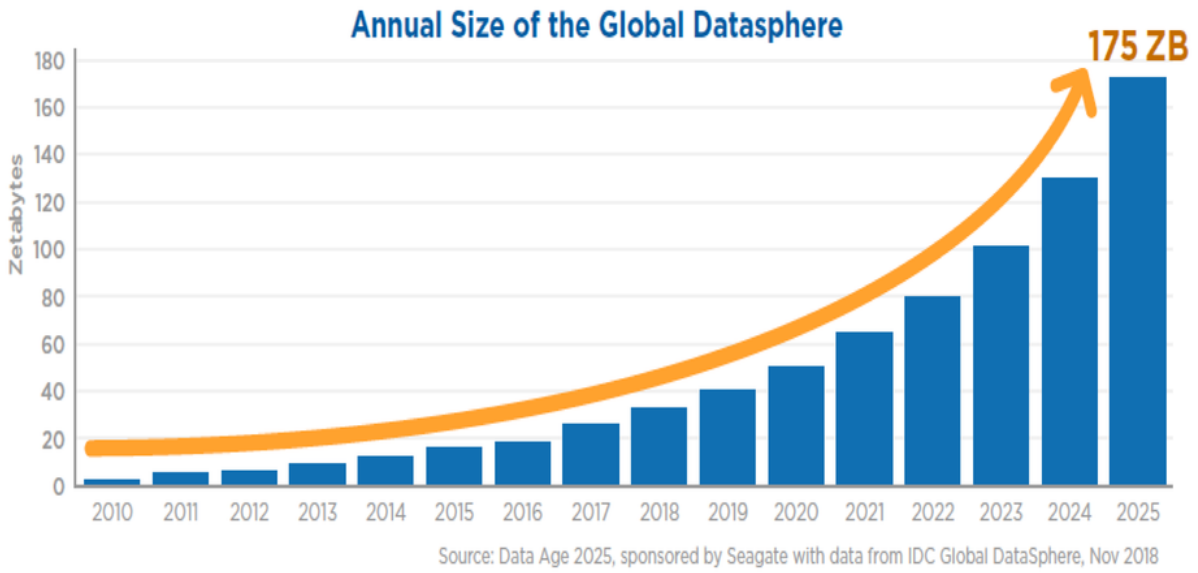


Fig. 1 Annual size of the global datasphere

Big data

The concept of big data was proposed to address and provide the solution to the information explosion. The big data involves data generated from different sources and with a lot of different characteristics. Basically, the three elements of big data are volume, variety and velocity as shown in Fig. 2. Besides volume, both of the variety and velocity are the essential components for big data. In another word, how to handle the real time or near real time data such as news is also a challenge in big data. Along with the big volume of real time or near real time data generated everyday, the more and more people feel anxiety and perplexity because they feel lost in the sea of information.

The basic idea here is to use an intelligent system or software agent to help to read the real time and big volume data and extract the most valuable information from it. It will save a lot of time for humans who thus can focus on the productive work. It is almost impossible for humans to read all of the news articles every day. With the help of intelligent systems or software agents, however, people can grab the important points or ideas as fast as possible. Eventually, human capabilities will be widely expanded.

