International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Text Analysis from Internet for Medical Observation

Akshada dighe[1], Revati Deshpande[2], Pooja Jadhav[3].

([#1] Assistant Professor, Department of Information Technology, Genba Sopanrao Moze College of Engineering) BE IT Students , Genba Sopanrao Moze College of Engineering)

Abstract-

The monitoring of global diseases and epidemics is an important duty of public health organizations. Their mission is to protect the public from serious health threats. To do so effectively, we need timely and reliable medical data from a variety of sources. To execute this,we offer a freely available system for monitoring disbease outbreaks by evaluating textual reports, usually in the form of news received from the Internet, In this paper. The system is built around two primary components: MedISys, which uses information retrieval technology, and PULS, which uses information extraction technology.

***Keywords*-- Information Retrieval, Information Extraction, multilinguality, medical intelligence, multi-document information aggregation**

I. Introduction

On a daily basis, professionals in a variety of professions must navigate through vast amounts of data from many sources. Most European Union (EU) countries have a national agency that constantly analyses the media for emerging dangers to public health in their country, as well as for recent health-related occurrences.

It got easier to identify relevant items and compile and manage them electronically as more news sources became available on the internet. At the same time, the number of available sources grew,it became vital to monitor news from neighbouring countries and key travel destinations as a result of increased travel and the resulting importation of contagious diseases.

Text analysis software that detects potentially relevant news articles can help these and other professional communities improve the pace and efficiency of their job, which is normally slow and repetitious.

Users can utilize the search function to create Boolean search term combinations that will filter items from vast collections. In addition to keyword-based filtering, the European Commission's Medical Information System, MedISys, aggregates statistics about query matches, allowing it to provide early- warning signals by spotting sudden increases in media reports about any Public Health-related issue and alerting interested user groups.

II Related work

Information retrieval and extraction have been extensively studied in recent decades, with a large body of literature on both issues. They're usually examined individually, with results reported in distinct venues, and they're classified as discrete problem areas because they use different approaches. Both IR and IE serve a user's information requirement conceptually, though at different levels. It is believed that in real-world scenarios, IR and IE may interact in a pipeline form, for example. The possibilities of tighter engagement are still largely unexplored.

1. Information Retrieval in MedISys

MedISys, the Medical Information System, collects public health reports in a variety of languages from a variety of Internet sources throughout the world, classifies them into hundreds of categories, finds trends across categories and languages, and notifies users. MedISys offers three levels of access: (1) open to the public, (2) restricted to Public Health experts outside the European Commission (EC), and (3) complete access within the EC. The MedISys public site3 provides a quantitative summary of the most recent studies on a wide range of diseases and disease types.

Collection and standardisation of Web documents

MedISys now monitors an average of 50,000 news pieces every day from 1400 news portals in 43 languages around the world, from commercial news providers such as 20 news agencies and Lexis-Nexis, and from roughly 150 specialist Public Health sites. To provide broad geographic coverage, the monitored sources were intentionally chosen with the goal of covering all major European news portals, as well as prominent news sites from across the world. Individual users can request the inclusion of extra news sources, such as local newspapers in their country, however these user-specific sources are often handled individually in order to ensure a balanced distribution of news sources and genres across languages.

