A thesis paper submitted to the Department of Computer Science and Engineering, Shah Jalal University of Science & Technology, in partial fulfillment of the requirement for the degree of Bachelor of Science (Engineering).

Supervisor

Md. Saiful Islam

Lecturer, Dept. of CSE

Shahjalal University of Science &

Technology, Sylhet-3114

Thesis Committee

Prof. Dr. Shahidur Rahman

Head, Dept. of CSE

Shahjalal University of Science &

Technology, Sylhet-3114.

5th May 2014

Thesis on Information Retrieval:

Predicting crime occurrence retrieving news from different online newspapers.

Department of Computer Science and Engineering

Shahjalal University of Science and Technology

Sylhet, Bangladesh

CSE 404 Report

Submitted By

Rajib Chandra Das (2009331008)

Md. Ruhul Amin (2009331011)

Md. Jumman Hossain (2009331054)

Acknowledgement:

We are grateful to our supervisor “Md. Saiful Islam” for giving us cordial support and direction to our working progress. We also thank to the author and websites from which we get so much knowledge and support for our thesis. And finally special thanks to Pipilika Bangla search engine for providing valuable data.

Abstract:

There have thousands of crime are happening daily all around. But people keep their heed only few of them. Less concern or less statistics both are increasing daily crime rates. Providing much statistics on crime will surely be significant for general people, police or tourists to make their travelling decision.

This research report describes practical, learning-driven area based future crime occurrence prediction mechanism. Our approach relies on different online Bangla newspapers from where the crime data has been collected to provide a map consisting crimes of different areas along with a future crime occurrence prediction model.

Table of Contents

1. Introduction

Constantly, there have lots of crime happening all around. Most crime is not reported to the police so there is lot of room for error. Law enforcement agencies can affect the amount of crime reported through aggressive interactions with citizens. Crime statistics are confusing and frequently misunderstood. There are criminologists who spend their professional lives investigating the complexity of crime data. National Institute of Justice release crime survey data for the country based on reported and unreported crime and does not offer crime statistics for states, metro areas or cities.

Most crime rankings are based on crimes per 1,000 residents which immediately creates an unfair playing field if you get thousands of tourists or workers per day. Those thousands of "outsiders" will inevitably commit crimes or inadvertently create opportunities for crime that would not exist in cities or states not getting a lot of tourists or daily workers.

So the bottom line is that crimes and crimes reported can and will differ for reasons having little or nothing to do with the quality of policing or crime control strategies.

Considering all of these cases we have come up with a solution which can give an approximation to people about the safety of a specific location with crime ranking if different areas. Our solution collects crime information from different online newspapers and point them on a local map with ranking.

2. Our goal

Our goal is to design a future crime occurring prediction system alongside area based crime ranking. Where, to get the crime occurrence of different locations have been collected after crawling different news from different Bangla online newspaper. This system also provides a map where different crime has been pointed according to their occurrence zone. This will definitely be helpful for general people, tourists to decide their outing and police to know where the crime is happening frequently.

3. Out Studies:

3.1. Web Crawler

A web crawler is a program that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks. Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. Moreover, they are used in many other applications that process large numbers of web pages, such as web data mining, comparison shopping engines, and so on. Despite their conceptual simplicity, implementing high-performance web crawlers poses major engineering challenges due to the scale of the web. In order to crawl a substantial fraction of the "surface web" in a reasonable amount of time, web crawlers must download thousands of pages per second, and are typically distributed over tens or hundreds of computers. Their two main data structures – the "frontier" set of

yet-to-be-crawled URLs and the set of discovered URLs – typically do not fit into main memory, so efficient disk-based representations need to be used. Finally, the need to be "polite" to content providers and not to overload any particular web server, and a desire to prioritize the crawl towards high-quality pages and to maintain corpus freshness impose additional engineering challenges.

3.2. How does a web crawler work

When a search engine's web crawler visits a web page, it "reads" the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich Meta tags. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. The website is then included in the search engine's database and its page ranking process.

Web crawlers may operate one time only, say for a particular one-time project. If its purpose is for something long-term, as is the case with search engines, web crawlers may be programed to comb through the Internet periodically to determine whether there has been any significant changes. If a site is experiencing heavy traffic or technical difficulties, the spider may be programmed to note that and revisit the site again, hopefully after the technical issues have subsided.

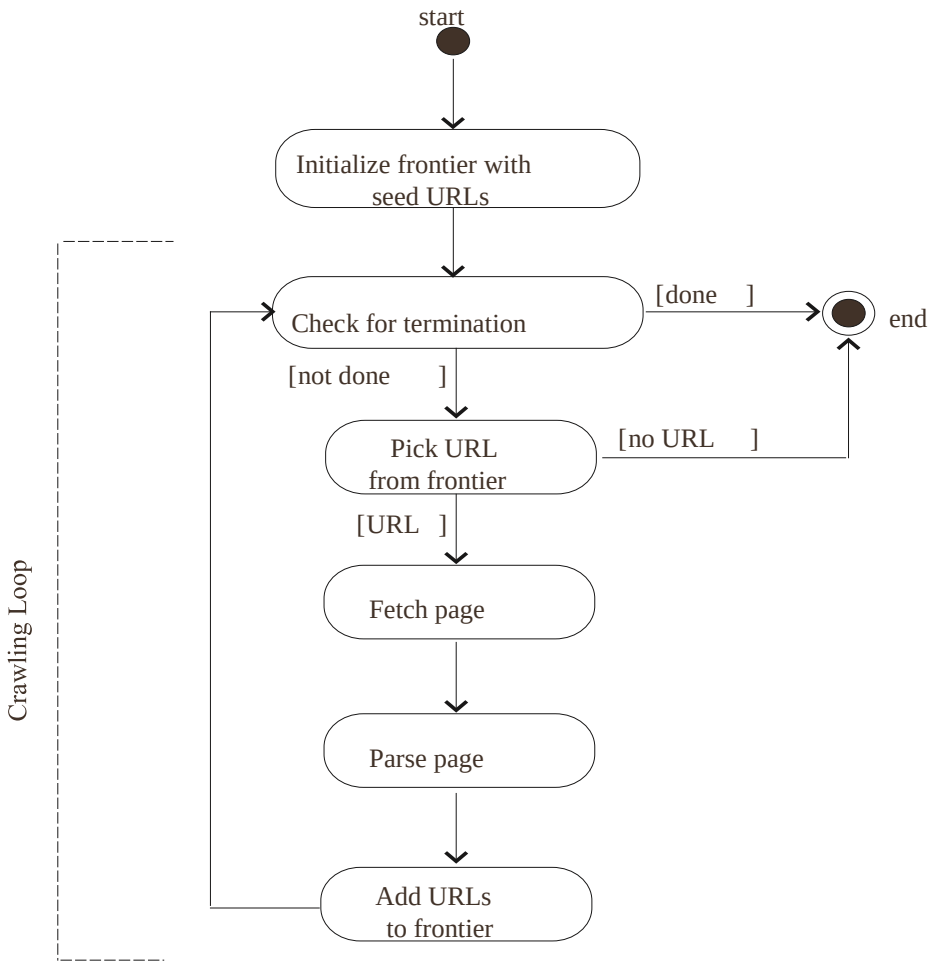


Fig. 1. Flow of a basic sequential crawler

3.3. Parsing

Parsing may imply simple hyperlink/URL extraction or it may involve the more complex process of tidying up the HTML content in order to analyze the HTML tag tree. Parsing might also involve steps to convert the extracted URL to a canonical form, remove stop words from the page's content, and stem the remaining words. These components of parsing are described next.

3.4. Stop-word listing and Stemming

When parsing a Web page to extract content information or in order to score new URLs suggested by the page, it is often helpful to remove commonly used words or stop words such as "it" and "can". This process of removing stop words from text is called stop listing. The stemming process normalizes words by conflating a number of morphologically similar words to a single root form or stem. For example, "connect," "connected," and

"connection" are all reduced to "connect." Implementations of the commonly used Porter stemming algorithm are easily available in many programming languages. One of the authors has experienced cases in the biomedical domain where stemming reduced the precision of the crawling results.

3.5. Keyword Extraction

Keyword extraction is an important technique for document retrieval, Web page retrieval, Document clustering, summarization, text mining, and so on. By extracting appropriate keywords, we can easily choose which document to read to learn the relationship among documents. A popular algorithm for indexing is the TF-IDF measure, which extracts keywords that appear frequently in a document, but that don't appear frequently in the remainder of the corpus. The term "keyword extraction" is used in the context of text mining, for example 15. A comparable research topic is called "automatic term recognition" in the context of computational linguistics and "automatic indexing" or "automatic keyword extraction" in information retrieval research.

3.6. Cosine Similarity

The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (TF-IDF) of each document, but the angle between the documents. What we have to do to build the cosine similarity equation is to solve the equation of the dot product for the $\cos\theta$:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\|\|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

And that is it, this is the cosine similarity formula. Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude.

3.7. TF-IDF

TF–IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.

Variations of the TF–IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. TF–IDF can be

successfully used for stop-words filtering in various subject fields including text summarization and classification.

3.8. Jaccard Similarity

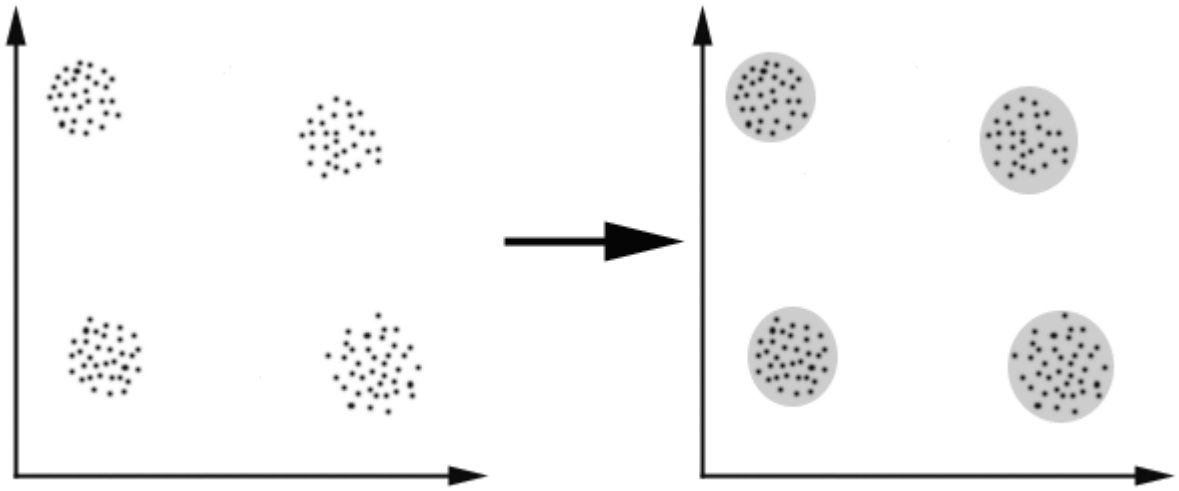
Jaccard similarity is statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

3.9. Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:



In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

3.10. K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the n data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

3.11. Naïve Bayes Classifier

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

3.12. Hidden Markov Model

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be considered the simplest dynamic Bayesian network.