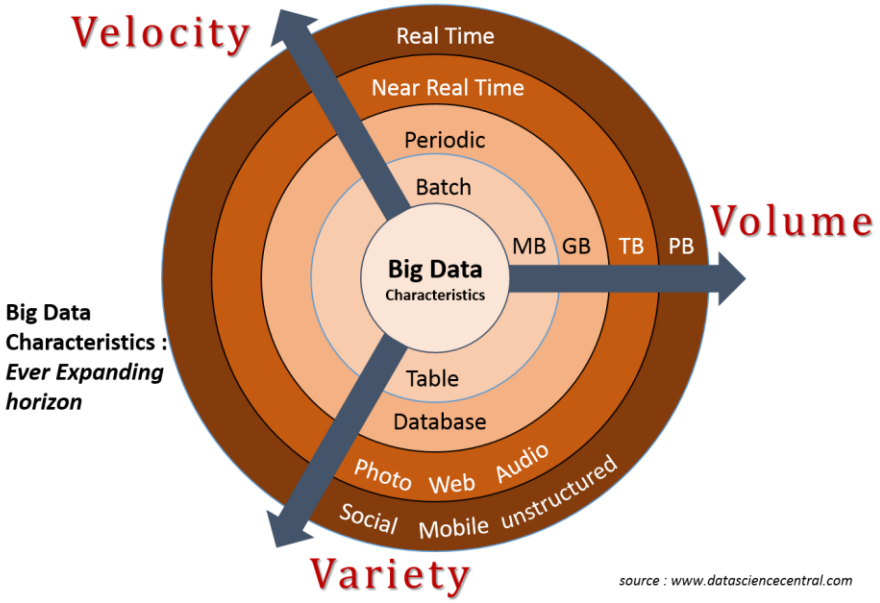


Fig.2 Three elements of big data

Hours spend in reading news

A survey done by the global web index as shown in Fig. 3 indicates that people spend 1.35 hours every day reading news. More specifically, people spend more time reading online news than news from print papers. It occupies almost 17% of the effective working hours by assuming 8 hours of effective working hours per person per day.

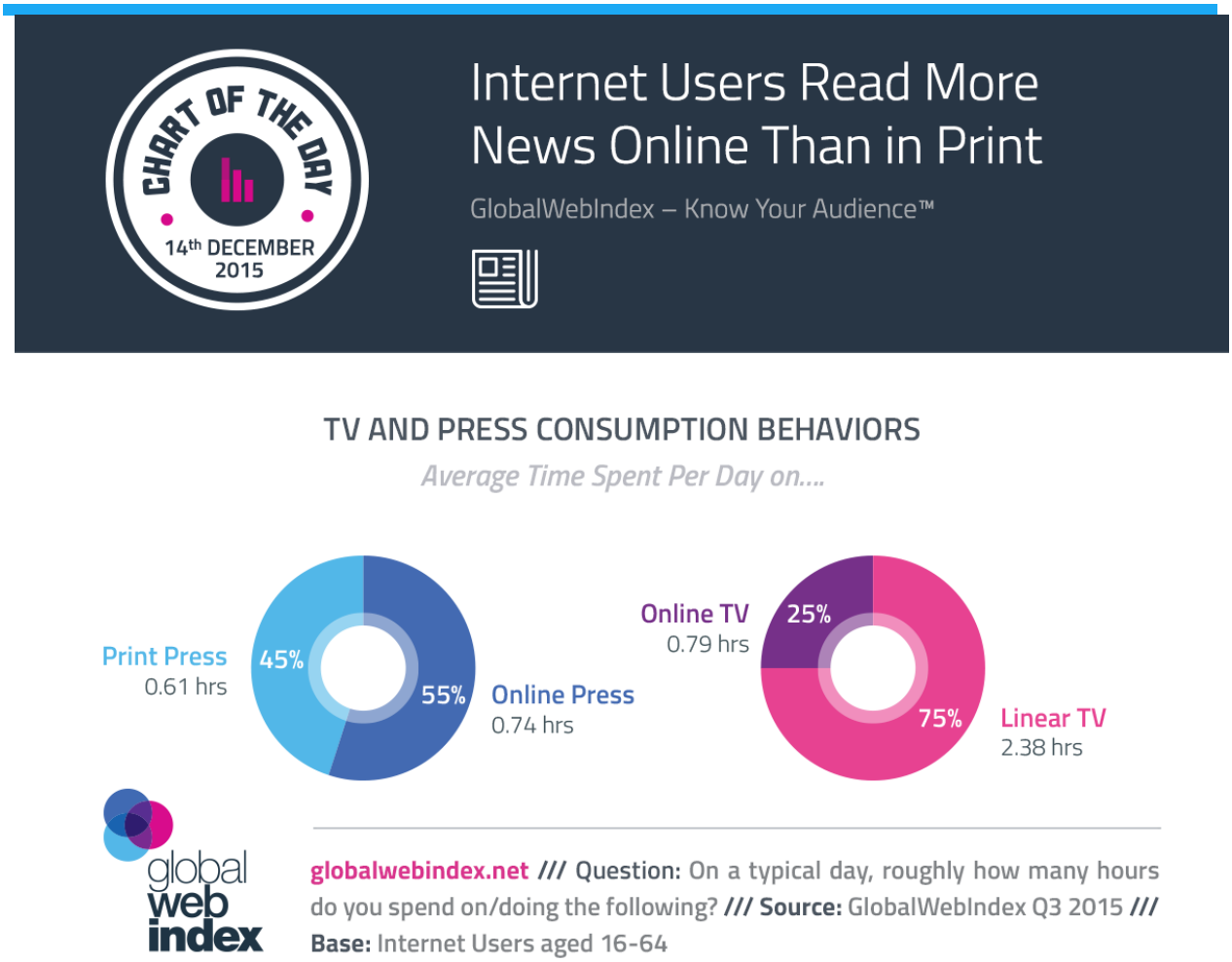


Fig. 3. A survey data about hours people spend in reading news and watch TV everyday

The news reading process is time consuming and most people have no choice but only read selective news articles. However, this will make people tend to form the biased impression of the information and cannot see the full picture of the information. Eventually, based on the biased information, people may probably make the wrong decision. Sometimes it could be an important decision to an event, to a people, or even to an organization/company.

