Data Mining | Ekantipur News

Sparsh Shrestha
SE201337
Gandaki College of Engineering & Science. Lamachaur-19, Pokhara.

The data mining task is undertaken for the Elective II course “Big Data Technologies” under the strict supervision of Er. Ekendra Lamsal (Lecturer).

Objectives met

1. To scrape available news title, date and its author from the site https://www.ekantipur.com.np by using the Google Sheets.
2. To implement an automated JS function in the script editor to run the scraping in a time triggered event mode in order to store the scraped data timely.
3. To implement a Full-Text search in the imported (*.csv) table data via a simple search module in a web page.

Briefings

Web Scraping is the technique to extract large amount of data from websites. Using Xpath various sector of the website can be accessed. Formula is created in Google Sheet using ‘=IMPORTXML()’ and implemented in Google Script. A trigger is set to run the script every hour and the data is stored in the Google Sheet. The following syntax is used to create the formula. ‘=IMPORTXML(URL, xpath_query)’ where URL refers to the URL of the webpage that contains the data of interest and, xpath_query is the query to reach the required data in the web page.

Documentation of the Google Script code is provided in the code file itself in the form of comments. To import the collected data in MySQL, the collected data is downloaded as comma-separated value (.csv) file. A database is created in MySQL Server. A Table is created in the database with MyISAM Storage Engine type. The downloaded .csv file is imported in the table created.

A web page is created and the database is connected to it. A form is created with a place to enter the search key and a button. When the button is clicked, the Full-Text search query is run in the database containing the data we collected.
The MySQL query for Full-Text Search looks like:

*SELECT * FROM thtlive WHERE MATCH(date, title, author) AGAINST('$userinput') IN NATURAL LANGUAGE MODE*