INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

A STRATEGY FOR NEAR-DEDUPLICATION WEB DOCUMENTS CONSIDERING BOTH DOMAIN &SIZE OF THE DOCUMENT

[1]Mr. J. MAHESH, [2]AKULA MANIHAS, [3]AVARU AASHISH, [4]ADDA VENKAT TARUN,

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

[2,3,4]BTech Student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

akulamanihas1@gmail.com, avaruaashish@gmail.com, venkattarunadda@gmail.com

Abstract: *The advice on the web is adopting to huge volumes, so an arduous affair to atom near-duplicate abstracts efficiently. The alike and near-duplicate abstracts are breeding a boundless botheration for seek engines, appropriately decelerate or access the number of confined answers. Elimination of near-duplicates save arrangement bandwidth and reduces the accumulator amount and advances the superior of seek indexes. It as well decreases the amount on the limited host that is confined such web documents. Server applications are as well benefited by identification of abreast duplicates. In this avant-garde approach, the crawled web certificate is taken and keywords are acquired and are compared with the keywords accessible in the athenaeum of the accurate domain, again an accommodation of certificate acceptance to an accurate area is absitively adjoin the number of keywords akin in that accurate domain. After selecting the domain, the admeasurement of the ascribe certificate is advised and the seek amplitude is bargain and calculations of affinity array are as well diminished. Thereafter the affinity account is affected with abstracts which are acceptance to that accurate area only. This access reduces seek amplitude thereby abbreviation the seek time.*

Keywords: *web content mining, deduplication of web content, Web Structured Mining, web usage mining.*

I. INTRODUCTION

Data deduplication is a process of eliminating the redundant data in a system, typically it is meant for improving effective storage utilization. During the deduplication process, it identifies an extra copy of already existing data in a data set /storage medium, delete the extra copy, leaving only one copy of the data to be stored.

Data deduplication is needed, for the following reasons [1].

Block-level deduplication is much more efficient and more reliable compared with file-level deduplication and it works at fixed-size block or a variable-sized block by eliminating duplicate blocks. It can eliminate chunks of data smaller than its file; these mechanisms are focused on the size of data like file, sub-file, chunk, block, byte and bit and it's not able to address the content of a dataset. So, this kind of algorithms is not able to provide an optimized solution for a specific scenario like content level deduplication of streaming data

The advice on the web is exponentially advanced in massive volumes and appeal to use this abundant advice calmly and effectively. The web consists of added no. of assorted copies of aforementioned agreeable. Some advice repositories are mirrored artlessly to accommodate back-up and admission reliability. The seek engine faces a huge botheration due to the all-inclusive bulk of advice and it leads to extraneous answers. The alike and near-duplicate abstracts accept produced an added aerial for the seek engines alarmingly affecting their performance. The apprehension of abreast alike abstracts has afresh become a claiming and a area of abundant interest. A lot of studies accept brought calm on the Apprehension of Near-Duplicate Documents[2]. Several methods and algorithms for Near- Duplicate Apprehension accustomed by advisers are available. Thus, partially or absolutely alike web abstracts frequently arise on the web. Some abundant accomplishments advice has been addressed apropos the associated areas which cover Web Mining, Web Scraping, Alike Abstracts and more. The adjustment of carrying Data mining on the web is alleged Web mining. Digging the web abstracts and advertent the patterns from it. Web mining is mainly disconnected into 3 audible groups as follow,

Web Content Mining

Web content mining can be acclimated for the acquisition of admired data, information, and acumen from a web page. Web anatomy mining helps to access admired abstracts arrangement from the anatomy of hyperlinks. Due to adverse and blemish of anatomy in web data, automatic analysis of new abstracts arrangement can be arduous to some extent. Web agreeable mining performs scanning and anticipation the text, images, and groups of web pages according to the agreeable of the ascribe (query), by alignment the account in seek engines.

Web Structured Mining

The web structured mining can be acclimated to acquire the hotlink anatomy of hyperlink. It is acclimated to actuate that the web pages are either affiliated by advice or absolute hotlink connection. The abstraction of anatomy mining is to present a structural arbitrary of the website and accompanying web pages[3].

Web Usage Mining

Web usage mining is active for anticipation the weblog abstracts (access advice of web pages) and helps to acquisition the user admission patterns of web pages. Web server registers a weblog almanac for every web page. Analysis of similarities in weblog annal can be accessible to access the abeyant consumers for e-commerce companies.

