Deep Analysis of Textual Data in Multiple formats using Hadoop Techniques

[1] ·K Sivaramakrishna, [2] K.Srinivasarao, [3] BV Satish
[123]*Asst.prof., ALIET*

***Abstract* — *the analysis of different types of text content in sending mails, social online journals, messages, gatherings and different types of printed correspondence constitutes what we call content analysis. Content analysis is material to most businesses: it can help divide a great of many messages; you can break down client's remarks and inquiries in gatherings; you can perform assessment investigation utilizing content investigation via evaluating productive or depressing impression of an organization, variety, otherwise product. Content scrutiny has likewise considered as content extraction, and is a subset of the Accepted Communication Handling (ACH) background, identified as the establishing twigs of simulated intellects, when an enthusiasm for understanding content initially created. Right now Content Investigation is frequently measured as the following stride in Big Data investigation. Content Investigation has various subsets: Content Extraction, Named Individual Identification, Semantic network commented on area's portrayal, and some more. A few methods are right now utilized and some of them have picked up a great deal of consideration, for example, Machine Learning, to demonstrate a semi supervised improvement of frameworks, yet they additionally introduce various restrictions which make them not generally the main or the best decision*

A wide range of machine robotized frameworks are producing extensive measure of information in various structures like truthful data, text content, and bio-metric information that develops the term Big Data. In this Research article we are exaextraction issues, difficulties, and use of these sorts of Big Data with the thought of enormous information measurements. Here we are talking about online networking information analysis, content based analysis, content information analysis, their issues and expected application zones. It will inspire scientists to address these issues of capacity, administration, and recovery of information known as Big Data.

Keywords* — *Big Data Investigation, content extraction, Textual Investigation, Information Measurements.

I. INTRODUCTION

Accepted Communication Handling (ACH)) is the reasonable field of Computational Linguistics, albeit a few creators utilize the terms reciprocally. Once in a while ACH has been viewed as a associate authority of simulated Intelligence, and all the more as of late it assembles at the center of Cognitive Calculations, since most psychological procedures are moreover comprehended or produced as characteristic dialect expressions.[2]

ACH is an extremely expansive theme, and incorporates a gigantic measure of subsets: Natural Speech Consideration, Natural Speech Creation, Knowledge Base Construction, Dialog Handling organizations (and intellectual Coach Methods in scholastic erudition frameworks), Speech handing, Data extraction – Text Extraction – Text Analysis, et cetera. It is processed in this manuscript in Text Investigation (TA).

Content Investigation is considered as the latest identifier given to Natural Speech Identification, Data and Text Extraction. Over the most recent couple of years another name has picked up prominence, Big Data, to allude for the most part to unstructured content (or other data sources), more regularly in the business instead of the scholarly territory, likely in light of the fact that unstructured free content records for 75% in a production setting, which includes tweets, online journals, wikipedias and reviews [1].No scholastic papers are covering this ideas is a fact.

II. DISCERNMENT AND METHODOLOGIES FOR TEXTUAL CONTENT ANALYSIS:

Content Investigation is an augmentation of information extraction, that is used to discover printed designs from enormous non-organized resources, rather than information put away in social databases. Content Investigation, otherwise called Smart Content Analysis, Data Extraction or Knowledge-Discovery in Text (KDT), alludes by and large to the way toward removing non-minor data and information from unstructured content. Content Investigation is like information extraction, with the exception of that information extraction apparatuses are intended to deal with organized information from databases, either put away all things considered or accordingly from processing them with unstructured information. Content Investigation can cover shapeless or semi-organized informational indexes, for example, messages, full-content records and HTML documents, online journals, daily paper articles, scholarly papers, and so on.

comprises of the accompanying strides and undertakings [6]

Beginning with a gathering of records, a content extraction device recovers a specific report and preprocess them by identifying arrangement and quality sets. At that point it would experience a content examination stage, infrequently rehashing strategies until data is extricated. The fundamental methodology in every one of the parts is to discover an example (from either a rundown or a past procedure) which coordinates a control, and afterward to apply the lead which clarifies the content. Every segment plays out a specific procedure on the content for example sentence division

