International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

Web based Content Extraction and Retrieval in Web Engineering

C.H. Sarada Devi, Dr.T. Kumanan,
Prameela Devi. Chillakuru

Abstract--- The fast and wide-ranging pervasion of data and information over the web possess a high dispersion of an enormous capacity of normal language textual possessions. Excessive attention has been evolved in the existing scenario for determining, distribution and retrieving of an enormous source of knowledge. For this purpose, processing enormous data capacities in a sensible time frame is an important challenge and a vital necessity in numerous commercial and exploration fields. Computer clusters, distributed systems and parallel computing paradigms are being progressively applied in the current years; subsequently they presented important developments for computing presentation in data-intensive contexts, like Big Data mining and analysis. NLP is one of the significant features which can be utilized for text explanation and first feature extraction from request area with high computational supplies; therefore, these responsibilities can have advantage over similar architectures. This study shows a discrete framework for running NLP tasks in a parallel fashion and crawling web documents. The system was found on Apache Hadoop environment, and on its equivalent programming paradigm, called MapReduce. Authentication is done using the explanation for extracting keywords and critical phrase from the web documents in a multi-node Hadoop cluster. The results of the proposed work shows increased storage capacity, increased speed in data processing, reduced user searching time and receives the accurate content from the large dataset stored in HBase.

Keywords--- Natural Language Processing, Hadoop, Text Parsing, Web Crawling, Big Data Mining, HBase.

I. INTRODUCTION

Big Data is well-defined as a group of datasets in which the size includes the complication to control and dispensation of information that brings challenges in using traditional tools (i.e., management of N-dimensional data circles by necessary text files and SQL files). Those issues provide Big Data monoliths [1] as abundant as the systems of secondary information (Pollock, 2013a) producing primary failure for greatest isolated and public data providers in which lesser amount of data do not include simpler management (Akers, 2013)[2]. There are several techniques used for the extraction of information using NLP. One of the best techniques used in information extraction from the text data is Text mining (Pande et al., 2016)[3]. Extraction of information is evaluated in terms of certain factors such as precision, slot error rate, F-measure and Recall. Eruption of efficient information gives rise to the technology of Information Retrieval (IR) and Information Extraction (IE) [4] (Sonit Singh, 2018). Natural language is taken as an input in IE systems, and it produces structured information specified by certain criteria which can be applied to a particular application. Data mining methods manage capacious datasets to mine major patterns from information; social networking sites offer huge amount of datasets for their practice and therefore they become best candidates to mine information with the tools used for data mining (Charu Virmani et al., 2017)[5]. Therefore, web mining or data mining offers crucial intelligence to a social network in order to create and relate accurately in user-friendly fashion. Knowledge representation occurs in web data mining based on Natural Language Processing described by (Yue Chen, 2010)[6]. In this, 400 sentences are extracted using Web new corpus to form the corpus test, and this type of information consists of a description of events, relationships, and object attributes. This model epitomizes the web document at the semantic level, and its knowledge structure has resilient scalability.

NLP is a theory-motivated range of computational methods targeted at programme examination and demonstration of human language. It enables computers to achieve an extensive variety of ordinary language associated tasks, at all stages, extending from analyzing and part-of-speech (POS) classification to machine conversion in addition to dialogue schemes. Deep learning designs in addition to procedures contain previously completed imposing signs of progress in domains such as processor vision and design gratitude. After this tendency, the current NLP investigation gradually concentrates on the application of innovative deep learning approaches (Mikolov et al., 2010; Mikolov et al., 2013)[7, 8] and deep learning approaches (Socher et al., 2013)[9]. Deep learning permits multilevel programmed feature representation learning. In contrast, traditional machine learning based NLP systems lies deeply on hand-crafted structures. Certain hand-crafted arrangements are time consuming and frequently imperfect.

II. RELATED WORK

Barbosa et al., (2015) deliberated different methods for the extraction of information [10]. And also explained the knowledge inferencing which comprises a combination of multiple types of data sources and extraction methods for verifying the existing and current knowledge. There are four inferencing techniques namely deep natural language processing using machine learning techniques, large-scale probabilistic reasoning, data cleaning using integrity techniques, and leveraging human expertise for domain knowledge extraction.

Manuscript received September 16, 2019.

C.H. Sarada Devi, Research Scholar, Meenashi Academy of Higher Education and Research, Chennai, India. (e-mail: saradhadevi6@gmail.com)

Dr.T. Kumanan, Professor/CSE, Meenashi Academy of Higher Education and Research, Chennai, India. (e-mail: kumananvetri@gmail.com)

Prameela Devi. Chillakuru, Research Scholar, Meenashi Academy of Higher Education and Research, Chennai, India. (e-mail: prameela246@gmail.com)

Retrieval Number: B10130982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1013.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

So, those methods can be used in various systems for the extraction of information in web sentences. Chandurkar et al., (2017) discussed a Question Answering (QA) system which combines the fields of Information Retrieving (IR) and NLP [11]. The QA system was based on target subject and question from the UI (User Interface). Here, the TREC 2004 dataset is taken to encode the XML document. The user can give his/her own question by adding a text box in the system. The precise DBpedia page is based on target subject, and Standford Dependency Parser which integrates the other routine for extracting the target or focusing subject in the question. But it totally depends on the python factoid question classifier.

Florence et al., (2015) proposed a system called summarizer which depends on semantic analysis of web documents [12]. The Cluster values are used by the system in Resource Description Framework (RDF) space to construct the summaries. This system can produce short and long summaries, based on the original length of the documents. At last, the clustering algorithm extracts subject, Object, and Verb (SOV), and it can be grouped together using extracted RDF triples. Then, for the final summary, the selected valuable sentences are extracted using Sentence Selection (SS) algorithm. Collobert et al. (2011) established that a modest in-depth learning outline overtakes most advanced procedures in numerous NLP tasks like Semantic Role Labelling (SRL), Named-Entity Recognition (NER) and POS tagging. Mean while, many complex deep learning founded procedures were planned to resolve strict NLP responsibilities [13].The study by Goldberg, (2016), simply accessed the fundamental ideologies focused on smearing neural networks to NLP in a tutorial method. Researchers believed that their work would give readers a more widespread impression of present researches in this domain [14].Azcarraga et al., (2012), examined that NLP methods are the base of language founded solutions which usually offer more specific consequences by statistical approaches; though, they are computationally very exclusive [15].

Pavithra et al., (2013) examined wrapper induction procedure to excerpt the data. Combination of XSLT in addition to DOM by means of XML are used to advance the system. XML procedures are very effective for applications which were based on web. As per the content, this technique offered efficient information extraction and chains the adjustment of wrappers. It improved both flexibility and usability [16].Sridevi et al., (2016) suggested a Viterbi process for building medical corpus data including the information extraction from the clinical text based on background that helped the clinician to prefer faster and enhanced verdicts. It also enhanced quality of the treatment [17].Wu Wei et al. (2013) proposed a novel extraction rule language which expressed the combined logic for data integration, direction-finding regarding information extraction and web page. A source data object is used to exemplify and wrap a web data region together with information records as well as items. The system allows the users to define complex target data entity through XML and it describes the target data entity structure and formerly approves integration scripts to convert and map information that is extracted from source to target data objects [18].

