ISSN 2395-1621

Semantic Analysis Of Text

Saheel Bobde, Saurabh Raj, Sangramsinh Pawar, Rishabh Raj

saheelbobde016@gmail.com,
saurabhraj262@gmail.com,
sangrampawar94@gmail.com,
rishabhshellac@gmail.com

Department of Computer Engineering

Sinhgad Institute of Technology and Science, Narhe.

ABSTRACT

Document classification is a problem in information and computer science. It is basically the process of categorizing documents in certain categories correctly. It is considered as one of the key techniques used for organizing the data by automatically assigning a set of documents into predefined categories based on their content. Recent advances in computer and technology resulted in an ever increasing set of documents. The need is to classify the set of documents according to the type. So, the classification is widely used to classify the text into different classes. This paper proposes a document classification system to identify the domain of the document. This classification is going to be performed by using Naive-Bayes approach which is one of the machine learning algorithms. It consists of a set of phases and each phase can be accomplished using various techniques. Selecting the proper technique that should be used in each phase affects the efficiency of the text classification performance.

ARTICLE INFO

Article History

Received: 25th May 2021

Received in revised form :

25th May 2021

Accepted: 27th May 2021

Published online :

27th May 2021

Keywords: Document Classification, IF-IDF, Files Search, Word Count.

I. INTRODUCTION

Document classification is basically the process of categorizing documents in certain categories correctly. By classifying, we are aiming to assign one or more classes to a document making it easier to manage and sort. This is especially useful for publishers, news sites, bloggers or anyone who deals with a lot of content like managing growing repositories of documents in an organization. By clustering and categorizing documents, the work done in these areas can be completed easily. The size and number of online and offline documents is increasing exponentially. The need for identifying groups of similar documents has also increased for either getting rid of multiple versions of same documents or extracting relevant set of documents from huge document repositories. Document is given as input to the system, then the system can do pre - processing steps. After that main words are extracted by using NLP and TF-IDF algorithm, and are matched by using Naıve Bayes. And then, the documents get classified. This project uses machine learning techniques for classification of documents. Machine Learning enables systems to recognize patterns on the basis of existing algorithms and data sets. Machine Learning undoubtedly helps people to work more creatively and efficiently. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience. In this project, by classifying document, one or more categories are assigned to a document, making it easier to manage and sort. This is especially useful for publishers, news sites, blogs or anyone who deals with a lot of content. Machine Learning uses different algorithms to train systems like Support Vector Machine (SVM), Naıve Bayes, K-nearest neighbour (KNN), Decision tree, K-means etc. In this system, Naıve Bayes algorithm is being used as by comparing different algorithms, Naıve Bayes shows highest efficiency and it is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Problem Statement -

To enhance the accuracy of document clusters incrementally by using supervised document-clustering technique that incorporates domain identification and its summarization.

Objectives-

- To build the model for supervised document clusters.
- To identify domain of given document.
- To generate the summary of keywords from the given document.

II. LITERATURE SURVEY

In the last chapter, a brief introduction about this project is given. It introduces the concept of document classification. Classification has become more important due to the growth of big data with which we could obtain huge data daily. Document classification process is often used in areas such as sentiment analysis, text summarisation, etc. And in this chapter, details about the reference papers that have been referred for this project are given.

2.1 Literature Survey Today, with the widespread use of the internet, data sharing has also increased in the right proportion. In addition, parallel to this increase in usage, the move of many documents to the internet enables the documents to be analyzed and information can be extracted from these documents. In order to be able to analyze the documents and to extract information as a result of these analyses, it is necessary to classify documents first. Document classification is the process of categorizing documents in certain categories correctly. Text and document classification processes are often used in areas such as sentiment analysis, text summarization, etc.

Ghanbarpour, H. Naderi [1] In this paper, an attribute-specific ranking method is proposed based on language models to rank candidate answers according to their semantic information up to the attribute level. This method scores answers using a model enriched with attribute-specific preferences and integrating both the structure and content of answers. The proposed model is directly estimated on the sub-graphs (answers) and is defined such that it can preserve the local importance of keywords in nodes.

Karl Severin, Swapna S. Gokhale Aldo Dagnino. [2] In this scheme supporting efficient ranked keyword search for achieving effective utilization of remotely stored encrypted data. Inside this structure, we use a feasible once-over to in addition enhance the intrigue suitability, and get the ostensibly debilitated constrain framework to cover get the chance to instance of the demand client. Security examination shows that our course of action can accomplish gathering of records and report, trapdoor confirmation, trapdoor unlinkability, and hiding access instance of the intrigue client.

Vidhya.K.A, G.Aghila [3] showed a safe multi-catchphrase arranged look design over encoded cloud information, which meanwhile underpins dynamic fortify operations like destruction and development of reports. Naive Bayes works well for the data characteristics with certain deterministic or almost deterministic dependencies that is low entropy distribution, however the fact is that algorithm work well even when the independence assumption is violated.

