International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018) pp. 131-136

An Automated MapReduce Framework for Crime Classification of News Articles Using MongoDB

K.Santhiya

Research Scholar, Department of Computer Applications
Bharathiar University, Coimbatore, Tamil Nadu, India.
Orcid Id: 0000-0003-1260-164X

V.Bhuvaneswari

Assistant Professor, Department of Computer Applications
Bharathiar University, Coimbatore, Tamil Nadu, India.
Orcid Id: 0000-0001-8000-1246

Abstract

Crime rate has increased globally in all countries and various abuses related to crime is reported in various sources such as video, audio, social blogs and article. Less percentage of crime are recorded legally for legal hearings. . Violence against crime is classified into various abuses for managing legal procedures. The current scenario the crime against women and children are found to be increasing day by day From olden days until now, women and children at all ages are affected with different sorts of abuses, is evident through news daily and social media. The problem of Crime analysis can be viewed in a Big Data Perspective as it encompasses one of the main V's of big data Volume, rate of crime recording can be equated to Velocity and the source of recording of crime represent the characteristic Variety. This paper a framework and methodology is proposed to classify crime against women and children reported in news dailies. An average of 5-6 crime cases against women and children has been reported in the newspapers in different form as unstructured text. The methodology encompasses a MapReduce framework to classify automatic document classification based on crime abuses in different groups using NoSQL document database MongoDB. The experimental results of the proposed methodology provide diagnostic analysis of crime abuses under different classification. The experimental analysis it is found that the proposed methodology is simple in storage, processing data with a minimum time in milliseconds. The outcome based on the classification of the dataset for the period considered for the year 2015-16 based on diagnostics analytics it is found that 81% of crimes were related to sexual abuse, and 12% of crimes were related to physical abuse in women and children.

Keywords: mapreduce, mongodb, classification, unstructured, big data, R, clustering, text mining

INTRODUCTION

The term "Information Explosion" was coined seventy years ago to quantify the tremendous increase in the volume of data. According to F.Rider in the year 1944, for every sixteen years, the American University Libraries tend to increase two-fold in size. If this growth rate persists, Rider predicted that the Yale library in 2040 tends to accumulate "approximately 0.2 billion volumes" (no.of bites for one book is around 10 MB, then 1 GB = 109Books [1]. 0.2 Billion Volumes equal to $0.2 \times 10^9 \times 10MB = 2 \times 10^{15}B = 2PB$.

According to study conducted byP.Lyman and H.R.Varian in 2000, found that in 1999, 1.5 Exabytes of unique information are produced world-wide. The same researchers conducted a similar study in 2003 and found that about 5 Exabytes of new information was generated in the world in 2002. However, such a history of sizing data volumes led to the evolution of the idea of "big data". The article written by M.Cox and D.Ellswath in 1977 was the first to make use of the term "Big Data". According to International Data Corporation (IDC) estimation, the size of data tends to exponentially increase every year after 2020.

A Bird's Eye on Big Data

Clearly big data definitions have evolved rapidly, with three defining dimensions, called data volume, velocity and variety. They are collectively termed as 3V's proposed by D.Laney in the year 2001[2]. The three V's are described below:

Volume: In accordance with the survey organized by IBM in 2012, if the data size exceeds beyond one terabyte, then it could be treated as big data [3]. Snapshots of huge volume of data generated are as follows: Hundreds of petabytes of data roughly equivalent to 100 million gigabytes are processed by Google per month. Amazon maintains a big bank of 152 million customer accounts [4]. Nearly 750 million pictures are uploaded to Facebook on a monthly basis.

Velocity: It refers to rapid and timely collection of data which in turn enhance the commercial value of big data. Few instances are APPLE receives about 47,000 APP downloads every minute. Every 60 seconds, consumers spend $272070 on web shopping.

Variety: It refers to data modalities in which the data are represented, since the data gets generated from multitude of sources. The different data modalities are structured, semi-structured and unstructured. According to Cuckier(2010), only 5% of structured data exists in the form of spreadsheets or relational databases. The rest 95% of data which encompasses

documents, audio, video, images, graphs and social media text messages are in unstructured form [5].