With the help of an intelligent system or software agent, it will analyze all the information for its importance and summarize the key points to the customers. It will not only save the time for people, but also help people understand the thing in a comprehensive and distinctive way, which will provide the basis for people to make the right decision.

In the current work, we will use the strength of an intelligent system or software agent to assist in the news reading. The objective is to let customers acquire/understand key points of the information effectively and comprehensively in 10 minutes everyday.

Business Case / Market Research/Product Plan

In this section, the business case, market research as well as product plan will be discussed and projected.

Business case

Problem:

Due to information explosion nowadays, people need to read more and more news to get a brief and comprehensive understanding of the current trend and the most controversial happenings in the world. Based on a survey from the global web index as shown in Fig. 3, which shows that people spend 1.35 hours every day reading news on average. More than half of them read news online.

Solution:

Our proposed solution is to provide a one-stop site for the customers to read comprehensive news intelligent summary content in short time (less than 10 minutes per day). The system will use an intelligent process or software agent to assist in news acquiring and summarization. In the product plan, we will support customized options for different tastes of the customers in the future.

Market research

So far we didn't find any similar product providing the same service. The close services providers which may be the potential competitors are news subscription or RSS. However, for either case, people may still need to subscribe to the news from websites one by one, or the pushed news threads are not well organized and without intelligent analysis to distinct the importance.

In our product, we will provide the news from different sources and use the intelligent language processing to summarize the content to a short summary. The news are selected based on the popularity and the sequence is arranged with the highest popularity on the top. In this way, customers can read the summary for each news to understand what's happening and know which news is more popular. As a result, the customers can read the news in a much more effective way.

Product plan

The product will be provided as a subscription service. The subscription fee for this release 1 product is $5/month. In the advanced version, more news resources and more coverage of areas and countries will be conducted. The subscription fee for this release 2 product is $7/month. In the future, we will add more customized options for the different tastes of the customers such as sport news, stock news, health news, technology news, entertainment news, Science and Nature news, etc. The subscription fee will increase accordingly to $10/month for pro service after release 2.

System Design / Model

System design

Near real time data acquisition process

The design of our system involves the web scraping with the IPA tool (TagUI), local AI with the natural language processing, as well as the front-end back-end web service and website design. As shown in Fig. 4, the TagUI (RPA) tool from the server directly communicates with the news website to fetch news data with web scraping. During this process, the intelligent robot process is designed to fetch the top news with highest popularity. After that, local AI is working on the server to do the natural language processing to summarize the news content. The news URL, ranking no. with popularity, news title, fetching time, summary content, etc. All will be saved into the database with SQLite.