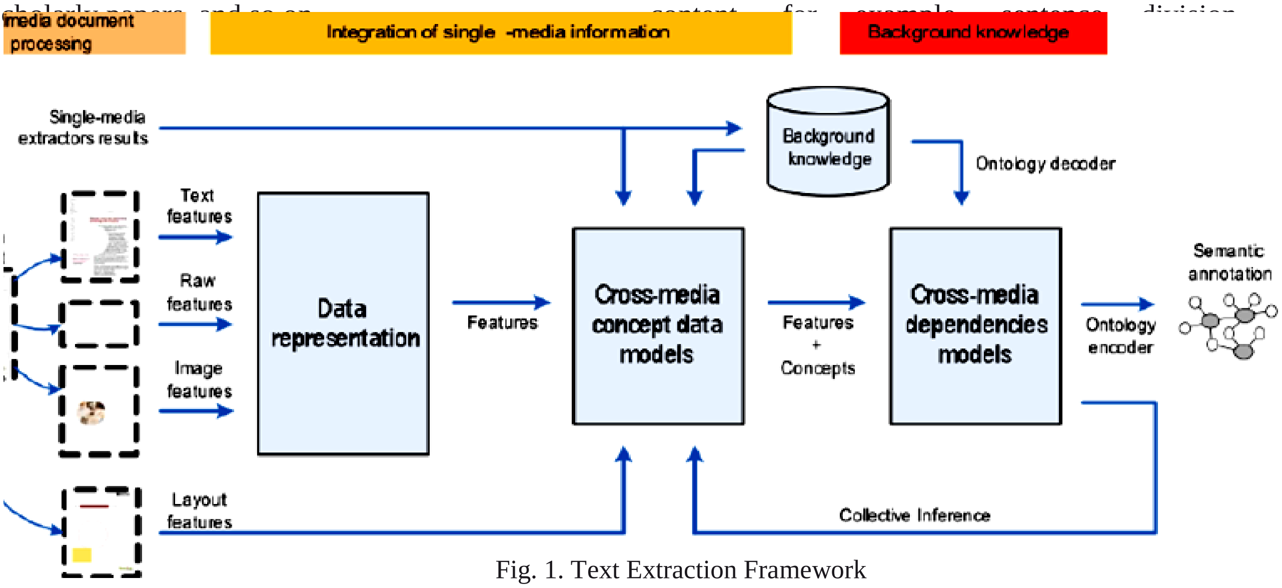


Fig. 1. Text Extraction Framework

Content Investigation is an inter penalizing background which depicts on data extraction, information extraction, machine erudition, insights and calculational etymology[1].

Content Investigation is picking up noticeable quality in numerous businesses, from advertising to back, on the grounds that the way toward extricating and dissecting extensive amounts of content can assist chiefs with understanding business sector flow, foresee results and patterns identify extortion and oversee chance. The multidisciplinary way of Text Investigation is vital to comprehend the unpredictable incorporation of various skills: PC engineers, language specialists, specialists in Law, BioMedicine or Finance, information researchers, analysts, bringing about that the innovative work approach is divided because of various customs, strategies and interests. A regular content examination application

The subsequent procedure gives "organized" or semi-organized data to be additionally utilized (e.g. Information Foundation Edifice, Ontology enhancement, and Contraption Knowledge calculation approval, uncertainty indicators for Problem and Response frameworks[8]).

A portion of the systems that has produced and utilized as a part of the content extraction procedure are data extraction, theme following, synopsis, arrangement, grouping, idea linkage, data perception, address replying, and profound learning.

A. INFORMATION EXTRACTION

Data extraction (IE) programming distinguishes key terminologies and associations inside contented. It does it by identifying for already defined arrangements in available information, a procedure more often than not called design coordinating, commonly in view of customary indications. The most mainstream type of IE is a substance acknowledgment (NER). NER tries to find and characterize nuclear components in content into preinstructed classes (generally coordinating preestablished ontologies). NER procedures remove elements, for example, the names of people, associations, areas, transient or spatial expressions, amounts, fiscal qualities, stock qualities, rates, quality or protein names, and so forth. These are a few apparatuses applicable to the undertaking: Apache OpenACH [2], Stanford Named Individual Identifier [3] [4], LingPipe [5].

B. THEME IDENTIFICATION AND DETECTION

Watchwords are an arrangement of noteworthy terms in a research manuscript which results an abnormal state portrayal of its substance to perusers. Distinguishing catchphrases from a lot of news information is exceptionally valuable which inturn can create a short outline of news articles. As content which is available online archives quickly increment in size with the development of WWW, watchword mining [6] has turned into the premise of a few content extraction applications, for example, web indexes, content order, rundown, and theme location. Manual watchword extraction is a to a great degree troublesome and tedious errand; truth be told, it is practically difficult to concentrate catchphrases physically if there should be an occurrence of news reports distributed in a solitary date because of their volume.

