TEXT MINING USING PYTHON

Madhura Anil Zende[1], Megha Bhaskar Tuplondhe [2], Shalan Baban Walunj[3],
Sujata Vasudev Parulekar[4]
[1,2,3,4]Student, Department of Computer Engineering,
Cummins College of Engineering for Women, Pune
Email: [1]madhura.zende06@gmail.com, [2]megha.tuplondhe@outlook.com,
[3]shalan.walunj@gmail.com, [4]sujata.parulekar10@gmail.com

Abstract:
Today, amount of data generated is increasing exponentially and gained popularity of internet allows us to make proper interpretation and processing of this data for organization betterment. Data generated from various sources is unfiltered and needs processing for its proper usage. This data is typically in unstructured format which makes its usage difficult. Python is widely used as a general purpose interpreted language. Different public application programming interfaces provided by python can be used for extracting data from various resources. Python provides inbuilt libraries and functions that can be directly used. MongoDB can be used for storing unstructured data.
Text mining involves discovery of unknown information by exploiting information from large data sets generated by various sources in from different textual resources. Text mining involves text extractions, analysis, filtering. Text mining discovers and presents knowledge, business rules, facts that are locked in textual form. This paper explains text mining using python to effectively address basics in text mining.

Key words: data mining, information retrieval, patterns, text mining.

1. Introduction:

Text mining is process of analyzing text to glean information that is useful. As compared with other type of data stored in databases, text is unstructured and very difficult to manage. In today's world text is the most common vehicle for communication. Text mining is discovery of unknown information that tries to extract pattern from large database. Approximately 90% of the world's data which daily generated is in unstructured formats. Text is just raw data. Text mining is generally refers to analyze large natural language text and detects usage patterns to extract useful information. Text mining in future may aspire experts to discover new scientific information by analyzing literature. Information extracted should be actionable. Text mining is used to discover knowledge through analysis of text.

2. Basics of Text Mining:

Information retrieval is first step in text mining. It is process of searching and recovering information from huge amount of stored data. Specific rules are there to retrieve information from text and that rules also defines what degree of retrieval should be considered. This is process of identifying and returning relevant information. Text categorization is text mining operation. It is the process of assignment of document to predefined categories. Collected documents should be classified as per requirement. Information extraction is part of text mining process. Information extraction includes identification of certain entities in the text, extraction of entities and representation of it in required format. Important task of text mining is to search for structured data inside document. Text mining also includes identification of trends in data-It is the process of knowing choice of people in today's world.[2]

3. Text Mining Vs Data Mining:

In text mining there is linguistic analysis. It performs different searching functions. In this main focus of search engines is on text search that is specifically focuses at text based web content. Text mining is done for extracting new knowledge from the mountains of text. Data mining is one step in whole process of knowledge discovery from data. Knowledge discovery from data process involves collection of important, valid, new, and useful knowledge from data. The main goal in text mining is to discover unknown information which is that no one yet knows.[3]

The patterns are extracted from data which is natural language text. The text data is free-form, unstructured and semi-structured data. Text includes reports, articles, emails, letters, etc. In data mining the data can be images, audio, videos, etc. Data mining involves case-based reasoning, data visualization, also the main uses of data mining are cross-selling, segmentation and profiling, response modeling.

4. Python functionalities for Text Mining:

Python Text mining package contains variety of useful function for text mining in Python. Its main focus on statistical text mining and makes it easy to create a term document from a collection of documents. This matrix can be then read into statistical package for further analysis. The package provides useful utilities for finding collection, computing the edit distance between word and make long document into smaller piece. Pattern is web mining module for the python programming language which provides tools for data mining Google, Twitter, Wikipedia API, DOM parser and natural language processing.

For web crawling there are the basics tools of urllibs and request. For extracting and Parsing the content python has regular expression handling in the re module and BeautifulSoup. Also there are specialized module for json, feeds and XML. Scrapy is a large framework for crawling and extraction. The NLTK package contains numerous natural language processing methods like sentence and word tokenization, part of speech tagging, chunking and classification. Natural Language Toolkit is a suite of libraries and programs for statistical natural language processing for python programming language. It provide easy to use interface to text processing libraries for classification, tokenization and parsing. NLTK is available for windows, MAC OS X and Linux.

5. Applications of text mining using python:

In today's world we are often face up by tasks which are complicated or very large to accomplish without using smart tools. In recent years, text mining has developed as an essential business tool for discovering the hidden patterns of data. The text mining is used in many sectors for review the large data sets. In publishing and media sector text mining used for production and the optimization of the information retrieving done by Extracting, Loading, and transporting the information. In the Banks and Financial market, CRM tool is used for to improve the customer communication management by re-sending the message by search engine asking question in natural language. In Hospital and Pharmaceutical companies used for classification and extraction of information from articles, scientific abstracts and patents.

The several types of applications are used is in the telecommunications industries: the most important aim of these industries are that all applications find an answer, from human resources management to market analysis, from customer opinion survey to spelling correction.

The Text Mining applications can be summarized as:

- Publishing and media.
- Telecommunications Industries
- Information technology Organization.
- Banks, insurance and financial markets.
- Government Offices
- Pharmaceutical companies and Hospitals.[1]

Conclusions:

Everyday vast amount of data is generated through various activities, so text mining is used to exploit the potential. Text mining is the process of deriving high-quality information from text. Information retrieval, text analysis, information extraction are all text mining operations. Text mining can help organization to gain insight from text based content. It is also useful to extract huge unstructured data, retrieve useful data and then store the structured data in the database. Automatic text mining allows people without any domain knowledge to get information from various documents.

Acknowledgement

We would like to express our profound gratitude and regard to Neeta Maitre and Prof. Sushila Shelke, for their exemplary guidance, constant encouragement and valuable feedback in the completion of this paper.

We are thankful to our colleagues who provided expertise that greatly assisted the research, even if they may not agree with all of the interpretations provided in this paper.

References

[1] [Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[2] M.A Hearst, "Text data mining: Issues, techniques, and the relationship to information access",*Presentation notes for UW/MS workshop on data mining*

[3] A. Tan, "Text Mining: The state of the art and the challenges" Available: print