Exponential Growth of Unstructured Data

According to IDC Estimation, 90% of global data has been generated in the last four years, out of which 80% of data are in the form of unstructured and the rest 10% is in structured form [6]. The Fig.1 shows how these disordered datasets has covered the entire globe over the years.

Figure 1: Growth of Unstructured data

Analyzing unstructured data to enhance their Return on Investment (ROI) are the need of hour for majority of big company partners like Google, IBM, Amazon, WallMart etc.,

Challenges of Analyzing Unstructured Data

In ancient days, data management and analysis system rely upon RDBMS. However such databases can handle only structured data other than semi-structured or unstructured data. It also utilizes expensive hardware, but still under performs when it is applied to massive volume and heterogeneity of big data. To overcome these limitations, the research community proposed some solutions from different perspectives. Out of which NoSQL databases emerged as a good choice for permanent storage and management of massive amount of heterogeneous datasets [7]. However, NoSQL databases otherwise called as non-traditional relational databases are becoming more familiar for the storage of big data. NoSQL databases offer certain features like schema less, non-relational, BASE Transaction, highly distributable, sharding and replication. Various types of NoSQL databases exist like key-value databases, column-oriented databases, document-oriented and Graph databases. In this paper, we make use of one of the document oriented databases called MongoDB for storing and automatic classification of huge volume of text documents related to crime or violence against women and children.

Crime Investigation

Direct or indirect physical (or) mental harm (or) cruelty to women and children can be defined as a semantic meaning of "Crime (or) Violence against Women and Children". Women at all ages have been subjected to different sorts of abuses as shown in Fig.2. In this paper we have outlined some of the innovative techniques that have been carried over to conduct research on violence against women and children.

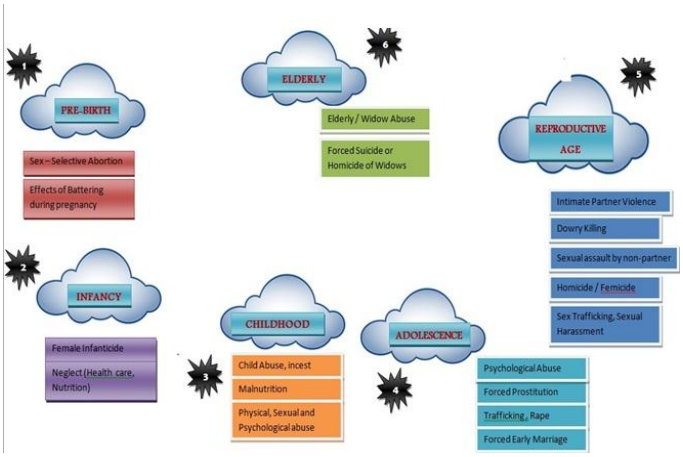


Figure 2: Life Cycle of Violence against Women

Objective

The objective of the proposed work is to build an automated MapReduce framework for classifying crime news articles using Document Oriented NoSQL database MongoDB.

The contributions of our work are:

(i) Acquisition of news articles from dailies and blog articles related to VAW and deploying in MongoDB.

(ii) Creation of domain specific corpus which helps in automatic classification of documents.

(iii) Implementation of MapReduce query in MongoDB to facilitate parallel processing of documents.

(iv) Integrating platforms R with MongoDB to perform document clustering and for effective visualization.

Organizations

The rest of the paper is arranged as follows. Section II explains the methodology and the framework used in this proposed work. In Section III describe the details of the experimental results. Finally, the conclusion and recommendations for future work are drawn in Section IV.

.

