TEXT MINING

Computer Organization & Programming

Class: CS 550 A

Name: Akshatha Vasant Hegde

CWID: 20009287

Text mining, also known as text data mining, the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. Text mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed or presented directly. Text mining employs a variety of methodologies to process the text such as Natural Language Processing (NLP) by applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms. The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.[2]

Since 80% of data in the world resides in an unstructured format, text mining is an extremely valuable practice within organizations. Text mining tools and natural language processing techniques allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality insights.[1]

Some of the main advantages of Data mining are:

- Scalability: with text mining it's possible to analyze large volumes of data in just seconds. By automating specific tasks, a lot of time that can be saved and used to focus on other tasks.
- Real-time analysis: thanks to text mining, we can prioritize urgent matters accordingly including, detecting a potential crisis, and discovering problems and emergencies in real time.
- Consistent Criteria: when working on repetitive, manual tasks people are more likely to make mistakes. They also find it hard to maintain consistency and analyze data

subjectively. Automating these tasks not only saves time but also allows more accurate results and assures that a uniform criterion is applied to every task.[3]

Text is a one of the most common data types within databases. Depending on the database, this data can be organized as:

- Structured data: This data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Structured data can include names, addresses, and phone numbers.
- Unstructured data: This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or formats like, video and audio files.
- Semi-structured data: This data is a blend between structured and unstructured data formats. Examples of semi-structured data include XML, JSON and HTML files.[1]

The terms, text mining and text analytics, are largely synonymous, but can have a more nuanced meaning.

Text mining identifies relevant information within a text and therefore, provides qualitative results. It combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.

Text analytics, however, focuses on finding patterns and trends across large sets of data, resulting in more quantitative results. It is usually used to create graphs, tables and other sorts of visual reports.

Choosing the right approach depends on what type of information is available. In most cases, both approaches are combined for each analysis, leading to more compelling results.[3]

There are different methods and techniques for text mining. Some very basic methods include word frequency, collocation and concordance.

Word frequency can be used to identify the most recurrent terms or concepts in a set of data. Finding the most mentioned words in unstructured text can be useful when analyzing customer reviews, social media conversations or customer feedback.

Collocation refers to a sequence of words that commonly appear near each other. The most common types of collocations are bigrams (a pair of words that are likely to go together) and trigrams (a combination of three words).

Concordance is used to recognize the context or instance in which a word or set of words appears. Human language can be ambiguous; the same word can be used in different contexts. Analyzing the concordance of a word can help understand its exact meaning based on context. [3]

Even though advanced techniques for text mining may seem complicated, it can be simplified into a few general steps.

The first step with text mining is gathering data. For example, to analyze conversations with users through a company's live chat, the first task would be to generate a document containing this data. Data can be internal (interactions through chats, emails, surveys, spreadsheets, databases, etc.) or external (information from social media, review sites, news outlets, and any other websites).

The second step is preparing the data. Text mining systems use several NLP techniques — like tokenization, parsing, lemmatization, stemming and stop removal — to build the inputs of the machine learning model.

Then for the text analysis itself, the two most common methods for text mining are text classification and text extraction.[3]

A few of the most common NLP preprocessing techniques used in text mining are tokenization, term frequency, stemming and lemmatization.

- Tokenization: Tokenization is the process of breaking text up into separate tokens, which can be individual words, phrases, or whole sentences.
- Term frequency: Term frequency tells you how much a term occurs in a document. Terms can be either individual words or phrases containing multiple words.
- Stemming: Stemming is the process of reducing words to their root form. For example, we would reduce the word *robotics* to the stem *robot*. The stem is usually a full word but does not need to be.
- Lemmatization: Lemmatization is a more complex approach to determining word stems. In lemmatization, we use different normalization rules depending on a word's lexical category (part of speech). This way, the stemmer can grasp more information about the word being stemmed and use that to group similar words more accurately.[4]

The advanced techniques for text analysis include text classification and text extraction.

Text classification is the process of assigning categories (tags) to unstructured text data. This essential task of Natural Language Processing (NLP) makes it easy to organize and structure complex text, turning it into meaningful data.

Text extraction is a text analysis technique that extracts specific pieces of data from a text, like keywords, entity names, addresses, emails, etc. By using text extraction, the hassle of sorting through their data manually to pull out key information can be avoided.

Most times, it can be useful to combine text extraction with text classification in the same analysis.[3]

Perhaps the most common end use case of text mining is text categorization. Text mining would be the first step for building a model that can categorize text into specific domains, such as spam versus non-spam emails, or detecting explicit content. Document classification is another common type of text categorization, especially for sorting news articles into categories such as domestic, international, sports, and lifestyle.

Other applications of text mining include document summarization, and entity extraction for identifying people, places, organizations and other entities. You can also use for sentiment analysis, to identify and extract subjective information from written natural language. Sentiment analysis is especially useful for businesses to detect what their customers are saying on internet forums and social media.[4]

Thus, we have seen that text mining (also at times referred to as *text analytics*) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.[2]

References

[1] IBM Cloud Education. (n.d.). *What is text mining?* IBM. Retrieved May 5, 2022, from https://www.ibm.com/cloud/learn/text-mining

[2] *What is text mining, text analytics and Natural Language Processing?* What is Text Mining, Text Analytics and Natural Language Processing? Linguamatics. (n.d.). Retrieved May 5, 2022, from https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

[3] *What is text mining? A beginner's guide*. MonkeyLearn. (n.d.). Retrieved May 5, 2022, from https://monkeylearn.com/text-mining/

[4] TELUS International. (2015, February 5). *What is text mining? applications & preprocessing techniques*. Warning. Retrieved May 6, 2022, from https://www.telusinternational.com/articles/what-is-text-mining