Jindal et al. (2013) technologically advanced an equivalent NLP system created on Learning Based Java (LBJ) model [19] Rizzolo and Roth, (2010), in addition to utilizing Charm++ by similar software design model [20]. Exner and Nugues, (2014) observed the Koshik multi-language NLP platform which has been intended for extreme scale-processing including the examination of unstructured natural linguistic documents dispersed upon a cluster which was contingent on Hadoop [21]. It supports numerous types of algorithms, like text tokenization, dependency parsers, and co-reference solver. Utilizing a Hadoop dispersed architecture, and its Map Reduce program design model is a competence benefit to professionally and effortlessly scale by addition of low-cost commodity hardware to the cluster. Barrio and Gravano, 2017 elucidated the extraction of information schemes which determine classified information in natural language script. Knowing controlled procedure permits abundant more affluent inquiring besides data mining compared to probable completed native linguistic text. Nevertheless, extraction of a material is a computationally costly mission besides enlightening a competence of an extraction procedure with text groups of precarious attention. The research concentrates on a particularly respected family of text groups, individually, the so-called deep-web text groups, whose insides aren’t crawl-able and are merely obtainable through enquiring. There is a very significant step for effective material extraction over deep-web text groups [22].Wang and Stewart 2015 studied geographic information science, modeling geographic dynamics found on spatiotemporal material mined from a Web, particularly unconstructed facts like online news reports. Consideration of spatiotemporal besides semantic data from a group of Web forms allows us to shape a rich exemplification of geographic details labeled in the text, taking where, when, or what proceedings have happened. This work inspects the part ontologies performance as a vital constituent in a procedure of semantic material abstraction. They showed the means by which ontologies could be utilized by combining with usual linguistic gazetteers in instruction to grammatical procedure material around danger spatiotemporal and proceedings supplement abstraction with semantics [23].

Match and Avdan, (2018) discussed a technique to remove difficulties (like typographical errors, spelling mistake, and incorrect arrangements) which disturb the attainment degree of a geocoding procedure. In current technique, the data of address is analyzed utilizing NLP approaches. Misspellings, abbreviations or omissions are impassive through both Match Rating Compute Method and Levenshtein Distance Procedure. As an outcome, the addresses are reorganized into an exact arrangement. To compute the technique belongings on consequences of geocoding procedure. A test dataset containing addresses of primary schools in Eskishehir is shaped.

The geocoding procedure is approved out with current sample of addresses set, both earlier and then the calibration procedure is useful [24].

Glauber and Claro, 2018 explained how to ensure all applicable primary research.

Even by smearing our inquiry threads in 5 files, merely an EMNLP discussion essences the utmost considerable principal volume of revisions; 630 additional opportunities existing merely one or else two studies.

Their concern was relevant revisions on un-indexed opportunities in the carefully chosen files.

To alleviate this restraint, they relate a specific inquiry threads in Google Scholar. Even restrictive the enquiry in current file, the studies amount is advanced related to the additional files [25].

III. PROPOSED METHODOLOGY

- Interactive data mining permits users to access the search for patterns from diverse angles.
- The data mining procedure must be interactive because it is not very easy to know what can be exposed within the database and challenge to store the huge amount of data and also low data processing speed.
- There are various sorts of data deposited in databases and data warehouses. It is unimaginable for one scheme to mine all these types of data. Consequently, the diverse data mining system must be interpreted for various classes of data.

3.1 Architecture Diagram

- Web Harvesting
- Retrieve and Response
- (JSoup Library)
- Preprocessing
- (NLP Tokenization Algorithm)
- Map-Reducer (Map reduce Algorithm)
- Mapper tasks
- (Tokenization Algorithm and Sorting)
- Reducer tasks
- (Searching Algorithm)
- Hadoop
- Database (Storage)
- User
- Web Page URL address as an input
- For web extraction

(Ex: www.google.com)

- The query in the front end

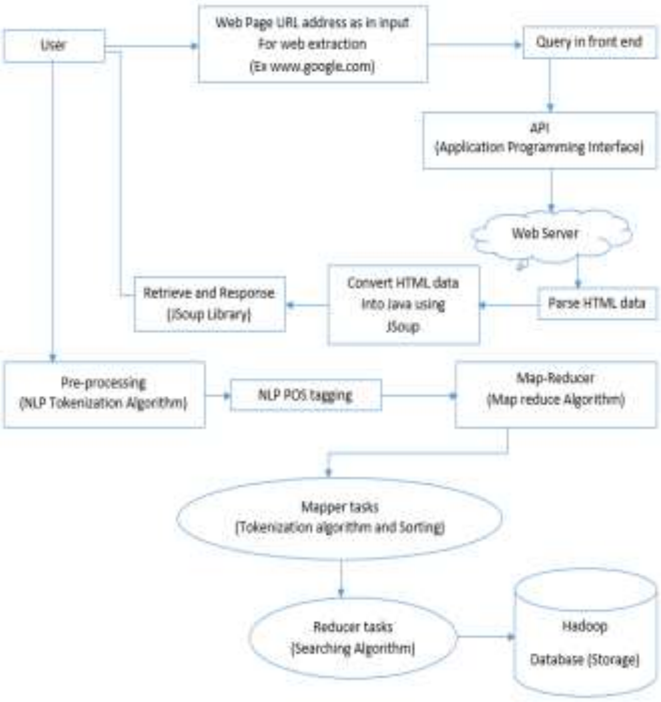


Figure 1: Architecture diagram

3.1.1. Description

The framework offered input name as an URL then extracts all types of contents from web pages, and is shown in the textbox. Parts-of-Speech (POS) extract the phrases from noun or verb clauses. It allocates data to each identified token. The Map-Reduce contains two crucial tasks, viz. Map and Reduce. Here, the distinct elements are shattered into a tuple. In this the map takes a specific group of data and converts the data into a variable set.

The guide or mapper's activity is to practice the information. For the most important part, the information is put away in the Hadoop record framework (HDFS) as a document or a registry. The Reducer's activity is to process the data that originates from the mapper. In the wake of preparing, it delivers another procedure of yield, which will be consumed in the HDFS.

Modules

- Dataset Collection
- Map Reducer
- Mapper tasks
- Reducer tasks
- Hadoop Storage

3.1.1.1. Module Description

a. Dataset Collection

- In this Module, the Input name is given as an URL then all kinds of contents are extracted from web pages and shown in the textbox. Web Harvesting (WH) and Web Scrawling are similar technologies in an increasingly popular technique utilized by websites to channel user’s searches to their website.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- The processed and extracted dataset are collected from the URL (uniform resource allocation).

b. Pre-Processing

- Tokenization is a stage which parts elongated strings of content into little pieces or tokens. Bigger lumps of content can possibly be tokenized into sentences; sentences can be tokenized into words, and so forth. Additionally, preparing is for the majority performed after a bit of composing is suitably tokenized.
- Tokenization is all duded as a content division or lexical investigation. Division is used to allude to the breakdown of a huge lump of content into pieces more immense than words (e.g., passages or sentences), while tokenization is held for the breakdown procedure which results solely in words.

