Classification of News Articles

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Computer Science Engineering

by

Bhawna Rawat (Y13UC068)
Pushpendra Khandelwal (Y13UC212)

Under Guidance of
Dr. Sakthi Balan Muthia

Department of Computer Science Engineering
The LNM Institute of Information Technology, Jaipur

Dec 2016

The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled "Classification of News Articles" submitted by Bhawna Rawat (Y13UC068), Pushpendra Khandelwal (Y13UC212) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by him/her at the Department Of Computer Science Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2016-2017 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this thesis is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

Date

Adviser: Dr. Sakthi Balan Muthia

To The LNMIIT

Acknowledgments

We have taken efforts in this project, However, it would not have been possible without the kind support and help of many individuals and organization. We would like to extend our sincere thanks to all of them.

We are highly indebted to members of The LNM Institute to Information Technology for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project. We would like to express our gratitude towards our families and members of LNMIIT for their kind co-operation and encouragement which help us in completion of this project.

We would like to express our special gratitude and thanks to our mentor Dr. Sakthi Balan for giving us such attention and time without whose supervision, this wouldn't have been possible. Our thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities.

Abstract

We intend to built an automation system for classification of news articles for the purpose of minimizing the manual workload of the user as our B. Tech Project (BTP). As a daily user of Internet, we come across thousand of news related to a vast horizon, it becomes tedious to filter out the relevant information for students specific to their academic needs. In many real-world scenarios, the ability to automatically classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers.In this project we have initially scraped the news from various e-newspapers using Beautiful Soup(Python). Beautiful Soup is a built in library in python that is used to scrap the data. The scarped data is further stored in a csv file. For supervised classification, the data at first is preprocessed and later will be analysed into various level of classification.

The data is broken into three categories training, validation and test. Before the data can be used for Supervised classification, preprocessing of the data has been done

Contents

List of Figures

Chapter *1*

Survey

1.1 Survey

A Survey was conducted on the library staff and students where the details of how currently the LNMIIT news feeds works. In the survey it was found out that the process is done manually and is time consuming which results in large information loss in different fields. Further the news feeds are currently restricted to academic sections only. The survey conducted on students illuminated the need to broken the horizon of academic field into information related to various test a held for higher Studies.

The Survey gave us the motivation to built an automation system for classification of news articles. Currently the system is thought to work in the academic field only keeping in mind the points detailed out in the survey.

The classification for the academic area has been design into various levels with academic being the root level, further classified into academic research and academic higher studies.The two sections are further broken down into various engineering streams and academic departments.

Chapter *2*

Introduction

Natural language processing is a powerful techniques for automatically classifying documents. These techniques are predicated on the hypothesis that documents in different categories distinguish themselves by features of the natural language contained in each document. Salient features for document classification may include word structure, word frequency, and natural language structure in each document.Our project looks specifically at the task of automatically classifying newspaper articles from the LNMIIT library e-subscriptions(The Times of India, The Economics Times, Financial Express, Hindustan Times, The Indian Express). The various subscriptions archives had a large number of articles which require classification into specific sections (Academic Research, Academic Higher Studies etc).
Our project is aimed at investigating and implementing techniques which can be used to perform automatic article classification for this purpose. At our disposal is a large archive of already classified documents so we are able to make use of supervised classification techniques. We randomly split this archive of classified documents into training, validation and testing groups for our classification systems (hereafter referred to simply as classifiers). This project experiments with different natural language feature sets as well as different statistical techniques using these feature sets and compares the performance in each case. Specifically, our project involves experimenting with feature sets for Naive Bayes Classification, Support Vector Machine, Precision, Recall, Accuracy and F-measure.
It is well known that to obtain a better performance larger data sets can be collected and different machine learning model can be trained. However,the challenging aspect of the classification is that even it cannot be specified by very large data set. Thus their is a need of a model which have sufficient prior knowledge in the form of a good and sufficient training data to compensate for the unknown data.

2.1 The Area of Work

Let us first talk about Extraction of News articles through Web scraping

Web Scraping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in system or to a database. Web scraping is done manually by user looking at the inspect element of web-page and extracting stuff of their use.

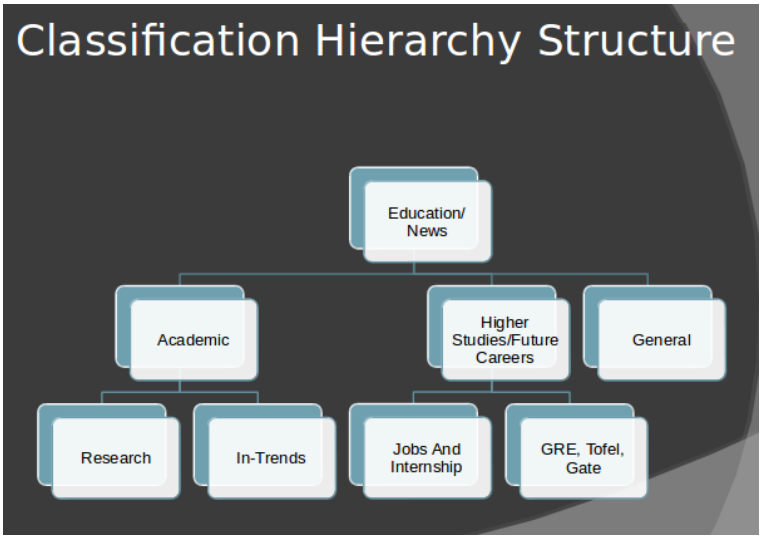

Figure 2.1

We have primarily used Beautiful Soup for Web Scraping. Beautiful Soup is a Python library for pulling data out of HTML and XML files providing Python idioms for iterating, searching, and modifying the parse tree. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

2.2 Problem Addressed

Now we will talk about the problem being worked upon, Beautiful Soup requires data in web page to be in a specific format. This restricts the extraction of news articles from some e-subscriptions(Hindustan Times) where data was present in unstructured manner. Since only Beautiful Soup is used, extraction of news articles from the above subscription has been minimal.

To comprehend the above problem similar titles or headings have been extracted from other newspapers.

Chapter *3*

Future Work

3.1 Future Work

3.1.1 Transforming unstructured data into structured data

The news articles extracted from the LNMIIT e-subscriptions are raw, unstructured data. The textual data needs to be preprocessed by removing the stopwords, white spaces, punctuation, conversion of uppercase to lower case and stemming of words.

3.1.2 Creating suitable Models

The structured will be used to built various models in Support Vector Machine(SVM) and naive bayes. Based on the various performance matrix a final model will be chosen for classification.

3.1.3 User Interface Software

A desktop built in software or browser plugin will be developed to make it easier for the users to search the various news articles related to their fields. The System will be automatically add news articles on a daily basis for continuous improvement in the training set.

Chapter *4*

Reference

1. https://economictimes.indiatimes.com/jobs/
2. http://timesofindia.indiatimes.com/science
3. http://economictimes.indiatimes.com/science/
4. http://timesofindia.indiatimes.com/indutrys/ https://www.crummy.com/software/BeautifulSoup/bs4/doc/
5. https://economictimes.indiatimes.com/academic/
6. https://economictimes.indiatimes.com/jobs/