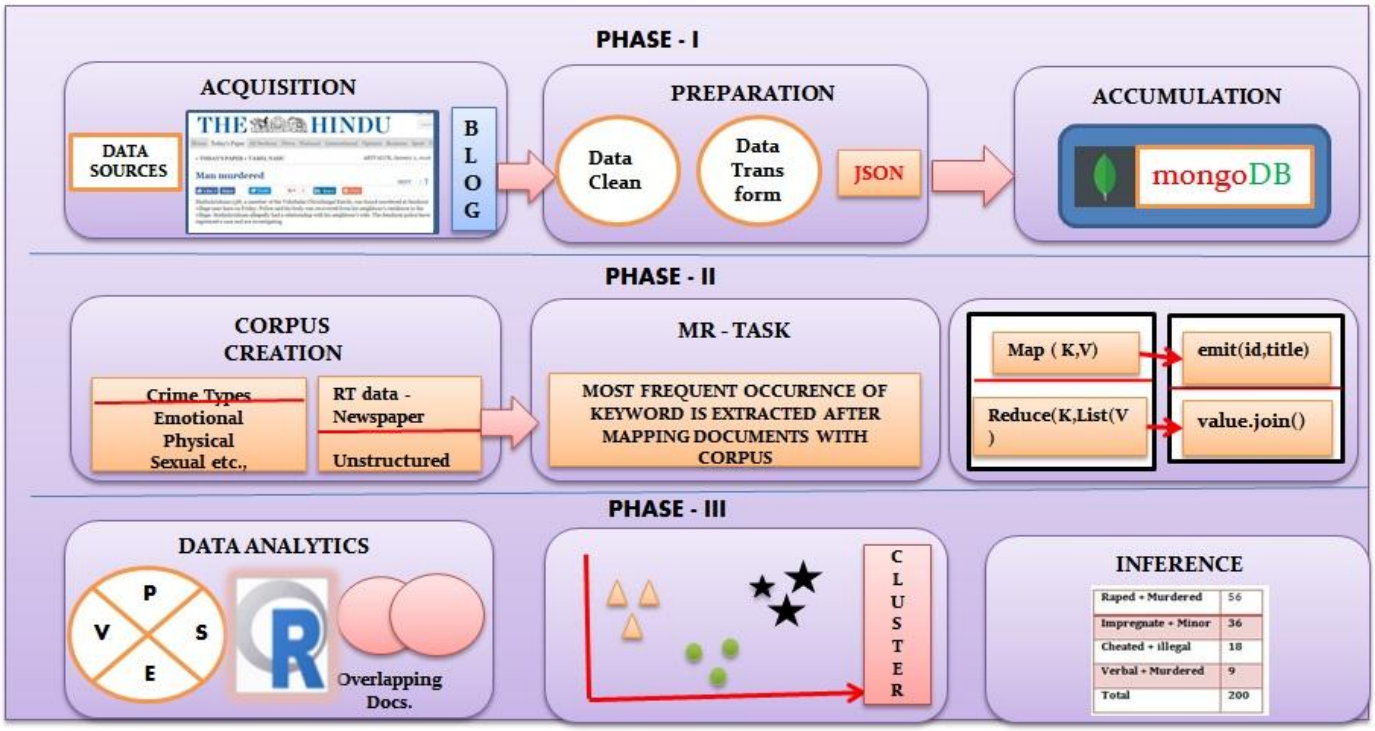


Figure 3: Crime Investigation Framework Using MongoDB

METHOD

The proposed work implements the MR framework given in Fig.3 for text indexing and automatically classifying the documents into appropriate clusters by matching it with domain-specific corpus. The crime classification framework consists of three phases. Phase 1 deals with data acquisition, preparation and dumping data. In the second phase we concentrate on corpus creation, performing mapreduce text for processing the text in two different functions. Phase 3 deals with performing data analytics task using machine learning algorithms such as clustering and classification are performed and visualized under R environment

Phase 1. Data Acquisition

In data acquisition phase, we have created a corpus of news articles and blog documents related to VAW. The data collected is unstructured in nature, is loaded in a NoSQL DBMS using JSON format. The section given below provides a state-of-art of MongoDB , a NoSQL document database.