A point following structure facilitates by keeping client summary and, in light of the available reports the user sees, calculates dissimilar records which are important to the client. Google offers a free subject following apparatus [7] that enables clients to pick watchwords and advises them when news identifying with those points ends up plainly accessible. NER procedures are additionally utilized as a part of upgrading point following and recognition by coordinating names, areas or regular terms in a given subject by speaking to similitudes with different archives of comparative substance [8]. Theme recognition is firmly related with Classification (see underneath).

C. CONTENT ILLUSTRATION

Content outline has a extended and productive convention in the concept of Content Investigation. It might be said content synopsis falls additionally under the class of Accepted Communication Handling (ACH). It helps in making sense of regardless of whether a elongated report which represents the customers troubles and qualities inspecting for supplementary data. With vast writings, content synopsis forms and outlines the record which is considered with client to peruse the main section. The way to synopsis is to decrease the extention and description of a report along with holding its principle focuses and general importance.

The methodologies mainly utilized by content rundown devices are stretch extraction. Imperative verdicts in an research article are factually subjected and positioned. Outline instruments may likewise intended for headings and different indicators of subtopics with a specific end goal to distinguish the key purposes of a report.

The techniques for outline can be ordered in two general gatherings:

• Shallow investigation, limited to the structured level of portrayal and attempt to extricate critical parts of the content;

• Deeper investigation, accept a semantics level of portrayal of the first content (regularly utilizing Information Retrieval strategies).

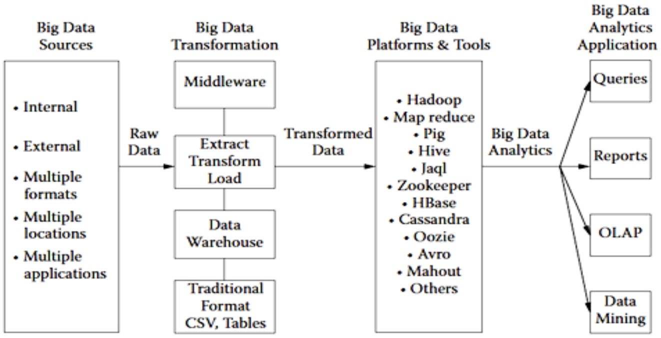


Fig. 2. Content Summarization

A. ARRANGEMENT OR CLASSIFICATION

Arrangement includes distinguishing the fundamental subjects of an archive by putting the report into a predescribed location of points (moreover as scientific classifications or ontologies). Order just tallies words that show up and, from the checks, distinguishes the primary themes that the archive covers. Classification frequently depends on connections recognized by searching for wide provisos, smaller provisos, comparable words, and associated terms. Arrangement instruments regularly have a strategy for positioning the archives all together of which reports have the most substance on a specific point [10]. Another technique is to speak to subjects as topical diagrams, and utilizing a level of closeness (or separation from the "reference" chart) to order archives under a given class [11].

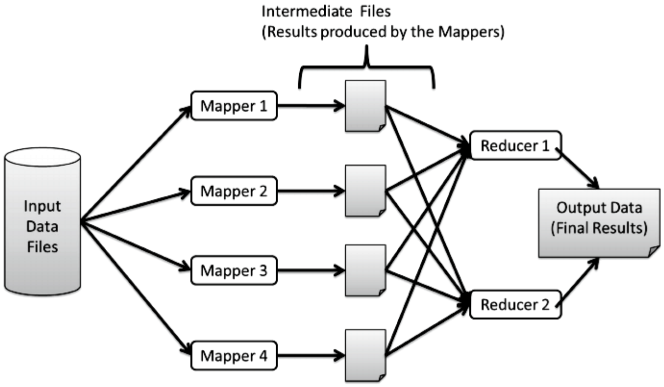


Fig. 3. Content Classification

B.DATA ASSEMBLING

Bunching method is wormed to gather comparable archives, yet it varies from order in that reports are grouped without the utilization of predefined themes. At the end of the day, while order suggests administered (machine) learning as in past information is utilized to dole out an offered archive to a given class, bunching is unsupervised learning: there are no already characterized points or classifications. Utilizing grouping, archives can show up in different subtopics, in this way guaranteeing a helpful record won't be precluded from query items (numerous ordering references). An essential grouping calculation makes a vector of points for each report and doles out the archive to a given theme bunch

