INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

NLP/TA FOR BIGDATA

SRIVANI BOBBA[1], SUMALATHA BANDARI[2]

1. Department of CSE, Hyderabad Institute of Technology & Management, Hyderabad, Telangana, India.
2. Department of IT, AGTI's Dr. Daulatrao Aher College of, Engineering. Karad, Maharashtra, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Natural Language Processing (NLP) is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language. Significant growth in the volume and variety of data is due to the accumulation of unstructured text data. Text analytics, through the use of natural language processing (NLP), holds the key to unlocking the business value within these vast data assets. In the era of big data, the right platform enables businesses to fully utilize their data lake and take advantage of the latest parallel text analytics and NLP algorithms. In such an environment, text analytics facilitates the integration of unstructured text data with structured data (e.g., customer transaction records) to derive deeper and more complete depictions of business operations and customers. This paper discusses about natural language processing/text analytics for Big Data and the Big Data Architectures and Framework (BDAF).

Keywords: Bigdata, Real-Time Text Analytics, Hadoop, Hive, Pig

Corresponding Author: MR. SRIVANI BOBBA

PAPER-QR CODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Srivani Bobba, IJPRET, 2015; Volume 3 (9): 177-185

INTRODUCTION

Natural Language Processing (NLP) is the scientific discipline concerned with making natural language accessible to machines. NLP addresses tasks such as identifying sentence boundaries in documents, extracting relationships from documents, and searching and retrieving of documents, among others. NLP is a necessary means to facilitate text analytics by establishing structure in unstructured text to enable further analysis.

Text analytics refers to the extraction of useful information from text sources. It is a broad term that describes tasks from annotating text sources with meta-information such as people and places mentioned in the text to a wide range of models about the *documents* (e.g., sentiment analysis, text clustering, and categorization).

Fortunately, today's consumers are quite willing to share their intents and sentiments via social media, if you can gather and process the information. Hence the rapidly developing field of social customer relationship management, or Social CRM."

Gardner continues, "Part of the equation for making Social CRM effective comes from properly capturing the natural language knowledge delivered through the many social channels available to users. Even that is but a first step to being able to gain ever-deeper analysis, however, and rapidly and securely making those insights available where they pay off best. This podcast brings together customer analytics services provider Attensity, with its natural-language processing technology, and HP Vertica, with Big Data analytics capabilities, to explain how to effectively listen to the social Web and rapidly gain valuable insights and actionable intelligence.

The interconnected world of web and mobile apps, distributed sensor networks and cloud computing clusters require a new breed of data capture and analytics infrastructure that can handle the increasing volume and velocity of data. The best way to get big data flowing in real-time is with middleware that takes care of message queuing and delivery so publishing applications and sensors can send data without worrying about where it needs to go or how it needs to get there. This entails the establishment and management of topics and queues, dynamic routing rules, and intelligent handling of fault conditions.

The NLP is difficult because the language is flexible, there is constantly new words, new meanings, different meanings in different contexts, language is subtle, the language is complex, and there are many hidden variables (knowledge to the world, knowledge of the context, knowledge of the techniques of human communication.

In this area Tera data offers Aster analytic solutions involving Attensity and they make it easy to handle large volumes of textual data, analyze them and give them meaning. Specifically they facilitate the application of linguistic principles to extract the context of entities and relationships, similar to what a human would; facilitate the automatic detection and extraction of entities such as name, place...; facilitate the use of custom classification rules to classify texts in content, sorted by relevance, and discover information. It is also to bring these historical data transactions or contacts, and understanding based on what customers have expressed on the web, what it is wrong or what they are interested in, to define communications, appropriate offers , or to identify customers, high-potential targets.