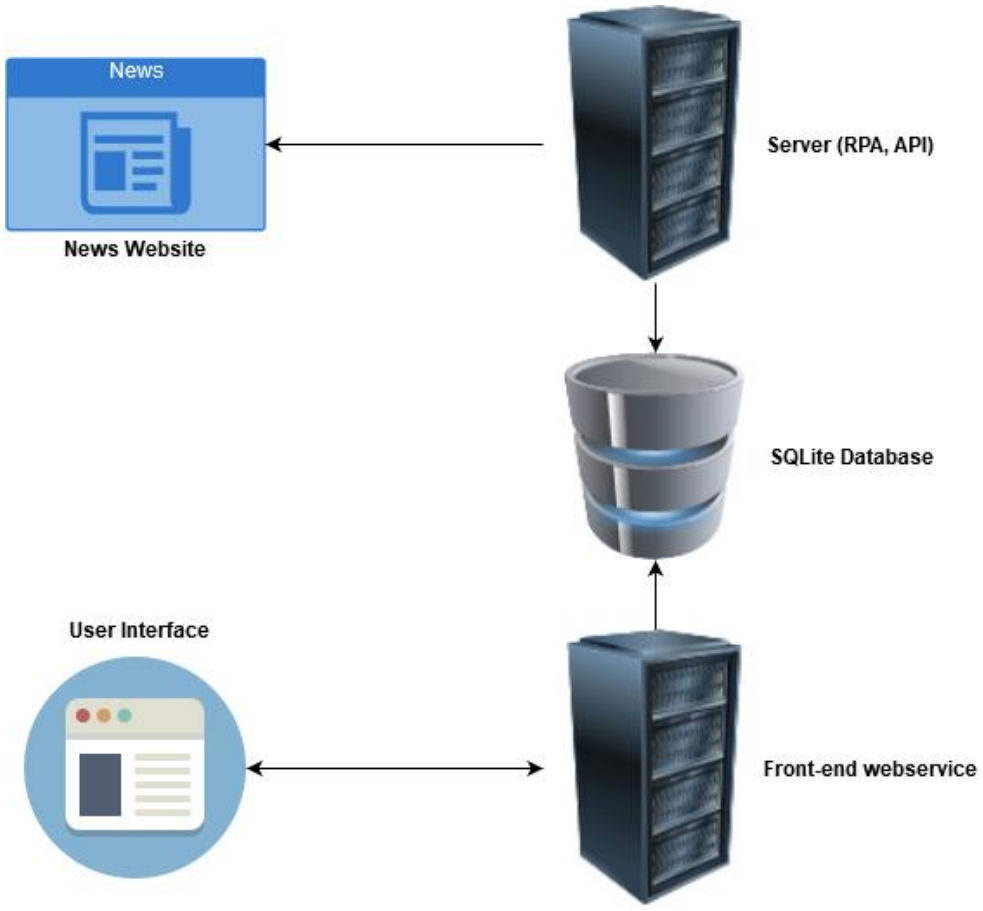


Fig. 4. Architecture of the system design

The time interval of the data acquisition process is designed to be 1 hour in frequency. There are two reasons, the first is typically the news website doesn't update as frequently as social media, the second is because the objective of this product is to help customers to improve the efficiency of acquisition of knowledge from news and articles, it has no reason to promote the customers to check the update so frequently.

Web Service design and process

The database structure of the product is shown in Fig. 5. As mentioned in the previous section, the information included in the database is news URL, ranking no. with popularity, news title, fetching time, summary content, etc. Besides these, there are 2 more tables to store the information of all the news website URL and reference (sequence) no. for news/site, respectively.

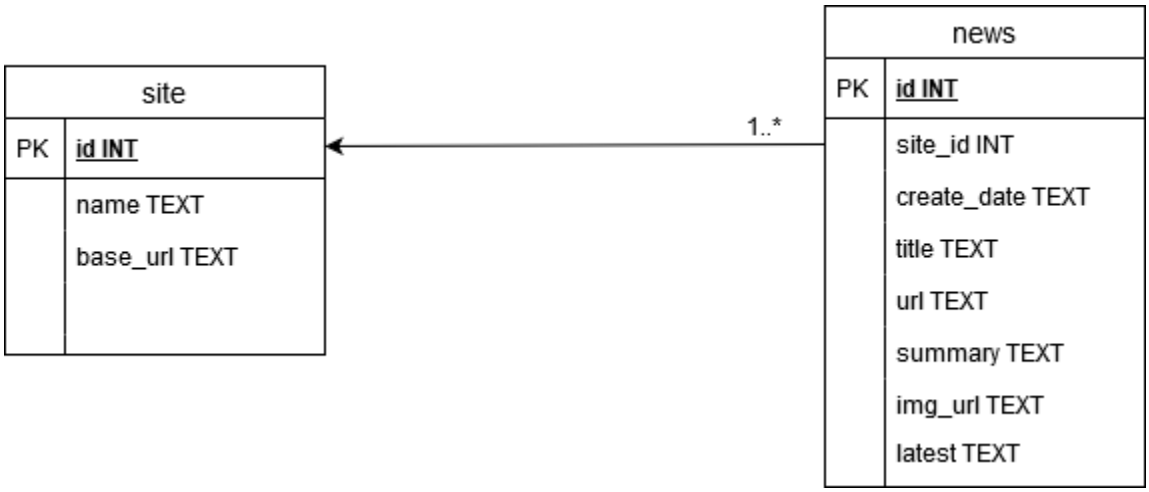


Fig. 5. Database structure

For web service, the corresponding data will be pulled from the database. The news title and news summary with popularity sequence will be shown to the user for reading. The news url will also be provided if the user would like to check more details. The details of the user guide and user interface will be shown and discussed in the section "Installation and User Guide".

