International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 10, ISSN No. 2455-2143, Pages 1-3
Published Online February 2019 in IJEAST (http://www.ijeast.com)

NLP BASED EVENT EXTRACTION FROM MAIL

Hrushikesh.R.Medhekar
Student
Department of Computer Engg,
Vidyalankar Institute of Technology
Mumbai, India.

Rohan.A.Pawar
Student
Department of Computer Engg,
Vidyalankar Institute of Technology
Mumbai, India.

Tanmay.D.Salekar
Student
Department of Computer Engg,
Vidyalankar Institute of Technology
Mumbai, India.

Vipul Dalal
Associate Professor
Department of Computer Engg,
Vidyalankar Institute of Technology
Mumbai, India.

Abstract - **Now-a-days natural language processing is an emerging technology that has led many to develop and implement various systems which involves interacting with the computer system and in turn the computer system process the input which would be human languages and produces the desired output. The biggest example of natural language processing would be google translate which can translate one language to other. We have observed that in today's world most of the professional communication takes place with the help of emails. If any event related information has to be sent than it is sent via emails. It may happen that the user may lose this event related mail in a pool of tremendous amount of emails. Hence we propose to implement a system for extracting events from mail using NLP. In this project we propose to implement a system that extracts events from text and this text can be anything from email or messages but for ease we have decided to use only emails for event extraction. So what our system would do is extract that important information from the email and then save it as a reminder. Later this reminder would be given to the concerned person before the actual event occurs. In the same way the events can be extracted for suppose say client meetings, workshops, inductions etc.**

Keywords: **Event extraction from text, Natural language processing, ANNIE, TF-IDF categorization, RAPIER, pattern matching.**

I. INTRODUCTION

Now-a-days the entire computer industry is implementing and moving towards a more "automated" approach to ease up things which used to take a sizable amount of time. Now this automated approach can be implemented by using lots of methodologies that are available to use may it be Machine learning, Artificial intelligence, Natural language processing, Deep learning, Neural networks etc. Using anyone of the above mentioned technologies we can make a system more smart in a way such that it will ease up things for us. The technology we will be using for implementing this project would primarily include various aspects of natural language processing otherwise known as NLP. Natural language processing is the ability of a computer program to understand human language as it is spoken. NLP is also a component of AI. Developing NLP applications is quite complex as compared to other applications. This is because normally users interact with computers using a programming language which is highly structured, clear, unambiguous and contains a limited set of functions which are typically understandable by the computer. However human speech is not always precise it can contain lots of complexities which might often include regional slang, context of what is being spoken and often the way it is written. We are using this technology to develop an innovative approach of extracting events that are described in the body of the mail and saving it so that the user will get a notification on the day of the event. We can say that it works as a sort of an automated remainder system. There are various ways of doing this by using techniques like TF-IDF categorization, Pearl based pattern matching, Naïve-bayes, Hidden markov model, Normal Pattern matching technique. The input message will be first tokenised using a tokeniser then POS tagging will be done using a POS tagger later we will perform a check for synonyms and finally the event will be extracted which in turn will be stored in a database for alerting the user on a specified date.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 10, ISSN No. 2455-2143, Pages 1-3
Published Online February 2019 in IJEAST (http://www.ijeast.com)

II. LITERATURE SURVEY

Recently, following the progress of wireless internet and smartphone devices, IPhones the amount of data on the web is dramatically increasing with no constraint to time or location. Also the use of artificial intelligence in our day to day life is increasing drastically. This would make devices and computer system more intelligent and smart. Also the amount of digital data is increasing day by day. In this paper we have suggested an innovative technique to extract events from emails. The problem of extracting knowledge from email has addressed in the past by applying various research efforts. Currently many email applications can be easily integrated with personal calendars. The most common application available for this is Microsoft outlook. This system successfully bridges the gap between a user's personal calendar and the emails that the user receives. However the user is still solely responsible for mapping the events to the calendar. In this project we provide an automated approach to map the events to calendars and later notifying the user about the events on the specified date. This project will implement pattern matching technique for retrieval of information regarding the event. Various other techniques available for doing this are:-

2.1 TF-IDF based categorization.
2.2 Pattern matching technique.
2.3 ANNIE (A nearly new information extraction system)
2.4 RAPIER (Rapid production of information extraction rules)

2.1 TF-IDF based categorization:-

TF-IDF is numerical statistic which is used to find the importance of word in the corpus. Term Frequency (TF) used to calculate the occurrence of a word in a given text relative to total number of documents [1]. Inverse Document Frequency (IDF) is, total number of documents relative to the number of documents that contains specific word.