2. TECHNIQUES FOR MINING VERY LARGE AND/OR STREAMING TEXT CORPORA

Analyzing large textual collections has develop increasingly challenging given the size of the data existing and the rate that more data is being created. Topic-based text summarization methods coupled with cooperative visualizations have offered promising approaches to address the challenge of evaluating large text corpora. As the text corpora and vocabulary grow larger, more topics require to be created in instruction to capture the significant latent themes and nuances in the corpora. However, it is tough for most of recent topic-based visualizations to represent large number of topics without being jumbled or illegible. To enable the representation and navigation of a large number of topics, we offer a visual analytics system Hierarchical Topic (HT).

High-Performance Text Mining contains three components for processing unstructured text data, which lead to the automatically generated term-by-document matrix that forms the foundation for computing SVD dimensions. These SVD dimensions constitute the numeric representation of the text document collection and are formatted to be directly used in predictive analysis that includes text-based insights. These three components are:

Document parsing, which applies natural language processing (NLP) techniques to extract meaningful information from natural language input. Specic NLP operations include document tokenizing, stemming, part-of-speech tagging, noun group extraction, default setting or stop/start list-denition processing, entity identication and multiword term handling.

Term handling, which supports term accumulation, term ltering and term weighting. This entails quantifying each distinct term that appears in the input text data set/collection, examining default or a customized synonym list, as well as ltering (removing terms based on frequencies or stop lists) and weighting the resultant terms.

Text processing control, which supports core and threading control processing, manages the intermediate results, controls input and output, and uses the results that are generated by document parsing and term handling to create the term-by-document matrix, stored in a compressed form.

Unstructured data comes in various formats: text, audio, video, images, and more. The constant streaming of data on social media outlets and websites means the velocity at which data is being generated is very high. The variety and the velocity of the data, together with the volume (the massive amounts) of the data organizations need to collect, manage, and process in real time, create a challenging task. As a result, the three emerging applications for text analytics will likely address the following:

- Handling big (text) data
- Real-time text analytics

2.1. Handling Big (Text) Data

Based on the industry's current estimations, unstructured data will occupy 90% of the data by volume in the entire digital space over the next decade. This prediction certainly adds a lot of pressure to IT departments, which already face challenges in terms of handling text data for analytical processes. With innovative hardware architecture, analytics application architecture, and data processing methodologies, high performance computing technology can handle the complexity of big data. Using sophisticated implementation methodologies such as symmetric multiprocessing (SMP) and massively parallel processing (MPP), data is distributed across computing nodes. Instructions are allowed to execute separately on each node. The results from each node are combined to produce meaningful results. This is a cost- effective and highly scalable technology that addresses the challenges posed by the three V s. (variety, velocity, and volume) of big data.

A typical text mining project involves the following tasks:

Data Collection: The first step in any text mining research project is to collect the textual data required for analysis.

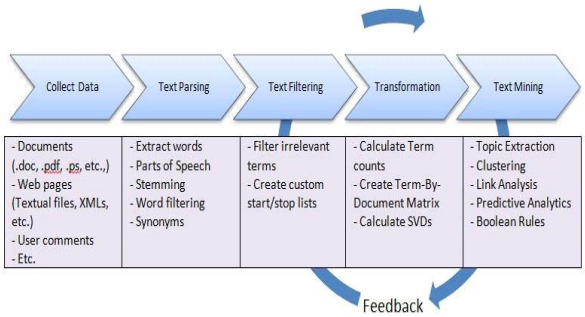
Text Parsing and Transformation: The next step is to extract, clean, and create a dictionary of words from the documents using NLP. This includes identifying sentences, determining parts of speech, and stemming words. This step involves parsing the extracted words to identify entities, removing stop words, and spell-checking. In addition to extracting words from

documents, variables associated with the text such as date, author, gender, category, etc., are retrieved. The most important task after parsing is text transformation. This step deals with the numerical representation of the text using linear algebra –based methods, such as latent semantic analysis(LSA), latent semantic indexing(LSI), and vector space model.

