THINK INDIA JOURNAL

ISSN:0971-1260
Vol-22-Issue-14-December-2019

Automatic Text Summarization Using NLTK

M SWAPNA, Assistant Professor, Department of Computer Science and Engineering, Matrusri Engineering College, Saidabad, Hyderabad, Telangana

G Sai Manish, B Rachana, N Sindhuja, Matrusri Engineering College, Saidabad, Hyderabad, Telangana.

Abstract - Maintaining data uniqueness is one of the important feature for many areas like in colleges and universities Plagiarism check and data uniqueness is one of the main criteria for preparing paper of paper publishing. In order to maintain plagiarism free content there is need of effective methods where in existing method user should rewrite entire content manually if there is many pages of content to be written then it takes lot of time which is time taking process . With advancement of machine learning and artificial intelligence we can develop a application which can automate process of content rewriting. In this project we are developing a application in python named article rewriter or plagiarism remover in python which will rewrite entire given content in short time. In this project Natural language processing is used in which text summarizer libraries are used. In this project we also compare output of different text summarizer algorithms**.**

1. INTRODUCTION

Plagiarism detection is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet have made it easier to plagiarize the work of others.

Detection of plagiarism can be undertaken in a variety of ways. Human detection is the most traditional form of identifying plagiarism from written work. This can be a lengthy and time-consuming task for the reader and can also result in inconsistencies in how plagiarism is identified within an organization. Text-matching software (TMS), which is also referred to as "plagiarism detection software" or "anti-plagiarism" software, has become widely available, in the form of both commercially available products as well as open-source software. TMS does not actually detect plagiarism per se, but instead finds specific

ISSN:0971-1260
Vol-22-Issue-14-December-2019

passages of text in one document that match text in another document.

Systems for text-plagiarism detection implement one of two generic detection approaches, one being external, the other being intrinsic External detection systems compare a suspicious document with a reference collection, which is a set of documents assumed to be genuine.[5] Based on a chosen document model and predefined similarity criteria, the detection task is to retrieve all documents that contain text that is similar to a degree above a chosen threshold to text in the suspicious document. Intrinsic PDSes solely analyze the text to be evaluated without performing comparisons to external documents. This approach aims to recognize changes in the unique writing style of an author as an indicator for potential plagiarism. PDSes are not capable of reliably identifying plagiarism without human judgment. Similarities are computed with the help of predefined document models and might represent false positives.

A study was conducted to test the effectiveness of plagiarism detection software in a higher education setting. One part of the study assigned one group of students to write a paper. These students were first educated about plagiarism and informed that their work was to be run through a plagiarism detection system. A second group of students was assigned to write a paper without any information about plagiarism. The researchers expected to find lower rates in group one but found roughly the same rates of plagiarism in both groups.

2. Literature Survey

Automatic Text Summarization [1] H. Dalianis,[2] M. Hassel, is the plan to get an important data from a huge amount of information. The amount of data accessible on internet is increasing every day so it turns space and time expanding matter to deal with such huge amount of information. So, managing that large amount of data is makes a major problem in different and real data taking care of uses. The Automatic Text Summarization undertaking makes the users simpler for various Natural Language applications, like, Data Recovery, Question Answering or content decreasing etc. Automatic Text Summarization assumes an inescapable part by creating significant and particular data from a lot of information. Filtering from heaps of reports can be troublesome and tedious. Without a summary or rundown, it can take minutes just to make sense of what the people will discuss in a paper or report. So the

ISSN:0971-1260
Vol-22-Issue-14-December-2019

Automatic Text Summarization that concentrates a sentence from a content record, figures out which are the most imperative, and returns them in a readable and organized way. Automatic Text Summarization is a piece of the field natural language processing, which is the manner by which the PCs can break down, and get importance from human dialect. Automatic Text Summarization that uses the classifier structure and its rundown modules to look over huge amount of reports and returns the sentences that are helpful for producing a summary. Programmed outline of content works by taking the overlapping sentences and synonymous or sense from wordnet most overlapping sentences are considered as high score words [3] H. Seo, H. Chung, H. Rim, S. H., Myaeng, S. Kim, [4] A. J. Cañas , A. Valerio, J. LalindePulido, M. Carvalho, M. Arguedas. The higher recurrence words are considering most worth. And the top most worth words and are taking from the content and sorted according to its recurrence and generate a summary. Lesk algorithm [5] S. Banerjee, T. Pedersen, [6]M. Lesk, is used for evaluating the waits for the input text using online semantic dictionary wordnet and it also uses the word sense disambiguation to identifying the most overlapping sentences in the input content that type of sentences are called equivocal words. Those types of words or sentences are having higher recurrences during the summarization. In numerous normal dialects, a word can speaks to numerous implications/sense, and such type of word is called a homograph. WSD is the route toward making sense of which sentiment a homograph is used as a piece of given setting. WSD is a long-standing issue in computational linguistics, and has a come bonafide application including machine elucidation, information extraction, and information recuperation. Gener-accomplice, WSD use the setting of a word for its sense disambiguation, and setting information can begin from either clarified/unannotated content or other learning resources, for instance, responsive view point word expert, parallel corpora.

Natural Language Processing

Natural Language Processing technique using the nltk for building a main stage for python projects to work with human dialect information. This gives the easier to-utilize by giving the interfaces to one or more than 40 corpora and lexicon assets, for libraries for characterization, for splitting paragraphs sentences, to get its original form of words, labeling, parsing, and vocabulary thinking,

ISSN:0971-1260
Vol-22-Issue-14-December-2019

and wrappers for modern thinking quality common dialect handling libraries, and for dynamic discourse discussion.