System Development & Implementation in tools

The system is developed with the implementation of tools: TagUI, API, SQLite, Local AI, Flask and AWS. As shown in Fig. 6, both TagUI and Api are used for news content acquisition. One reason to use both methods is to compare the performance for them. Most of our news content is retrieved with the TagUI method in essence. The content is stored in the database with SQLite. Based on this data, local AI will do the Natural language processing to get the summary for each article.

Fig. 6. System Development & Implementation in tools

The web service is built with the Flask, and using AWS. AWS is one the most popular cloud platforms to construct websites and provide a variety of services for both the personal and enterprise users.

Appendix of report: Installation and User Guide

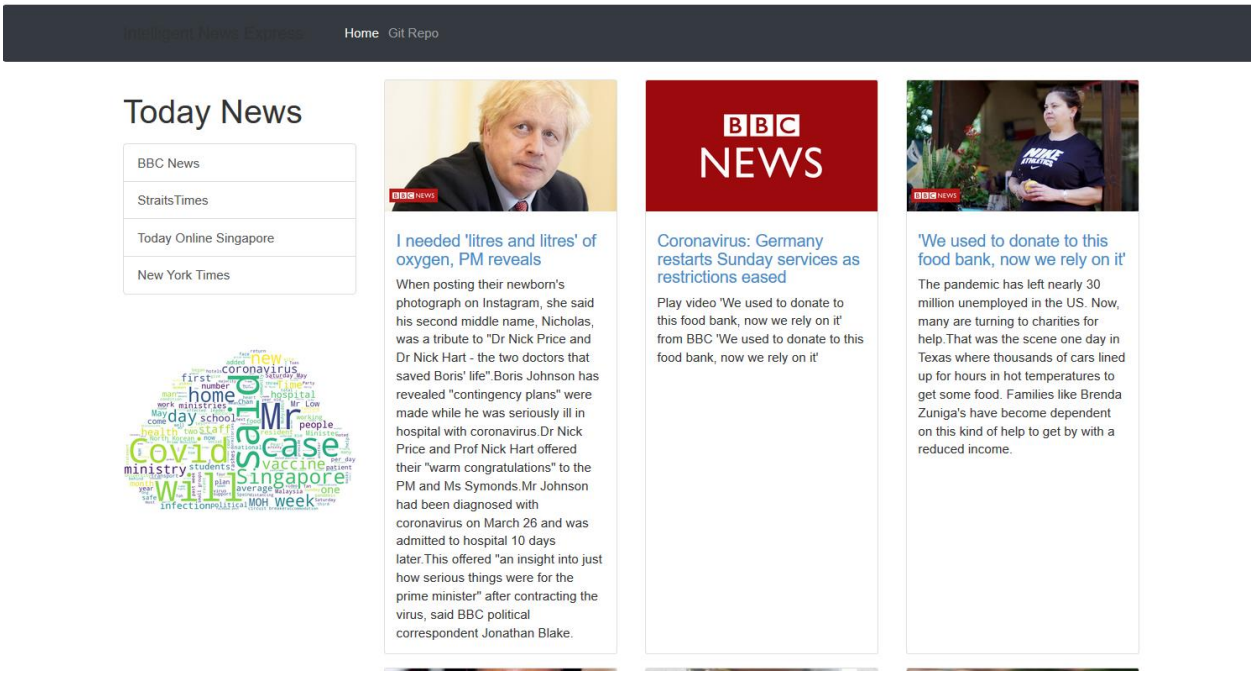
Installation

1. Clone git repo: https://github.com/SkyDB/ISA-CA1-Intelligent-News-express.git
2. Install python 2.7
3. pip install apscheduler==2.1.2 sqlite3 wordcloud tagui flask request pandas pyteaser pprint pillow
4. RPA server:
    a. Open terminal and cd to project root folder
    b. In scheduler.py, sched.add_cron_job(job_function, hour='0-23') defines that the RPA job will run every hour. If you want to test, you could change it to sched.add_cron_job(job_function, minute='0-59') to let it run every minute. Do take note that in console it may pop up some warning but it's okay since it is because the maximum of running instance is 1, the current job is not finished yet
    c. $ python scheduler.py
    d. Step 6 may take several minutes. Once it is done, news data will be stored in ./db/news.db. You could download DB Browser for SQLite (https://sqlitebrowser.org/) to check the raw data.
5. Web application
    a. Open another terminal and cd to proejct root folder/web
    b. $ python main.py
    c. Open browser and go to url: http://localhost:5000/