With the exponential growth of online information, effectively and efficiently harnessing the vast volumes of data available on the web has become a pressing challenge. The web comprises numerous copies of the same content, often mirrored for backup and accessibility purposes, posing difficulties for search engines in providing relevant results. The presence of duplicate and near-duplicate documents has significantly impacted search engine performance. Consequently, the detection of near-duplicate documents has emerged as an area of great interest and complexity. Various researchers have proposed methods and algorithms for

Near-Duplicate Document Detection, addressing the frequent occurrence of partially or completely duplicated web data. One key technique for analysing web data is Web Mining, which involves extracting valuable information and insights from web pages. Web mining can be classified into three distinct groups: Web Content Mining, Web Structured Mining, and Web Usage Mining[4].

Web Content Mining focuses on extracting valuable data, information, and patterns from web pages. It utilizes web form mining to identify patterns from hyperlinks and performs text scanning and categorization to arrange search engine results according to query content. Web Structured Mining involves analysing the link structure of web pages to determine their interconnectedness. By providing a structural summary of websites and associated web pages, structured mining facilitates a better understanding of the underlying information architecture. Web Usage Mining aims to predict user access patterns by analysing weblog data, which records the access information for each web page. By identifying similarities within weblog records, this technique can help e-commerce companies identify potential consumers and enhance their targeting strategies. In this abstract, we delve into the challenges posed by near-duplicate documents and highlight the importance of web mining in extracting valuable insights from web data. By leveraging the principles of web content mining, web structured mining, and web usage mining, researchers and practitioners can enhance their data mining endeavours and maximize the benefits of the abundant information available on the web.

II. LITERATURE SURVEY

Charikar's simhash method for dimensionality abridgement is advised to admit near-duplicate abstracts which map top dimensional vectors to small-sized fingerprints. A web page is angry into a set of appearance area anniversary affection is apparent with its weight G.S Manku et al. supplemented the idea of feature weight to random projection. Features are affected application accepted Information Retrieval methods like tokenization, case folding, and stop-word abatement stemming and byword detection. With simhash, high-dimensional vectors are adapted into f-bit finger-print area f is small-sized fingerprints. The cryptographic assortment functions like SHA-1 or MD5 aftermath assorted assortment ethics for the two abstracts with individual byte aberration but simhash will assortment them into agnate hash-values as Hamming Distance is small. According to Charikar's this adjustment with 64-bit fingerprints appears to plan abundant in convenance for an athenaeum of 8B web pages.

V. A. Narayana et al. had presented an adjustment for abreast alike apprehension of web pages in web crawling. After accepting a new web page from web crawler, the arrangement extracts the agreeable of that page into abounding tokens and calculates its affinity account with abounding assorted absolute documents. A certificate would be advised a near-duplicate web page if its affinity account was greater than the beginning that had been predefined. This adjustment affords high-grade seek engine superior and the aeroembolism anamnesis amplitude for repositories and seek amplitude for classifying abreast alike web documents.

In this paper, the Poisson process Algorithm, which is a stochastic process that counts the number of occurrences of a specified event over a period of time is proposed. Here, the Poisson process is introduced into the content-level data deduplication for the streaming data. The major complexity of deduplication on streaming data is its behaviour, which means that streaming data are dynamically changing over a time period (randomness), it’s very difficult to handle, because it needs a more intelligent way like Poisson Process (Stochastic) to carry out the deduplication at content level of data which saves crucial data process time and brings forth effective deduplication mechanism.

Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. Increasingly, matched data are becoming important in many application areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining. With the increasing size of today's databases, the complexity of the matching process becomes one of the major challenges for record linkage and deduplication. In recent years, various indexing techniques have been developed for record linkage and deduplication. They are aimed at reducing the number of record pairs to be compared in the matching process by removing obvious nonmatching pairs, while at the same time maintaining high matching quality.

III. PROPOSED SYSTEM

An innovative idea is advanced to finding near-duplicate web documents i.e., considering both the size of the input document and domain belongs to has been considered. The repository is completely divided into 5 Domains as, Software Engineering, Mechanical Engineering, Civil Engineering, Electrical Electronics Engineering, and Biological Science The Domains are further divided into 3 chunks which are as, Size 1_64 KB, Size 65_128 KB and Size 129 KB 7 The whole repositories are joined to the central repository by u_id which is the primary key in the size repository. The newly crawled web document is compared with all available domains. After the domain is decided, the size of the input document is considered and a similarity score is calculated. By this process, 1 domain repository out of 5 domain repositories and 1 size repository out of 3 size repositories are searched, thus reducing the search space by 1/15[1/5(domains)* 1/3(size)]. All the u_id's which are belonging to the particular repository is considered in the key repository while testing the duplicate detection process.