MongoDB – At a glance

MongoDB is an open-source NoSQL DBMS widely used in application domains for online travel and shopping service by Expedia, a location-bases SNS by Foursquare and also used by global security company McAfee to handle huge transactions [8]. MongoDB is ranked fourth among all DBMSs, statistics given by DB-Engines, the best DBMS ranking web-site. MongoDB and relational databases differ each other in terms of storage, implementation and other functional concepts. Some of the terminologies used to describe them also vary. The corresponding terms to 'Table', 'Row' and 'Document' in a relational database are collection, column and field respectively, in a MongoDB database. In table 1 we have presented the comparison of querying data with respect to MongoDB and RDBMS.

Data Model and non-fixed Schema

MongoDB is schema less and stores documents as Binary JSON (BSON) objects, which is similar to that of object [9]. Every document in the collection has an _id field and it acts as a primary key. JSON is an open standard format based on java script language [10]. In JSON, every object is represented as an unordered collection of name/value pairs and it begins with { (left brace) and terminates with } (right brace). As there is no predefined schema in MongoDB, documents made in any form JSON documents can be easily inserted. In other words, the data models in MongoDB can be changed flexibly. Load Balancing and Fail-Over mechanisms can be handled automatically by MongoDB as it supports horizontal expansion with automatic sharing to distribute data among thousands of nodes.

Phase 2. Map Reduce Text Processing

A corpus related to different types of abuses is created as a dictionary by discussing with legal expert. The corpus contains collections related to terms such as rape, murder, suicide, arrest, assault etc., MapReduce tasks is created in MongoDB to relate the terms in the corpus (dictionary) and the document corpus news articles. The map task split the input document in the form of keys and values, thus matches

all the documents with that of the corpus. The reduce task joins all the values corresponding to a particular key, thus emits all the documents with a specific key. The document attributes represented in JSON notation are indexed to perform search for a specific text from the document collection.

Phase III. Data Analytics and Visualization

In this phase, we have classified and clustered the documents using MR framework. Before applying classification algorithms, certain level of text preprocessing has to be carried out like stemming, stop word removal and transformation of text into vector format. In this phase, we have integrated and configured our Hadoop Ecosystem,integrating with R. The output of the MR framework is classified using naïve bayes classifier in R. The classified documents are further grouped to identify the overlapping of several abuses for a specific crime using Kmeans clustering algorithms.

RESULTS

This section describes the experimental results of the work. A detailed view on the data corpus used in this work is presented.

Dataset

The dataset collection of 3000 news articles downloaded from the Hindu archive available in the hyperlink: http://thehinduarchive.com. The text documents are transformed to JSON notation and stored in MongoDB. Fig. 4 indicates the JSON notation of one particular news document after loading it into MongoDB.

Domain-Specific Corpus

The Domain-Specific Corpus created in this work consist of VAW. Violence strikes women from all kinds of backgrounds and of all ages irrespective of location.

Table 1 : Difference between MongoDB and SQL Query Syntax

Query Operations	**MongoDB**	**SQL**
CREATE DATABASE	Use database_name	Create database database_name;
CREATE	>db.tst.insert({name:"value",course:”value” age:”value”})	Create table tst(name nVarchar(20),course nVarchar(20), age Integer(2))
INSERT	>db.tst.insert({name:"suba",course:"mcom",age:"24"})	Insert into tst(name,course,age) values ('suba,','mcom','24);
READ	db.tst.find({"_id": ObjectId("5806226d0a1365374e618944")}).pretty()	Select * from tst;
UPDATE	>db.tst.update({"_id": ObjectId("580625e00a1365374e618948")},{$set:{name:"karthi"}})	Update tst set name='karthi' where _id = 'ObjectId("580625e00a1365374e618948” '
DELETE	>db.tst.remove({"_id": ObjectId("58062d0e0efd7f09461e81ca")})	Delete from tst where _id = 'ObjectId("580625e00a1365374e618948” '
DROP	>db.tst.drop()	DROP table tst;

```
{
        "_id" : ObjectId("582f72a00234c058b0f0ffee"),
        "request" : {
                "options" : [
                        "_=1479374223052",
                        "callback=jQuery111108312358129695349_1479374223050",
                        "format=jsonp"
                ],
                "pageUrl" : "http://www.thehindu.com/todays-paper/tp-national/tp
-tamilnadu/man-held-for-misbehaving-with-girls/article9289679.ece",
                "api" : "analyze",
                "version" : 3
        },
        "humanLanguage" : "en",
```