c. WH-NLP-POS Tagging Algorithm in Big data

- Web harvesting program detects websites that comprise particular content directed at a particular web harvest inquiry.
- The web data is downloaded by the program of web harvesting, including a link which will direct the user to the companies' website.
- This data is then indexed by well-known search engines such as Yahoo and Google to assist the data to be easily accessed by further searches.
- Increasing exposure by the usage of web harvesting is important for a business to expand and grow online. The key factor in increasing online business revenue is the usage of web harvesting. This will in turn be a consequence in an increased probability that an online search will direct a potential customer to your website.
- As an outcome, it has become significant for a development in business that they pursue outside assistance in their efforts to increase their web presence. To the end, businesses must seek out the most cost-effective and user friendly web harvesting software obtainable to ensure that their efforts are rewarded with positive outcomes every time.
- The meaning of 'web scraping' and 'web harvesting' are similar, but 'web harvesting' is often called crawling multiple sites and extracts a specialized set of data. It can also be called directed web crawling.
- One of the valuable tools for web harvesting is FMiner. Multiple site links can be added as the starting URL in a file comprising extraction project and configure the extracted outcome to be saved into the same database. Then, using the "schedule" and "incremental" functions, the program can harvest page contents from multiple sites continuously and incrementally.

d. NLP

- To find relevant information and/or summarize the content of documents in large volumes of information for collective insight, natural language processing for big data can be leveraged automatically.
- Natural Language Processing (NLP) is a method to analyse readable text that is generated by humans for language processing, artificial intelligence and translation.
- In order to analyse the text accurately and precisely, there are some methods for NLP in order to deal with the challenges such as the collection and storage of the text corpus, and analysis. NLP practices also gain benefits and experience through research in linguistics, artificial intelligence, machine learning, computational statics and other sciences.
- However, at recent times, because of the explosion of information, the utilization of traditional NLP faces many challenges such as the volume of unstructured and structured data, accuracy of the results and velocity of processing data.
- In addition, there are so many slangs and indefinite expressions used on social media networking sites, which give NLP pressure to examine the meanings, which may also be difficult for some people.
- Furthermore, nowadays, people heavily depend on search engines such as Google and Bing (which use NLP as their core technique) in their daily study, work, and entertainment.
- All of the above mentioned factors encourage computer scientists and researchers to find more robust, efficient and standardized solutions for NLP.

e. NLP Tokenization

- Tokenization may be described as the process of splitting the text into smaller parts called tokens, and this is deliberated as a crucial step in NLP.
- The process of slicing the given sentence into smaller parts (tokens) is called as tokenization. In general, the given raw text is tokenized based on a group of delimiters.
- Tokenization is used in tasks such as processing searches, spell-checking, identifying parts of speech, document classification of documents, sentence detection etc.
- Simple Tokenizer − Tokenizes the available raw text using character classes.
- Whitespace Tokenizer − Uses whitespaces in order to tokenize the given text.
- Tokenizer ME − Converts raw text into separate tokens. It uses Maximum Entropy to make its decisions.

Steps

- Tokenize the sentences
- Print the tokens
- Instantiating the respective class.

f. POS Tagging

- The Parts of Speech of a given sentence can be detected using OpenNLP and then can be printed. Instead of full name of the parts of speech, OpenNLP uses short forms of each part of speech. The following table shows the different parts of speeches detected by OpenNLP and their meanings.
- To tag the parts of speech of a sentence, OpenNLP uses a model, a file named en-posmaxent.bin. This is a predefined model that is trained to tag the parts of speech for the given raw text.

Steps

- Loading the model
- Tokenizing the sentence
- Instantiating the POSTaggerME class
- Generating the tags
- Printing the tokens and the tags

g. Map-Reducer

In a large dataset, a programming framework called map reduce is allowed for distributed and parallel processing in a distributed environment. It performs well without considering the issues namely fault tolerance, reliability, etc. So, map reduce

At the end, the Map Reduce gives a flexibility to create code logic deprived of considerate regarding system design concerns.

Let the basic unit of data in Map-Reduce computations be, in which keys including the values are continually only binary strings. The Map-Reduce program contains the sequence of reducers and mappers. Let the input be a multiset of pairs which is indicated through input to execute a program. For r=1, 2, …., R, do:

- For Execute Map, let be a multiset of pairs output through , i.e.,
- For Shuffle, let be a multiset of values for each k, then.
- For Execute reduce, the sequence of tuples such as, are generated by the reducer. Let be the multi-set of pairs output by , i.e.,

h. Mapper tasks

- Mapper in Hadoop takes each record created by the Record Reader as input. Then processes each document and create key-value pairs.
- This key-value pair is entirely unlike the input pair.
- The mapper output is called intermediate output which is saved on the local disk. Mapper doesn't save its production on HDFS, as it is temporary data and stored on HDFS to create multiple copies.

i. Reducer tasks

Reducer is a phase in Hadoop which comes after Mapper stage. The yield of the mapper is provided as the contribution for Reducer which proceeds and creates another arrangement of creation, which will be put away in the HDFS. Reducer first handles the in-between values for individual key produced as a result of the map function in addition to further providing an output.

Then, a map in addition to reduce phases run in slots implies that each node could run beyond one Map or else Reduce task in matching; generally, the slots quantity is correlated through cores quantity present in a specific node. Here, the Hadoop implementation time in all the phases are:

$$t_{total} = t_{map} + t_{shuffle} + t_{reduce} \qquad (1)$$

The product of a particular map task and quantity of map tasks for each and every node is called as the time of the Map phase .phase.

$$t_{map} = \frac{W_{map} * C_{U_{map}}}{T} * \frac{n_{map}}{p} \qquad (2)$$

$$t_{reduce} = \frac{W^{in}_{U_{reduce}} * C_{U_{reduce}}}{T} * \frac{n_{reduce}}{p} \qquad (3)$$

The output data size product of a single map task is the quantity of enduring shuffle tasks divided by means of a network bandwidth,

$$t_{shuffle} = \frac{W^{out}_{U_{map}} * (n_{map} \bmod p)}{B} \qquad (4)$$

The performance model of map reduce in Hadoop is given by,

$$t_{total} = \frac{W_{map} * C_{U_{map}}}{T} * \frac{n_{map}}{p} + \frac{W^{out}_{map} * (n_{map} \bmod p)}{n_{map} * B} + \frac{W^{in}_{U_{reduce}} * C_{U_{reduce}}}{T * p} \qquad (5)$$

Where - output information of a map phase, - number of map tasks, -single reduce task.

