WIKIPEDIA
The Free Encyclopedia

Text mining

Text mining, **text data mining** (**TDM**) or **text analytics** is the process of deriving high-quality information from text. It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources."[1] Written resources may include websites, books, emails, reviews, and articles. High-quality information is typically obtained by devising patterns and trends by means such as statistical pattern learning. According to Hotho et al. (2005) we can distinguish between three different perspectives of text mining: information extraction, data mining, and a knowledge discovery in databases (KDD) process.[2] Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via the application of natural language processing (NLP), different types of algorithms and analytical methods. An important phase of this process is the interpretation of the gathered information.

A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted. The document is the basic element when starting with text mining. Here, we define a document as a unit of textual data, which normally exists in many types of collections.[3]

Text analytics

Text analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation.[4] The term is roughly synonymous with text mining; indeed, Ronen Feldman modified a 2000 description of "text mining"[5] in 2004 to describe "text analytics".[6] The latter term is now used more frequently in business settings while "text mining" is used in some of the earliest application areas, dating to the 1980s,[7] notably life-sciences research and government intelligence.

The term text analytics also describes that application of text analytics to respond to business problems, whether independently or in conjunction with query and analysis of fielded, numerical data. It is a truism that 80 percent of business-relevant information originates in unstructured form, primarily text.[8] These techniques and processes discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing.

Text analysis processes

Subtasks—components of a larger text-analytics effort—typically include:

- Dimensionality reduction is important technique for pre-processing data. Technique is used to identify the root word for actual words and reduce the size of the text data.
- Information retrieval or identification of a corpus is a preparatory step: collecting or identifying a set of textual materials, on the Web or held in a file system, database, or content corpus manager, for analysis.
- Although some text analytics systems apply exclusively advanced statistical methods, many others apply more extensive natural language processing, such as part of speech tagging, syntactic parsing, and other types of linguistic analysis.[9]
- Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: people, organizations, place names, stock ticker symbols, certain abbreviations, and so on.
- Disambiguation—the use of contextual clues—may be required to decide where, for instance, "Ford" can refer to a former U.S. president, a vehicle manufacturer, a movie star, a river crossing, or some other entity.[10]
- Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, quantities (with units) can be discerned via regular expression or other pattern matches.
- Document clustering: identification of sets of similar text documents.[11]
- Coreference: identification of noun phrases and other terms that refer to the same object.
- Relationship, fact, and event Extraction: identification of associations among entities and other information in texts.
- Sentiment analysis involves discerning subjective (as opposed to factual) material and extracting various forms of attitudi sentiment, opinion, mood, and emotion. Text analytics techniques help analyze sentiment at the entity, concept, or topic l opinion holders and objects.[12]
- Quantitative text analysis is a set of techniques stemming from the social sciences where either a human judge or a com or grammatical relationships between words in order to find out the meaning or stylistic patterns of, usually, a casual pers of psychological profiling etc.[13]
- Pre-processing usually involves tasks such as tokenization, filtering and stemming.

Information extracti
task of automatical structured informa unstructured and/or machine-readable docun electronically represe Typically, this involv human language texts natural language proc Recent activities ir document processing annotation and content e images/audio/video/doc be seen as information ex

Applications

Text mining technology is now broadly applied to a wide variety of government, research, and business needs. All these gr for records management and searching documents relevant to their daily activities. Legal professionals may use text m example. Governments and military groups use text mining for national security and intelligence purposes. Scientific res mining approaches into efforts to organize large sets of text data (i.e., addressing the problem of unstructured d