Text Filtering: In a corpus of several thousands of documents, you will likely have many terms that are irrelevant to either differentiating documents from each other or to summarizing the documents. You will have to manually browse through the terms to eliminate irrelevant terms. This is often one of the most time-consuming and subjective tasks in all of the text mining steps. It requires a fair amount of subject matter knowledge (or domain expertise). In addition to term filtering, documents irrelevant to the analysis are searched using keywords. Documents are filtered if they do not contain some of the terms or filtered based on one of the other document variables such as date, category, etc.

Text Mining: This step involves applying traditional data mining algorithms such as clustering, classification, association analysis, and link analysis. As shown in Dig1, text mining is an iterative process, which involves repeating the analysis using different settings and including or excluding terms for better results.

Fig 1: Text Mining Process Flow



2.2. Real-Time Text Analytics

Another key emerging focus area that is being observed in text analytics technology development is real-time text analytics. Most of the applications of real-time text analytics are addressing data that is streaming continuously on social media. Monitoring publicactivity on social media is now a business necessity. Less companies want to track news feeds and blogposts for financial reasons. Government agencies are relying on real-time text analytics that collect data from in numerate sources on the web to learn about and predict medical epidemics, terrorist attacks, and other criminal actions.

However, real time can mean different things in different contexts. For companies involved in financial trading by tracking current events and news feeds, real time could mean milliseconds. For companies tracking customer satisfaction or monitoring brand reputation by collecting customer feedback, real-time could mean hourly. For every business, it is of the ut most importance to react instantly before something undesirable occurs.

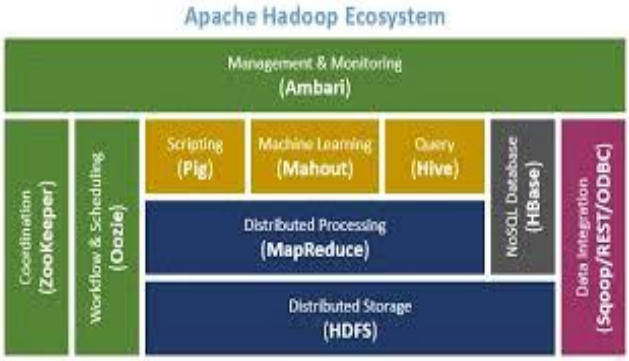
The future of text analytics will surely include the next generation of tools and techniques with increased usefulness for textual data collection, summarization, visualization, and modeling. Chances are these tools will become staples of the business intelligence (BI) suite of products in the future. Just as SAS Rapid Predictive Modeler today can be used by business analysts without any help from trained statisticians and modelers, so will be some of the future text analytics tools. Other futuristic trends and applications of text analytics are discussed by Berry and Kogan (2010).

3. ARCHITECTURES ANDFRAMEWORKS

3.1. Hadoop Framework

Apache Hadoop is a frame work that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each providing computation and storage. Rather than rely on hardware to deliver high-availability, the framework itself is designed to detect and handle failures at the application layer, thus delivering a highly-available service on top of a cluster of computers, each of which may be rone to failures.

Fig 2: Hadoop Ecosystem



HDFS (storage) and Map Reduce (processing) are the two core components of Apache Hadoop. The most important aspect of Hadoop is that both HDFS and map Reduce are designed with

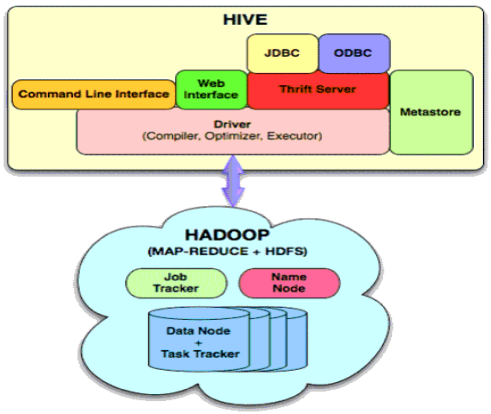
each other in mind and each are co-deployed such that there is a single cluster and thus provides the ability to move computation to the data not the other way around. Thus, the storage system is not physically separate from a processing system.