User guide

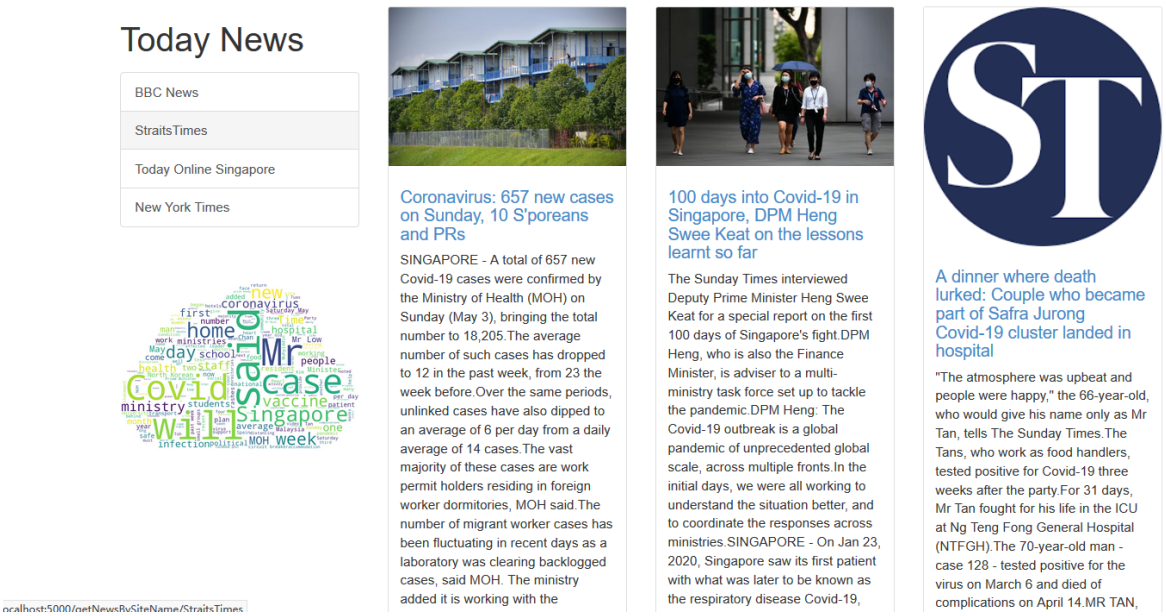
Follow the installation guide and start both RPA server and web application, open browser and go to url: http://localhost:5000/

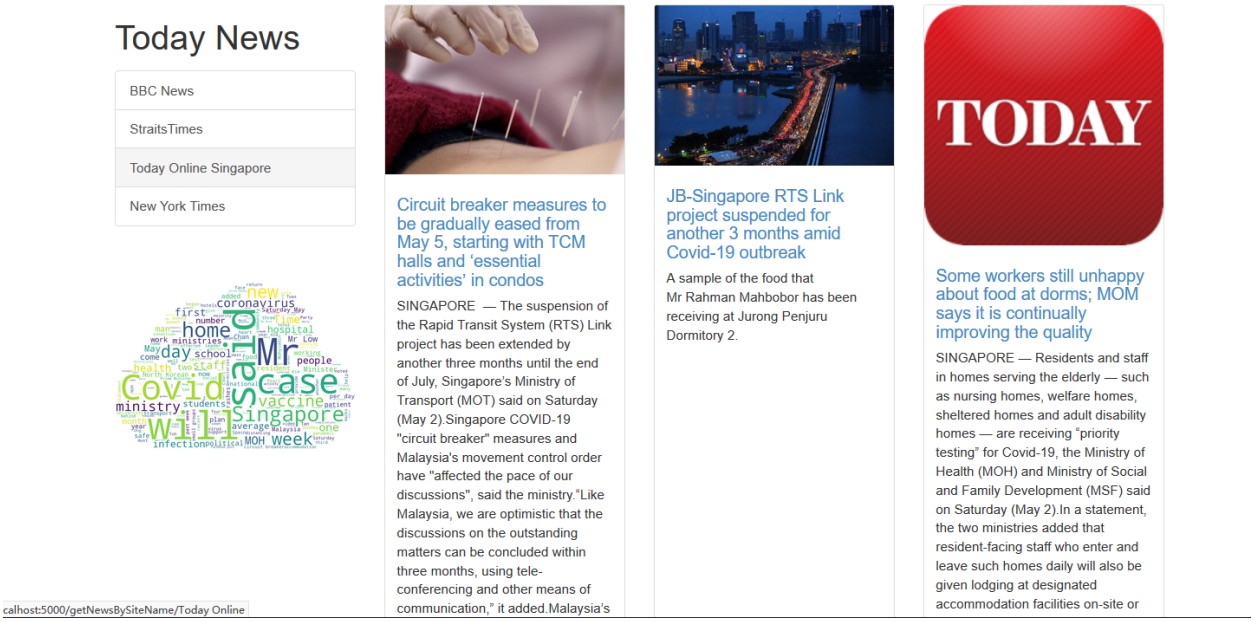
Main page includes news titles, images and a summary of the news.