Pawar Supriya, Dr. S. A. Ubale [4] proposed a gainful multi-catchphrase break even with word ask for over blended cloud information by recovering best k scored records. The vector space model and TFIDF demonstrate are utilized to gather record and question time. The KNN calculation used to scramble record and demand vectors and develop a unique tree called Balanced M-way Search (BMS) Tree for asking for and propose a Depth First Search Technique (DFST) figuring to complete reasonable multi-catchphrase proportionate word arranged search for. The effectiveness and precision of DFST estimation are addressed with a case, BMS tree, it takes sub-straight time multifaceted nature.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré [5] It has proposed a paradigm for the programmatic creation of training sets called data programming in which users express weak supervision strategies ,which are programs that label subsets of the data, but that are noisy and may conflict, which gives high quality results.. This course of action handles gathering sort structure to part the report record into D Domains and R Ranges. The Domain depends on upon the length of the watchword; the Range parts inside the space in context of the fundamental letter of the catchphrase. An intelligent model is utilized to search for over the encoded recorded watchword that takes out the data spillage.

III. PROPOSED ARCHITECTURE

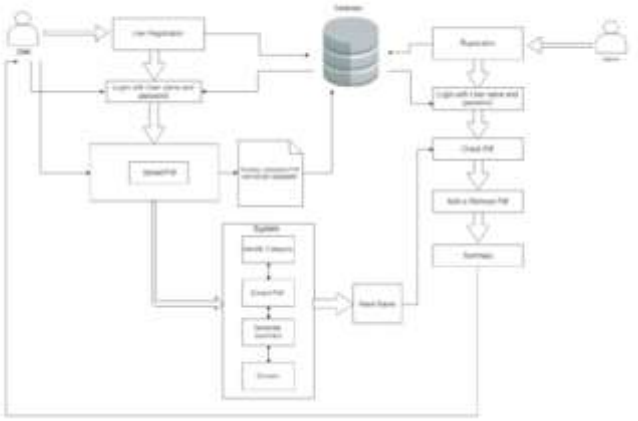

Fig 3.1. System Architecture

In the figure 3.1, the system architecture describes the overall flow of the system. This system is useful for the early classification of the post. The user who will use this system needs to first register into the system. The details will be stored in the database. After registration, the user will log in to the system using the login page. The algorithm used in the system is Core NLP for text mining i.e. for removing stop words, after mining TFIDF is used for extracting main words and Naïve bayes is used for classification of documents based on their content.

There are two main approach for classification of documents:

A. Extract keywords:

Firstly the PDF will be uploaded then keywords are extracted and compared with the keywords that are stored in the database, for this process TF-IDF algorithm has been used.

TF-IDF (Term Frequency –Inverse Document Frequency) is used to convert a document into structured format. It is a numerical to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval. The TFIDF value increases

proportionally to the number of times a word appears in the document.

TF-IDF:
Step1: Clean data / preprocessing - Clean data(standardise data),Normalize data(all lower case),lemmatize data(all words to root words).
Step2: Tokenize words with frequency
Step3: Find TF for words
Step4: Find IDF for words
Step5: Vectorize vocab

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the document length (aka. the total number of terms in the document) as a way of normalization often divides the term frequency:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

NAÏVE BAYES:
Naive Bayes classifier calculates the probability of an event in the following steps:
Step 1: Calculate the prior probability for given class labels
Step 2: Find Likelihood probability with each attribute for each class
Step 3: Put these value in Bayes Formula and calculate posterior probability.
Step 4: See which class has a higher probability, given the input belongs to the higher probability class

IV. ACKNOWLEDGEMENT

We take this opportunity with great pleasure to express our deep sense of gratitude towards our guide Mrs. Prajakta Ambekar for her valuable guidance and incessant encouragement and co-operation extended to us during this project work.

V. CONCLUSION

The system will identify the domain of the document, classify the document and make a short summary of keywords. Due to the use of this system, identification of domain of input document will be very easy and fast which is useful for business or education purposes

REFERENCE

1] A. Ghanbarpour, H. Naderi, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2018.An Attribute-Specific Ranking Method Based on Language Models for Keyword Search over Graphs.

2] Karl Severin, Swapna S. Gokhale Aldo Dagnino. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), pp: 978-1-7281-2607-4.Keyword-Based Semi-Supervised Text Classification

3] Vidhya.K.A, G.Aghila (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010. A Survey of Naïve Bayes Machine Learning approach in Text Document Classification

4] Pawar Supriya, Dr. S. A. Ubale.International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue VII, July 2017. Multi-Keyword Top-K Ranked Search over Encrypted Cloud Using Parallel Processor.

5] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré Stanford University, pages 3567–3575, 2016. Data Programming:Creating Large Training Sets, Quickly.