Dataflow Diagram

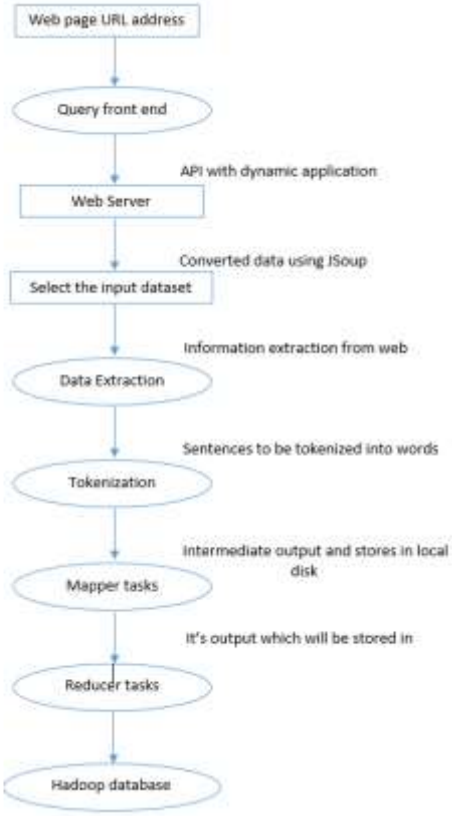
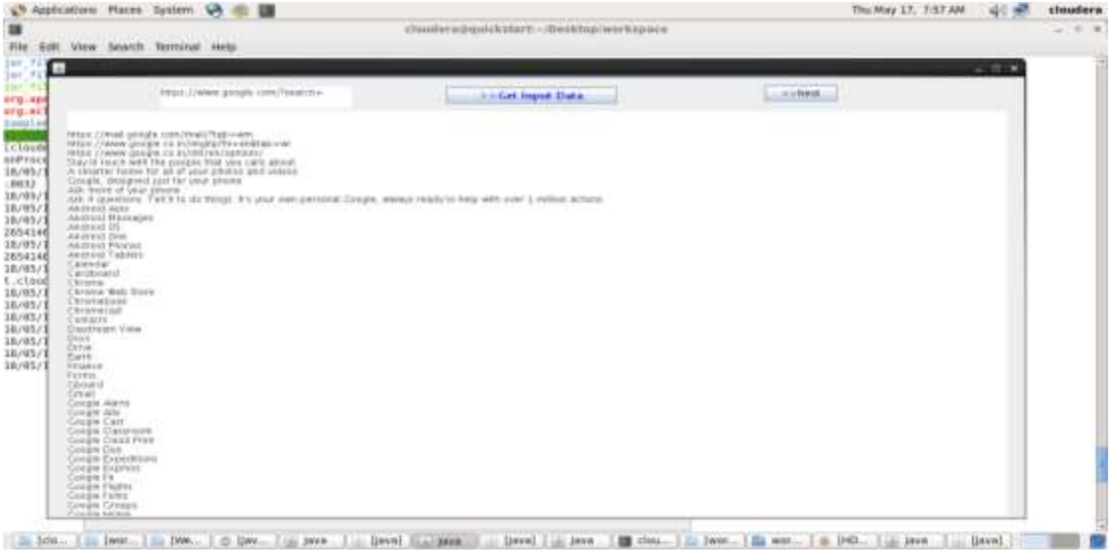

Web page URL address
Query front end
API with dynamic application
Web Server
Converted data using JSoup
Select the input dataset
Information extraction from web
Data Extraction
Sentences to be tokenized into words
Tokenization
Intermediate output and stores in local disk
Mapper tasks
It's output which will be stored in
Reducer tasks
Hadoop database

Figure 2: Data flow diagram

IV. RESULTS AND DISCUSSION

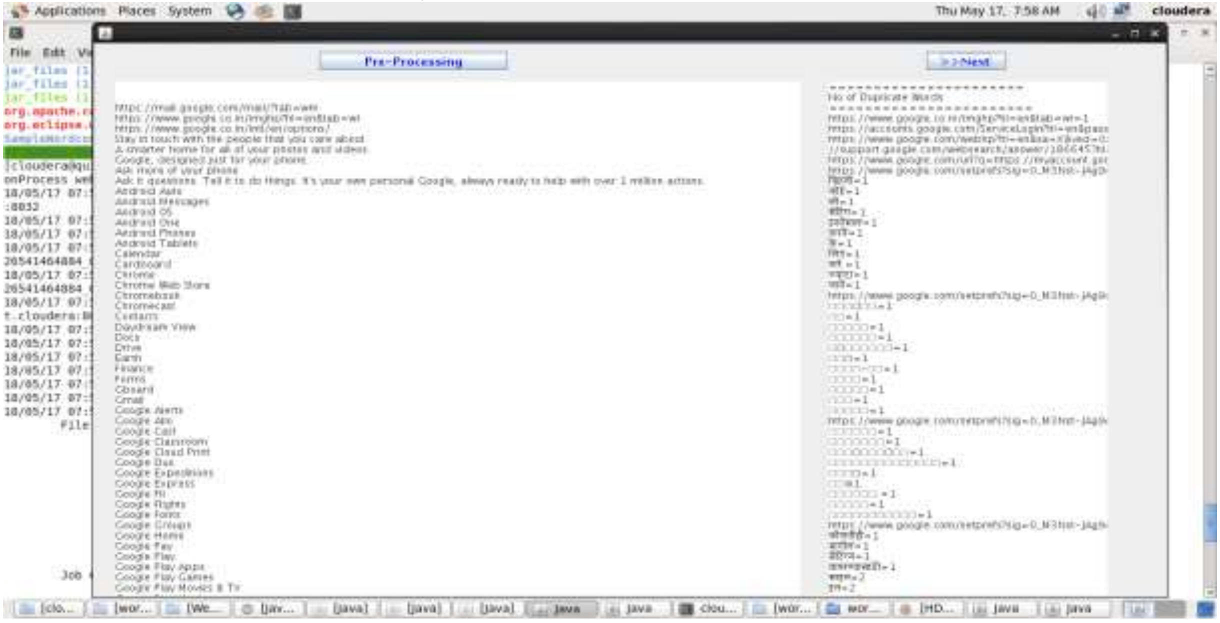
The performance assessment of projected system scalability is displayed in a method comparable to the prior authentication. On the other hand, research prolonged the dataset beginning from 10,000 to 20,000 web page and documents and assessed the time of dispensation for Keyword Extractor Module (KEM) implementation on the text contented of a deliberated dataset that was studied before through the Nutch-based crawler. Figure 2 displays the Web Extraction method. Along these lines, no further exterior network admission will be crucial for keywords in addition to extraction of key phrase. In this approach reduce blockages besides additional concerns are not contingent on procedure parallelization. Figure 3 displays the Pre-Processing method. Various developments have described the probable solutions for these overall enhancements to a code besides a test formation, with an esteem to the version projected in Nesi et al., (2015), both in relationships of time presentations in addition to scalability. The Hadoop cluster design considered for tests was evaluated on several formations, varying from 2 to 5 nodes. Each and every node is a Linux 8-cores workstation by means of Hadoop HDFS. To evade errors linked to data integrity and disappointments outstanding to decommissioning in addition to re-commission of cluster nodes, Hadoop permits completing a rebalance of deposited blocks amongst the cluster active nodes [26].


Applications Places System
Thu May 17, 7:57 AM
cloudera
cloudera@quickstart:~/Desktop/workspace
File Edit View Search Terminal Help
https://www.google.com/?search=
>>Get Input Data
>>Next

```
https://mail.google.com/mail/?tab=wm
https://www.google.co.in/imghp?hl=en&tab=wi
https://www.google.co.in/intl/en/options/
Stay in touch with the people that you care about.
A smarter home for all of your photos and videos.
Google, designed just for your phone.
Ask more of your phone.
Ask it questions. Tell it to do things. It's your own personal Google, always ready to help with over 1 million actions.
Android Auto
Android Messages
Android OS
Android One
Android Phones
Android Tablets
Calendar
Cardboard
Chrome
Chrome Web Store
Chromebook
Chromecast
Contacts
Daydream View
Docs
Drive
Earth
Finance
Forms
Gboard
Gmail
Google Alerts
Google Allo
Google Cast
Google Classroom
Google Cloud Print
Google Duo
Google Expeditions
Google Express
Google Fit
Google Flights
Google Fonts
Google Groups
```

Figure 3: Upload of input Dataset

In this procedure, the Input is given as an URL then all categories of contents are extracted from web pages and displayed in a textbox. Using NLP, the obtained data are pre-processed and structured. Keyword in the document is extracted. Communication between the keyword and the words inside the document is identified. Using the neural network between the keyword and the words inside, the report is determined. The text is selected, only if the model is detected. Figure 4 shows the Testing Content Extraction.