2.2 Pattern matching technique:-

Pattern Matching addresses the problem of finding all occurrences of a pattern string in a text string. Pattern matching algorithms have many practical applications [1]. Computational molecular biology and the world wide web provide settings in which efficient pattern matching algorithms are essential. New problems are constantly being defined. Original data structures are introduced and existing data structures are enhanced to provide more efficient solutions to pattern matching problems.

2.3 ANNIE:-

It was developed at the University of Sheffield and it stands for- A nearly new information extraction system. The ANNIE pipeline is composed of the tokenizer, the gazetteer, the sentence splitter, the part of speech tagger, and the named entity transducer [1].Its main use was to be used a named entity recognizer. Tags of people, location, dates, parts of speech and sentence boundaries are provided to us by ANNIE. The output body of the mail comprises of these XML tags which wraps the recognized tags. The components mentioned above are used independently of one another which can be used at different point of time to provide different annotations.

2.4 RAPIER:-

RAPIER's learning program learns rules that use constraints on words and on part of speech tags [4]. It trains on a training set where each training example is a collection of three files: (a) the original email, (b) the original email after passing it through a sentence splitter and part of speech tagger, (c) a filled event template containing the date, time, location, and title of the event contained in the email.

III. PROPOSED SYSTEM

Basic steps for extracting the events from emails are mentioned below:-

1.Downloading and reading unread mails.
2.Preprocessing of mails.
3.Event extraction module.
4.Save the extracted data in database.
5.Notifying the user regarding the event on the specified date.

1. Downloading and reading unread mails:-

Once the user logs in, the system will first check for unread mails and download these mails and store them in .txt format into a fixed folder. This is the location from where the text of the mails would be accessed for further processing. This data is then retrieved in the python file for event extraction.

2. Preprocessing of mails:-

In this step the system will check the validity of the mails. If the mail is not an event related mail then that mail will be discarded and no further action will be taken on that data. However if the mail has relevant information regarding any event then it will be processed. It will search date time pattern in the text and if such a pattern is found then it will change the date time to our predefined pattern. Here we tag the text using regex to find the date and time mentioned in the textual data. This data is then checked to find what date and time will be considered in our format and store the data in database.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 10, ISSN No. 2455-2143, Pages 1-3
Published Online February 2019 in IJEAST (http://www.ijeast.com)

3. Event extraction module:-

In this module the event is extracted. After processing the mail the system will check if the event is stored in our system if yes it will check whether the event mentioned in the mail is seminar, meeting, workshop or a birthday invitation. It will classify the mail according to the data present in the mail.

4. Save the extracted data in the database:-

The data extracted from the mail would be saved to the database. Later this extracted data that we got as a result of processing the mail would be used for notifying the user. The extracted data would include the date, time and venue of the event.

5. Notifying the user regarding the event on a specified date:-

The date, time and venue that we got from processing the mails would be sent to the user via SMS text. Once the system is able to extract date, time and venue of the event that will happen in the future then it will be sent to the users registered number. It is mandatory that the user should have not enrolled them to the DND service if they have their DND service enabled then they will not receive the SMS regarding the event.

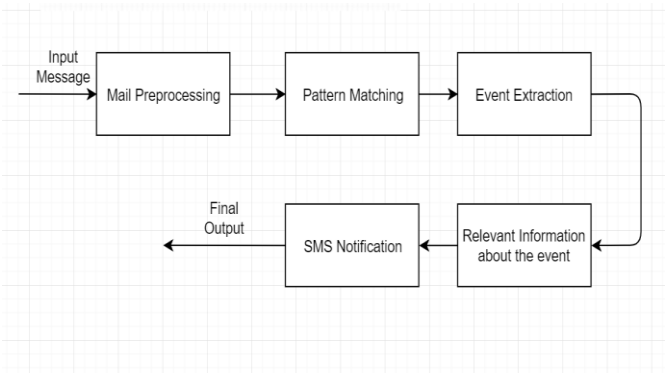

Fig: Block diagram of the system

IV. CONCLUSION

Now-a-days people get lots of emails regarding various events that will occur in future. We can even integrate machine learning and artificial intelligence in this project which will help to successfully identify mail containing event information from other emails like promotional events etc. However this functionality is not implemented in our project. A standalone android application can also be created that would perform same functionality of extracting events from emails containing event related information.

V. REFERENCE

[1] Black A.J., Ranjan N.(2004). Automated Event Extraction from Email.
[2] Hogenboom F.,Frasincar F.,Kaymak U., and Jong-An F. An Overview of Event Extraction from Text.
[3] Almgran M. and Berglund J.(2000).Information Extraction of Seminar Information.
[4] Nath A.,Krishnanath V.,Mathew K.,Pranay T.,Gopi S. NLP Based Event Extraction from Text Messages.
[5] Ritter A.,Etzioni O.M.,Clark S. Event extraction from Twitter.
[6] Cybulska A., and Vossen P.Historical Event Extraction from Text.