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of

states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

3.13. Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

4. Implementation

4.1. News crawling

For crawling the news we have used Pipilika search engines crawler. News contains the title of the news, domain name, published date and location in some cases. The crawled news looks like this:

<Index>
<filePath>F:\Data WareHouse\small repository\Crawler_Data\www.amadershomoy2.com\12-7-2012-12-48-16</filePath>
<byteInfo>1 3700861728 3700861791 3700861791 3700870240</byteInfo>
<indexed>true</indexed>
<TITLE> রাজধানীতে ছিনাতাইকারীর ছুরিকাঘাতে নিহত ১, আহত ১</TITLE>
<CONTENT> রাজধানীতে ছিনাতাইকারীর ছুরিকাঘাতে নিহত ১, আহত ১ ||
অসধফবৎঝযড়সডু.ঈড়স (আমাদের সময়.কম) জুলাই ১২, ২০১২, বৃহস্পতিবার : আষাঢ় ২৮, ১৪১৯ । আপডেট বাংলাদেশ সময় রাত ১২:০০ অৎপযরাব আজজ€্বকর পাতাসমঞ্চহ
রাজধানীতে ছিনাতাইকারীর ছুরিকাঘাতে নিহত ১, আহত ১ নিজস্ব প্রতিবেদক আমাদের সময়.কম
ইসমাইল হোসেন ইমু ও জোনায়েদ মানসুর : রাজধানীর মহাখালীতে ছিনাতাইকারীর ছুরির আঘাতে মাছ ব্যবসায়ী নিহত। তার নাম হাফিজ উদ্দিন (৪০) । এ ঘটনায় ইউসুফ (১৫) নামে আরেক জন আহত হয়েছেন। আজ ভোর সাড়ে পাচঁটার দিকে এ ঘটনা ঘটে। ইউসুফকে ঢাকা মেডিকেল কলেজ হাসপাতালে ভর্তি করা হয়েছে। ইউসুফ হাসপাতালে সাংবাদিকদের জানায়, তারা দু""জনই গাজীপুর এলাকার মাছ ব্যবসায়ী। ভোরে কাওরান বাজারের উদ্দেশ্যে গাজীপুর থেকে রওনা হলে মহাখালী ফ্লাইওভারের নিচে পৌছঁালে একটি সাদা মাইক্রোবাস তাদের গতিরোধ করে। মাইক্রোবাস থেকে এক সাদা পোশাকধারী ছুরি বের করে তাদের জিম্মি করে। টাকা ও মোবাইল চাইলে তারা দিতে রাজি না হওয়ায় দুর্বৃত্তরা তাদের দু""জনকে এলোপাতাড়ি ছুরিকাঘাত করে পালিয়ে যায়। এ ব্যাপারে শিল্পাঞ্চল থানায় পুলিশ আজ বেলা সাড়ে ১১টা পর্যন্ত এ ঘটনা জানেনা বলে আমাদের সময়.কমকে

```
        জানায়। বিস্তারিত আসছে---------- স্থানীয় সময়: ১২.১১ ঘন্টা, ১২ জুলাই ২০১২ বদরুল
        বোরহান / </CONTENT>
                <CATEGORY> অন্যান্য</CATEGORY>
                <CITY> ঢাকা গাজীপুর</CITY>
                <DOMAIN> www.amadershomoy2.com</DOMAIN>
                <DATE> 201207120642</DATE>
                <URL>
        http://www.amadershomoy2.com/content/2012/07/12/middle0103.htm/</URL>
                <TYPE> news</TYPE>
                <PATH> F:\Data WareHouse\small
        repository\Crawler_Data\www.amadershomoy2.com\12-7-2012-12-48-
        16</PATH>
                <BYTE_INFO> 1 3700861728 3700861791 3700861791
        3700870240</BYTE_INFO>
</Index>
```

4.2. Parsing the crawled news

For parsing relevant data from the crawled news we have used Jsoup 1.7.1 library. We have got more than 500 MB crawled news Data from our research lab. Every news was stored with syntax of XML and they were well formatted in packet by packet. Here is the example of a news packet.

```
<Index>
        <byteInfo>1 11315933 11315980 11315980 11342687</byteInfo>
        <indexed>true</indexed>
        <TITLE>নরসিংদীতে ছাত্রলীগ নেতা খুন</TITLE>
        <CONTENT>নরসিংদী জেলা ছাত্রলীগের এক নেতাকে কুপিয়ে হত্যা করেছে দুর্বত্তরা। মঙ্গলবার বিকালে
                শিবপুর উপজেলার পুটিয়া বাজার এলাকায় এ হামলায় নিহত শাহিন কাদির মাহিন ওরফে
                মাইনুদ্দিন (৩০) জেলা ছাত্রলীগের সাংগঠনিক সম্পাদক বলে পুলিশ জানিয়েছে। সদর
                থানার ওসি মো. দুজ্জামান জানান, দুপুর ২টার দিকে মাহিন পুটিয়ার বাজারে নিজ ব্যবসা
                প্রতিষ্ঠানে যান। সেখান থেকে বিকালে ঘোড়াদিয়া এলাকায় বাড়ি ফেরার পথে কয়েকজন
                সন্ত্রাসী তাকে কুপিয়ে গুরুত্বর আহত করে। নরসিংদী জেলা সদর পাতালে নেয়ার পর
                কর্তব্যরত চিকিৎসক তাকে মৃত ঘোষণা করেন।
        </CONTENT>
        <CATEGORY>CRIME POLITICS DESH</CATEGORY>
        <CITY>ঢাকা</CITY>
        <DOMAIN>http://www.kalerkantho.com</DOMAIN>
        <DATE>201404250000</DATE>
        <URL>http://www.kalerkantho.com/feature/ronger-mela/2014/04/24/76173</URL>
        <TYPE>news</TYPE>
</Index>
```

From this news packet we need title, content and date. But we parsed all these news features for our performance measurement. Here we parsed title, content, category, city, domain, date, URL and type. Here we just became careful about the date and we parsed date feature manually. Here is the source code for parsing news packet.

```
Void ParseDocument(String data)
{
   Document doc = Jsoup.parse(data, "", Parser.xmlParser());
   Elements el = doc.select("Index");
   String title,content,category,city,domain,date,url, type;

   for (Element e : el) {
      title = (e.select("TITLE").text());
      content = (e.select("CONTENT").text());
      category = (e.select("CATEGORY").text());
      city = (e.select("CITY").text());
      domain = (e.select("DOMAIN").text());
      date = (e.select("DATE").text());
      url = (e.select("URL").text());
      type = (e.select("TYPE").text());
      date = PurifyDateString(date);
      //Do whatever I want
   }
}

private String PurifyDateString(String date) {
  if (date.contains("-"))
  {
     String[] str = date.split("[ -]+");
     dd = str[2] + "-" + str[1] + "-" + str[0];
  }
  else
  {
     dd += date.substring(0, 4);                dd += "-";
     dd += date.substring(4, 6);                dd += "-";
     dd += date.substring(6, 8);
  }
  return dd;
}
```

4.3. Indexing root words and stop words

Bengali has as many as 1, 00,000 unique words, of which 50,000 are considered 'Totsomo' (তৎসম- direct re-borrowings from Sanskrit), 21,100 are 'Todbhobo' (তদ্ভব- native words) and the rest are 'Bideshi' (বিদেশী foreign borrowings) and 'Deshi' (দেশী- Austroasiatic borrowings). All Bengali words are categorized into 5 parts of speech: Noun, Adjective, Pronoun, Conjunction and Verb. Nouns and Pronouns are inflected for case, including nominative, objective, genitive (possessive) and locative. The case marking pattern for each noun being inflected depends on the noun's degree of animacy. When a definite article such as – টা-ta (Singular) or গুলা-gula (plural) is added, as in the tables below, nouns are also inflected for number.

Singular noun inflection

	Animate	**Inanimate**
Nominative	ছাত্রটা chhatrô-ṭa the student	জুতাটা juta-ṭa the shoe
Objective	ছাত্রটা**কে** chhatrô-ṭa-**ke** the student	জুতাটা juta-ṭa the shoe
Genitive	ছাত্রটা**র** chhatrô-ṭa-**r** the student's	জুতাটা**র** juta-ṭa-**r** the shoe's
Locative	-	জুতাটা**য়** juta-ṭa-**y** on/in the shoe

Plural noun inflection

	Animate	**Inanimate**
Nominative	ছাত্র**রা** chhatrô-**ra** the students	জুতাগুলা/জুতোগুলো juta-gula/juto-gulo the shoes
Objective	ছাত্র**দের(কে)** chhatrô-**der(ke)** the students	জুতাগুলা/জুতোগুলো juta-gula/juto-gulo the shoes
Genitive	ছাত্র**দের** chhatrô-**der** the students'	জুতাগুলা/জুতোগুলো**র** juta-gula/juto-gulo-**r** the shoes'
Locative	-	জুতাগুলা/জুতোগুলো**তে** juta-gula/juto-gulo-**te** on/in the shoes

A document has a list of words containing useless words and useful words. For example we may look over a sample Bengali Article:

Sample Bengali Article: ঢাকার বনানী এলাকায় এক বিকাশ প্রতিনিধিকে কুপিয়ে ৭ লাখ টাকা ছিনতাই করেছে দুর্বৃত্তরা। রোববার দুপুর দেড়টায় রাজধানীর সবুজবাগ থানার বাসাবো ওয়াসা রোডে জিন ইন্টারন্যাশনাল নামের বিকাশ এজেন্সির বিক্রয় প্রতিনিধি এনামুল হককে (৪০) কুপিয়ে সাত লাখ টাকা ছিনতাই করে। তিনি ঢাকা মেডিকেল কলেজ হাসপাতালে (ঢামেক) চিকিৎসাধীন রয়েছেন। এনামুল হক জানান, ওই স্থানে ৫/৬ দুর্বৃত্ত তার গতিরোধ করে ধারালো অস্ত্র দিয়ে কুপিয়ে হাতে থাকা টাকার ব্যাগটি নিয়ে যায়। ব্যাগে সাত থেকে আট লাখ টাকা ছিল বলে তিনি জানান। ঢামেকের কর্তব্যরত চিকিৎসক জানান, এনামুল হকের হাতে, বুকে ও পিঠে ধারালো অস্ত্রের আঘাত আছে। জিন ইন্টারন্যাশনালের সুপারভাইজার তানভির নেওয়াজ খান জানান, খবর পেয়ে প্রথমে এনামুল হককে বাসাবো জেনারেল হাসপাতালে নিয়ে যাওয়া হয়। পরে ভাল চিকিৎসার জন্য তাকে ঢামেকে স্থানান্তর করা হয়। সবুজবাগ থানার ভারপ্রাপ্ত কর্মকর্তা (ওসি) বাবুল মিয়া এ প্রসঙ্গে জানান, মারামারির ঘটনা শুনেছি। ছিনতাই কীনা খতিয়ে দেখা হচ্ছে।

Here we have found out a set of stop words: {“এক”, “নামের”, “করে”, “তিনি”, “রয়েছেন”, “জানান”, “তার”, “করে”, “দিয়ে”, “থাকা”, “নিয়ে”, “যায়”, “ছিল”, “বলে”, “আছে”, “প্রথমে”, “যাওয়া”, “হয়”, “পরে”, “করা”, “শুনেছি”, “কিনা”, “ভাল”, “দেখা”, “হচ্ছে”}. And the rest words are considered as main word (Keyword). We have studied more than 500 news articles and stored more than 1000 stop words in a text file.