C.TEXT INVESTIGATION IDENTIFIED DISTURBANCES

With regards to TA, Big Data is just a monstrous volume of composed dialect information. Be that as it may, wherever does the wilderness recline among analytics of Big Data and Small Data. At hand a process is evolving truth: while just 15 years prior a content capacity of 160 million words was viewed as colossal, at present no less than 8.500 million utterance datasets are accessible. Niether exclusively which questions essentially about its extent, additionally about excellence and genuineness: information from online networking is brimming with clamor and contortion. All data clusters have these issues however they are all the more possibly genuine for extensive datasets in light of the fact that the PC is a mediator and the human master don't see them specifically, similar to the case in little datasets. In this manner, information purging procedures devour huge endeavors and frequently after the purifying, the accessibility of data to prepare frameworks is insufficient to get solid expectations, as occurred in the Google Flu Movements fizzled analyze [27].

The motivation behind is that numerous enormous data clusters are not the yield of utensils intended to deliver substantial and solid information for examination, and furthermore in light of the fact that information purging is about (for the most part subjective) choices on the applicable plan highlights. Another key issue is the entrance to the information. Much of the time, the scholastic gatherings have no entrance to information from organizations, for example, Google, Twitter or Facebook. For example, Twitter just makes a little part of its tweets accessible to the general population with its APIs. Also, the tweets accessible don't take after a given example (they are a "different pack") so it is hard to land at a decision relating to their indications. As an outcome, the imitation of investigations is practically inconceivable, while the underneath resources and the hidden innovation are not freely accessible. Boyd and Crawford [29] go advance: restricted admittance to Big Data makes new computerized isolates, the Big Data rich and the Big Data poor. Customer needs the way to gather them, and the mastery to break down them. Strikingly, little yet very much curated accumulations of dialect information (the conventional corpora) offer data that can't be deduced from huge datasets [30].

Step by step instructions to get a handle on the metaphorical employments of dialect, fundamentally incongruity and similitude, is likewise an outstanding issue to legitimately comprehend content. Basically, the client's expectations are concealed on the grounds that the surface importance is distinctive to the basic significance. As an outcome, the terms have to be deciphered in setting and through additional etymological information, a reality that has been difficult for people, which is significantly complex for equipment. Step by step instructions to make an interpretation of a given allegory into another dialect is to a great degree troublesome. A few assessments ascertain that allegorical dialect is around 15-20% of the aggregate substance in web-based social networking discussions.

III.PROPOSED METHODOLOGY

Step 1. Inputrequest ought to as .JAR document t which includes Driver source code, Mapper relative code and Reducer code.

Step 2. Job hunter doles out the mapper errands by following the business foundation from the .JAR document on the all the accessible assignment handlers.

Step 3. when all the assignment trackers are finished with mapper processes, they propel a similar status back to Job Tracker.

Step 4. With all the errand trackers do with mapper stage, at that point work tracker starts sort and rearrange stage on all the mapper yields.

Step 5. After finishing sort and rearrange, work hunter starts
reducer stage on all accessible assignment hunters.

Step 6. If all assignment hunters do with compressing stage, they refresh a similar condition back to the occupation trailer.

Mapper and Reducer are client driven stages. Mapper class yield filename is "part-m-00000" and Reducer class yield filename is "part-r-00000". Work Tracker and Task Tracker are the two daemons which are totally in charge of MapReduce handling.

IV.EXAMPLES OF TEXTUAL INVESTIGATION APPLICATIONS

We will quickly audit two conspicuous territories of use of Content Investigation, with a substantial business affect: (1) Medical Investigation – grouping of research articles of medicinal substance, and (2) Legal Investigation – Information retrieval from legitimate writings.

A. Health Check Investigation – Categorization of research manuscripts or therapeutic substance Biomedical content extraction or BioACH exhibits some interesting information sorts. Their average writings are digests of logical papers, and additionally therapeutic reports. The primary errand is to group papers by a wide range of classifications, to bolster a database (like MEDLINE). Different applications incorporate ordering records by ideas, normally based or identified with ontologies or representing "translational research," that is, utilizing essential organic investigation to educate scientific practice (for example, naturally extraction of medication cooperations, or quality relationship with infections, or transformations in proteins).