3. OVERVIEW OF THESYSTEM

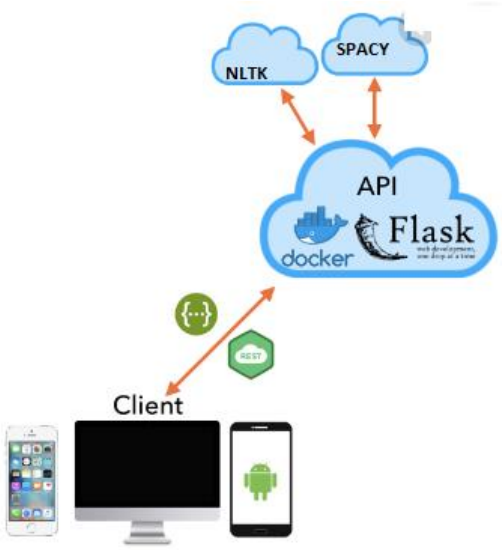

Fig 3.1 System Architecture

3.1 EXISTING SYSTEM

- In existing system in order to remove plagiarism for content manual process was involved in which user should understand each meaning of sentence and rewrite entire content with own words. Which is time taking process.

3.2 DISADVANTAGE OF EXISTINGSYSTEM

- In order to write content uniquely we need to do it manually which is time taking process and lot of human resource is required.
- Time and cost for plagiarism was high.

3.3 PROPOSEDSYSTEM

- In this project NLP is used for understanding in put text or data from URL and then summarize text and rewrite entire content in short time by using text summarizer algorithm..

3.4 ADVANTAGES OF PROPOSED SYSTEM

- Using NLP technique it is easy to convert text to unique way using summarization technique.
- Time and cost consumed for plagiarism if very less.

4. OUTPUT RESULTS

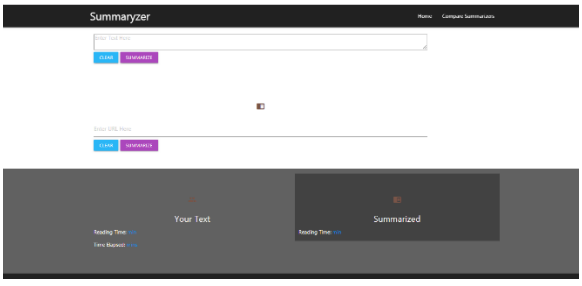

Fig 4.1:Home Page

ISSN:0971-1260
Vol-22-Issue-14-December-2019

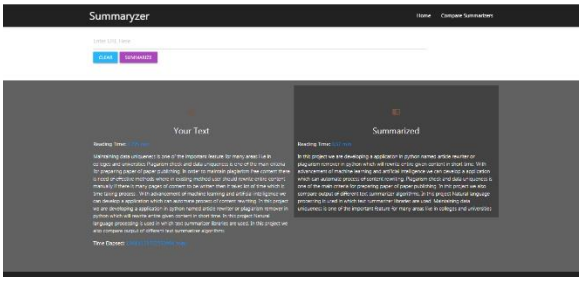

Fig 4.2: Text Summarizer

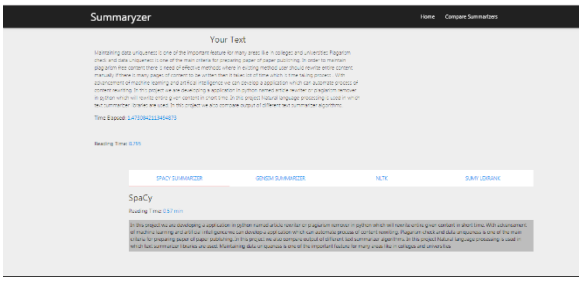

Fig 4.3: Output Text Results

5. CONCLUSION

Millions of people who also have the internet at their fingertips are wondering exactly the same thing as you right now: How can I make money online? How can I get search engine exposure for my website (or blog)? What will give me a leg up on the competition? Thankfully you have already arrived at the answer to all these questions. Article Rewriter Tool is available for free to make your online business as successful as possible, with minimal effort on your part.

The most common way for people to find products or services online is to use search engines, especially Google, Bing or Yahoo. These search engines have certain criteria for giving websites more (or less) opportunity to be returned in search results. The way to obtain dependable, long term search engine optimization is to post as much quality content to your website as possible. The more unique readable text your site contains, the more logical area search engines will have to index and thus refer people to your site. More quality content means more opportunities for your website or blog to receive traffic from major search engines.

7. REFERENCES

[1] H. Dalianis, "SweSum – A Text Summarizer for Swedish," Technical report TRITA-NA-P0015,IPLab-174, NADA, KTH, October 2000.D.

[2] M. Hassel,"Resource Lean and Portable Automatic Text Summarization. PhD thesis, Department of Numerical Analysis and Computer Science," Royal Institute of Technology, Stockholm, Sweden 2007.

[3] H. Seo, H. Chung, H. Rim, S. H., Myaeng, S. Kim, "Unsupervised word sense disambiguation using WordNet relatives," Computer Speech and Language, Vol. 18, No. 3, pp. 253-273, 2004.

[4] A. J. Cañas , A. Valerio, J. Lalinde-Pulido, M. Carvalho, M. Arguedas, "Using WordNet

ISSN:0971-1260
Vol-22-Issue-14-December-2019

for Word Sense Disambiguation to Support Concept Map Construction," String Processing and Information Retrieval, pp. 350- 359, 2003.

[5] S. Banerjee, T. Pedersen,"An adapted Lesk algorithm for word sense disambiguation usingWordNet," In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February, 2002.

[6] M. Lesk,"Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC, 1986.