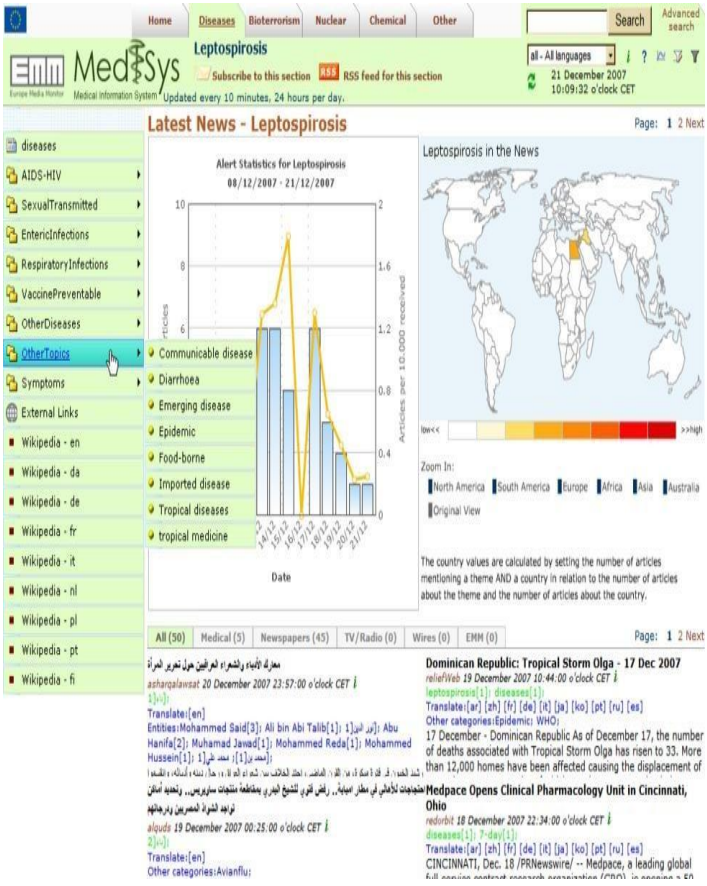
MedISys retrieves RSS feeds when they are accessible. RSS (Really Simple Syndication) is an XML format with specified tags that is commonly used for news and other documents distribution. Scraper software searches for links on pre-defined Web pages, such as those that display the most recently published articles, for additional sources. Using sophisticated transform-mations, the scraper generates an RSS feed from these pages automatically. These conversions are site-specific; they are currently created and maintained manually, one for each news site.

MedISys enables the selection of articles on any topic using Boolean search terms or lists of search phrases with positive or negative weights, as well as the establishment of an acceptance threshold. The user can specify that search words appear within a specific distance (number of words) and that wild cards be used. When dealing with highly inflected languages, wildcards are essential. An alert is the name given to each of these subject definitions. Alerts are multilingual, which implies that search word combinations might come from a variety of languages. Specialist users can construct their own subject-specific alert definitions in addition to the general alerts pre-defined by the JRC's team of engineers. Users are responsible for ensuring that their notifications are accurate and full. The JRC created a specific algorithm that allows the system to scan incoming articles in real time for thousands of alert definitions. The RSS file is updated with information about the warnings identified in each article.

Detection of early-warning trends across languages and news sources

MedISys' alert definitions are multilingual, allowing the mention of a disease or symptom to be recognised in a variety of languages. MedISys keeps track of all disease alerts for each country, i.e., it keeps track of all documents mentioning a certain disease and country throughout a two-week period. By comparing the statistics for the last 24 hours with the two-week rolling average, an alerting mechanism detects a rapid spike in the number of reports for a certain category and country. The higher the alert level, the more articles there are for a specific category-country combination relative to the predicted number of articles (i.e., the two-week average). Figure 2 illustrates a MedISys graph with the greatest alert level combinations at any given time. The alarmlevels high, medium, and low are indicated by the colour codes red (leftmost bars), yellow (middle bars), and blue (rightmost bars).

Distribution of the extracted information to the



MedISys users

MedISys' Web interface can be used to keep up with the current trends and find publications about diseases and regions. RSS feeds are accessible for each page, allowing users to integrate the findings into their own applications.
In Internet Explorer, pattern matching is the algorithm that ensures that keywords appear only in relevant contexts. Users who are interested in specific disease scenarios—outbreaks and epidemics, vaccination programmes, and so on—as opposed to users who want to monitor documents that reference diseases in a larger context—will find this useful.

The University of Helsinki created PULS, or Pattern-based Understanding and Learning System well as a link to the URL where the full news text can befound.

Because keyword-based queries may activate on pages that are off-topic yet mention the alerts in unrelated contexts, IE helps to improve precision change and publish newsletters that can be shared through email or mobile phones (SMS). The title and first few words of each article are displayed, as news from specific sources or countries and quickly Besides the accuracy of the MedISys filtering and categorisation, an important issue for users is multiple reporting: due to the high number of independent news sources, MedISys captures many reports that readers of one or a few news sources would miss, but the flip-side of the coin is multiple reporting. This causes extra work for the users and makes monitoring daily news a time-consuming task. The solution to this problem lies in the aggregation of reports into larger units. MedISys and PULS use different approaches to aggregation, which are not currently integrated.