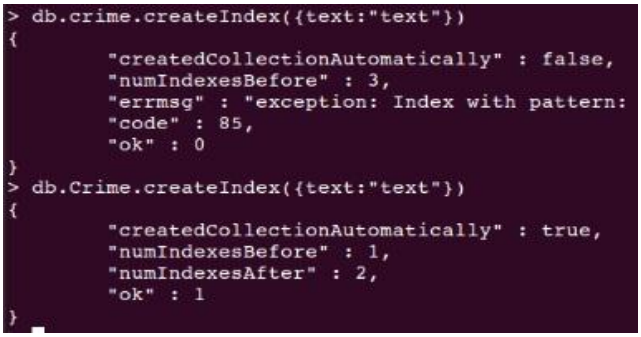

Figure 4: JSON Notation loaded in MongoDB

The corpus consists of VAW classified as Dating Violence, Domestic and Intimate Partner Violence, Emotional Abuse,

Human Trafficking, Same-sex Relationship Violence, Sexual assault and abuse, Stalking, Violence against immigrant and refugee women, Violence against women at work, Violence against women with disabilities.A Corpus is framed comprising of all legal terms that comes under the above mentioned types of violence by seeking opinions from legal experts.

Text Indexing

The text indexing in the methodology is facilitated using regular expression in MongoDB. The attribute has to be indexed in the collection at the time of creation. Text Indexing capabilities are not supported by MongoDB versions below 3.0. Fig. 5 shows the text index created for Crime collection.



```
> db.crime.createIndex({text:"text"})
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 3,
        "errmsg" : "exception: Index with pattern:
        "code" : 85,
        "ok" : 0
}
> db.Crime.createIndex({text:"text"})
{
        "createdCollectionAutomatically" : true,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
```