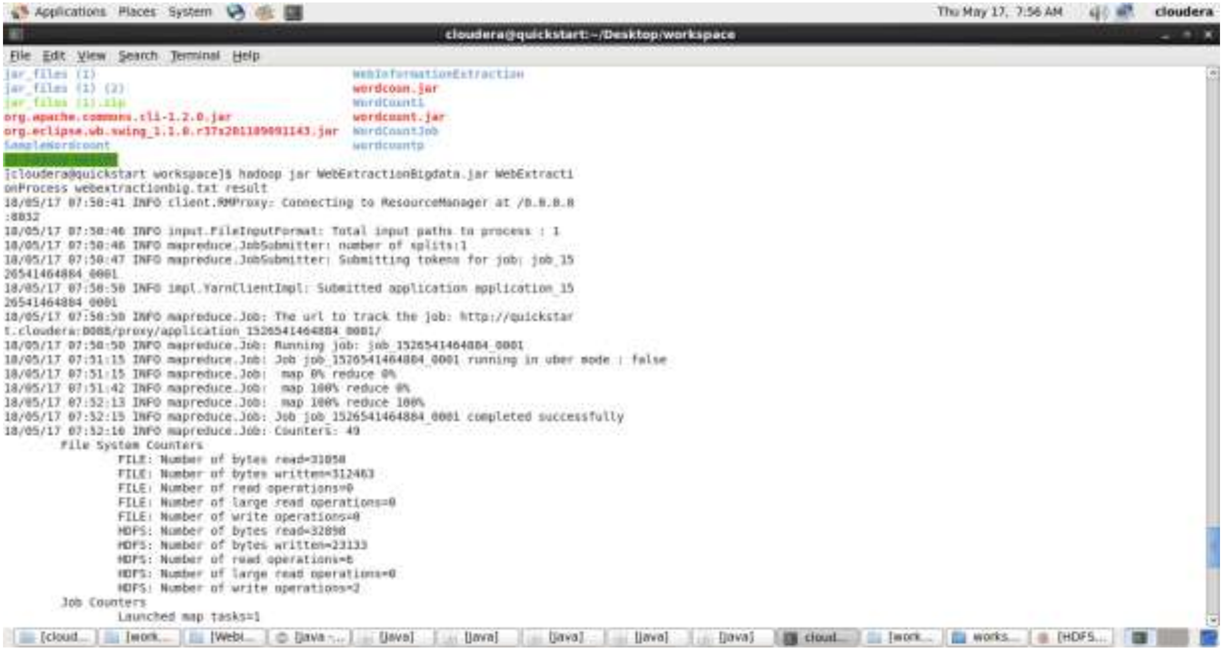


```
https://mail.google.com/mail/?tab=wm
https://www.google.co.in/imghp?hl=en&tab=wi
https://www.google.co.in/intl/en/options/
Stay in touch with the people that you care about.
A smarter home for all of your photos and videos.
Google, designed just for your phone.
Ask more of your phone.
Ask it questions. Tell it to do things. It's your own personal Google, always ready to help with over 1 million actions.
Android Auto
Android Messages
Android OS
Android One
Android Phones
Android Tablets
Calendar
Cardboard
Chrome
Chrome Web Store
Chromebook
Chromecast
Contacts
Daydream View
Docs
Drive
Earth
Finance
Forms
Gboard
Gmail
Google Alerts
Google Allo
Google Cast
Google Classroom
Google Cloud Print
Google Duo
Google Expeditions
Google Express
Google Fit
Google Flights
Google Fonts
Google Groups
Google Home
Google Pay
Google Play
Google Play Apps
Google Play Games
Google Play Movies & TV
```

```
========================
No of Duplicate Words
========================
https://www.google.co.in/imghp?hl=en&tab=wi=1
https://accounts.google.com/ServiceLogin?hl=en&pass
https://www.google.com/webhp?hl=en&sa=X&ved=0
//support.google.com/websearch/answer/186645?hl
https://www.google.com/url?q=https://myaccount.goo
https://www.google.com/setprefs?sig=0_M3Nst-jAg0
[illegible]
```

Figure 4: The Pre-Processing method

The model of MapReduce provides speculative carrying out of tasks, besides it is intended to offer redundancy to cope fault tolerance. As a result, along these lines, the Job Tracker may have the requirement to postpone fizzled or else exterminated assignments, besides this may influence the time of implementation of the whole procedure. In this way, for execution correlation, the best preparing circumstances have been chosen among a few test occasions that have been led for every hub setup. With current method, the quantity of efforts for redoing unsuccessful or exterminated tasks invented must be reduced. For the whole examination dataset, approximately 3.5 million documents have been extracted. As a term of examination, running the same CiteSeerExtractor, a RESTful API application on similar corpora dataset on a single non-Hadoop workstation took around 115 hours. Figure 5 illustrates the Map-Reducer process. A probable clarification to this critical presentation opening can be the fact that the cypher of Java for the standalone CiteSeerExtractor, a RESTful API application is not optimized for multi-threading. Despite the fact that the Map-Reduce adaptation executed in Hadoop can be an advantage of MapReduce configuration parameters that characterizes the highest number of guide and lessens errand spaces to proceed all abusing multi-center innovation even on a single hub bunch. Figure 6 displays the Hadoop Storage data set.