Storing root word is not so easier like stop word. There are more than 40,000 words are used in bangladeshi news articles. In order to compute root word from a sample word we have to study about bangla grammartical rules (পদ, প্রকৃতি, প্রত্যয়, উপসর্গ, কারক, বিভক্তি, সন্ধি-বিচ্ছেদ) and also required some natural language processing. But that was not our concern. So we stored more than 4,500 words and corresponding root word in a text file. The root word map example is shown below:

পুলিশে পুলিশ
পুলিশকে পুলিশ
পুলিশের পুলিশ
পুলিশও পুলিশ
পুলিশদের পুলিশ
পুলিশরা পুলিশ
পুলিশদেরকে পুলিশ
খেলায় খেলা
খেলার খেলা
খেলতে খেলা
খেলাকে খেলা

4.4. Extracting Top words from news

We have studied about more than 500 news articles and observed that every Pronoun(সর্বনাম) and Conjunction(অব্যয়) are stop word**s**. Pronouns and Conjunctions are not used to define a news category. Pronouns and Conjunction are added to stop word list. We can also add several verb(ক্রিয়া) and Adjective(বিশেষন) words to stop word list. Actually we need all the Nouns (বিশেষ্য), several verbs (ক্রিয়া) and adjectives (বিশেষন).

We picked top words from about 250 news articles. First of all, we counted term frequency for every term. The term which is not a stop word and exist more than 5 times we picked it as a top word. Here is the sample term frequencies of our news data.

পুলিশ -->crime = 200 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 9 Total: 209
দল -->crime = 24 -->sports = 61 -->entertainment = 1 -->technology = 1 -->others = 31 Total: 118
ডাকাত -->crime = 72 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 72
মামলা -->crime = 55 -->sports = 0 -->entertainment = 2 -->technology = 2 -->others = 2 Total: 61
শিশু -->crime = 32 -->sports = 0 -->entertainment = 2 -->technology = 0 -->others = 25 Total: 59
খুন -->crime = 55 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 55
অভিনয় -->crime = 1 -->sports = 0 -->entertainment = 50 -->technology = 0 -->others = 0 Total: 51
হাসপাতাল -->crime = 44 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 6 Total: 50
ছিনতাই -->crime = 50 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 50
কর্মকর্তা -->crime = 34 -->sports = 2 -->entertainment = 0 -->technology = 5 -->others = 9 Total: 50
হত্যা -->crime = 49 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 49
অভিযোগ -->crime = 40 -->sports = 0 -->entertainment = 1 -->technology = 0 -->others = 5 Total: 46
ম্যাচ -->crime = 1 -->sports = 44 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 45
আহত -->crime = 41 -->sports = 0 -->entertainment = 2 -->technology = 0 -->others = 2 Total: 45
বিশ্বকাপ -->crime = 0 -->sports = 40 -->entertainment = 1 -->technology = 0 -->others = 0 Total: 41
নিহত -->crime = 37 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 3 Total: 40
গোল -->crime = 0 -->sports = 35 -->entertainment = 0 -->technology = 0 -->others = 2 Total: 37
লাশ -->crime = 34 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 1 Total: 35
ধর্ষণ -->crime = 30 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 2 Total: 32
মেয়ে -->crime = 20 -->sports = 2 -->entertainment = 9 -->technology = 0 -->others = 0 Total: 31
স্ত্রী -->crime = 24 -->sports = 2 -->entertainment = 2 -->technology = 0 -->others = 2 Total: 30
খেলা -->crime = 3 -->sports = 23 -->entertainment = 2 -->technology = 0 -->others = 2 Total: 30
আসামি -->crime = 30 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 30
রিয়াল -->crime = 0 -->sports = 27 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 27
কোচ -->crime = 0 -->sports = 27 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 27
ফাইনাল -->crime = 0 -->sports = 25 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 25
গুলি -->crime = 24 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 24
অভিনেত্রী -->crime = 0 -->sports = 0 -->entertainment = 22 -->technology = 0 -->others = 2 Total: 24

মাঠ	-->crime = 4	-->sports = 16	-->entertainment = 1	-->technology = 1	-->others = 0	Total: 22
পয়েন্ট	-->crime = 2	-->sports = 14	-->entertainment = 1	-->technology = 0	-->others = 5	Total: 22
দুর্বৃত্ত	-->crime = 22	-->sports = 0	-->entertainment = 0	-->technology = 0	-->others = 0	Total: 22
গান	-->crime = 0	-->sports = 0	-->entertainment = 18	-->technology = 0	-->others = 4	Total: 22
অস্ত্র	-->crime = 21	-->sports = 1	-->entertainment = 0	-->technology = 0	-->others = 0	Total: 22
হামলা	-->crime = 18	-->sports = 0	-->entertainment = 0	-->technology = 1	-->others = 2	Total: 21
মাইক্রোসফট	-->crime = 0	-->sports = 0	-->entertainment = 0	-->technology = 21	-->others = 0	Total: 21
অভিনেতা	-->crime = 0	-->sports = 0	-->entertainment = 20	-->technology = 0	-->others = 0	Total: 20
জয়	-->crime = 1	-->sports = 13	-->entertainment = 1	-->technology = 0	-->others = 3	Total: 18
খেলোয়াড়	-->crime = 0	-->sports = 18	-->entertainment = 0	-->technology = 0	-->others = 0	Total: 18
মৃত্যু	-->crime = 13	-->sports = 0	-->entertainment = 2	-->technology = 0	-->others = 2	Total: 17

We pick a term as a top word which is more frequent in news articles.

4.5. Categorizing the news

We used Naïve Bayes text classification for news article categorization. The probability of a document d being in class c is computed as

Here, $P(c) = \frac{\text{Number of document in Category,c}}{\text{Number of Total Document}}$

nd = Number of term in Document d

$P(t_k|c) = \frac{\text{Term Frequency in Category,c}}{\text{Term Frequence in all Documents}}$

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}},$$

Now we will compute probability for every category individually. Maximum value will be defined by specific category.

$$c_{map} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c).$$

In this equation many conditional probabilities are multiplied, one for each position $1<=k<=n_d$. This can result in a floating point underflow. It is therefore better to perform the computation by adding logarithms of probabilities instead of multiplying probabilities. The class with the highest log probability score is still the most probable. So the equation will be changed to,

$$c_{map} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c)].$$

To eliminate zeros, we use add-one technique, which simply adds one to each count.

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

Here is the source code of implementation of naive bayes clustering:

```
private String Naive_Bayes_Clustering(String document)
{

   double Likelihood_Probability, Posterior_Probability;
   double Prior_Probability[] = new double[5];
   Occurance O;
   String words[] = document.split(" ");
   for(int i=0;i<Prior_Probability.length;i++)
   {

     Prior_Probability[i]= Math.log10(documents_kount[i]/
                           total_documents);

     for(String term:words)
     {
        if(rawData.useless.contains(term)) continue;
        if(rawData.rootWord.containsKey(term))
            term = rawData.rootWord.get(term);

        if(preeData.TermFrequency.containsKey(term))
             O = preeData.TermFrequency.get(term);
        else
             O = new Occurance();

        Likelihood_Probability = (O.kounter[i]+1)*1.0 /
                                 (O.total + 1)*1.0;


      Prior_Probability[i]+=Math.log10(Likelihood_Probability);
     }

   }
   int indx=0;
   double argMax=-(1<<30);
```

```
    for(int i=0;i<Prior_Probability.length;i++)
    {
       if(Prior_Probability[i]>argMax)
       {
          argMax  = Prior_Probability[i];
          indx=i;
       }
    }

    return getNameOf(indx);
}
private String getNameOf(int id)
{
    if(id==0) return "crime";
    if(id==1) return "sports";
    if(id==2) return "entertainment";
    if(id==3) return "technology";
    return "others";

}
```

Now, Data for parameter estimation example for Naïve Bayes Clustering:

Contents	**Category**
১৪ বছর আগে চট্টগ্রামের বহদ্দারহাটে ছাত্রলীগের গাড়িতে হামলা চালিয়ে আট জনকে হত্যার য়ে মৃত্যুদণ্ডে দণ্ডিত চার আসামির সবাই আপিলের রায়ে খালাস পেয়েছেন।	Crime
রাজধানীতে দুর্ধর্ষ ডাকাতি । বাসার দুই দারোয়ানকে বেঁধে পার্কিং করা একটি পালসার ব্রান্ডের মোটরসাইকেল নিয়ে যায় ও একটি প্রাইভেটকারের (ঢাকা মেট্রো-গ-৩৫-৩৫৯৯) যন্ত্রাংশ খুলে নিয়ে যায়।	Crime
চুয়াডাঙ্গার দামুড়হুদায় গভীর রাতে বাসায় ঢুকে আমজাদ হোসেন (৪৫) নামের এক ব্যক্তিকে কুপিয়ে হত্যা করেছে সন্ত্রাসীরা। শুক্রবার রাতে উপজেলার জয়রামপুর গ্রামের চৌধুরী পাড়ায় এ ঘটনা ঘটে। নিহতের লাশ শনিবার সকাল ৭টার দিকে উদ্ধার করে ময়নাতদন্তের জন্য চুয়াডাঙ্গা সদর হাসপাতালের মর্গে পাঠিয়েছে পুলিশ।	Crime
টি-টোয়েন্টি বিশ্বকাপের চ্যাম্পিয়ন শ্রীলঙ্কা এখনো পর্যন্ত টি-টোয়েন্টি বিশ্বকপের সবগুলো ম্যাচেই জয়লাভ করেছে ভারত। আইসিসি টি২০ বিশ্বকাপের ফাইনালে ভারতকে ৬ উইকেটে হারিয়ে শিরোপা নিজেদের ঘরে নিয়েছে শ্রীলঙ্কা।	Sports
নতুন একটা সোনাক্ষি বলিউডে প্রধান আলোচ্য এখন একটাই- ওজন কমিয়েছেন সোনাক্ষি সিনহা! এবং প্রেম করছেন শহিদ কাপুরের সঙ্গে। 'ডেইলি টেলিগ্রাফ' অবলম্বনে সোনাক্ষিকে নিয়ে লিখেছেন	Entertainment

শাকিল ফারুক লোকের কৌতূহলের সীমারেখা থাকে না আসলে। নইলে ভারতের এমন নির্বাচনী ডামাডোলে কংগ্রেস-বিজেপি-আম আদমি পার্টিকে সরিয়ে, সোনাক্ষি সিনহার ওজন প্রসঙ্গ কি আর আলোচনার মুখ্য বিষয় হয়ে উঠতে পারে! সোনাক্ষি যত বিরক্তি নিয়েই বলুন, 'এ নিয়ে কথা বলতে আগ্রহী নই'- আলোচনা চলবেই।	
অঅ-অ+ টিভি পর্দায় জাহিদ হাসানের নানা রূপ। পুরনো রূপ ভেঙে তিনি আবার নতুন রূপে হাজির হয়েছেন 'নজিরবিহীন নজির আলী' নাটকে। লিখেছেন মাহ্‌বুব হাসান জ্যোতি 'আমি ভাই ভিন্ন ধরনের চরিত্রের কাঙাল'- আলাপচারিতার শুরুতেই বললেন জাহিদ হাসান। আরমান ভাইয়ের চরিত্র ছাপিয়ে তিনি এখন নজর আলী হয়ে উঠছেন।	Entertainment
বিশ্বকাপের সঙ্গে ফ্রান্সের কোচের ভাবনায় ইউরো ইউরোপের সবচেয়ে বড় ফুটবল আসরের আয়োজক ফ্রান্স। তাই এখন থেকেই ভবিষ্যত সাফল্যের কথা ভাবতে হচ্ছে ফরাসি এই কোচকে।	Sports
অর্থনৈতিক উন্নতি ও নৈতিক দায় ভোরবেলা যাচ্ছিলাম কমলাপুর স্টেশনে। রাস্তায় খুব কম যানবাহন। তবু এক সিগন্যালে একটু থামতে হলো বাঁ দিক থেকে কয়েকটি গাড়ি ক্রসিং পার হয়ে ডান দিকে আসায়। একটি পিকআপ পাশ ঘেঁষে থামতে বাধ্য হলো। অত ভোরে কোনো বোকা চালকও লালবাতি মানেন না। পিকআপে দুটো চটের বস্তা। একটি বস্তার মুখের দিকে সামান্য ফাঁক দিয়ে দেখা গেল একটি মরা মুরগির পা ও পাখনা। চল্লিশ-পঞ্চাশ বছর আগে হলে মনে করতাম মরা মুরগি কুড়িয়ে কেউ ডাস্টবিনে ফেলে দিতে নিয়ে যাচ্ছে। এখন মনে পড়ল অন্য কথা। আমার কর্তব্য ছিল গাড়িটিকে আটকে পুলিশকে খবর দেওয়া। তা না করতে পারায় গ্লানি ও অপরাধ বোধ করি।	Others
গ্যালাক্সি এস৫ মিনি আসছে গ্যালাক্সি এস৫ স্মার্টফোনটির একটি মিনি বা ছোট সংস্করণ আসছে। গ্যালাক্সি এস৫ এর এ সংস্করণটি হবে পানি-রোধী। স্যামসাং নিউজিল্যান্ডের অফিশিয়াল ওয়েবসাইটে এ তথ্য জানানো হয়েছে। অবশ্য মিনি সংস্করণটির তথ্য এখনও আনুষ্ঠানিকভাবে ঘোষণা করেনি দক্ষিণ কোরিয়ার প্রতিষ্ঠানটি।	Technology

Now here is the data structure for calculating Term Frequency from learned data.

