(12) PATENT APPLICATION PUBLICATION
(19) INDIA
(22) Date of filing of Application :24/06/2023

(21) Application No.202311042454 A
(43) Publication Date : 21/07/2023

(54) Title of the invention : SYSTEM AND METHOD FOR SCRAPING NEWSPAPER ARTICLES FROM A WEBSITE

(51) International classification :A47G 291200, B42D 070000, G06F 169535, G06F 169550, G07F 110400
(86) International Application No :NA
Filing Date :NA
(87) International Publication No : NA
(61) Patent of Addition to Application Number :NA
Filing Date :NA
(62) Divisional to Application Number :NA
Filing Date :NA

(71)**Name of Applicant :**
1)Chitkara University
Address of Applicant :Chitkara University, Chandigarh-Patiala National Highway, Village Jhansla, Rajpura, Punjab - 140401, India. Patiala ---------- ----------
2)Bluest Mettle Solutions Private Limited
Name of Applicant : NA
Address of Applicant : NA
(72)**Name of Inventor :**
1)MISHRA, Saket
Address of Applicant :ODC-4, Panchshil Tech Park, inside Courtyard by Marriott premises, Hinjewadi Phase - 1, Pune - 411057, Maharashtra, India. Pune ---------- ----------
2)PANDEY, Sakshi
Address of Applicant :ODC-4, Panchshil Tech Park, inside Courtyard by Marriott premises, Hinjewadi Phase - 1, Pune - 411057, Maharashtra, India. Pune ---------- ----------
3)KUKREJA, Vinay
Address of Applicant :Chitkara University, Chandigarh-Patiala National Highway, Village Jhansla, Rajpura, Punjab - 140401, India. Patiala ---------- ----------

(57) Abstract :
The present invention discloses a system (100) and method (200) for scraping a plurality of newspaper articles from a website. The system includes a processor (102) that includes a web scraping module (106) that identifies and extracts a plurality of URLs of the newspaper articles from the website. The extracted news articles are retrieved and stored in a memory. Utilizing natural language processing techniques, the system extracts textual content, images, keywords, and metadata from the stored newspaper articles and formats them accordingly. The formatted newspaper articles are transmitted to a computing device (110). The system further allows for the identification and selection of specific articles based on predefined criteria, data extraction from the selected articles, multi-threaded article scraping from multiple websites, cross-lingual news scraping through language translation and localization, and generation of reports based on the analyzed newspaper articles.

No. of Pages : 26 No. of Claims : 10