Cross- document aggregation

.

Integration

This section details the integration of MedISys and PULS, with the goal of demonstrating that the whole is more than the sum of its parts, even at this early stage. Between MedISys and PULS, an unique RSS tunnel has been put up. PULS now only processes papers written in English. MedISys uses the tunnel to send papers to PULS that it categorises as relevant to the medical sector. The documents are currently delivered in plain text format. This is in addition to the standard processing on the MedISys side, which includes monitoring running averages for all alerts, etc. Every 10 minutes, a batch of documents newly discovered on the Web is sent. On the PULS side, the IE system analyses all documents received from MedISys and sends back structured information extracted from the documents across the tunnel (also at 10 minute intervals). While both sites are running in real-time, this communication is asynchronous.

Evaluation—Summary of Results

Over twenty Public Health authorities use the restricted site, the customisable Regional News Service view of the data, and the automatically generated notifications, which are now accessible by an average of 1,700 unique users per day. The user response has been overwhelmingly good. Users frequently request more news sources or alert categories while providing comments. Heavy MedISys users occasionally indicate a desire for more thorough news filtering to decrease the amount of articles that mention an alert out of context. Unwanted noise, for example, is deemed news about a celebrity in the context of which an illness is stated. Using automatically trained classifiers to differentiate relevant from irrelevant news (e.g., sports, film, etc.) is one technique to address this problem. Another alternative is to fine-tune the binary alert definitions, add new search phrases, and apply weights to the filter. Because alert definitions are updated on a regular basis, this tuning of alert definitions is a continuous process. The third solution, as shown in this chapter, is to combine the Information Retrieval approach employed in MedISys with a subsequent Information Extraction phase.

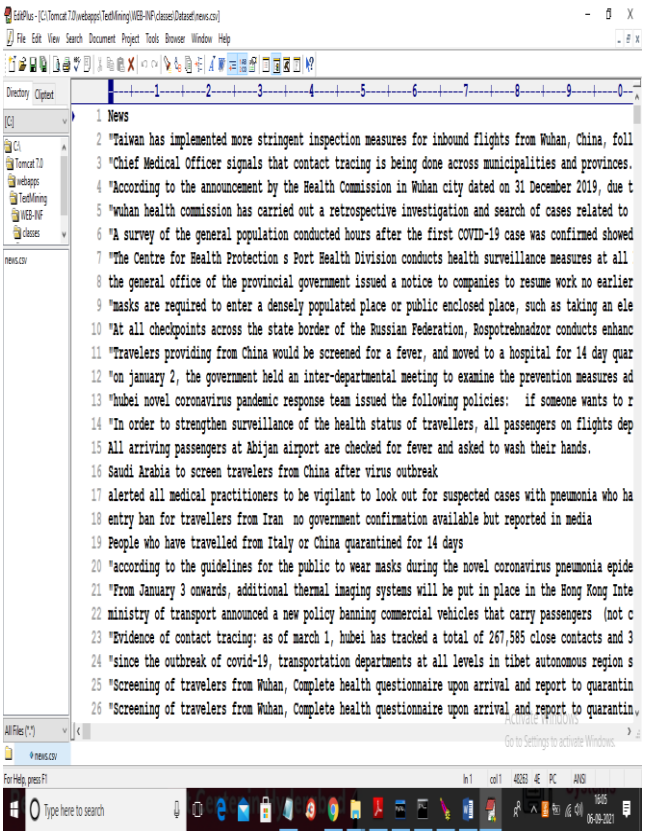
Conclusion and future work

The merger of the two previously separate systems MedISys and PULS has resulted in a stronger application that provides users with complementary functions via a unified user interface. The integration of IR in MedISys and IE in PULS provides extra benefits for infectious disease outbreaks, which are covered by both systems: To begin, PULS' computationally intensive procedures just need to be applied to the MedISys-pre-filtered document collection. Second, the medical event extraction patterns operate as a secondary filter, allowing only disease outbreak reports to be identified. MedISys is intended to capture not just disease outbreak reports, but also illness-related news stories. PULS's event detection helps users interested in disease outbreaks minimise the quantity of reports by filtering out just under three-quarters of incoming reports, with roughly 14% being mistakenly filtered relevant reports.