সোনাক্ষি -->crime = 0 -->sports = 0 -->entertainment = 4 -->technology = 0 -->others = 0 Total: 4
বিশ্বকাপ -->crime = 0 -->sports = 4 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 4
ভারত -->crime = 0 -->sports = 2 -->entertainment = 1 -->technology = 0 -->others = 0 Total: 3
গ্যালাক্সি -->crime = 0 -->sports = 0 -->entertainment = 0 -->technology = 3 -->others = 0 Total: 3
হত্যা -->crime = 2 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2
শ্রীলঙ্কা -->crime = 0 -->sports = 2 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2
রূপ -->crime = 0 -->sports = 0 -->entertainment = 2 -->technology = 0 -->others = 0 Total: 2
রাত -->crime = 2 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2

মরা -->crime = 0 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 2 Total: 2
মন -->crime = 0 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 2 Total: 2
ফ্রান্স -->crime = 0 -->sports = 2 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2
পুলিশ -->crime = 1 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 1 Total: 2
তথ্য -->crime = 0 -->sports = 0 -->entertainment = 0 -->technology = 2 -->others = 0 Total: 2
টোয়েন্টি -->crime = 0 -->sports = 2 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2
জাহিদ -->crime = 0 -->sports = 0 -->entertainment = 2 -->technology = 0 -->others = 0 Total: 2
চরিত্র -->crime = 0 -->sports = 0 -->entertainment = 2 -->technology = 0 -->others = 0 Total: 2
গাড়ি -->crime = 1 -->sports = 0 -->entertainment = 0 -->technology = 0 -->others = 1 Total: 2
কোচ -->crime = 0 -->sports = 2 -->entertainment = 0 -->technology = 0 -->others = 0 Total: 2

Here is a new document:

ডাকাতি মামলার আসামিকে কুপিয়ে হত্যা
ঢাকার কেরানীগঞ্জে বাবুল ওরফে হক বাবুল (৩২) নামের এক ব্যক্তিকে কুপিয়ে হত্যা করেছে দুর্বৃত্তরা।

Here is the calculated probabilities using this algorithm:

crime: -0.47712125471966244
sports: -3.112605001534575
entertainment: -3.112605001534575
technology: -3.4136349971985562
others: -3.4136349971985562

Category Calculated: crime (As it is the Maximum Probability)

4.6. Extracting locations and dates

For ranking the locations based on the crime scene it is obvious to find the find the exact crime locations. To do this first we store the locations. Here, we map all the Thana and district's name and store them on the database. It'll be more accurate if we store the corresponding union name. We find the location at the time of finding the root word and stop word. Here is the location finding algorithm.

Database:

location_id	Name	Parent	type	x	y
85	মঙ্গলা	বাগেরহাট	thana	706	2084
84	বাঘেরহাট	বাগেরহাট	thana	763	1845
83	চিতলমারী	বাগেরহাট	thana	801	1796
82	কচুয়া	বাগেরহাট	thana	709	1957
81	ইন্দুরকানিমোরলগঞ্জ	বাগেরহাট	thana	774	1951
80	বাগেরহাট	খুলনা	district	757	1916
79	লোহাগড়া	নড়াইল	thana	706	1613
78	কালিয়া	নড়াইল	thana	704	1693
77	নড়াইল	খুলনা	district	670	1638
76	হরিনাকুন্ড	ঝিনাইদহ	thana	471	1404
75	শৈলকূপা	ঝিনাইদহ	thana	537	1395
74	মহেশপুর	ঝিনাইদহ	thana	368	1558
73	কোটচাঁদপুর	ঝিনাইদহ	thana	442	1495
72	কলীগঞ্জ	ঝিনাইদহ	thana	507	1535
71	ঝিনাইদহ	খুলনা	district	491	1479
70	দামুরহুদা	চুয়াডাঙ্গা	thana	325	1439
69	জীবননগর	চুয়াডাঙ্গা	thana	371	1495
..	..	..	..	..	..
..	..	..	..	..	..
100	আলমডাঙ্গা	চুয়াডাঙ্গা	thana	384	1362

Here is the algorithm of location finding:

FindLocation(Doc)

1. A = set of all stored locations;
2. V = ∅;
3. L = ∅;
4. Foreach word ∈ Doc
    a. word = Findroot(word);
    b. If word ∈ A
        i. Loc = getLoc(word);
        ii. V = {V ∪ Loc};
        iii. Continue;
5. Foreach location ∈ V
    a. If location.type = "thana"
        i. L = {L, locatio0n}

Two Sample Documents:

1. দোহার উপজেলায় প্রবাসী নুরুল ইসলাম মাঝির বাড়িতে ডাকাতি হয়েছে। গত শনিবার রাতে উপজেলার নারিশা ইউনিয়নের ঝনকি গ্রামে এ ডাকাতি হয়। রাত ২টার দিকে ১২/১৫ জনের একটি সংঘবদ্ধ ডাকাত দল দেশীয় অস্ত্র নিয়ে উপজেলার নারিশা ইউনিয়নের ঝনকি গ্রামের কুয়েত প্রবাসী নুরুল ইসলাম মাঝির বাড়িতে হানা দেয়। ডাকাত দল বসত বাড়ির মূল গেটের কাঠের দরজা ভেঙে ভেতরে প্রবেশ করে। সেসময়ে পরিবারের সকলকে অস্ত্রের মুখে জিম্মি করে হাত বেঁধে আলমারিতে থাকা নগদ ২ লাখ ৫০ হাজার টাকা, ১০ ভরি স্বর্ণালঙ্কার, ২টি মোবাইল সেট লুটে নেয়।
2. রাজনগরে ডাকাতি, আহত ৩ রাজনগর (মৌলভীবাজার) প্রতিনিধি | ৩০ মার্চ ২০১৪, রবিবার, ৯:৩৬ রাজনগরে একদিনের ব্যবধানে আবারও দুর্ধর্ষ ডাকাতির ঘটনা ঘটেছে। ৮-১০ জনের ডাকাতদল অস্ত্রের মুখে জিম্মি করে ১৩ ভরি স্বর্ণালঙ্কার, নগদ ১ লাখ ৫০ হাজার টাকাসহ বিভিন্ন মালামল লুট করে নিয়ে যায়। ডাকাতদের হামলায় মহিলাসহ ৩ জন আহত হয়েছেন। আহতরা হলেন রন্টু পাল (৫০) তার স্ত্রী গীতা রানী পাল (৪০) ও ছেলে রনি পাল (১৫)। আহতদের বিভিন্ন মৌলভীবাজার ২৫০ শয্যার হাসপাতালে ভর্তি করা হয়েছে। শুক্রবার গভীর রাতে দক্ষিণ টেংরা গ্রামের রন্টু পালের বাড়িতে মুখোশ পড়া ৮-১০ জনের একদল ডাকাত হানা দেয়।

Here the location found is:

News Id	Thana	District
Document 1	দোহার	ঢাকা
Document 2	রাজনগর	মৌলভীবাজার

4.7. Finding similarity of different news

To find the similarity between different news, first we have calculated the Term Frequency (TF) of the inputted news, then the Inverse Document Frequency (IDF) and finally the cosine similarity. Measurement of cosine similarity ensures the similarity between different news**.** Similarity helps to distinguish the similar type crime in the particular area. We also use this cosine similarity to find the same news published in different newspapers. For finding the same news first we calculate the cosine similarity but cosine similarity is not enough for this. We also find the occurrence date and the location where this crime scene happened. If the date and location is same and similarity value is greater than a threshold value then we decide that this documents are same. Suppose we have several news having words like "ছিনতাই", "খুন", "ডাকাতি" and here we have calculated the TF value of those words in the below documents.

1. দোহার উপজেলায় প্রবাসী নুরুল ইসলাম মাঝির বাড়িতে ডাকাতি হয়েছে। গত শনিবার রাতে উপজেলার নারিশা ইউনিয়নের ঝনকি গ্রামে এ ডাকাতি হয়। রাত ২টার দিকে ১২/১৫ জনের একটি সংঘবদ্ধ ডাকাত দল দেশীয় অস্ত্র নিয়ে উপজেলার নারিশা ইউনিয়নের ঝনকি গ্রামের কুয়েত প্রবাসী নুরুল ইসলাম মাঝির বাড়িতে হানা দেয়। ডাকাত দল বসত বাড়ির মূল গেটের কাঠের দরজা ভেঙে ভেতরে প্রবেশ করে। সেসময়ে পরিবারের সকলকে অস্ত্রের মুখে জিম্মি করে হাত বেঁধে আলমারিতে থাকা নগদ ২ লাখ ৫০ হাজার টাকা, ১০ ভরি স্বর্ণালঙ্কার, ২টি মোবাইল সেট লুটে নেয়।
2. রাজধানী রামপুরার বনশ্রীতে দুর্ধর্ষ ডাকাতির ঘটনা ঘটেছে। মঙ্গলবার দিবাগত গভীর রাতে বনশ্রীর এফ ব্লকে ৪ নম্বর রোডের ২৩ নম্বর বাসায় এ ঘটনা ঘটে। জানা যায়, গভীর রাতে ৫/৭ জনের ডাকাত দল ওই বাসার কলাপসিবল গেট খুলে ভেতরে প্রবেশ করে। বাসার দুই দারোয়ানকে বেঁধে পার্কিং করা একটি পালসার ব্রান্ডের মোটরসাইকেল নিয়ে যায় ও একটি প্রাইভেটকারের (ঢাকা মেট্রো-গ-৩৫-৩৫৯৯) যন্ত্রাংশ খুলে নিয়ে যায়।

Similarity between this two documents is: 73.27995279592422%

4.8. Finding and removing the same news.

A single news can be published in different newspapers. For calculating the exact crime occurrence, we've to remove the repetitive news form the sample data. To calculate this we use cosine similarity for matching the document and then find the location and published date. If the crime occurrence date and locations are same and similarity is greater than a threshold value then we can assume that this two documents are same. Suppose we have N documents. Now doc_i and doc_j two same news from different news source. If the similarity between this documents $S_{(i, j)}$ is greater than 60% and the location of loc_i and loc_j is same and there publishing date $date_i$ and $date_j$ is same then this two document is similar. If there are N documents then the complexity is O (n^2). Because we need to calculate all pair similarity.