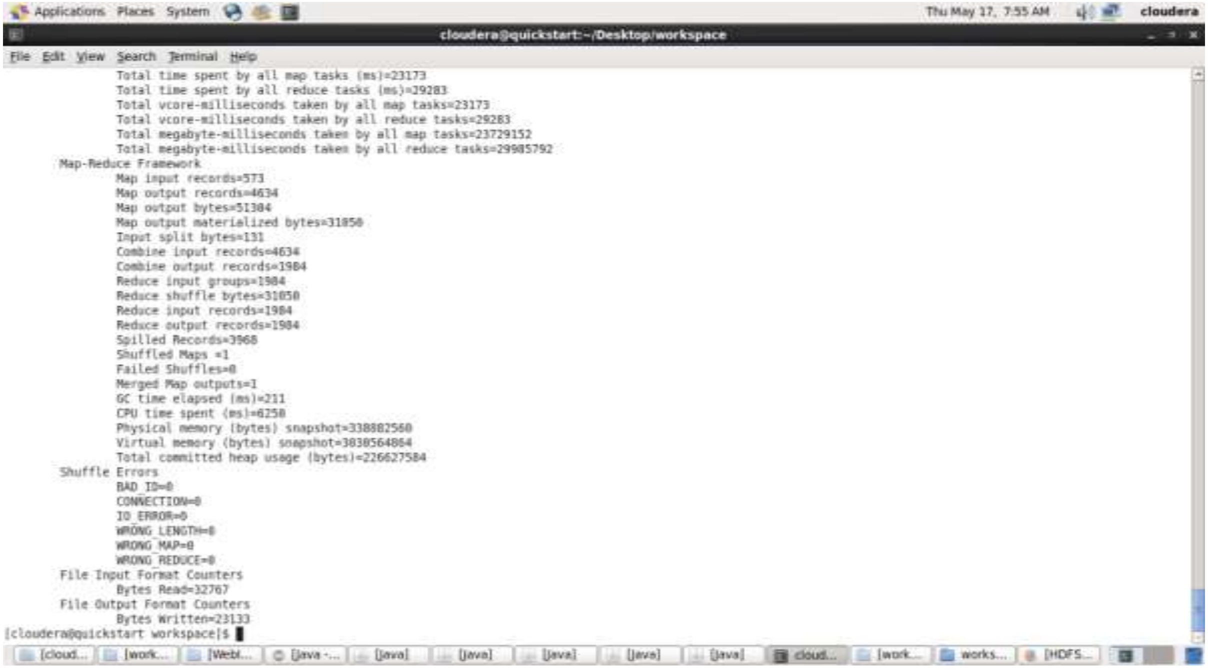

Applications Places System
Thu May 17, 7:56 AM
cloudera
cloudera@quickstart:~/Desktop/workspace
File Edit View Search Terminal Help
[cloud... [work... [Webl... [Java -... [Java] [Java] [Java] [Java] [Java] cloud... [work... works... [HDFS...

```
jar_files (1)                                      WebInformationExtraction
jar_files (1) (2)                                  wordcoun.jar
jar_files (1).zip                                  WordCount1
org.apache.commons.cli-1.2.0.jar                   wordcount.jar
org.eclipse.wb.swing_1.1.0.r37x201109091143.jar    WordCountJob
SampleWordcount                                    wordcountp

[cloudera@quickstart workspace]$ hadoop jar WebExtractionBigdata.jar WebExtracti
onProcess webextractionbig.txt result
18/05/17 07:50:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
18/05/17 07:50:46 INFO input.FileInputFormat: Total input paths to process : 1
18/05/17 07:50:46 INFO mapreduce.JobSubmitter: number of splits:1
18/05/17 07:50:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
26541464884_0001
18/05/17 07:50:50 INFO impl.YarnClientImpl: Submitted application application_15
26541464884_0001
18/05/17 07:50:50 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1526541464884_0001/
18/05/17 07:50:50 INFO mapreduce.Job: Running job: job_1526541464884_0001
18/05/17 07:51:15 INFO mapreduce.Job: Job job_1526541464884_0001 running in uber mode : false
18/05/17 07:51:15 INFO mapreduce.Job:  map 0% reduce 0%
18/05/17 07:51:42 INFO mapreduce.Job:  map 100% reduce 0%
18/05/17 07:52:13 INFO mapreduce.Job:  map 100% reduce 100%
18/05/17 07:52:15 INFO mapreduce.Job: Job job_1526541464884_0001 completed successfully
18/05/17 07:52:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=31050
                FILE: Number of bytes written=312463
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=32898
                HDFS: Number of bytes written=23133
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
```

Figure 5: The Testing of Content Extraction

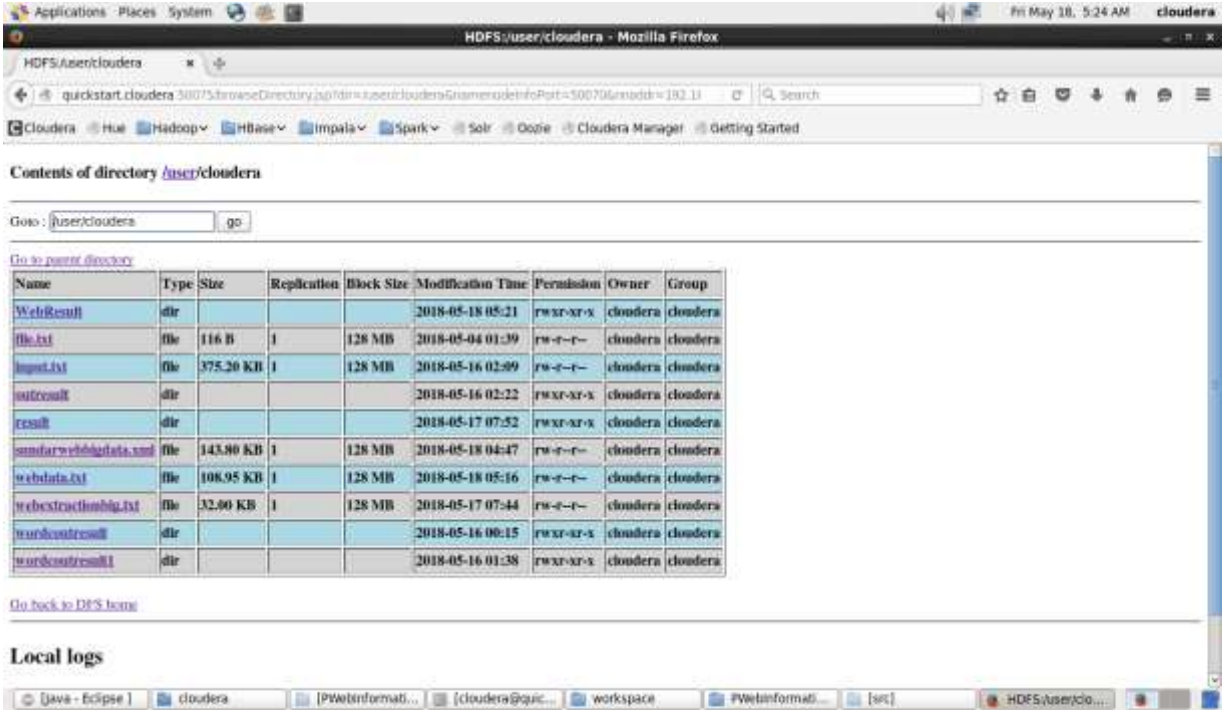
The MapReduce contains two essential tasks, namely Map and Reduce. The map consists of a group of data and changes it into an alternative group of data, where separate essentials are fragmented down into the tuple.


Applications Places System
Thu May 17, 7:55 AM
cloudera
cloudera@quickstart:~/Desktop/workspace
File Edit View Search Terminal Help
[cloud... [work... [Webl... [Java -... [Java] [Java] [Java] [Java] [Java] cloud... [work... works... [HDFS...

```
                Total time spent by all map tasks (ms)=23173
                Total time spent by all reduce tasks (ms)=29283
                Total vcore-milliseconds taken by all map tasks=23173
                Total vcore-milliseconds taken by all reduce tasks=29283
                Total megabyte-milliseconds taken by all map tasks=23729152
                Total megabyte-milliseconds taken by all reduce tasks=29985792
        Map-Reduce Framework
                Map input records=573
                Map output records=4634
                Map output bytes=51384
                Map output materialized bytes=31050
                Input split bytes=131
                Combine input records=4634
                Combine output records=1984
                Reduce input groups=1984
                Reduce shuffle bytes=31050
                Reduce input records=1984
                Reduce output records=1984
                Spilled Records=3968
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=211
                CPU time spent (ms)=6250
                Physical memory (bytes) snapshot=338882560
                Virtual memory (bytes) snapshot=3030564864
                Total committed heap usage (bytes)=226627584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=32767
        File Output Format Counters
                Bytes Written=23133
[cloudera@quickstart workspace]$ █
```

Figure 6: The Map-Reducer Process

The map or mapper's work is to develop the input data. Generally, the contribution data is present in the procedure of directory or file and stored in the Hadoop file system (HDFS).