The ACH strategies incorporate biomedical elements acknowledgment, design acknowledgment, and mechanism learning for extricating structured relations with ideas. Biomedical elements acknowledgment comprises of perceiving and ordering substance identities in biomedical spaces, for example, proteins, qualities, illnesses, medications, organs and medicinal claims to fame. An assortment of lexical assets are accessible in English and different dialects, and additionally a wide gathering of commented on corpora (as GENIA) with structured and applied associations between substances. In spite of their accessibility, no single asset is sufficient nor exhaustive since new medications and qualities are found continually. This is the fundamental test for BioACH.

Totally three methodologies for extricating relations between elements are available:

- Linguistic-dependent methodologies: the thought is to utilize parsers to get a handle on meaningful structures and guide them into structured portrayals. They are normally in light of lexical assets and their principle disadvantages are the wealth of equivalent words and spelling varieties for substances and ideas.

- Pattern-based methodologies: these strategies make utilization of an arrangement of examples for potential connections, characterized by space specialists.

- Machine Learning-based methodologies: from commented on writings by human specialists, these strategies extricate relations in new accumulations of comparable writings. Their principle weakness is the prerequisite of calculating cost and preparing and testing on a lot of manual-labeled information. To enlarge the mining framework to other kind of information or dialect involves new individual exertion in explanation.

Friedman et al. proposed a study on the best in class and projected in BioACH, supported by the US National documents of drugs. This testimony recognizes that "the most huge frustrating component for quantifiable ACH is unavailability of vast scale de-distinguished quantifiable structure, which are required for preparing and assessment."

B. Authorized Investigation – Knowledge mining from legitimate writings .Individual territory receiving a considerable measure of consideration concerning the items of common sense of Content Investigation is that regarding to the data retrieving from writings with legitimate substance. All the more particularly, suit information is brimming with references to judges, legal advisors, parties (organizations, open associations, et cetera), and licenses, assembled from a few a large number of pages containing a wide range of Intellectual possessions (IP) case data. This has offered ascend to the expression authorized Investigation, since examination assists in finding designs with importance covered up in the archives of information. What it intends to legal advisors is the blend of bits of knowledge originating from base up information with high-low expert and knowledge originated in statutes, directions and incite sentences. This spots target information at the middle rather than the purported narrative information.

The fundamental issue is that lawful literary data is communicated in normal dialect. While a scan can be made for the string offended party, there are no looks for a string that speaks to a person who tolerates the part of offended party. To make dialect on the maze more important and organized, extra substance must be combined to the base substance, where the Semantic Web (semantic parts' labeling) and Natural Language Processing play out their commitment.

We begin with information, the corpus of writings, and afterward a yield, writings commented on with XML labels, JSON labels or different systems. Nonetheless, getting from a corpus of printed data to clarified yield is a requesting undertaking, blandly alluded to as the information procurement bottleneck [43]. This errand is extremely requesting on assets (particularly labor with enough mastery to prepare the frameworks) and it is likewise exceptionally information escalated since whoever is doing the comment must comprehend what and how to comment on learning identified with a given space.

Handling Natural Language (NL) to bolster such luxuriously commented on records displays some innate issues. NL underpins the greater part of the accompanying, in addition to additional equipments:

(1) Inherent or Assumed data – "When did you quit taking medications?" (Assumes that the individual is addressed regarding taking medications sooner or later previously);

(2) The same frame with various logically subordinate implications .Normally a legitimate examination framework will comment on components of enthusiasm, so as to distinguish a scope of specific snippets of data that would be pertinent to lawful experts, for example,

- Case reference
- Names of gatherings
- Roles of gatherings, which means offended
- Identities of judges
- Identities of lawyers
- Roles of lawyers, which means the side they speak to

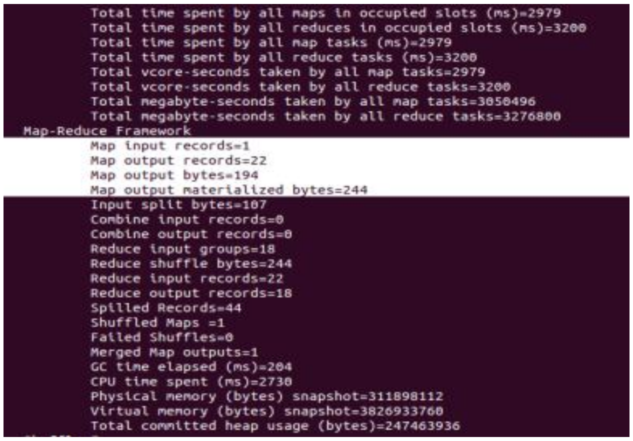
  (Offended party or litigant)
- Final choice
- Cases referred to
- Nature of the case, which means utilizing watchwords to group the case as far as subject (e.g., criminal attack, protected innovation, and so on.)