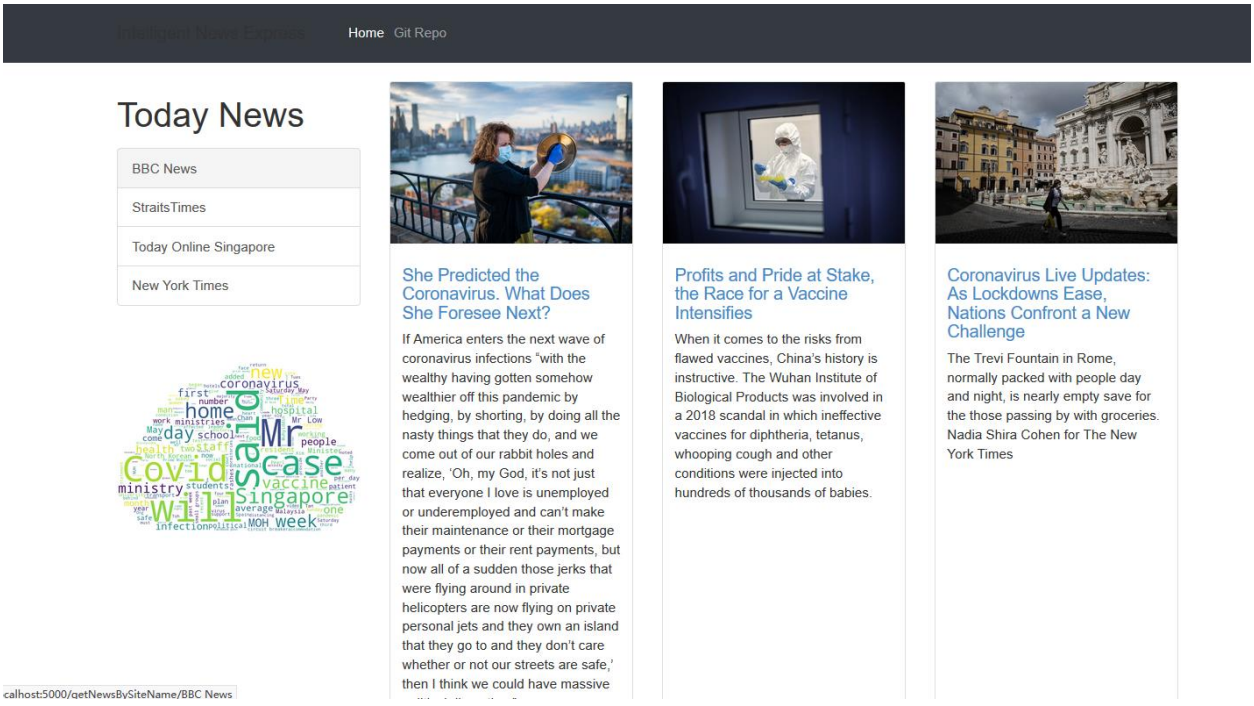


Click the side panel to switch news sources. Then the web application will run a sql query to fetch all latest news in the database, which the RPA server fetch from the news website.