Contents of directory /user/cloudera

Goto : /user/cloudera go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
WebResult	dir				2018-05-18 05:21	rwxr-xr-x	cloudera	cloudera
file.txt	file	116 B	1	128 MB	2018-05-04 01:39	rw-r--r--	cloudera	cloudera
input.txt	file	375.20 KB	1	128 MB	2018-05-16 02:09	rw-r--r--	cloudera	cloudera
outresult	dir				2018-05-16 02:22	rwxr-xr-x	cloudera	cloudera
result	dir				2018-05-17 07:52	rwxr-xr-x	cloudera	cloudera
similarwebbigdata.xml	file	143.80 KB	1	128 MB	2018-05-18 04:47	rw-r--r--	cloudera	cloudera
webdata.txt	file	108.95 KB	1	128 MB	2018-05-18 05:16	rw-r--r--	cloudera	cloudera
webextractionbig.txt	file	32.00 KB	1	128 MB	2018-05-17 07:44	rw-r--r--	cloudera	cloudera
wordcountresult	dir				2018-05-16 00:15	rwxr-xr-x	cloudera	cloudera
wordcountresult1	dir				2018-05-16 01:38	rwxr-xr-x	cloudera	cloudera

Go back to DFS home

Local logs

Figure 7: The Hadoop Storage

Hadoop directs the Map and Reduce responsibilities to the suitable servers in the group. The outline achieves all the particulars of data-passing like delivering responsibilities, confirming task completion, and copying data from the group among the nodes. Most significant computation takes place on nodes on original recordings that decreases the network circulation. After the achievement of the specified responsibilities, the group gathers and reduces the data to form a suitable consequence, and directs it back to the Hadoop server.

Web Information Extraction validation results

1. Precision

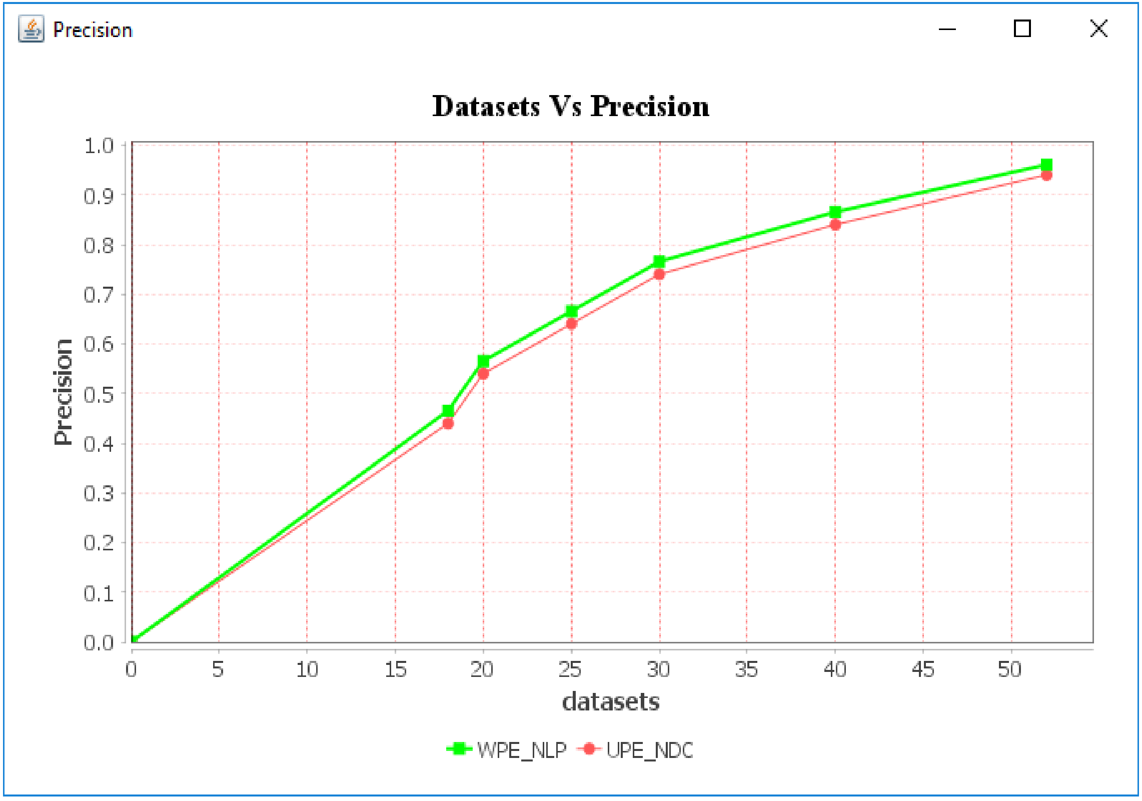


Figure 8: Datasets compare with precision

Description

Our proposed WPE_NLP methods have the highest precision compare than Existing UPE_DC method

2. Recall

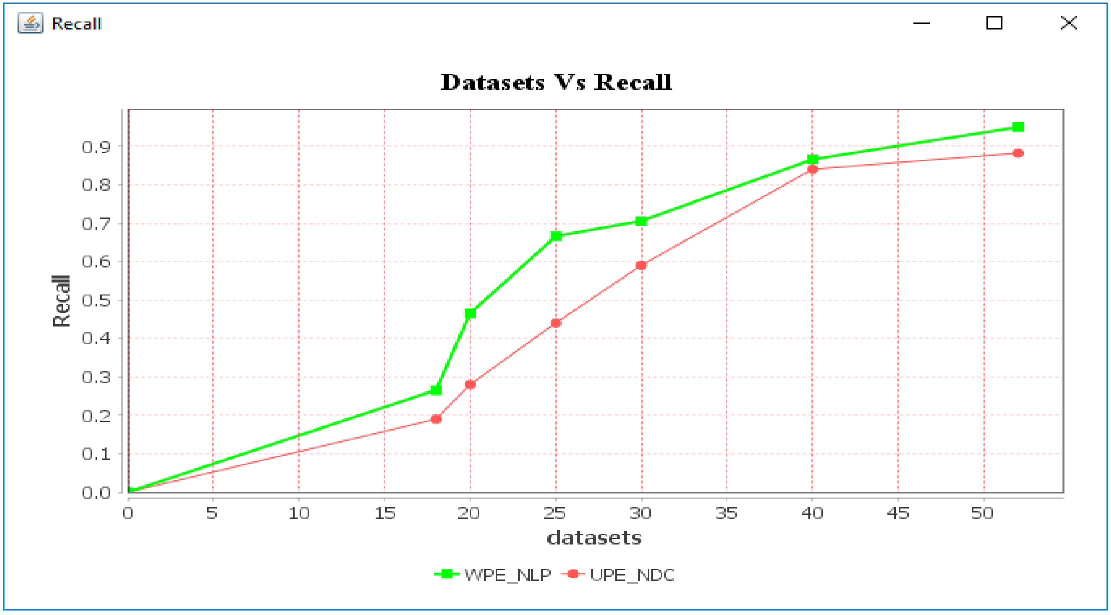


Figure 9: Datasets compare with recall

Description

Our proposed WPE_NLP methods have the highest Recall compare than Existing UPE_DC method