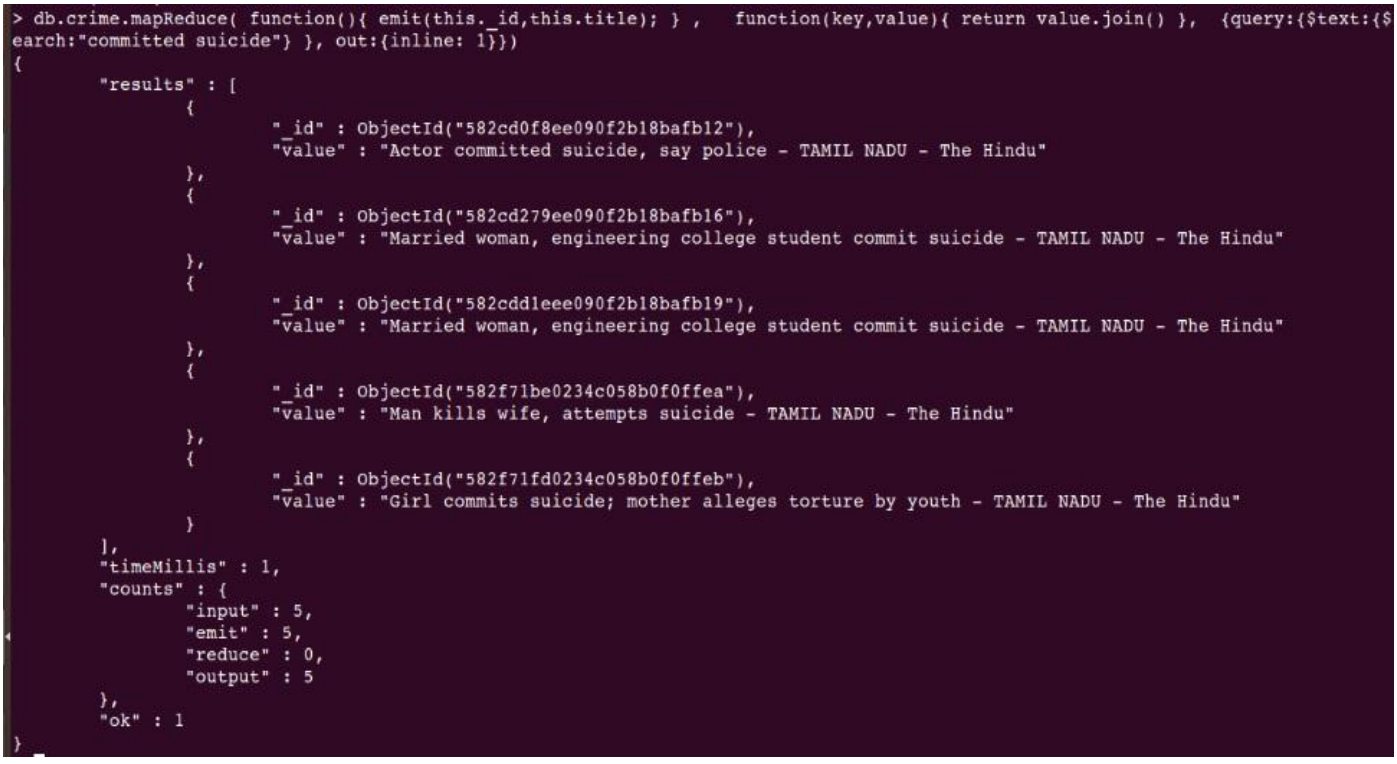

Figure 5: Text Index Creation for attribute text

Text Processing Using MR

The MR snapshot for text searching is given in Fig. 6. The snapshot provides a view to search the keyword listed in the corpus say "Committed Suicide" as a map term and presents the corresponding reduce task by aggregating the document corresponding to document id and news title for the value.

Data Analytics Task Using R

The resultant documents obtained from the Reduce task are imported into R environment using an API rmongodb.



```
> db.crime.mapReduce( function(){ emit(this._id,this.title); } ,   function(key,value){ return value.join() },  {query:{$text:{$
earch:"committed suicide"} }, out:{inline: 1}})
{
        "results" : [
                {
                        "_id" : ObjectId("582cd0f8ee090f2b18bafb12"),
                        "value" : "Actor committed suicide, say police - TAMIL NADU - The Hindu"
                },
                {
                        "_id" : ObjectId("582cd279ee090f2b18bafb16"),
                        "value" : "Married woman, engineering college student commit suicide - TAMIL NADU - The Hindu"
                },
                {
                        "_id" : ObjectId("582cdd1eee090f2b18bafb19"),
                        "value" : "Married woman, engineering college student commit suicide - TAMIL NADU - The Hindu"
                },
                {
                        "_id" : ObjectId("582f71be0234c058b0f0ffea"),
                        "value" : "Man kills wife, attempts suicide - TAMIL NADU - The Hindu"
                },
                {
                        "_id" : ObjectId("582f71fd0234c058b0f0ffeb"),
                        "value" : "Girl commits suicide; mother alleges torture by youth - TAMIL NADU - The Hindu"
                }
        ],
        "timeMillis" : 1,
        "counts" : {
                "input" : 5,
                "emit" : 5,
                "reduce" : 0,
                "output" : 5
        },
        "ok" : 1
}
```


Figure 6: Sample Output of MR Query executed under MongoDB

Table 2 provides the text preprocessing for the documents obtained from the reducer in R.