The current state of integration can be summarised as follows: neither system fully utilises the other's information aggregation methods. The MedISys categorization of news articles could be relevant for PULS's analysis, but it has yet to be used. Although the systems' taxonomies overlap, they have not yet been fully merged. These and other concerns will be addressed in future research.

Acknowledgements

MedISys would not have been possible without the contributions of members of the Web Mining and Intelligence team over the years. The National Center of Excellence "Algodan" of the Academy of Finland funded some of the work on PULS (Algorithmic Data Analysis). Text Mining from the Web for Medical Intelligence In this paper author is describing about two tools called Medisys and Puls where Medisys will read news articles and then detect disease outbreaks and Puls will extract information from such news. In propose paper author is combining two tools to build medical intelligence application which analyses web documents text to extract disease outbreaks in neighboring cities or countries. By extracting disease outbreaks information country can take timely decision to fight such diseases. In propose paper author is using Medisys too l to read news documents from WEB and then handover such information to Puls. Puls will extract information from such news by applying Text Mining techniques such as NLP. NLP will apply PART of SPEECH on Text data to identify diseases and countries names. Puls will hold all those diseases and countries information and then display in online application. Online users can access Puls website to gather disease outbreaks in neighbouring countries and cities. To implement this project we have used NEWS documents from WEB and below screen shots showing those news articles.



In above screen you can see all NEWS articles used in this project to extract and retrieve disease outbreaks in different countries and cities.

To implement this project we have used following software's

1) Stanford NLP: using this API we can tokenize all news articles and then remove stopwords and special symbols
2) Apply POS: using this module we will apply POS on clean news data to extract countries and disease names. All countries and disease names will be marked with NER (Names Entity Recognition)
3) All news articles will be converted to TF-IDF (term frequency – inverse document frequency) vector to count countries with outbreaks diseases.

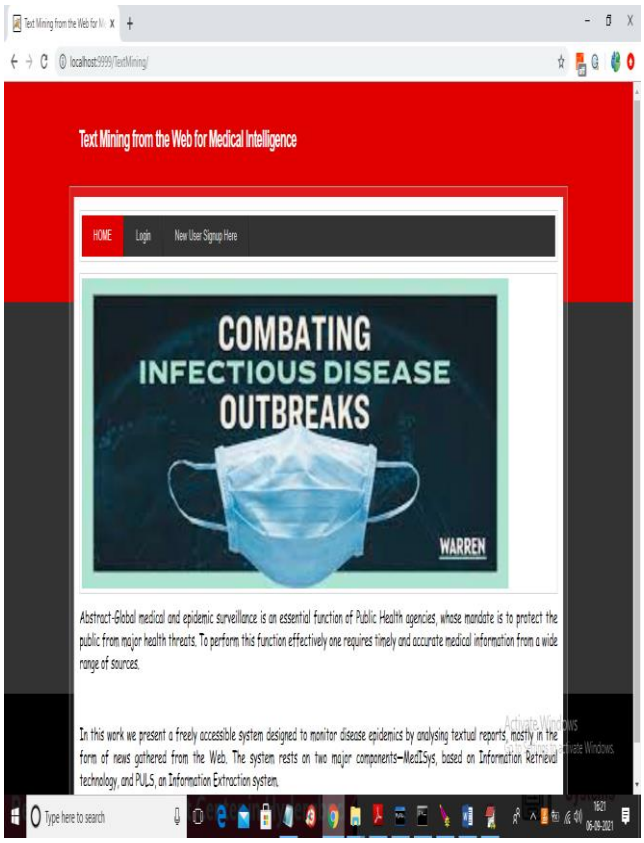
To implement this project we have designed following modules