The business ramifications of authorized Investigation have begun a packed branch of literary Big Data submissions. A few organizations have profited from a beneficial souk, for example, LexisNexis, concentrated on contribution of expectations on potential medicinal negligence issues to particular lawyers. In recent times, LexisNexis has gained Lex Machina [44], an organization that mines primarily case information around IP data.

Consistently, Lex Machina's crawler separates information (and records archives) from a few U.S Law vaults. The flatterer naturally catches each docket occasion and downloads input case archives. It changes over the archives by visual disposition acknowledgment (OCR) to identified content and supplies every one as a PDF document. At the point when the flatterer finds a specify of a copyright, it gets data related to the copyright from the copyrights and Trademarks Office (CTO) site. The flatterer conjures Lexpressions, an exclusive lawful content order motor. The ACH innovation groups cases and dockets and determines substance names (utilizing a NER motor). A procedure of segementation of the data separated is achieved by specific lawyers to guarantee great information. The organized content indexer then plays out an information purging operation to request every one of the information and stores it for hunt. Lex Machina's electronic submission empowers clients to run look questions that convey simple data recovery of the significant docket passages and records.

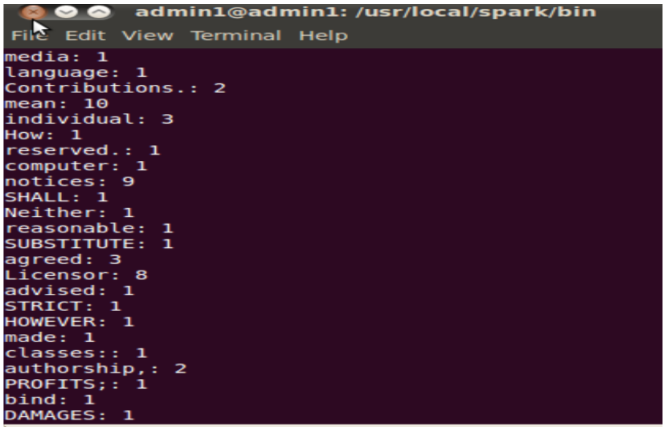
V. RESULTS & DISCUSSIONS

MapReduce execution for Word Count algorithm computation.The result is the rundown of words with the number of emergence of every utterance. The below figure shows the pattern of research data on the local Apache Hadoop. The inquire about pattern indicates,the the majority of the examination ponders are done specifically spaces. The research work demonstrates that in Zoology space, the greater part of centered research zones are vermin, composition, Cestode, and Fishes. It demonstrates that in the Botany space, predominantly look into focus is on Airospora, crop harvest, Fungi and Aerobiological. Results are conceivable utilizing a MapReduce WordCount calculation.



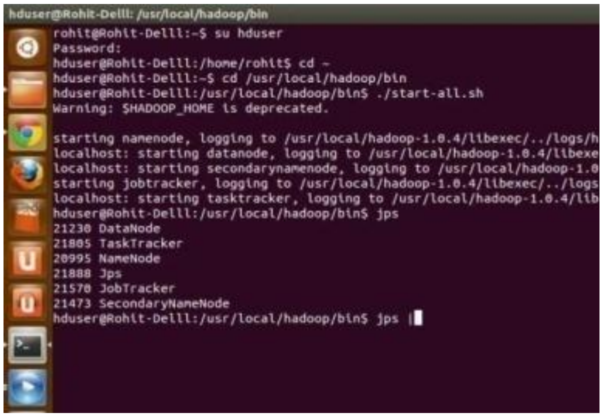
```
        Total time spent by all maps in occupied slots (ms)=2979
        Total time spent by all reduces in occupied slots (ms)=3200
        Total time spent by all map tasks (ms)=2979
        Total time spent by all reduce tasks (ms)=3200
        Total vcore-seconds taken by all map tasks=2979
        Total vcore-seconds taken by all reduce tasks=3200
        Total megabyte-seconds taken by all map tasks=3050496
        Total megabyte-seconds taken by all reduce tasks=3276800
  Map-Reduce Framework
        Map input records=1
        Map output records=22
        Map output bytes=194
        Map output materialized bytes=244
        Input split bytes=107
        Combine input records=0
        Combine output records=0
        Reduce input groups=18
        Reduce shuffle bytes=244
        Reduce input records=22
        Reduce output records=18
        Spilled Records=44
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=204
        CPU time spent (ms)=2730
        Physical memory (bytes) snapshot=311898112
        Virtual memory (bytes) snapshot=3826933760
        Total committed heap usage (bytes)=247463936
```

