Engineering and Technology Journal e-ISSN: 2456-3358
Volume 06 Issue 06 June-2021, Page No.-936-941
DOI: 10.47191/etj/v6i6.08, I.F. – 6.39

Knowledge Extraction System from English Newspaper

Kangan Maria[1], Prof. Sridhar Ranganathan[2]

[1]SCOPE-School of computer science Vellore Institute of Technology, Chennai, Tamil Nadu
[2]Vellore Institute of Technology, Chennai, Tamil Nadu

ABSTRACT: Web has tens of millions of files on any subject matter which can be from any field. It may be very tough for anybody to examine loads of documents to apprehend knowledge approximately any occasion. The aim of my research is to apply extraction technique that can be used for tracking topics. Creating a User Interface (UI) where user can communicate with Application. Extract records from newspapers, articles which can be in e-form as well as from newspapers. Information plays an important role in the human society. Information can be related to many important issues in finance, social science, and marketing. The system is based on the concept that will extract text and altering the structure of web page and the contents which are not relevant like ads and user comments will be excluded.

INDEX TERMS: Text Extraction, information retrieval, NLTK, Generating Summary, search engine, fetching records, Flask framework, MySQL

I. INTRODUCTION

Internet is a platform that helps in disseminating information produced daily towards many receiver platforms. Internet is a platform that helps in sharing, understanding different contents. Expanding number of individuals requesting for online hotspot for their day-by-day news. Large number of agencies publish the international and national news each day. Summarizing the contents of data gives the critical extracted information. In this paper, we focused on gathering related information from numerous sources which are available on the web and imposing approaches to perform information extraction, for presenting the information in suitable and user's easily readable format. Information gathered from numerous sources available on the web is important and necessary to provide complete knowledge of a particular domain. In association to news, the user will always have a consistently to peruse news information from more than one site but will have different perspective of showing data. Data will be the same on particular topic but will be shown in different forms. Sometimes high relevant topics are emerged on web, at that point users need to visit numerous locales to find out about their advantage of themes. In order to provide relevant information to users, it is necessary that each and every topic should be based on their interest. In this way, the projected news data extraction framework should gather information from more than one paper site to fulfill the user needs. This can be accomplished by methodically and totally crawling news and data. There is now a growing interest in searching online for information getting knowledge from structured data, and using the Internet to answer structured fact-finding queries. In the field of Information extraction Summarization is yet another important feature. Contents of web page need to be carefully explored and functional, neat, convenient and helpful sentences from the Web Page are to be inferred to give a summed-up perspective on the substance of the web news page. The neat data is to be separated and expressed to the customers. When setting up a summary, explanation connections between the sentences are to be thought of. These sentences are extracted from the Web pages and represented to the users. The relevant content with meaningful sentences is provided.

II. RELATED WORK

A few scientists have extended various strategies in data social affair and extraction frameworks. We proposed a method to improve the extraction of significant substance of the website page without wiping out principal content is removed. At that point by utilizing content length, the length of anchor text and the quantity of punctuations denotes the principal content is extracted. The data extraction is done in three stages - first and foremost the page is normalized to eliminate pointless labels, next valuable substance is searched for and afterward refinement of the separated substance is finished. The first one is that information create needs. From huge chunks of text, we summarize the data. The second point of view is that advertisements have an effect on individual's reading behaviors. We have to separate that part from our relevant data. The third point is to clean the text and use this text to create summaries.