SYSTEM ARCHITECTURE

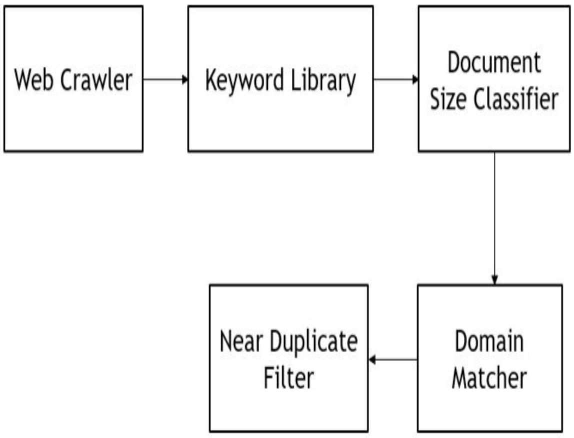


Fig.1 System architecture

Approaches to machine learning are continuously being developed. For our purposes, we'll go through a few of the popular approaches that are being used in machine learning at the time of writing

K-Nearest Neighbor

the k-nearest neighbour algorithm is a pattern recognition model that can be used for classification as well as regression. Often abbreviated as k-NN, the k in k-nearest neighbor is a positive integer, which is typically small. In either classification or regression, the input will consist of the k closest training examples within a space. We will focus on k-NN classification. In this method, the output is class membership. This will assign a new object to the class most common among its k nearest neighbors. In the case of k = 1, the object is assigned to the class of the single nearest neighbor. Let's look at an example of k-nearest neighbor. In the diagram below, there are blue diamond objects and orange star objects. These belong to two separate classes: the diamond class and the star class

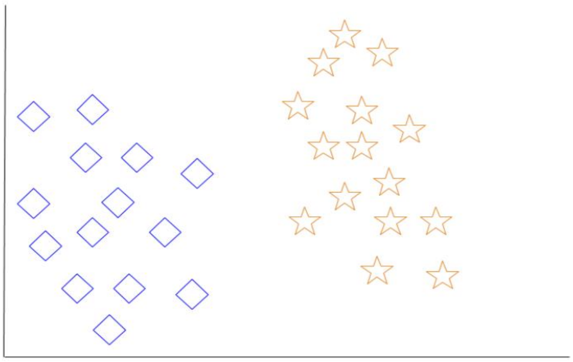


Fig.2K-NN

Decision Tree Learning

For general use, decision trees are employed to visually represent decisions and show or inform decision making. When working with machine learning and data mining, decision trees are used as a predictive model. These models map observations about data to conclusions about the data's target value. The goal of decision tree learning is to create a model that will predict the value of a target based on input variables. In the predictive model, the data's attributes that are determined through observation are represented by the branches, while the conclusions about the data's target value are represented in the leaves. When "learning" a tree, the source data is divided into subsets based on an attribute value test which is repeated on each of the derived subsets recursively. Once the subset at a node has the equivalent value as its target value has, the recursion process will be complete. Let's look at an example of various conditions that can determine whether or not someone should go fishing. This includes weather conditions as well as barometric pressure conditions

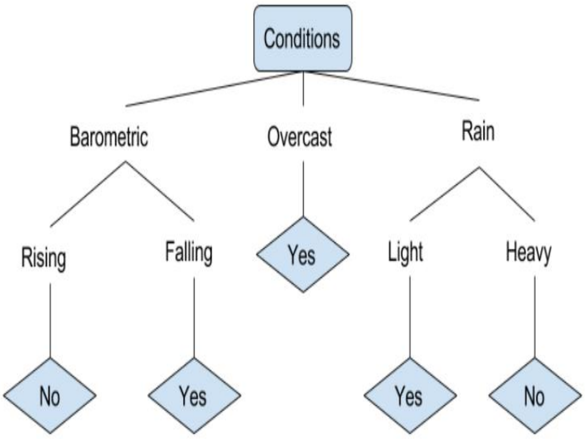


Fig.3 Decision Tree

In the simplified decision tree above, an example is classified by sorting it through the tree to the appropriate leaf node. This then returns the classification associated with the particular leaf, which in this case is either a Yes or a No. The tree classifies a day's conditions based on whether or not it is suitable for going fishing. A true classification tree data set would have a lot more features than what is outlined above, but relationships should be straightforward to determine. When working with decision tree learning, several determinations need to be made, including what features to choose, what conditions to use for splitting, and understanding when the decision tree has reached a clear ending.