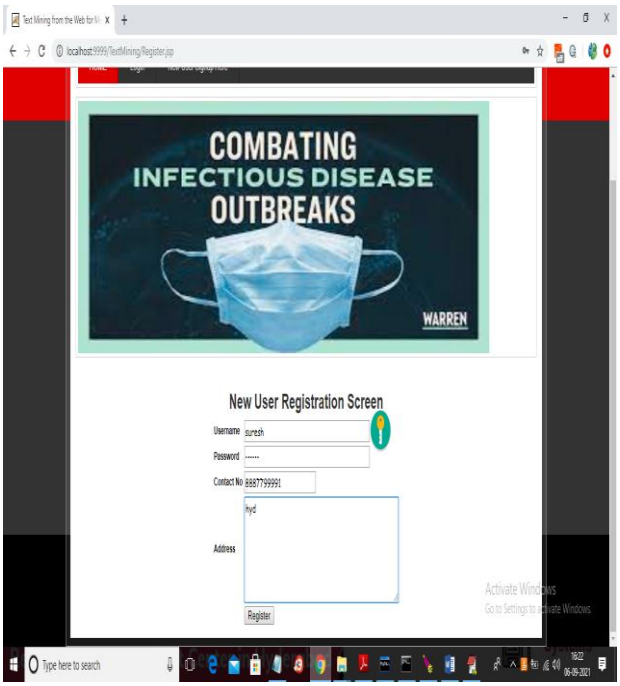
1) New User Signup: using this module new user can signup with application

2) User Login: using this module user can login to application

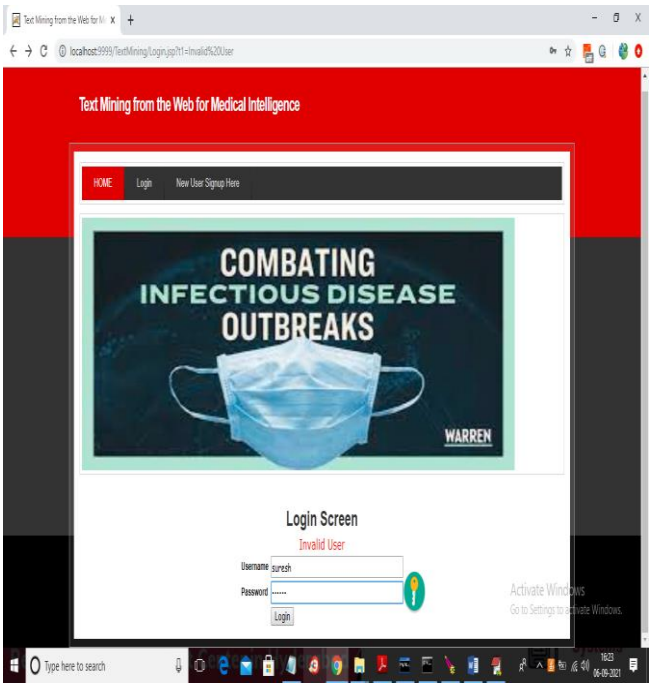
3) MedISys Process News: using this module we will read data from all news articles and then clean it from stop words and special symbols.
4) PULS Information Retrieval & Extraction: Puls will apply filter on news data to extract diseases and countries data and then retrieve all results and display to user.
5) Event Detection Graph: using this module we will plot event detection graph which contains information about countries and diseases outbreaks.

To run this project you need to install JAVA, MYSQL and TOMCAT server and then copy content from WEB-INF/DB.TXT file and then paste in MYSQL to create database and then put 'TextMining' folder inside Tomcat web-apps folder and then start tomcat server and enter URL in browser as 'http://localhost:server_port_no/TextMining to get below output screen

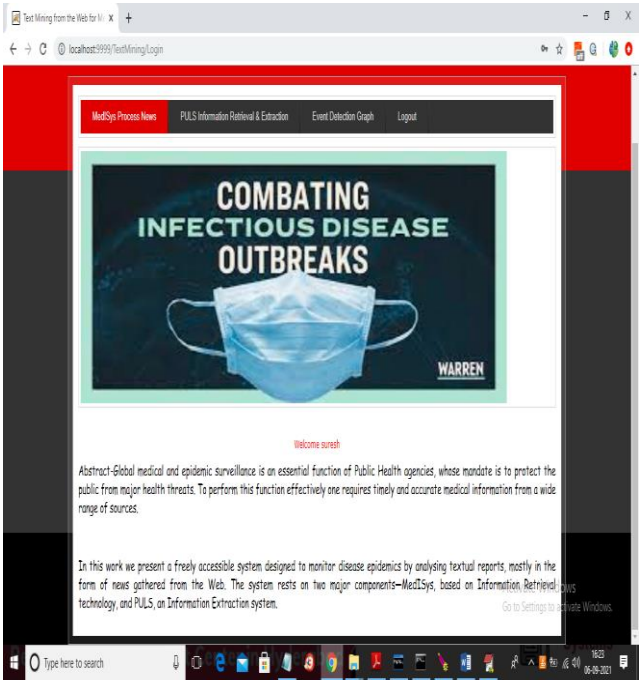




In above screen new user can enter signup details and then click on 'Register' button to complete signup and to get below screen.