We proposed a framework for analyzing relevant information in newspaper. We first segment and extract relevant data, and then store it in our repository i.e. our database created in MySQL. We show results of extracting information from these newspaper results, and find a few interesting advancements that can be applied in many possible applications. The extracted information could be used utilized to populate records. The extracted information from the web pages should contain taken out labels that contain less data substance and keeps just rich labels. Continuing with the upside of the way that there are outrageous and high number of information Web locales, a great deal of approaches have been anticipated for separating the news story substance from news. Exploiting the way that there are an incredible and expanding number of information Web destinations, a ton of approaches have been proposed for separating the news story substance from the news Web locales. A few methodologies utilize the manual extraction and a few methodologies depend on self-loader. Example- Many web data crawlers to extract the news content from the general news pages ultimately. The news destinations involve various types of Site pages. Other than the news pages, there are various non-news pages, similar to the blog, shopping, environment, advancement, business list and shockingly same pages with different URLs. Moreover, these word pages are gotten out in the various areas of information locales. The news locales are slithered to discover a lot more pages as could really be expected; however, it is difficult to perceive and acquire every one of the new pages rapidly from an enormous number of Web pages. One of the approaches is to use NLTK tool for data extraction.

III. PROPOSED SYSTEM

In Fig 1 one can see the system architecture is portrayed. As seen from the figure, the framework's functionalities can be comprehensively clarified when different task are done. One of the approaches is to use NLTK tool for data extraction. Text summarization is a subdomain of Natural language Processing (NLP) that arrangements with extracting outlines from huge lumps of writings. There are two driving sorts of methods which are utilized for text summarization: one is NLPbased strategies and other one is deep learning-based techniques. We will see a straightforward NLP-based strategy for text outline. We will simply use Python's NLTK library for summarizing Web news articles. To detect the contents of a news story, we evaluate the relevant data between the headline of news and each line on the news page. Our method is suitable to all types of news RSS feeds and is not dependent on the style of the news page.

- Beautiful soup is useful python library for web scraping.
- Another useful and important library that we need to parse XML and HTML is the lxml library.
- NLTK (Natural language toolkit) is a vital stage for building Python projects to work with human language information. It gives simple to-utilize interfaces to more than fifty corpora and lexical assets like WordNet, along the edge of an assortment of text measure libraries for characterization, tokenization, stemming, labeling, parsing, and etymology thinking, coverings for weapons-grade data handling libraries, and a brimming with life conversation discussion. It consists various steps:

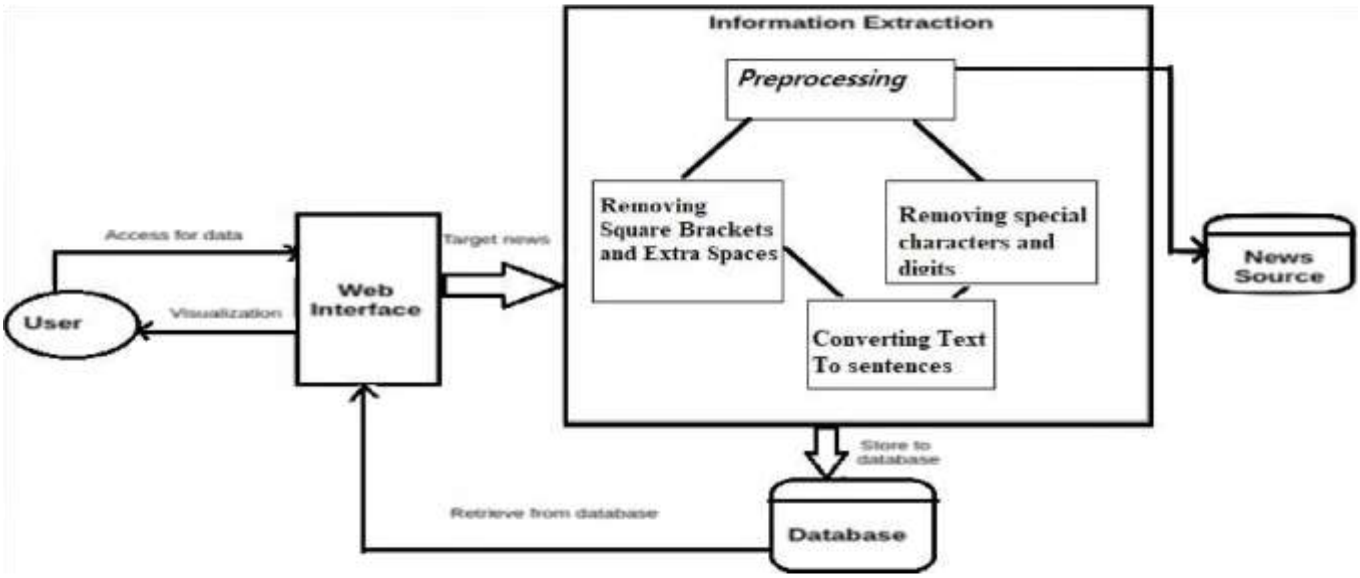
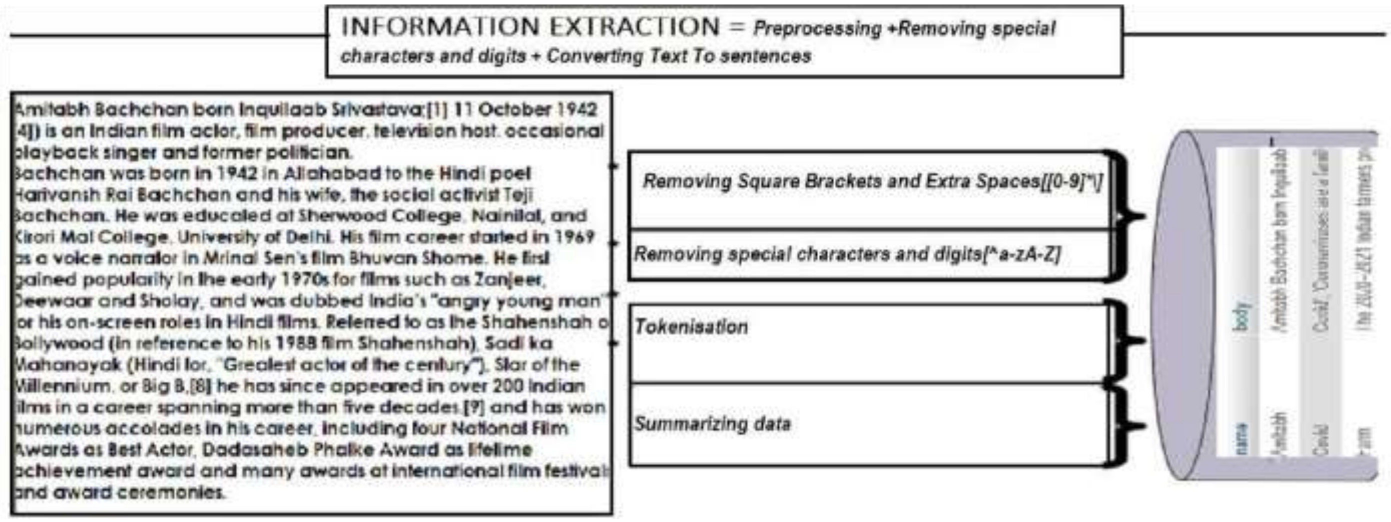


Fig 1. System Architecture

- **Pre-processing**-The first pre-processing step is to eliminate references from the article. Wikipedia, references are encased in square sections. The following content will eliminate every one of the square sections and will supplant the subsequent numerous spaces by a solitary space.
- **Removing Square Brackets and Extra Spaces-** The article text object contains text without sections and brackets. In any case, we would prefer not to eliminate whatever else from the article since this is the first article. We will not remove various numbers, punctuation marks, and uncommon characters from this substance since we will use this substance to make outlines and weighted word frequencies will be supplanted in this article.
- **Removing special characters and digits-** Now we have two objects article text, which contains the original article and clean text which contains the formatted article. Now special characters and digits will be removed which are not required.
- **Sentences to be converted** -Now, after we have preprocessed the information. Then, we need to tokenize the article into sentences. We will utilize the content item for tokenizing the article to sentence since it contains full stops. The clean doesn't contain any accentuation and subsequently can't be changed over into sentences utilizing the full stop as a boundary. Once data is extracted and data is ready to summarized and stored into the repository beforehand the user submits the query conversion. On stored data user preferred a query conversion, query retrieval When the user search in the system performs different query operation for the data it is interested, result will be presented. The system includes of the following parts: -
  - Web pages retrieval
  - Information Extraction
    - Collecting data
    - Preprocessing
    - Removing unwanted data
  - Summarizing
  - Detailed View Generation
  - Presenting information