1. পাঞ্জাবকে দিয়ে জয়খরা কাটাল মুম্বাই, স্পোর্টস ডেস্ক, বাংলানিউজটোয়েন্টিফোর.কম

   Decrease font Enlarge font

   মুম্বাই: ওয়াংখেড়েতে ফিরে আইপিএলের সপ্তম আসরে ষষ্ঠ ম্যাচে এসে জয়ের দেখা পেল মুম্বাই ইন্ডিয়ান্স। গত আসরে নিজেদের মাঠে অজেয় দলটি হারাল এবারের টুর্নামেন্টে প্রথম পাঁচটিতেই জেতা কিংস ইলেভেন পাঞ্জাবকে।

   কিংস ইলেভেন পাঞ্জাব: ১৬৮/৫ (২০ ওভার)

   মুম্বাই ইন্ডিয়ান্স: ১৭০/৫ (১৯.১ ওভার)

   ফল: মুম্বাই জয়ী পাঁচ উইকেটে

   পাঞ্জাবের ছুড়ে দেওয়া ১৬৯ রানের লক্ষ্যে নেমে শুরুতে চোখ ধাঁধানো কোনো ইনিংস খেলেনি গতবারের চ্যাম্পিয়নরা। শেষ তিন ওভারে তাদের প্রয়োজন ছিল ৪১ রান। কিন্তু কাইরন পোলার্ড ও আদিত্য তারের শেষ সময়ের ঝড়ে পাঁচ বল বাকি থাকতে জয় পেল তারা।

   পোলার্ড ১২ বলে দুটি করে চার ও ছয়ে ২৮ রানে অপরাজিত ছিলেন। তারে খেলেছিলেন ছয় বলে একটি করে চার ও ছয়ে ১৬ রানের হার না মানা ইনিংস। এর আগে ২৩ রানের মধ্যে দুটি উইকেট হারালেও চিদাম্বরম গৌতম ও অধিনায়ক রোহিত শর্মার ব্যাটে এগিয়ে যায় মুম্বাই। চিদাম্বরম ২৯ বলে ৩৩ ও রোহিত ৩৪ বলে চারটি চার ও দুটি ছয়ে ৩৯ রান করেন। এটাই সেরা ইনিংস। এছাড়া কোরি এন্ডারসন ৩৫ রানের দ্বিতীয় সেরা ব্যাটিং করেন। ২৫ বলে তিনটি চার ও দুটি ছয়ে সাজানো ইনিংস খেলে ম্যাচসেরা হয়েছেন নিউজিল্যান্ডের এই তারকা।

   সন্দীপ শর্মা ও রিশি ধাওয়ান পাঞ্জাবের পক্ষে দুটি করে উইকেট নেন।

   এর আগে টস জিতে ব্যাট করতে নেমে ২৪ রানের মধ্যে দুটি উইকেট হারিয়ে বিপদে পড়েছিল পাঞ্জাব। তবে রিদ্ধিমান সাহা ও গ্লেন ম্যাক্সওয়েলের ব্যাটে লড়াই করার মতো সংগ্রহ করে দলটি। রিদ্ধিমান ৪৭ বলে চারটি চার ও তিনটি ছয়ে অপরাজিত ছিলেন ৫৯ রানে। ম্যাক্সওয়েল ২৭ বলে পাঁচটি বাউন্ডারি ও দুটি ওভার বাউন্ডারিতে ৪৫ রানে আউট হন।

   ছয় ম্যাচে এটি প্রথম হার পাঞ্জাবের। বাংলাদেশ সময়: ২০৩৪ ঘণ্টা, ৩ মে ২০১৪

2. পোলার্ড মালিঙ্গায় মুম্বাইয়ের প্রথম জয়

   স্পোর্টস ডেস্ক, বিডিনিউজ টোয়েন্টিফোর ডটকম

   নিজেদের মাঠে খেলতে নেমেই হারের গণ্ডি থেকে বেরিয়ে এল মুম্বাই ইন্ডিয়ান্স। কাইরন পোলার্ডের শেষমুহূর্তের ঝড়ো ব্যাটিংয়ে এবারের আসরের সবচেয়ে সফল দল কিংস ইলেভেন পাঞ্জাবকে ৫ উইকেটে হারিয়েছে তারা। তবে পাঞ্জাব ইনিংসের শেষ দিকে অসাধারণ বল করা লাসিথ মালিঙ্গার অবদানও কম নয়।

   ছয় ম্যাচে গতবারের চ্যাম্পিয়ন মুম্বাইয়ের এটা প্রথম জয়। সমান সংখ্যক ম্যাচে পাঞ্জাবের এটা প্রথম হার। এই হারে রান রেটে পিছিয়ে পড়ে চেন্নাই সুপার কিংসের কাছে শীর্ষ স্থান হারিয়েছে তারা।

   শনিবার পাঞ্জাবের ৫ উইকেটে গড়া ১৬৮ রানের লক্ষ্য ৫ বল হাতে রেখেই অতিক্রম করে যায় মুম্বাই। শেষ ৩ ওভারে জয়ের জন্য স্বাগতিক দলের প্রয়োজন ছিল ৪১ রানের। ১২ বলের ‘ক্যামিও’ ইনিংসে দুটি করে ছক্কা ও চার মেরে অনায়াসে দলকে জয়ের বন্দরে পৌছে দেন ক্যারিবীয় অলরাউন্ডার পোলার্ড। ২৮ রানে অপরাজিত ছিলেন তিনি।

মুম্বাইয়ের ওয়াংখেড়ে স্টেডিয়ামে লক্ষ্য তাড়া করতে নেমে ২৩ রানের মধ্যে ২ উইকেট হারিয়ে শুরুতেই বিপদে পড়ে গিয়েছিল স্বাগতিকরা। তবে উইকেটরক্ষক চিদাম্বরম গৌতম ও অধিনায়ক রোহিত শর্মার ৪১ বলে ৪৭ রানের জুটিতে সে ধাক্কা সামলে ওঠে তারা। ২৯ বলে ৩৩ রান করেন গৌতম।

তারপরও রোহিত শর্মা ও কোরি অ্যাডারসনের ব্যাটে ভর করে জয়ের পথেই ছিল তারা। কিন্তু পরপর দুই ওভারে তাদের বিদায়ে আবারো শঙ্কায় পড়ে যায় দলটি।

রোহিত করেন ৩৪ বলে ৩৯ রান আর অ্যান্ডারসনের ব্যাট থেকে আসে ২৫ বলে ৩৫ রান।

তবে পোলার্ড আর আদিত্য তারের ১৫ বলে অবিচ্ছিন্ন ৪৪ রানের ঝড়ো ম্যাচজয়ী জুটিতে সহজেই কাঙ্ক্ষিত জয় মেলে মুম্বাইয়ের। ৬ বলে ১টি করে চার ও ছক্কায় অপরাজিত ১৬ রান করেন তারে।

পাঞ্জাবের পক্ষে দুটি করে উইকেট নেন সন্দীপ শর্মা ও রিশি ধাওয়ান। এর আগে টস জিতে ব্যাট করতে নেমে ২৪ রানের মধ্যে দুই উদ্বোধনী ব্যাটসম্যানকে হারিয়ে ধাক্কা খায় পাঞ্জাব। তবে গ্লেন ম্যাক্সওয়েলের আক্রমণাত্মক ব্যাটে বিপদ কাটিয়ে ওঠে আসরের সবচেয়ে সফল দলটি। উইকেটরক্ষক ঋদ্ধিমান সাহার সঙ্গে ৫০ বলে ৬৯ রানের জুটি গড়েন প্রথম তিন ম্যাচে টানা তিনটি অর্ধশতক করা ম্যাক্সওয়েল। স্পিনার হরভজন সিংয়ের বলে বেন ডাঙ্কের হাতে ক্যাচ দিয়ে ফেরার আগে ৪৫ রান করেন তিনি। ২৭ বলের ইনিংসে ৫টি চার ও ২টি ছক্কা মারেন তিনি।

তবে ম্যাক্সওয়েল ফিরলেও ৫৯ রান করে অপরাজিত ছিলেন সাহা। তার ৪৭ বলের ইনিংসটি ৪টি চার ও ৩টি ছক্কায় সাজানো। ৩৪ রান দিয়ে ২ উইকেট নেন মুম্বাইয়ের হরভজন সিং।

The similarity is: 92.30530291471328
Location Found in Doc1: null;
Location Found in Doc2: null;
Date published: 03-05-2014
Document 1 Source: http://www.banglanews24.com/beta/fullnews/bn/287186.html
Document 2 Source: http://bangla.bdnews24.com/cricket/article781418.bdnews

Result:

Is source same: No

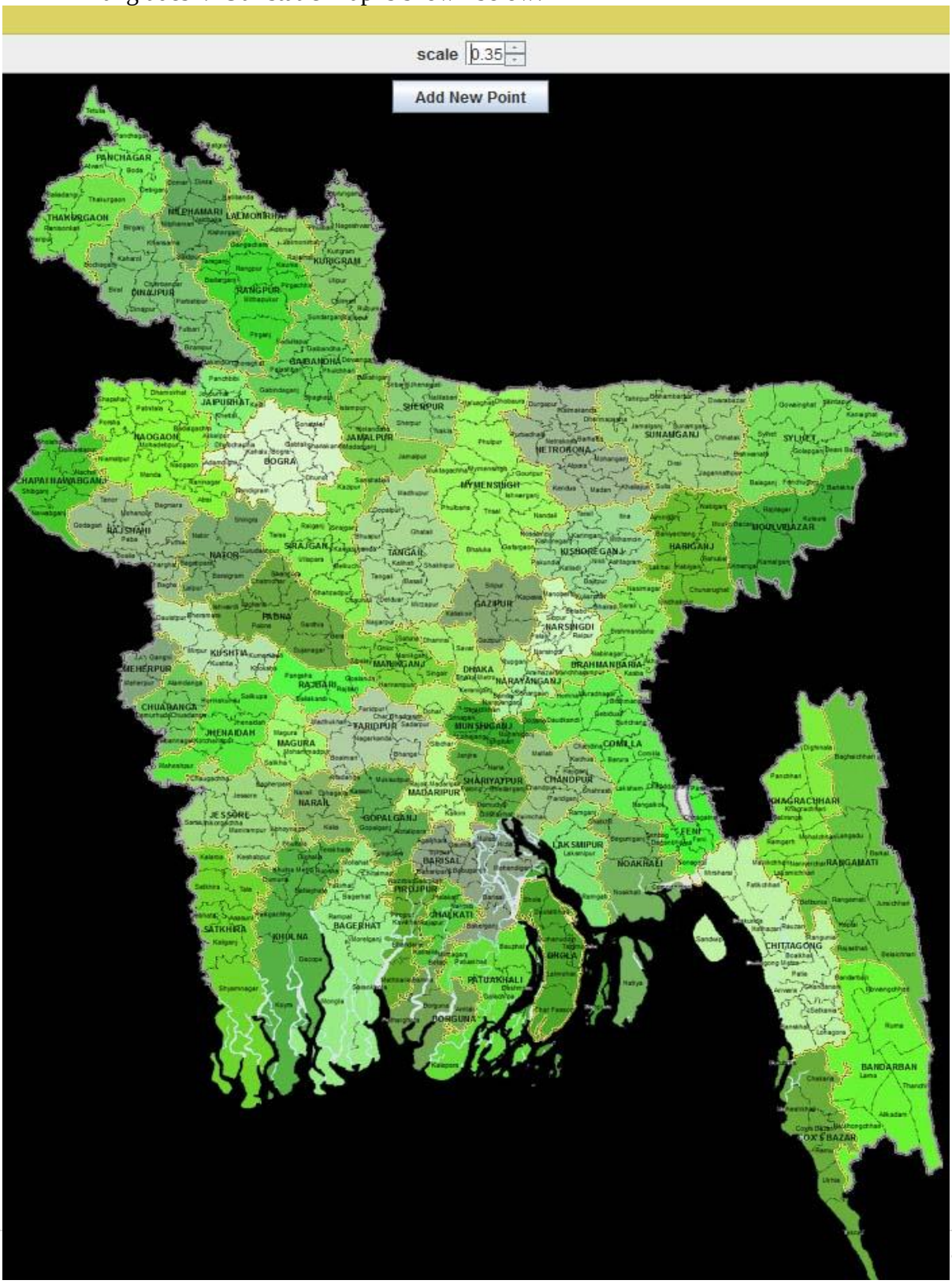
Result:

Is the Document Same: Yes;

5. Mapping the extracted data.

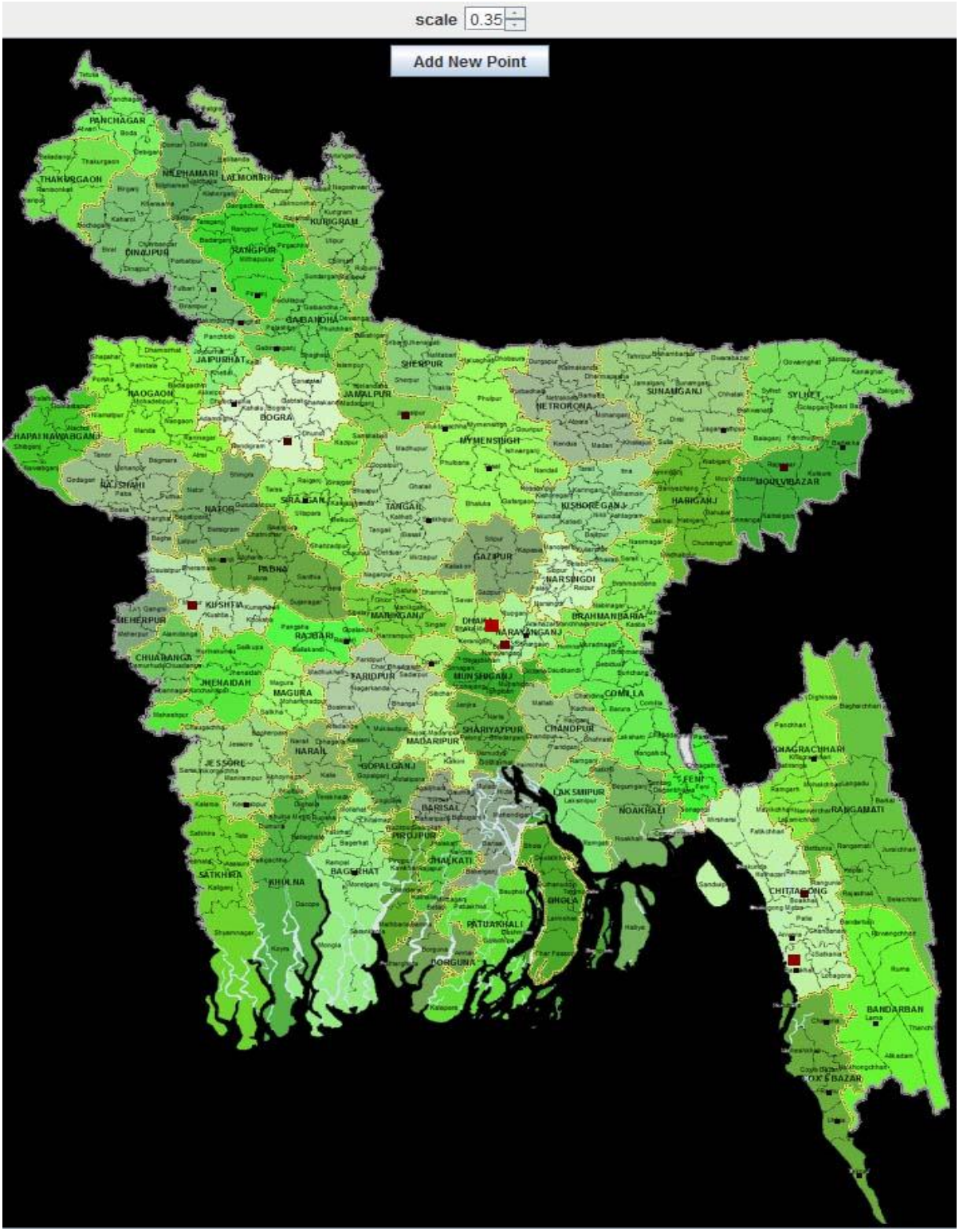
5.1. Designing the map

We have a static High Quality Map of Bangladesh which have a resolution 2000 x 2731 (BangladeshMap.png). This map shows us all Divisions, Districts and Thanas of Bangladesh. Our static Map is shown below:

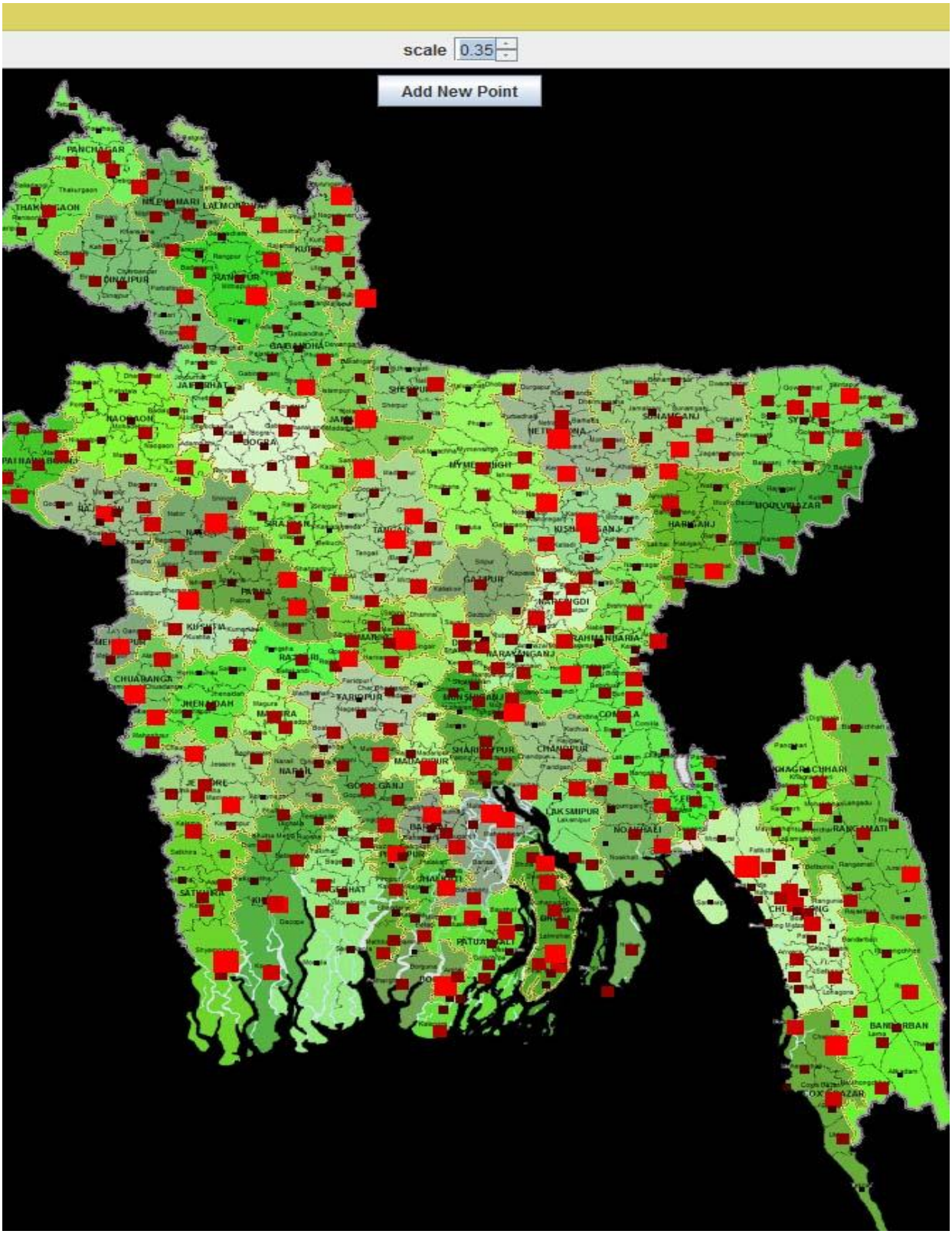


5.2. Plotting the data on map based on crime categories and location

We have stored the list of all location of Bangladesh consisting Division, District and Thana. The co-ordinate of all location relative to this map is also stored in a table. Now when we find a location of a crime we add dynamically a dot to the specific location. This feature is shown below:



When we find frequent location for different crime news, we just make it a larger dot than the previous with a limit width. We also changed the color of dot redder. Here is the face of the map for huge crime statistics.



6. Measuring crime occurrence probability

The crime data collected from RAB website for predicting future crime occurrence. Some sample crime statistics of few months of a year shown below:

CRIME REPORT on January 2013

Battalion	Dacoity	Hijack	Murder	Woman & Child Abuse	Abduction	Car Stealing	Shots	Arms Related	Narcotics Related	Theft Cases	Road Accident	Bomb Blastin	Other	Total
RAB-1	4	2	16	71	5	15	0	5	159	79	18	0	370	744
RAB-2	1	5	3	13	9	24	0	1	80	72	0	2	104	314
RAB-3	0	3	3	24	2	12	0	4	117	51	0	0	186	402
RAB-4	1	1	9	67	5	7	0	1	199	67	0	0	270	627
RAB-5	3	7	48	141	45	17	0	8	492	71	0	0	999	1831
RAB-6	4	0	12	154	11	1	0	19	306	41	0	5	935	1488
RAB-7	2	0	8	79	17	8	0	27	269	45	6	0	604	1065
RAB-8	3	1	3	111	0	0	0	1	18	7	0	0	833	977
RAB-9	12	2	39	242	11	5	0	1	381	30	6	0	1138	1867
RAB-10	2	10	8	21	5	12	0	3	286	16	0	0	149	512
RAB-11	6	0	33	145	3	4	0	9	531	139	0	0	1039	1909
RAB-12	4	0	27	106	8	0	0	13	285	6	0	0	723	1172
Total	42	31	209	1174	121	105	0	92	3123	624	30	7	7350	12908

CRIME REPORT on February 2013

Battalion	Dacoity	Hijack	Murder	Woman & Child Abuse	Abduction	Car Stealing	Shots	Arms Related	Narcotics Related	Theft Cases	Road Accident	Bomb Blastin	Other	Total
RAB-1	2	5	16	66	2	14	0	10	152	71	14	0	383	735
RAB-2	2	5	3	15	1	26	0	4	71	39	0	4	124	294
RAB-3	0	4	3	19	11	22	0	3	89	30	0	0	110	291
RAB-4	4	2	17	67	11	4	0	1	188	52	0	7	258	611
RAB-5	1	7	37	165	35	11	0	9	358	58	0	0	1071	1752
RAB-6	2	0	19	111	3	6	0	13	186	18	0	2	1039	1399
RAB-7	3	0	6	79	21	7	2	16	215	43	6	0	557	955
RAB-8	3	0	6	123	2	0	0	3	133	5	0	0	869	1144
RAB-9	10	3	32	245	9	4	0	2	323	28	4	0	1058	1718
RAB-10	0	1	6	21	1	6	0	2	194	28	0	0	138	397
RAB-11	6	0	34	145	1	12	0	9	420	109	0	0	1018	1754
RAB-12	2	0	25	106	4	0	0	6	306	53	0	0	766	1268
Total	35	27	204	1162	101	112	2	78	2635	534	24	13	7391	12318

RAB ID	Area
Rab-01	Dhaka
Rab-02	Dhaka
Rab-03	Dhaka
Rab-04	Dhaka
Rab-05	Rajshahi
Rab-06	Khulna
Rab-07	chittagong
Rab-08	Barishal
Rab-09	Sylhet
Rab-10	Dhaka
Rab-11	Narayanganj
Rab-12	Sirajganj
Rab-13	Rangpur

There have the crime statistics of previous year consisting crime zones, crime types and incident times. We have designed a statistical approach based on previous crime statistics to predict future crime occurring probability.