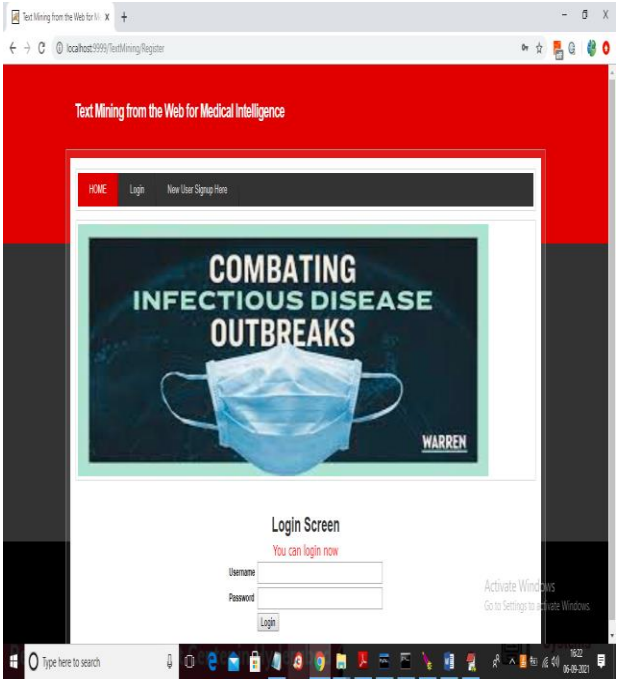


In above screen user is logged in and after login will get below screen.

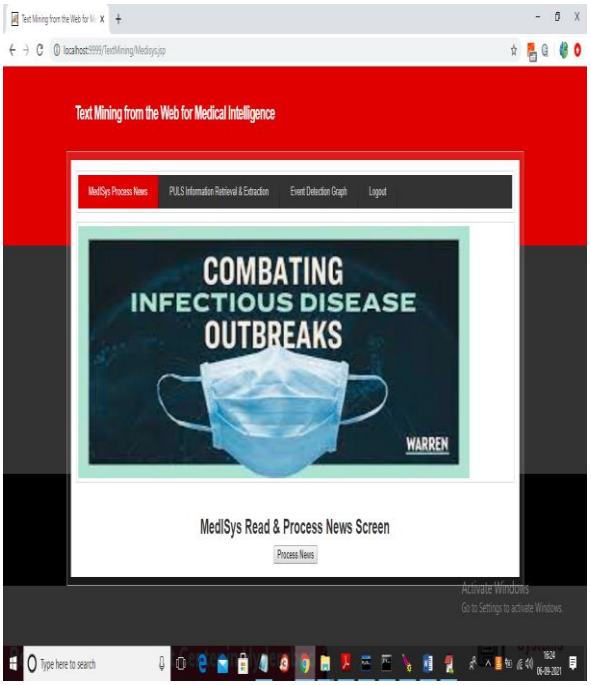


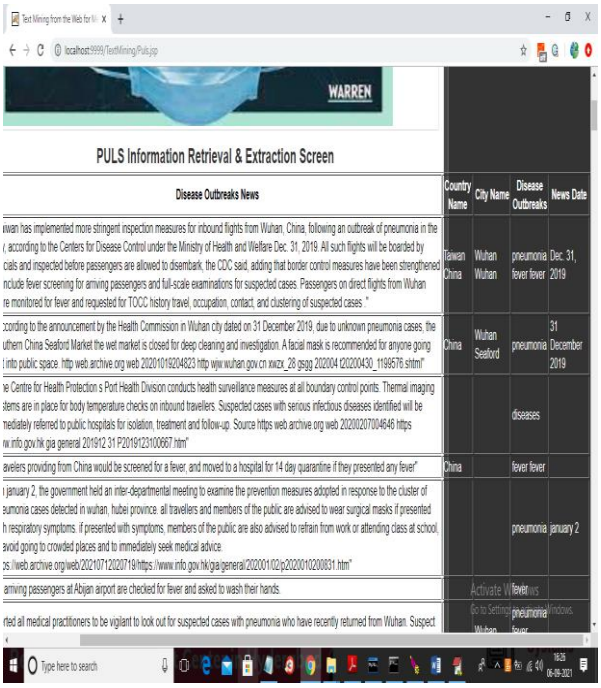
In above screen signup process completed and user can login now to get below screen

.