IV. EXPERIMENTAL SETUP

In the proposed framework, we need to give an ease of use through which user can determine the specific interest of information he is keen on. Likewise, the subject of his interest can be in technology field, economy field and many more. According to this information the system. will gather information required by the user and present it in a summarized form. This kind of information is possible when we will apply some tools and approaches. Focusing on crawling information from unstructured data to form structured data but in summarized form. The crawling is done occasionally to have the option to serve customer's request on current data and the crept data are to be put away in a data archive utilizing which we will produce a customized see for the customer.



mitabh Bachchan born Inquilaab Srivastava;[1] 11 October 1942 4]) is an Indian film actor, film producer, television host, occasional layback singer and former politician.
achchan was born in 1942 in Allahabad to the Hindi poet arivansh Rai Bachchan and his wife, the social activist Teji achchan. He was educated at Sherwood College, Nainital, and irori Mal College, University of Delhi. His film career started in 1969 s a voice narrator in Mrinal Sen's film Bhuvan Shome. He first ained popularity in the early 1970s for films such as Zanjeer, eewaar and Sholay, and was dubbed India's "angry young man" or his on-screen roles in Hindi films. Referred to as the Shahenshah o ollywood (in reference to his 1988 film Shahenshah), Sadi ka ahanayak (Hindi for, "Greatest actor of the century"), Star of the illennium, or Big B,[8] he has since appeared in over 200 Indian ilms in a career spanning more than five decades,[9] and has won umerous accolades in his career, including four National Film wards as Best Actor, Dadasaheb Phalke Award as lifetime chievement award and many awards at international film festival nd award ceremonies.

INFORMATION EXTRACTION = Preprocessing +Removing special characters and digits + Converting Text To sentences
Removing Square Brackets and Extra Spaces[[0-9]*\]
Removing special characters and digits[^a-zA-Z]
Tokenisation
Summarizing data
name
body
Amitabh
Amitabh Bachchan born Inquilaab
Covid
Covid, 'Coronaviruses are a famil
The 2020–2021 Indian farmers pr

Fig. 2. Information Extraction Module

This data is planned in a manner so framework can be utilized for required data explicitly on account of user statements. With this, the necessary task can be accomplished effectively and productively. The design step of the framework to gets all required information, for example, the sites to be summed up, labels, where to store the data, how to be utilize the data and so forth When this task is done, the separated data is put away in the Database. The repository helps in quick and proficient recovery of substance that matches any sort of user inquiry. Information is extracted, summarized, stored into the repository. Other tools we have used for storing data and presenting the summarized information using framework. The web app contains two sections, the front-end which is created with HTML, CSS and materialize.css and the back-end which is created with flask microframework and also connected with database. Most of the processing and functionality are done for building back-end. A structure's is a code library that makes a designer's life simpler when building web applications by giving reusable code to normal activities. Furthermore, linguistic research for unstructured text does not take advantage of HTML/XML tags and layout formats used in online documents. As a consequence, for the Web, less linguistically intensive methods have been established using wrappers, which are collections of highly accurate rules that extract the content of a specific page. Flask is a web application framework which is lightweight. It is named as a micro- framework since it doesn't need specific tools or libraries. Side tabs are utilized for single page web applications or to show various substance. Python and MySQL workbench ought to be introduced in the framework. You can utilize Visual studio or some other code editorial manager to deal with the application. Here, we imported the requests library just as the solicitation object from Flask. The previous is utilized to send outside HTTP GET requests to get the particular client gave URL inside the Flask application. Then, we added variables to catch the two capture and results, which are passed into the template. Flask MySQL is a Flask extension that permits you to get to a MySQL database.