Fig.4 Performing Mapreducing operation


admin1@admin1: /usr/local/spark/bin
File Edit View Terminal Help

```
media: 1
language: 1
Contributions.: 2
mean: 10
individual: 3
How: 1
reserved.: 1
computer: 1
notices: 9
SHALL: 1
Neither: 1
reasonable: 1
SUBSTITUTE: 1
agreed: 3
Licensor: 8
advised: 1
STRICT: 1
HOWEVER: 1
made: 1
classes:: 1
authorship,: 2
PROFITS;: 1
bind: 1
DAMAGES: 1
```

Fig.5 Result of text analysis



```
rohit@Rohit-Delll:~$ su hduser
Password:
hduser@Rohit-Delll:/home/rohit$ cd ~
hduser@Rohit-Delll:~$ cd /usr/local/hadoop/bin
hduser@Rohit-Delll:/usr/local/hadoop/bin$ ./start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop-1.0.4/libexec/../logs/h
localhost: starting datanode, logging to /usr/local/hadoop-1.0.4/libexe
localhost: starting secondarynamenode, logging to /usr/local/hadoop-1.0
starting jobtracker, logging to /usr/local/hadoop-1.0.4/libexec/../logs
localhost: starting tasktracker, logging to /usr/local/hadoop-1.0.4/lib
hduser@Rohit-Delll:/usr/local/hadoop/bin$ jps
21230 DataNode
21805 TaskTracker
20995 NameNode
21888 Jps
21570 JobTracker
21473 SecondaryNameNode
hduser@Rohit-Delll:/usr/local/hadoop/bin$ jps |
```

Fig.6 Hadoop checking

VI. FUTURE WORK

The innovations around content investigation are right now being connected in a few enterprises, for example, assumption and sentiment examination in media, back, medicinal services, promoting marking or buyer markets. Experiences are extricated from the conventional endeavor information sources, as well as from on the web and web-based social networking, since more the overall population has ended up being the biggest generator of content substance (simply envision web based informing frameworks like Whatsapp or Telegram).

The present condition of content investigation is extremely sound, however there is space for development in zones, for example, client familiarity, or societal regulation in. This bears great guarantees for both logical trialing and specialized development alike: Multi-lingual investigation is encouraged by mechanism learning (ML) and progress in mechanism interpretation; client encounter, statistical surveying, and buyer bits of knowledge, and computerized examination and media estimation are upgraded through content examination; other than the eventual fate of profound learning in ACH, since quite a while ago settled dialect building approaches scientific categorizations, parsers, lexical and semantic systems, and syntactic-administer frameworks will proceed as bedrocks in the zone; feeling investigation, emotional states intensified of discourse and content and additionally pictures and outward appearance investigation; new types of supratextual interchanges like emojis need their own way to deal with concentrate semantics and touch base at significant examination; semantic pursuit and information diagrams, discourse examination and synchronous machine interpretation; and machine-composed substance, or the ability to create articles (and email, instant messages, rundowns, and interpretations) from content, information, principles, and setting, as caught beforehand in the examination stage.

VII.CONCLUSION

Content Investigation, with its extended and impressive track, is a region in steady advancement. It stands at the focal point of Big Data's assortment vector, that of amorphous data, particularly with societal interchanges, where substance is created by a huge number of clients, substance comprising of pictures as well as the vast majority of the circumstances literary remarks or all out articles. Data communicated by methods for writings includes heaps of learning about the humanity and regarding the substances in this humanity and in addition the connections among them. That learning about the world has as of now considered to use so as to make the subjective applications, similar to IBM's Watson and IP soft's Amelia, that will communicate with people growing their capacities and helping them perform better. With expanded correspondence, Text Investigation will be extended and it will be expected to deal with the commotion and the insignificant from the truly critical data.