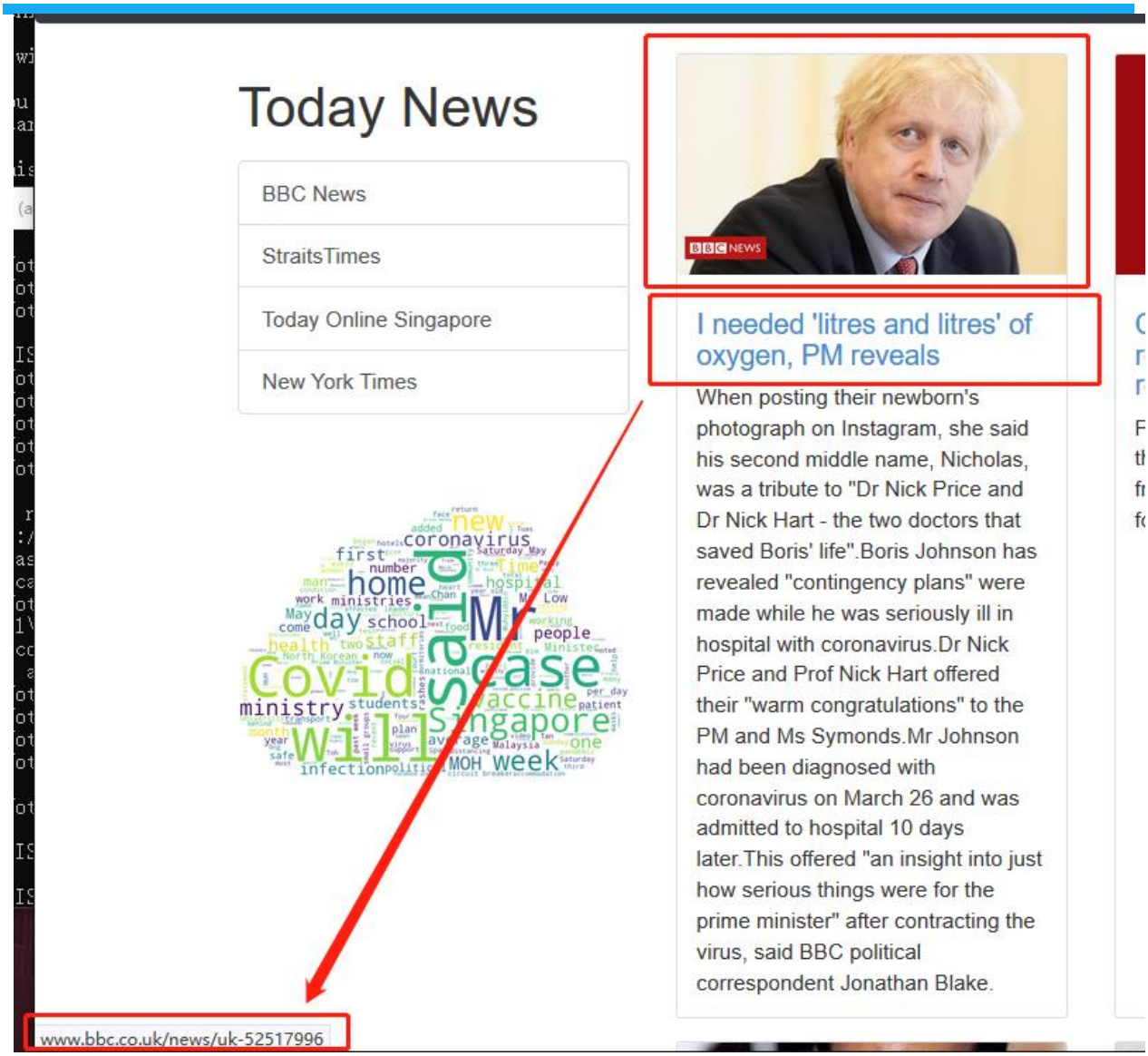
RPA server will update news every hour







Click the news title or news images to go to the original source.


Today News
BBC News
StraitsTimes
Today Online Singapore
New York Times
www.bbc.co.uk/news/uk-52517996
BBC NEWS
I needed 'litres and litres' of oxygen, PM reveals
When posting their newborn's photograph on Instagram, she said his second middle name, Nicholas, was a tribute to "Dr Nick Price and Dr Nick Hart - the two doctors that saved Boris' life".Boris Johnson has revealed "contingency plans" were made while he was seriously ill in hospital with coronavirus.Dr Nick Price and Prof Nick Hart offered their "warm congratulations" to the PM and Ms Symonds.Mr Johnson had been diagnosed with coronavirus on March 26 and was admitted to hospital 10 days later.This offered "an insight into just how serious things were for the prime minister" after contracting the virus, said BBC political correspondent Jonathan Blake.

https://www.bbc.com/news/uk-52517996

BBC Sign in News Sport Reel Worklife Travel Future Mo

NEWS

Home Video World Asia UK Business Tech Science Stories Entertainment & A

UK England N. Ireland Scotland Wales Politics

Coronavirus: Johnson reveals 'contingency plans' made during treatment

5 hours ago Share

Coronavirus pandemic

Boris Johnson has revealed "contingency plans" were made while he was