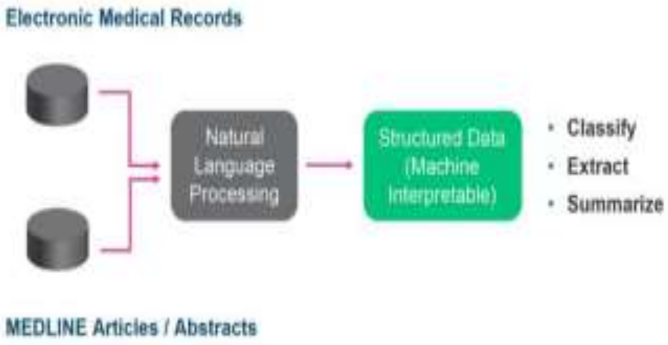

Electronic Medical Records
Natural Language Processing
Structured Data (Machine Interpretable)
Classify
Extract
Summarize
MEDLINE Articles / Abstracts

Fig.3 Extraction Module

Other major Internet news services (such as AltaVista News or Google News) show clusters of similar items, making it easy for readers to access all pieces on a particular topic. These services, however, do not provide summaries, so readers seeking a rapid overview of a topic must choose between reading a representative article in full or searching through all articles

V. RESULTS AND DISCUSSION

The figures 4.9(a),(b),(c),Fig 5 and Fig 6 are the results which is generated by the our model when user enter any query or during querying process.