VIII. REFERENCES

[1] Xerox Corporation (2015): *http://www.xrce.xerox.com/Research-Development/Industry-Expertise/Finance* (accessed 26 December 2015)

[2] Apache Opennlp (2015): http://opennlp.apache.org/ (accessed 19 December 2015)

[3] Doug cutting, Marco nicosia, "About Hadoop" http://lucene.apache.org/Hadoop/about.html.

[4] J. R. Finkel, T. Grenager, and C. Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). (online reading*: http://nlp. stanford.edu/~manning/papers/gibbscrf3.pdf*)

[5] Chakraborty, G., Pagolu, M. & Garla, S (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Publishing.

[6] S. Lee and H. Kim (2008). "News Keyword Extraction for Topic Tracking". Fourth International Conference on Networked Computing and Advanced Information Management, IEEE.

[7] Google Alerts (2016): http://www.google.com/alerts (accessed 10 January 2016)

[8] Seung Jin sul, AndreyTovchigrechko, "Parallelizing BLAST and SOM algorithms with Mapreduce-MPI library" 25th IEEE International Symposium on Parallel and Distributed Processing, IPDPS, 2011. [9] ATLAS Project (2013): http://www.atlasproject.eu/atlas/project/task/5.1 (accessed 10 January 2016)

[10] G. Wen, G. Chen, and L. Jiang (2006). "Performing Text Categorization on Manifold". 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, IEEE.

[11] H. Cordobés, A. Fernández Anta, L.F. Chiroque, F. Pérez García, T. Redondo, A. Santos (2014). "Graph-based Techniques for Topic Classification of Tweets in Spanish". *International Journal of Interactive Multimedia an Artificial Intelligence*.

[12] T. Theodosiou, N. Darzentas, L. Angelis, C.A. Ouzonis (2008). "PuReD- MCL: a graph-based PubMed document clustering methodology". *Bioinformatic*s 24.

[13] Q. Lu, J. G. Conrad, K. Al-Kofahi, W. Keenan (2011). "Legal document clustering with built-in topic segmentation", Proceedings of the 20th ACM international conference on Information and knowledge management.

[14] P. Cowling, S. Remde, P. Hartley, W. Stewart, J. Stock-Brooks, T. Woolley (2010), "C-Link Concept Linkage in Knowledge Repositories". AAAI Spring Symposium Series.

[15] C-Link (2015): *http://www.conceptlinkage.org/* (accessed 10 December 2015)

[16] Y. Hassan-Montero, and V Herrero-Solana (2006). "Improving Tag-Clouds as Visual Information Retrieval Interfaces", I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006.

[17] Wordle (2014): http://www.wordle.net/ (accessed 20 December 2015) [18] M. A. Hearst (2009) "Information Visualization for Text Analysis", in *Search User Interfaces*. Cambridge University Press (online reading: http://searchuserinterfaces.com/book/)

[19] D3.js (2016): *http://d3js.org/* (accessed 20 January 2016)

[20] Gephi (2016) *https://gephi.org/* (accessed 20 January 2016)

[21] L. Hirschman, R. Gaizauskas (2001), "Natural language question answering: the view from here", Natural Language Engineering *7*. CambridgeUniversityPress [22] OpenEphyra

[23] N. Schlaefer, P. Gieselmann, and G. Sautter (2006). "The Ephyra QA system". *2006* Text Retrieval Conference (TREC).

[24] YodaQA (2015): *http://ailao.eu/yodaqa/* (accessed 5 January 2016)

[25] P. Baudis (2015) "YodaQA: A Modular Question Answering System Pipeline". POSTER 2015 — 19th International Student Conference on Electrical Engineering. (online reading: *http://ailao.eu/yodaqa/yodaqa- poster2015.pdf*)

[26]DL4J (2015*): http://deeplearning4j.org/textanalysis.html* (accessed 16 December 2015)

[27]Google–Word2vec(2013): *http://arxiv.org/pdf/1301.3781.pdf* (accessed 20 December 2015)

[28] D. Lazer, R. Kennedy, G. King, and A. Vespignani (2014). "Big data. The parable of Google Flu: traps in big data analysis." Science, 343(6176).

[29] D. Boyd, and K. Crawford (2011). "Six Provocations for Big Data". A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. (Available at SSRN: *http://ssrn.com/abstract=1926431 or http:// dx.doi.org/10.2139/ssrn.1926431*)