We want to predict the crime occurring probability of a specific zone in a specific month. Our crime predictions parameters are:
C_{ZM} = Number of crimes in a specific zone in a specific month.
C_{TM} = Number of crimes in all zones in a specific month.
C_Z = Number of crimes in a specific zone in all months.
C_T = Total number of crimes.

Let, we have a list of zone, Z = {ঢাকা, সিলেট, কুমিল্লা, চট্টগ্রাম,...,রংপুর}
List of month, M = {January, February,… ,December}
List of year, Y = {2000….2013}
Idx1 = index of specific month.
Idx2 = index of specific zone.

We can define ,$C_{ZM} = \sum_{i=1}^{Yn} C(i, idx1, idx2)$

$C_{TM} = \sum_{i=1}^{Yn} \sum_{k=1}^{Zn} C(i, idx1, k)$

$C_Z = \sum_{i=1}^{Yn} \sum_{j=1}^{Mn} C(i, j, idx2)$

$C_T = \sum_{i=1}^{Yn} \sum_{j=1}^{Mn} \sum_{k=1}^{Zn} C(i, j, k)$

So, Probability of Next Crime Occurrence in a Specific Zone and Specific Month

$P_{ZM} = (C_{ZM} / C_{TM}) * (C_Z / C_T)$. When ($C_{ZM}$, C_{TM}, C_Z, C_T)>0

Month/ Location	2012 Jan	2012 Feb	2012 Mar	2012 Apr	2012 May	2012 Jun	2012 Jul	2012 Aug	2012 Sep	2012 Oct	2012 Nov	2012 Dec
কক্সবাজার	1	4	0	1	6	2	6	3	2	0	8	0
কিশোরগঞ্জ	3	1	2	1	2	1	3	1	3	0	1	0
কুমিল্লা	2	0	0	0	4	1	2	0	0	0	0	0
কুষ্টিয়া	1	1	1	1	5	2	2	1	1	2	0	0
কুড়িগ্রাম	1	2	1	0	13	1	1	1	2	3	3	0
খাগড়াছড়ি	0	0	1	3	5	2	2	2	2	2	2	0
খুলনা	0	1	1	2	6	2	2	0	2	0	0	0
গাইবান্ধা	1	1	1	2	1	1	2	1	0	1	0	0
গাজীপুর	5	4	11	7	33	7	3	5	5	9	5	0
গোপালগঞ্জ	4	2	5	3	5	5	4	6	5	6	4	0
চট্টগ্রাম	7	7	11	8	33	12	13	12	9	13	15	0
চাঁদপুর	1	0	0	0	0	1	0	0	1	1	1	0
চুয়াডাঙ্গা	0	0	0	0	2	0	1	0	0	0	0	0
জামালপুর	2	2	3	2	7	3	4	3	2	3	5	0
জয়পুরহাট	0	0	0	0	1	0	0	0	1	0	0	0
ঝালকাঠি	2	0	1	0	0	0	1	0	0	1	1	0
ঝিনাইদহ	0	0	0	1	0	0	0	0	1	1	1	0
টাঙ্গাইল	2	2	1	1	7	2	1	2	1	1	1	0
ঢাকা	23	25	26	21	182	27	22	21	20	17	23	0
দিনাজপুর	11	12	13	10	72	7	13	8	9	6	9	0
নওগাঁ	1	0	4	1	3	0	2	2	1	2	0	0
নরসিংদী	1	0	5	1	15	4	5	2	1	1	1	0
নাটোর	0	0	2	2	0	1	0	0	1	0	0	0
নারায়ণগঞ্জ	1	5	4	0	20	3	2	4	4	2	0	0
নীলফামারী	0	0	1	0	3	0	2	0	0	0	0	0
নেত্রকোনা	2	1	6	1	6	2	3	5	4	0	1	0
নোয়াখালী	2	0	0	0	6	0	1	1	3	0	0	0
নড়াইল	3	1	3	1	5	1	2	1	4	2	2	0
পঞ্চগড়	0	0	0	0	1	0	1	0	0	0	0	0
পটুয়াখালী	0	0	0	0	4	1	0	0	2	2	0	0

পাবনা	1	0	1	1	5	0	1	0	2	0	0	0
পিরোজপুর	0	0	0	0	0	1	1	0	0	0	0	0
ফরিদপুর	0	0	0	0	3	1	0	1	0	0	0	0
ফেনী	0	0	0	0	1	1	0	1	0	0	0	0
বগুড়া	3	6	2	1	10	3	5	4	7	3	5	0
বরগুনা	2	0	2	1	6	0	2	2	1	0	2	0
বরিশাল	1	0	4	2	3	1	2	1	1	1	1	0
বাগেরহাট	1	2	0	0	2	0	1	0	1	0	1	0
বান্দরবন	1	4	0	3	8	1	2	2	2	1	1	0
ব্রাহ্মণবাড়িয়া	2	3	1	0	15	1	1	3	2	3	1	0
ভোলা	0	0	0	0	1	1	1	2	1	1	0	0
মাগুরা	1	0	0	0	3	0	1	0	2	0	0	0
মাদারীপুর	1	0	0	0	4	0	2	0	1	0	0	0
মানিকগঞ্জ	1	0	0	1	1	0	0	0	0	0	0	0
মুন্সিগঞ্জ	1	0	1	1	4	0	3	2	2	1	1	0
মেহেরপুর	5	2	4	3	10	5	6	7	6	6	6	0
মৌলভীবাজার	0	1	2	1	2	2	2	0	0	0	0	0
ময়মনসিংহ	1	1	1	1	8	1	3	3	1	0	1	0
যশোহর	1	0	0	0	0	0	0	0	0	1	0	0
রংপুর	0	1	0	0	3	0	0	1	0	1	0	0
রাঙামাটি	1	0	1	1	3	0	1	0	0	1	0	0
রাজবাড়ী	0	0	0	0	0	0	2	0	0	0	0	0
রাজশাহী	7	4	6	4	6	7	4	5	6	7	3	0
লক্ষ্মীপুর	1	1	1	2	5	1	2	2	1	1	2	0
লালমনিরহাট	0	0	0	1	0	0	0	0	0	0	0	0
সাতক্ষীরা	3	2	4	1	11	6	2	2	1	0	1	0
সিরাজগঞ্জ	1	2	2	1	4	1	0	0	0	1	3	0
সিলেট	19	10	22	12	47	25	16	26	19	19	14	0
সুনামগঞ্জ	0	1	1	1	1	1	2	0	1	0	0	0
হবিগঞ্জ	1	1	1	0	4	3	3	1	3	1	1	0

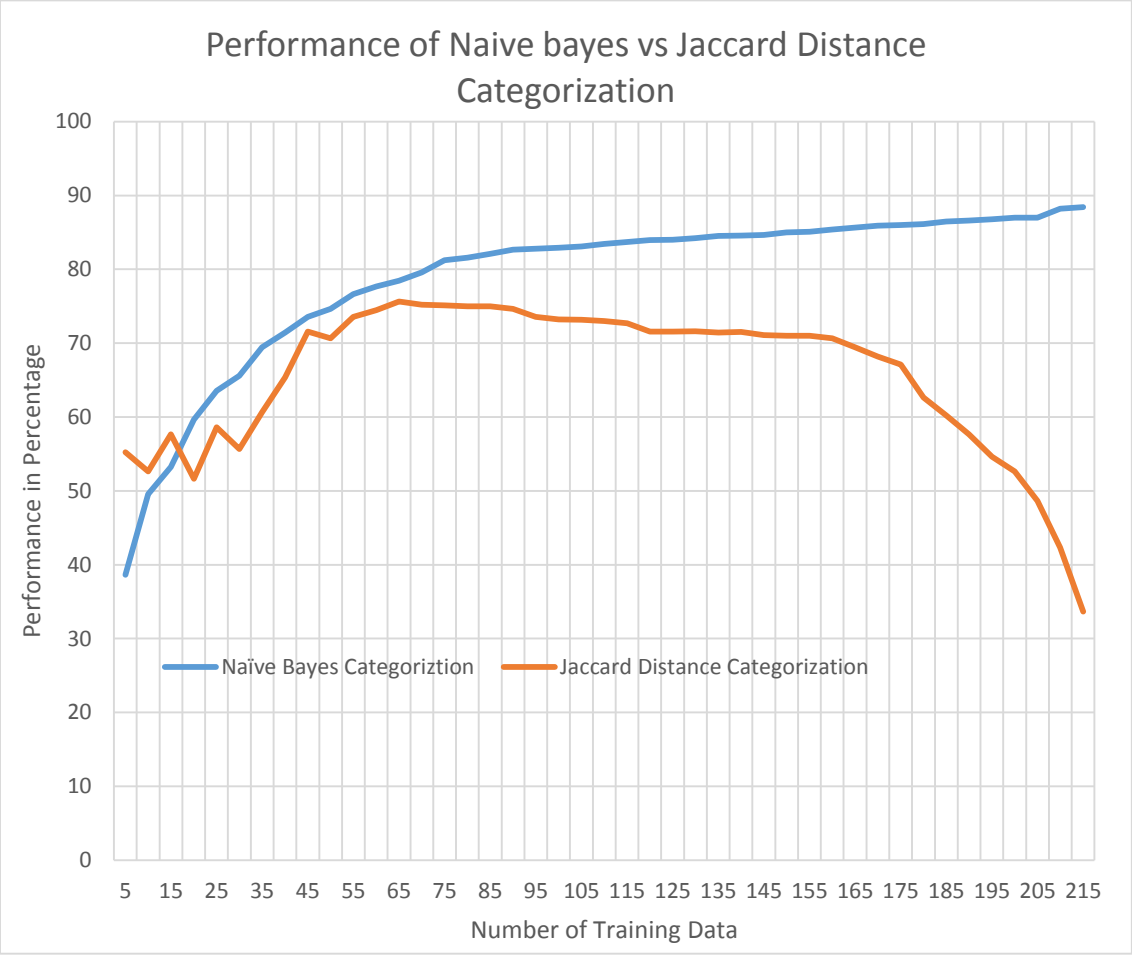
Month/ Location	2011 Jan	2011 Feb	2011 Mar	2011 Apr	2011 May	2011 Jun	2011 Jul	2011 Aug	2011 Sep	2011 Oct	2011 Nov	2011 Dec
কক্সবাজার	5	1	5	1	2	0	1	5	3	2	3	0
কিশোরগঞ্জ	1	2	1	2	1	0	1	0	2	6	1	0
কুমিল্লা	0	1	0	0	0	0	0	1	1	1	0	0
কুষ্টিয়া	2	0	2	3	0	1	1	1	2	1	1	0
কুড়িগ্রাম	0	1	0	2	1	1	1	2	5	3	0	0
খাগড়াছড়ি	3	2	2	2	2	1	1	4	3	7	1	0
খুলনা	0	1	0	1	0	0	1	0	0	1	0	0
গাইবান্ধা	0	1	1	0	0	1	2	0	1	0	0	0
গাজীপুর	5	11	11	10	3	2	8	7	7	5	6	0
গোপালগঞ্জ	3	2	4	5	3	8	6	3	7	4	5	0
চট্টগ্রাম	12	8	17	7	5	11	11	10	11	9	13	0
চাঁদপুর	1	0	1	0	0	0	1	0	0	0	0	0
চুয়াডাঙ্গা	0	1	0	0	1	0	0	0	0	0	0	0
জামালপুর	5	3	1	5	2	1	2	5	5	2	3	0
জয়পুরহাট	0	0	0	0	0	0	0	0	0	0	0	0
ঝালকাঠি	1	0	2	1	0	1	2	1	1	0	0	0
ঝিনাইদহ	1	0	0	0	0	1	0	0	0	0	0	0
টাঙ্গাইল	0	2	2	0	1	0	0	0	0	1	1	0
ঢাকা	27	24	29	21	30	19	31	18	35	20	24	0
দিনাজপুর	9	8	8	9	8	7	12	8	11	10	6	0
নওগাঁ	1	1	1	1	1	2	1	1	0	0	1	0
নরসিংদী	4	0	3	5	2	5	0	1	4	3	2	0
নাটোর	0	0	0	0	0	0	1	0	1	0	0	0
নারায়ণগঞ্জ	4	3	4	1	3	3	6	1	4	2	2	0
নীলফামারী	1	0	1	4	1	1	1	0	2	1	1	0
নেত্রকোনা	2	2	2	2	1	3	1	2	1	1	2	0
নোয়াখালী	0	0	6	0	0	0	0	1	3	2	2	0
নড়াইল	1	3	2	1	1	1	2	2	3	6	1	0
পঞ্চগড়	0	0	0	0	0	1	0	0	0	0	1	0