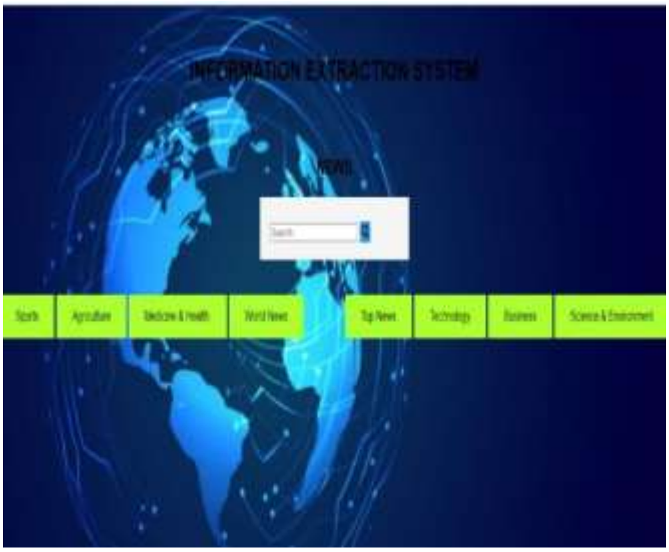


Fig. 4.9 (a) User interaction web page

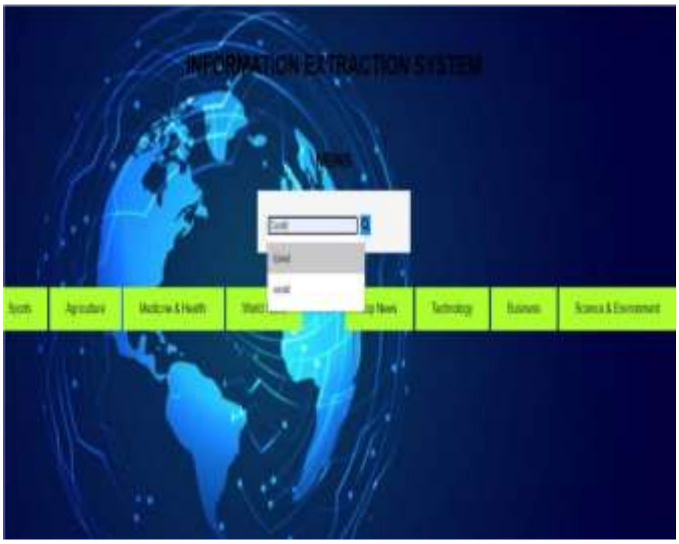
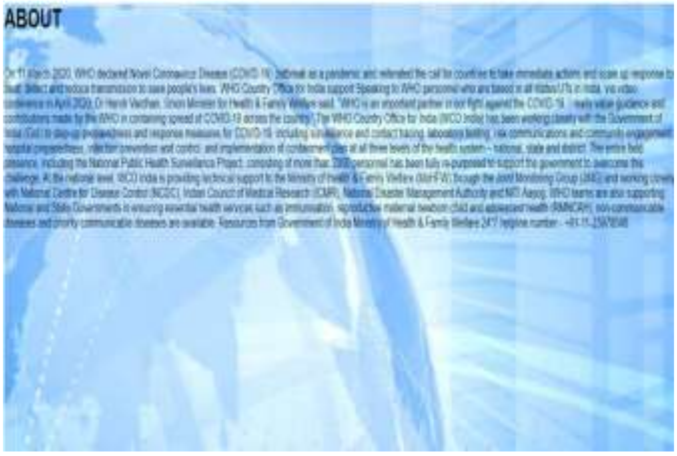
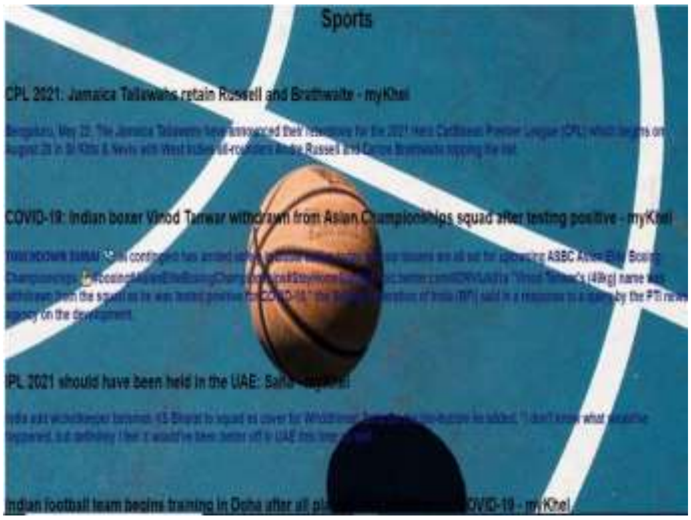


Fig. 4.9 (b) Search Screen web page


ABOUT

On 11 March 2020, WHO declared Novel Coronavirus Disease (COVID-19) outbreak as a pandemic and reiterated the call for countries to take immediate actions and scale up response to treat, detect and reduce transmission to save people's lives. WHO Country Office for India support Speaking to WHO personnel who are based in all States/UTs in India, via video conference in April 2020, Dr Harsh Vardhan, Union Minister for Health & Family Welfare said, "WHO is an important partner in our fight against the COVID-19. I really value guidance and contributions made by the WHO in containing spread of COVID-19 across the country." The WHO Country Office for India (WCO India) has been working closely with the Government of India (GoI) to step-up preparedness and response measures for COVID-19, including surveillance and contact tracing, laboratory testing, risk communications and community engagement, hospital preparedness, infection prevention and control, and implementation of containment plans at all three levels of the health system – national, state and district. The entire field presence, including the National Public Health Surveillance Project, consisting of more than 2000 personnel has been fully re-purposed to support the government to overcome the challenge. At the national level, WCO India is providing technical support to the Ministry of Health & Family Welfare (MoHFW), through the Joint Monitoring Group (JMG) and working closely with National Centre for Disease Control (NCDC), Indian Council of Medical Research (ICMR), National Disaster Management Authority and NITI Aayog. WHO teams are also supporting National and State Governments in ensuring essential health services such as immunization, reproductive maternal newborn child and adolescent health (RMNCAH), non-communicable diseases and priority communicable diseases are available. Resources from Government of India Ministry of Health & Family Welfare 24/7 helpline number - +91-11-23978046

Fig 4.9 (c) Summarized content web page



Sports

CPL 2021: Jamaica Tallawahs retain Russell and Brathwaite - myKhel

Bengaluru, May 22: The Jamaica Tallawahs have announced their retentions for the 2021 Hero Caribbean Premier League (CPL) which begins on August 28 in St Kitts & Nevis with West Indies all-rounders Andre Russell and Carlos Brathwaite topping the list.