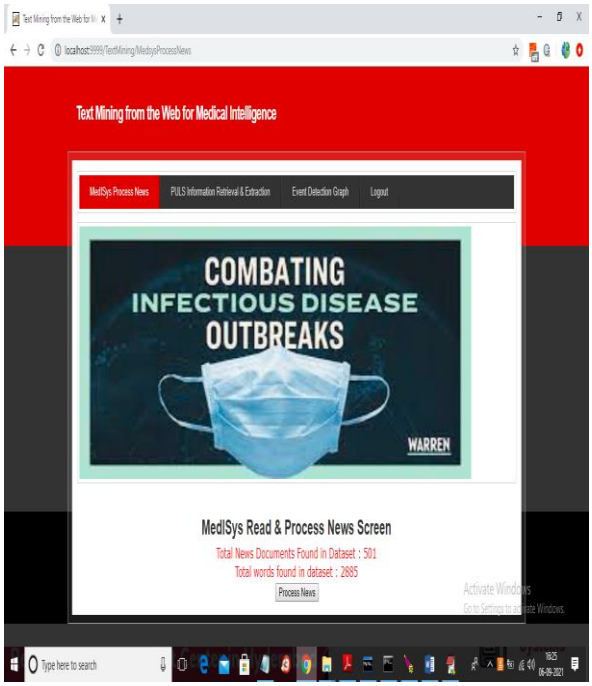


Now in above screen user can click on 'MedISys Process News' link to read and process all NEWS articles and to get below screen.