3. F-Measure

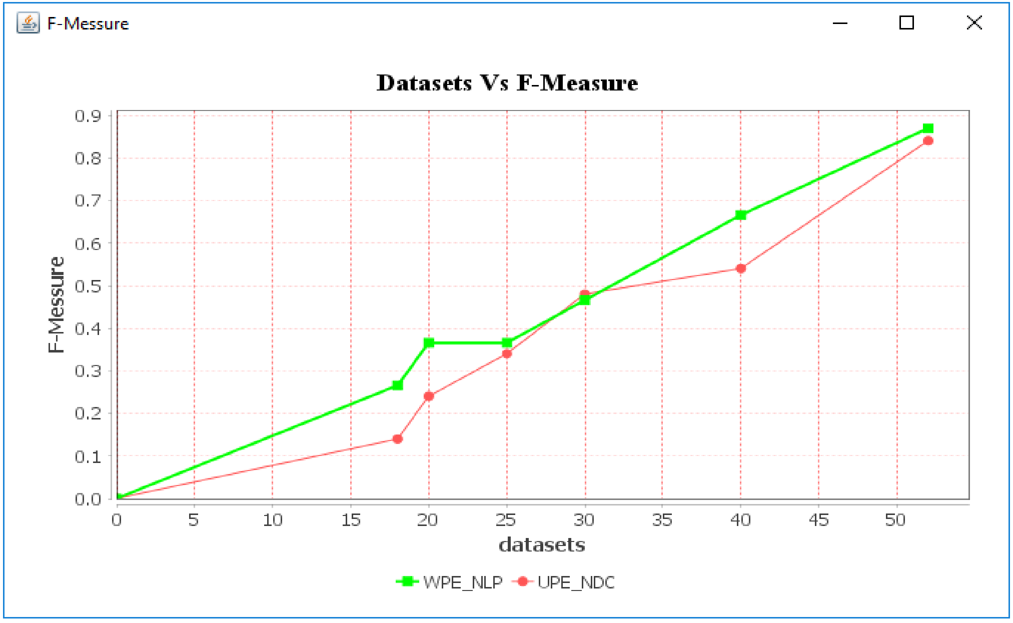


Figure 10: Datasets compare with F-Measure

Description

Our proposed WPE_NLP methods have the highest F-Measure compare than Existing UPE_DC method

V. CONCLUSION AND FUTURE WORK

In this exploration, discoveries indicate datasets from the web and the advantages of Big Data. It is pivotal to identify the manner which guarantees utilization, administration in addition to re-utilization of Sources of information, containing web URL Information, besides the nation over to assemble helpful URL and administrations. It is vital to evaluate the optimal approach Information Extraction using Neural Natural Language Processing (IE-NLP) to prefer on behalf of filtering along with scrutinizing the information. For the enhanced analytic processing, Hadoop with MapReduce may be utilized. In current paper, it is concluded that the fundamentals of MapReduce software is designed by the open source Hadoop background. This outstanding context of Hadoop raises the handling of vast information over the process of distribution and reacts very fast.

The future of big data offers different varieties of data similar to unstructured, semi-structured and structured data processing.

These data were originated from all over the place like videos, pictures, sensors, social media, and transactions, etc. and so on. In the future, the development of this data is probably expected soon.

REFERENCES

1. Pollock, R., 2013a. Forget Big Data; Small Data is the Real Revolution d Open Knowledge Foundation Blog. http://blog.okfn.org/2013/04/22/forget-big-datasmall-data-is-the-real-revolution.
2. Akers, K.G., Feb. 2013. Looking out for the little guy: small data curation. Bull. Am. Soc. Inf. Sci. Technol. 39 (3), 58-59.
3. Pande, V., & Khandelwal. Information Extraction Technique: A Review. IOSR Journal of Computer Engineering (2016) 16-20.
4. Sonit Singh, Natural Language Processing for Information Extraction, IEEE publications (2018) 1-24.
5. Virmani, C., Pillai, A., & Juneja, D. Extracting Information from Social Network using NLP. International Journal of Computational Intelligence Research (13) (4) (2017) 621-630.
6. Chen, Y. Natural Language Processing in Web data mining. IEEE 2nd Symposium on Web Society, 2010, 388-391.
7. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111– 3119.
8. Mikolov, T., M. Karafiat, L. Burget, J. Cernock ´ y, and ` S. Khudanpur, "Recurrent neural network based language model." in Interspeech, vol. 2, 2010, p. 3.
9. Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, 2013, p. 1642.
10. Barbosa, D., Wang, H., & Yu, C. (2015). Inferencing in information extraction: Techniques and applications. 2015 IEEE 31st International Conference on Data Engineering, 1534- 1537.
11. Chandurkar, A., & Bansal, A. (2017). Information Retrieval from a Structured Knowledgebase. 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 407- 412.
12. Florence, A., & Padmadas, V. (2015). A summarizer system based on a semantic analysis of web documents. 2015 International Conference on Technologies for Sustainable Development (ICTSD), 1-6.
13. Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.
14. Goldberg, Y. "A primer on neural network models for natural language processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.
15. Azcarraga, A., M. David Liu, and R. Setiono, "Keyword Extraction Using Backpropagation Neural Networks and Rule Extraction," in Proc. of IEEE World Congress on Computational Intelligence (WCCI), Brisbane, Australia, June 2012.
16. Pavithra, Monisa, & Ramya. (2013). A Design of Information Extraction System. International Journal of Advanced Research in Computer Science, 4 (8), 109-111.
17. Sridevi, & Arunkumar. (2016). Information Extraction from Clinical Text using NLP and Machine Learning: Issues and Opportunities. International Journal of Computer Applications, 11-16.
18. [18] Wei, W., Shi, S., Liu, Y., Wang, H., Yuan, C., & Huang, Y. Extraction Rule Language for Web Information Extraction and Integration. 2013 10th Web Information System and Application Conference (2013), 65-70. doi:10.1109/wisa.2013.21
19. Jindal, P., D. Roth, and L.V Kale, "Efficient Development of Parallel NLP Applications," Tech. Report of IDEALS (Illinois Digital Environment for Access to Learning and Scholarship), 2013.
20. Rizzolo, N., and D. Roth, "Learning-Based Java for Rapid Development of NLP Systems." In Proc. of the International Conference on Language Resources and Evaluation (LREC), 2010.
21. Exner, P. and Nugues, P., "KOSHIK - A Large-scale Distributed Computing Framework for NLP," in Proc. of the International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), pp. 463-470, 2014.
22. Pablo Barrio and Luis Gravano. Sampling strategies for information extraction over the deep web. Information Processing and Management 53 (2017) 309–331.
23. Wei Wang and Kathleen Stewart. Spatiotemporal and semantic information extraction from Web news reports about natural hazards. Computers, Environment and Urban Systems 50 (2015) 30–40
24. Dilek Küçük Match and Uğur Avdan. Address standardization using the natural language process for improving geocoding results. Computers, Environment and Urban Systems, 2018.
25. Rafael Glauber, Daniela Barreiro Claro, A Systematic Mapping Study on Open Information Extraction, Expert Systems with Applications (2018), doi: 10.1016/j.eswa.2018.06.046
26. Nesi, P., G. Pantaleo, and G. Sanesi, "A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction from Web Pages and Documents," in Proc. of 21st Int. Conf. on Distributed Multimedia Systems (DMS2015), 2015.