COVID-19: Indian boxer Vinod Tanwar withdrawn from Asian Championships squad after testing positive - myKhel

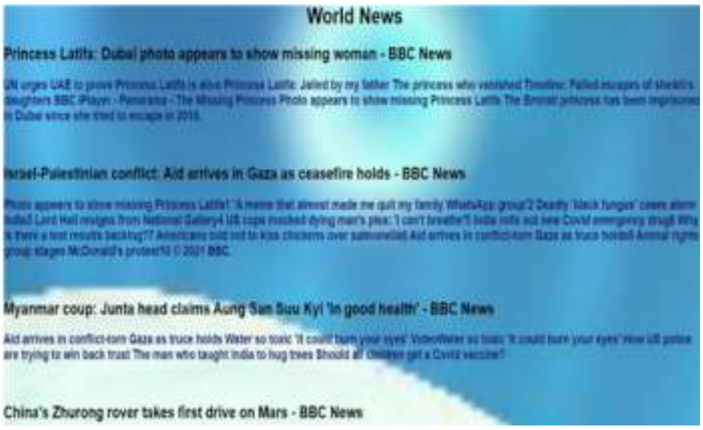
[illegible] contingent has arrived [illegible] all set for upcoming ASBC Asian Elite Boxing Championships [illegible] #BoxingAsianEliteBoxingChampionships [illegible] pic.twitter.com/[illegible] "Vinod Tanwar's (49kg) name was withdrawn from the squad as he was tested positive for COVID-19," the [illegible] Federation of India (BFI) said in a response to a query by the PTI news agency on the development.

IPL 2021 should have been held in the UAE: Sa[illegible]

India add wicketkeeper batsman KS Bharat to squad as cover for Wriddhiman [illegible] he added, "I don't know what would've happened, but definitely I feel it would've been better off in UAE this time [illegible]

Indian football team begins training in Doha after all pl[illegible]OVID-19 - myKhel

Fig 5. Sports News web page



World News

Princess Latifa: Dubai photo appears to show missing woman - BBC News

UN urges UAE to prove Princess Latifa is alive Princess Latifa: Jailed by my father The princess who vanished Timeline: Failed escapes of sheikh's daughters BBC iPlayer - Panorama - The Missing Princess Photo appears to show missing Princess Latifa The Emirati princess has been imprisoned in Dubai since she tried to escape in 2018.

Israel-Palestinian conflict: Aid arrives in Gaza as ceasefire holds - BBC News

Photo appears to show missing Princess Latifa1 'A meme that almost made me quit my family WhatsApp group'2 Deadly 'black fungus' cases alarm India3 Lord Hall resigns from National Gallery4 US cops mocked dying man's plea: 'I can't breathe'5 India rolls out new Covid emergency drug6 Why is there a test results backlog?7 Americans told not to kiss chickens over salmonella8 Aid arrives in conflict-torn Gaza as truce holds9 Animal rights group stages McDonald's protest10 © 2021 BBC.

Myanmar coup: Junta head claims Aung San Suu Kyi 'in good health' - BBC News

Aid arrives in conflict-torn Gaza as truce holds Water so toxic 'it could burn your eyes' VideoWater so toxic 'it could burn your eyes' How US police are trying to win back trust The man who taught India to hug trees Should all children get a Covid vaccine?

China's Zhurong rover takes first drive on Mars - BBC News

Fig 6. World News web page

These are the significant boundaries dependent on which the framework is evaluated since this is essentially a recovery framework. It is based on major parameters and specification based on which web framework is estimated. Fig4.9(a), (b), (c) shows the characteristics of queries that were run on the system .