The Hadoop framework transparently provides both reliability and data motion to applications. Hadoop implements a computational paradigm named Map Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the computer nodes, providing very high aggregate bandwidth across the cluster. Both Map reduce and the distributed file system are designed so that node failures are automatically handled by the framework. It enables applications to work with thousands of computation-independent computers and petabytes of data. The entire Apache Hadoop "platform" is now commonly considered to consists of the Hadoop kernel, Map Reduce and Hadoop Distributed File System(HDFS), as well as a number of related projects-including Apache Hive, Apache HBase, and others.

3.2. Hive Architecture

Hive is a data warehousing infrastructure based on the Hadoop. Hadoop provides massive scale out and fault tolerance capabilities for data storage and processing (using the map-reduce programming paradism) on commodity hardware. Hive is not designed for online transaction processing and does not offer real-time queries and row level updates. It is best used for batch jobs over large sets of immutable data (like web logs).

Fig 3: HIVE Architecture

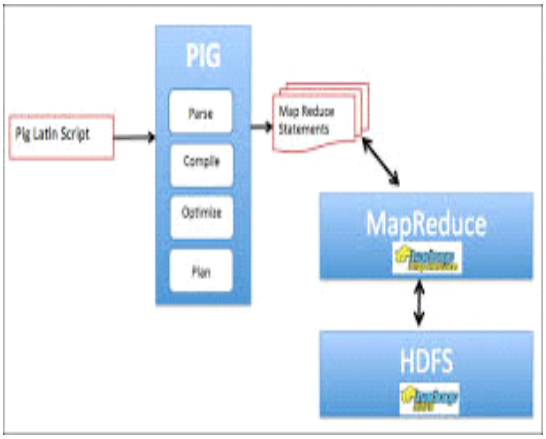


Also allows traditional map/reduce programmers Hive is designed to enable easy data summarization, adhoc querying and analysis of large volumes of data. It provides a simple query language called Hive QL, which is based on SQL and which enables users familiar with SQL to ad-hoc querying, summarization and data analysis easily. At the same time, Hive QL to be able to plug in their custom mappers and reducers to do more sophisticated analysis that may not be supported by the built-in capabilities of the language. Hive store the table schema in a database of its own called as "Derby" which supports single user and usually goes for MySql for more than one user.

3.3. Pig Architecture

Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Fig 4: PIG Architecture



Pig's language layer currently consists of a textual language called Pig Latin. The following are the features of the Pig Latin.

Ease of programming: it is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand and maintain.

Optimization opportunities: The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

Extensibility: Users can create their own functions to do special-purpose processing.

Pig Latin queries execute in a distributed fashion on a cluster. The current implementation compiles Pig Latin programs into Map-Reduce jobs, and executes them using Hadoop cluster.

4. CONCLUSION

High-Performance Text Mining addresses the ever-increasing need to process greater amounts of unstructured text data in less time to improve predictive modeling. The run-time gains affiliated with processing data in a high-performance environment ensures that no big data is too big. Natural Language Processing needs Big data, so that there is a need to work on combining NLTK and Hadoop to create our Big Data NLP architecture.

5. REFERENCES

1. http://www.informationweek.com/big-data/big-data-analytics/natural-language-processing-big-datas-role/d/d-id/1113826

2. http://datacommunitydc.org/blog/2013/05/big-data-and-nlp

3. http://semanticweb.com/natural-language-processing-big-data-making-sense-consumer-behavior_b43907

4. http://www.solacesystems.com/solutions/bigdata

5. http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

6. http://www.bigdatanews.com/profiles/blogs/big-data-natural-language-processing

7. http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/text-mine-your-big-data-106554.pdf

8. http://support.sas.com/publishing/pubcat/chaps/65646.pdf

9. pivotal.io/data-science-pivotal/features/text-analytics-and-natural-language

10. http://www.datacommunitydc.org/blog/2013/05/big-data-and-nlp/