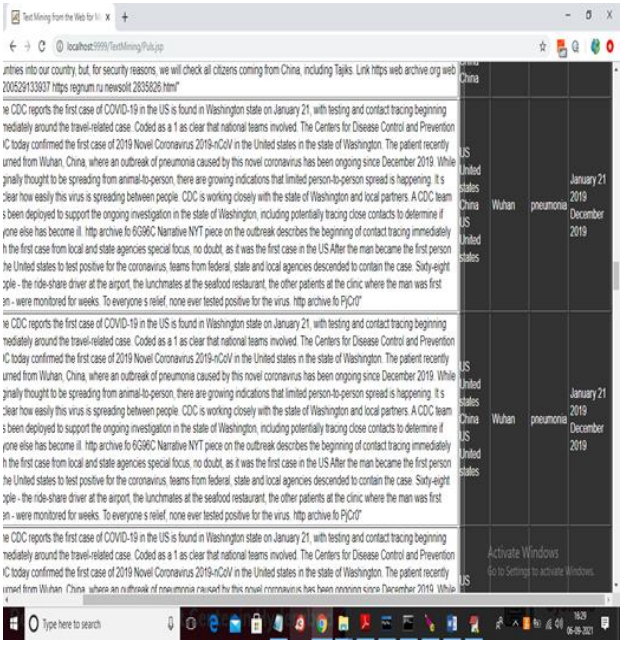


In the above screen click on 'process news' button to get below output

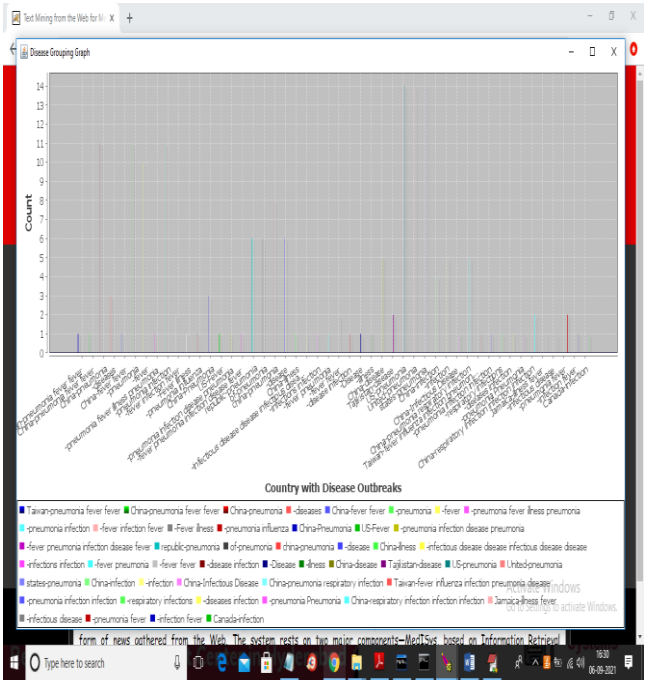


In above screen we can see dataset contains total 501 NEWS articles and all articles contains total 2885 words. Now news data is ready and now click on 'PULS Information Retrieval & Extraction' link to allow Puls to extract and retrieve disease outbreaks from News articles

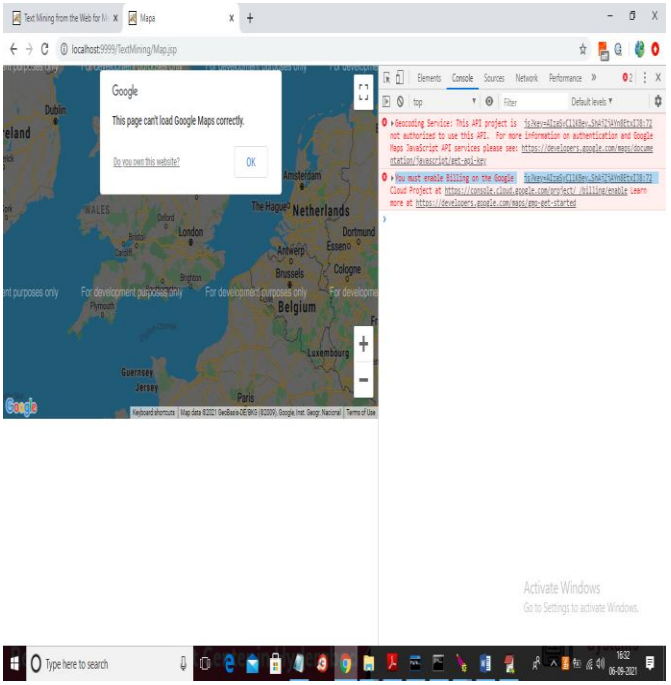
In above screen in first column we can read News Articles and in second column and third we can see names of countries and cities. In 4th column we can see disease outbreak names and in 5th column we can see DATE of news published. So by using this paper concept and by combining MEDISYS and PULS users can extract and retrieve diseases information easily. Now scroll down above screen to view all extracted information



Now in above screen click on 'Event Detection Graph' link to get below graph



In above graph x-axis represents names of countries and cities of disease outbreaks and y-axis represents number (count) of NEWS articles which publish that news articles. Here in am trying to display map also but GOOGLE MAP asking for money to enable that features so I am ignoring it.



In above screen in red colour text you can see Google asking to enable billing to access this MAP API

References

[1]R. Gaizauskas and A. Robertson, "Coupling information retrieval and information extraction: A new text technology for gathering information from the web," in Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet, Montreal, Canada, 1997.

[2]Defence Advanced Research Projects Agency, "Information extraction task: scenario on management succession," in Proc. 6th Message Understanding Conf. (MUC-6). Columbia, MD: Morgan Kaufmann, 1995.

[3]A. Robertson and R. Gaizauskas, "On the marriage of information retrieval and information extraction," in Information retrieval research 1997: Proceedings of the 1997 annual BCS-IRSG colloquium on IR research, Aberdeen, Scotland, J. Furner and D. Harper, Eds. London: Springer-Verlag, 1997.

[4]S. Doan, Q. Hung-Ngo, A. Kawazoe, and N. Collier, "Global Health Monitor—a web-based system for detecting and mapping infectious diseases," in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2008.

[5]C. Freifeld, K. Mandl, B. Reis, and J. Brownstein,"HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports," J Am Med Inform Assoc, vol. 15, pp. 150–157, 2008.