পটুয়াখালী	0	1	1	0	1	1	1	0	0	1	0	0
পাবনা	0	1	1	0	1	3	2	0	2	4	1	0
পিরোজপুর	1	0	0	0	0	1	0	0	0	1	0	0
ফরিদপুর	2	0	1	2	1	0	0	0	0	2	1	0
ফেনী	0	0	0	0	0	0	0	0	2	1	0	0
বগুড়া	6	5	4	6	2	3	1	3	5	5	2	0
বরগুনা	0	2	0	2	2	0	0	0	0	0	0	0
বরিশাল	2	0	1	4	1	1	0	2	3	3	0	0
বাগেরহাট	1	0	0	0	0	1	0	0	1	0	1	0
বান্দরবন	1	5	4	2	3	2	1	1	2	3	2	0
ব্রাহ্মণবাড়িয়া	0	3	4	0	2	2	1	2	1	1	4	0
ভোলা	0	2	0	1	1	0	0	0	1	1	1	0
মাগুরা	1	0	0	0	0	0	0	1	0	0	0	0
মাদারীপুর	0	0	1	0	0	1	0	0	0	0	0	0
মানিকগঞ্জ	0	0	0	1	0	0	0	0	0	0	0	0
মুন্সিগঞ্জ	0	1	1	1	0	1	2	0	0	0	1	0
মেহেরপুর	4	2	3	8	4	7	5	4	8	6	6	0
মৌলভীবাজার	0	0	0	0	0	0	1	1	1	1	0	0
ময়মনসিংহ	3	1	4	0	0	2	2	1	1	1	1	0
যশোহর	0	0	0	0	0	0	0	0	0	0	0	0
রংপুর	2	0	3	1	1	2	0	0	1	0	0	0
রাঙামাটি	2	0	0	1	1	1	1	0	0	1	1	0
রাজবাড়ী	0	0	0	0	0	1	0	0	1	0	0	0
রাজশাহী	4	5	3	4	3	4	2	7	5	7	3	0
লক্ষ্মীপুর	0	1	0	2	0	0	0	0	2	0	1	0
লালমনিরহাট	0	0	0	0	0	1	0	0	0	0	0	0
সাতক্ষীরা	1	1	2	4	0	2	5	4	3	2	1	0
সিরাজগঞ্জ	2	0	0	2	0	2	0	4	2	1	1	0
সিলেট	23	17	18	18	16	22	13	15	20	20	13	0
সুনামগঞ্জ	0	0	0	1	2	2	0	0	0	0	1	0
হবিগঞ্জ	1	1	1	1	0	0	2	0	2	0	1	0

With this table data and using our formula, we can roughly predict the next crime occurrence probability of a specific area in a specific time.

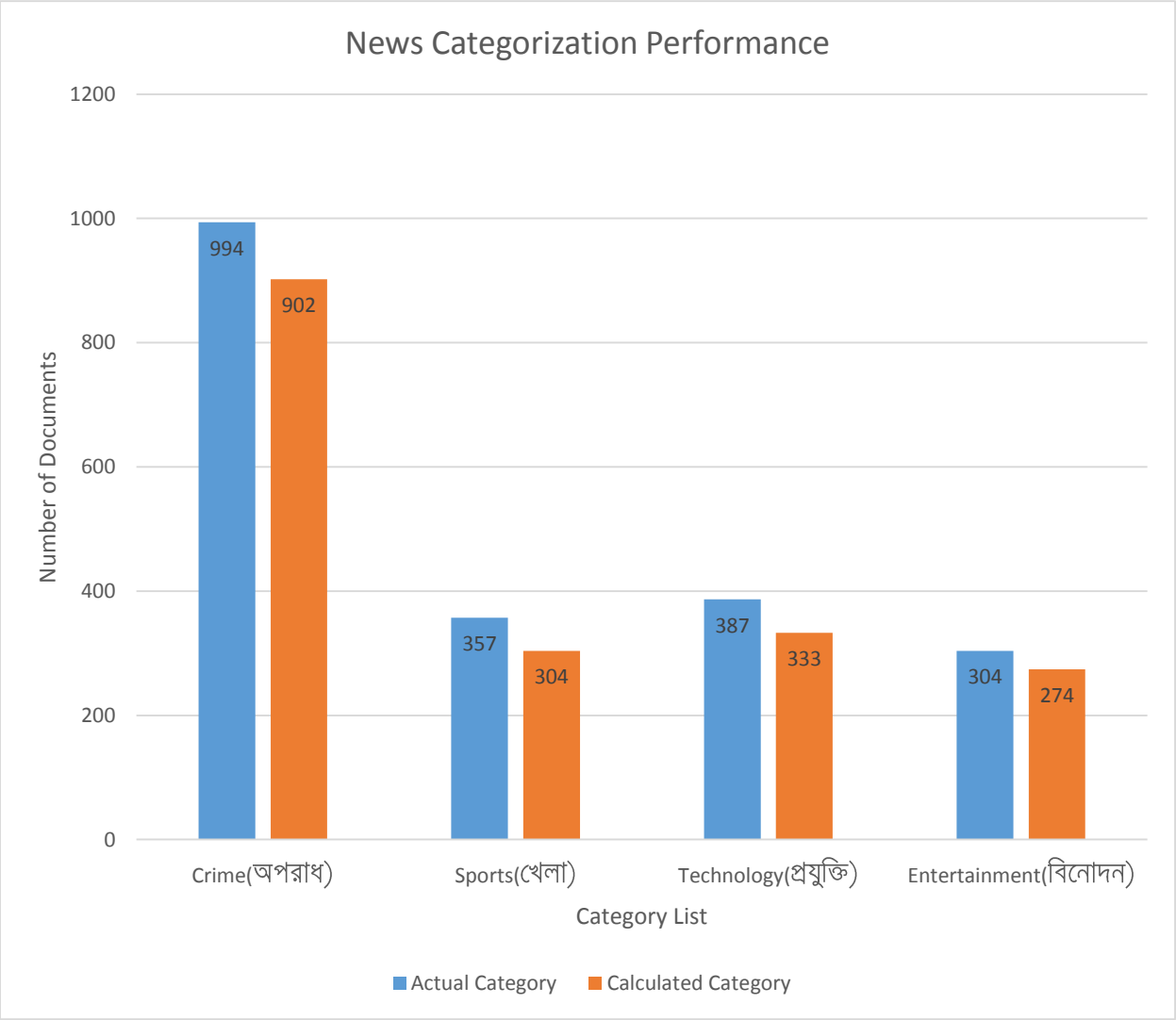
7. Performance Analysis

7.1 Performance of Naive bayes vs Jaccard Distance Categorization:



As we see from the graph, when number of training data increases performance of Jaccard Distance method decreases on the other hand Naïve Bayes clustering performs better with increased training data. That's why we selected Naïve Bayes Clustering Method to categorize news.

7.2 News Categorization Performance using Naïve Bayes Clustering:



Now we can calculate Performance from this graph for different news category.

Accuracy for Crime related News: 902/994*100 = 90.744 %

Accuracy for Sports related News: 304/357*100 = 85.154 %

Accuracy for Technology related News: 333/387*100 = 86.046 %

Accuracy for Entertainment related News: 274/304*100 = 90.131 %

Now Average Accuracy for Naïve Bayes Categorization = (90.744+85.154+86.046+90.131)/4

= 88.018%

8. Limitations

- We were required about 30000 root words but we worked with only 5000, so there have an option improve the root word finding algorithm.
- Here we do not finding any keywords actually, to categorize the news we have taken only the specific top words. Better accuracy can be gained through finding the key words.
- To find the locations we have find only up to Thana but if there have any union that haven't identified yet.
- We have used a static map to show the crimes of specific locations, but it can be pointed better through using Google map.

9. Future work

- There have some important scope to develop our approach, like:
- Finding the keywords dynamically with designing a smart algorithm.
- Developing a dynamic algorithm for root word finding.
- We have tried to design a better crime prediction algorithm, but it's possible to design a better approach using machine learning techniques.

10. Conclusion

To reduce the number of crime occurrence it's required to predict the crime prone zone. Unfortunately there haven't any previous work been done to predict crime of a specific location with retrieving news from different Bangla online newspapers. Though our developed approach is not giving perfect result but expecting this will surely be helpful for government, general people, police or tourists to decide their outing location.

11. References

[1] Y. MATSUO, M. Ishizuka: Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information.

[2] Gautam Pant, Padmini Srinivasan and Filippo Menczer: Crawling the Web

[3] MARC NAJORK, Microsoft Research, Mountain View, CA, USA: Web Crawler Architecture

[4] Robert J. KUHNS: A News Analysis System

[5] Introduction to Machine Learning, Second Edition by Ethem Alpaydın.

[6] Pattern Recognition and Machine Learning by Christopher M. Bishop.

[7] Probabilistic Graphical Models by Daphne Koller and Nir Friedman.

[8] Machine Learning by Tom M. Mitchell.

[9] Some Notes on Applied Mathematics for Machine Learning by Christopher J.C. Burges.

[10] Programming Collective Intelligence by Toby Segaran.

[11] Google, Wikipedia and some relevant websites on internet.

~End of the Thesis Report~