Table 2 : POS Tagging for particular news

Token	Tag	Lemma	Lttr	Wcl
Man	NN	man	3	noun
Held	VBN	held	4	verb
For	IN	for	3	preposition
sexually	RB	sexually	8	adverb
abusing	VBG	abusing	7	verb
daughter	NN	daughter	8	noun

Table 3: Classification of Crimes under different Categories

S.No	Types of Crime	No.of Documents
1	Sexual Harassment	**2386**
2	Raped + Murdered	**282**
3	Impregnate + Minor	**195**
4	Cheated + illegal relation	**39**
5	Verbal abuse	**23**
6	Sexual torture + Committed Suicide	**44**

After preprocessing the collection, the documents are classified using Naïve Bayes Classifier. The table 3 presents the no. of documents classified based on keys emitted in the reducer phase.Fig.7 shows the documents grouped under each classification for the key-value pair. The cluster labeled as 1 represents the document classified for the key 'sexual harassment'. The cluster labeled as 2 and 3 represents the documents classified for the key 'verbal abuse' and 'raped and murdered'. The cluster labeled as 4, 5 and 6 represents the documents grouped under 'impregnate minor', 'cheated illegal relation' and 'sexual torture commit suicide' respectively. The cluster overlaps indicates overlap of documents grouped under different crime abuses.

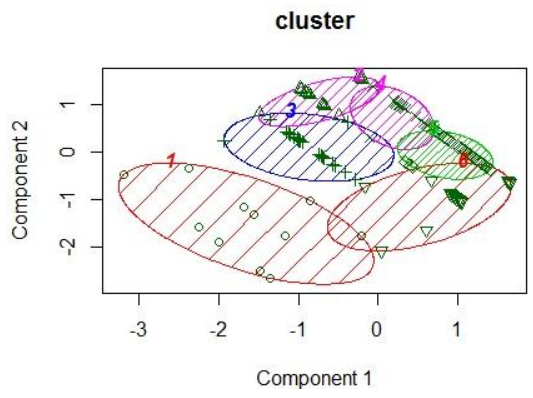


Figure 7: Crime categories plotted in Clusters

DISCUSSION

The framework VAW used in this work helps to group documents automatically based on the crime abuses. The Unstructured documents when processed using conventional databases and text processing requires huge amount of processing time. The contribution of our framework efficiently reduces the processing time as massive parallel processing architecture is used to execute the methodology. The outcome of this work when applied on huge volume on real dataset provide an insight of crime abuses which can be analyzed under the categories of age and location against crime. This also helps to do an analysis on legally registered crimes which will be implemented as our future work.

CONCLUSION AND FUTURE WORK

This work deals with crime investigation by considering articles from the newspapers and blogs using MR framework implemented on MPP Hadoop Ecosystem. The experimental result is found good for automatic classification and grouping of documents under various crime abuse categories. To process all the 3000 documents it took only 72 milliseconds which is very minimum. A Generic framework and methodology will be developed to classify crime categories automatically for including data multitude of sources like tweets and whatsapp data. This work can also be enhanced by deploying Hadoop under multi-node cluster setup and MongoDB under sharded cluster mode.

REFERENCES

[1] Mayer Schonberger and V,Cuckier K, "Big data : a revolution that will transform how we live,work and think", 2013.

[2] Doug Laney, "3d data management : controlling data volume, velocity and variety", Appl. Delivery Strategies Meta Group, 2001.

[3] IBM Data Growth and Standards : http://www.ibm.com/developerworks/xml/library/x-datagrowth/index.html?ca=drs , 2012.

[4] K.Cukier, "Data,data everywhere", Economist,2010, pp. 3-16.

[5] Amir Gandomi, Murtaza HaiderTed, "Beyond the hype : big data concepts, methods and analytics," International Journal of Information Management, pp. 137-144, 2015.

[6] Min Chen, Shiwen Mao, Yunhao Liu, "Big data : a survey", Mobile Netw Appl, Springer, 2014, pp.171-209.

[7] Catell R, "Scalable Sql and nosql data stores", ACM Sigmod Record, 2011,pp. 12-27.

[8] Jongseong Yoon, Doowon Jeog, "Forensic investigation framework for the document store nosql dbms : mongodb as a case study," Digital Investigation, Elsevier Publications, May 2016.

[9] Crockford D, The application/json media type for javascript object notation (JSON) , 2006.

[10] Chodorow K, MongoDB : the definitive guide. O'Reilly Media Inc, 2013.