Based on the results, it can be noted that user is getting required result based on his query. Then again, questions including specific news have more exactness than review. This is because of the utilization of word recurrence match as a standard to channel the words. Increasing in number of matching contents, the users demand increases as well. The system deals with proper and accurate information extraction so that user should get proper results according to them. Fig 5 and 6 shows the sports and world-based news whenever user interact with web page. These web pages contain data which is in summarized form

VI. CONCLUSION

We explored a good thought for mining data from web news pages. The proposed strategy joins the related data accessible from different news sources in the web, cleans them, searches for semantic relations between the page content and the chase question to give just request huge information to the customer. Going digital and helping users to interact with application. For better searching, organizing, and analyzing data connecting the structured and unstructured worlds. There are many applications that rely on the programmed extraction of design from unstructured information. Starting with fundamental named entity recognition systems research in the natural language community, the topic has grown to include a veritable community of researchers from machine learning, databases, and other fields and information retrieval. There has been a ton of work done on numerous pieces of the data extraction issue, for example, essential quantifiable and rule-based models, constructions and plans for managing extraction pipelines, execution improvement, and weakness the chiefs. In the underlying segment, we zeroed in on focus models for the extraction of sub-stances and associations through rule-based and quantifiable models. We introduced various types of rules for element extraction and for settling clashes. Then after working on Machine Learning model, we came to next part.XML based URL are extracted with the help of model we have created. This algorithm in Machine Learning model helps us to build flask application. Not only this HTML helps to connect with web pages, CSS helps to get styling our web pages, and MySQL helps to store data and retrieving data.

REFERENCES

1. Thomas Schloz and Stefan Conrad, "Extraction of Statements in News for a Media Response Analysis", Springer Verlag Berlin Heidelberg 2013
2. Basanta Chaulagain ,Bhuwan Bhatt, Bishal Chaulagain,Dip Kiran Pradhan Newar," Casualty Information Extraction from News Article and Its Analysis", Department of electronics Computer Engineering Lalipur, Nepal August, 2018
3. Ralph Grishman,"Information Extraction: Capabilities and Challenges ", International Winter School in Language and Speech Technologies Rovira i Virgili University Tarragona, Spain, January 21, 2012
4. Hao Han, Tomoya Noro and Takehiro Tokuda, "An Automatic Web News Article Contents Extraction System Based on RSS Feed",Journal of Web Engineering, Vol. 8, No. 3 2009
5. Sungjick Lee, Han-joon Kim,"Automatic Keyword
6. Extraction from News Articles Using TF-IDF Model", International Journal of Computer Applications (0975 – 8887) Volume 166 – No.6, May 2017
7. Chengzhi Z ,Huilin W et al, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems, Volume 4, issue 3, (2008).
8. Sunita Sarawagi, "Information Extraction", Foundations and Trends! R in Databases Vol. 1, No. 3 (2007) 261–377 !c 2008 [8] Mohammad Kamel

,"Sentimental Content Analysis and Knowledge Extraction from News Articles",University of Mashhad, Iran,9 Aug 2018

9. Raymond J. Mooney and Un Yong, "Text Mining with Information Extraction" Proceedings of the 4th International MIDP Colloquium, (2003) [10] Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", AAAI Technical Report WS-99-11.
10. Jasmeen Kaur, Vishal Gupta, "Effective Approaches For Extraction of Keywords", IJCSI International Journal of Computer Science, Volume 7, Issue 6, (2010).
11. Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "A New Approach to Key phrase Extraction Using Neural Networks", IJCSI International Journal of Computer Science Issues, Volume 7, Issue 2 No 3, (2010)
12. Monisha Kanakaraj1 and Sowmya Kamath," NLP based Intelligent News Search Engine using Information Extraction from e-Newspapers "S.2 Department of Information Technology,Conference Paper · December 2014.
13. JOAN FIGUEROLA HURTADO ,"Automated System for Improving RSS Feeds Data Quality",Edinburgh Napier University, August 2014