Process:

Upload web document Dataset: using this module we will upload dataset to application

1) Pre-process dataset: using this module we will read all images from dataset and then apply pre-processing technique such as resizing image,

.

2) Top 10 words & graph: using this module to show the 10 words and graph

3) Find near de- duplication documents: using this module we will plot De duplication of documents.

4) find near de-duplication images: using this module to open de documentation images.

Support for YOLO/DarkNet has been added recently. We aregoing to use the OpenCV dnn module with a pre-trained YOLO model for detecting common object

IV. RESULTS

In near duplicates project student is asking to get near duplicates images also apart from text documents. So, u can put all similar or related or duplicates images inside 'images' folder like below screen shots.

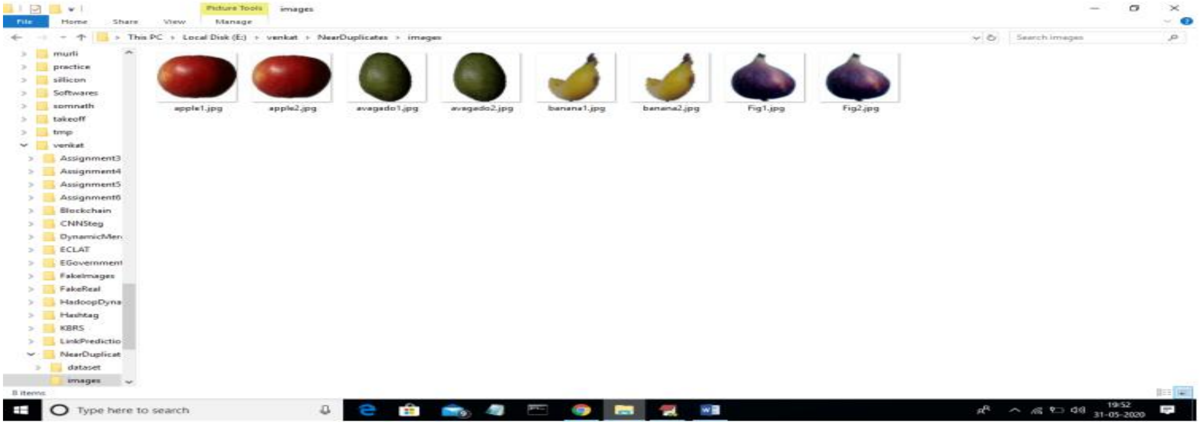


Fig.4 Image Dataset

In above screen I put some related fruit images inside 'images' folder and you too can put your own images also. Now run code by double click on 'run.bat' to get below scree

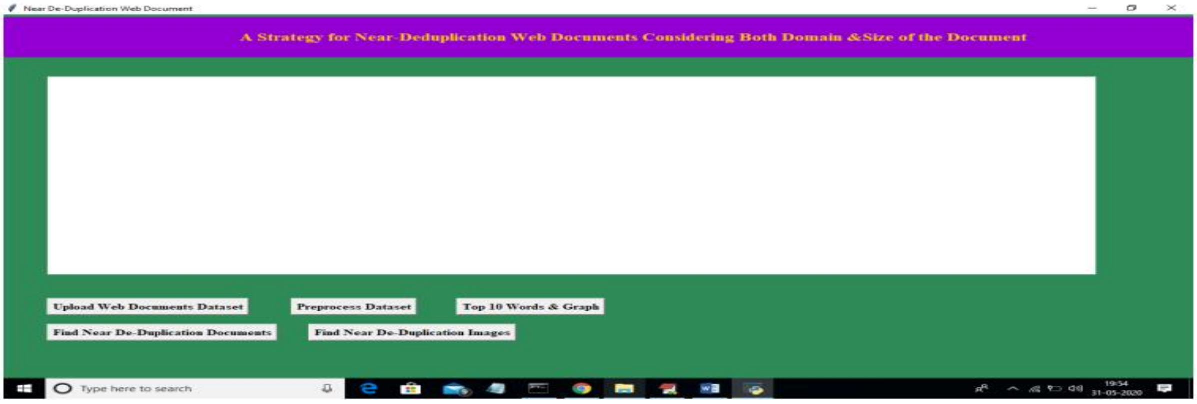
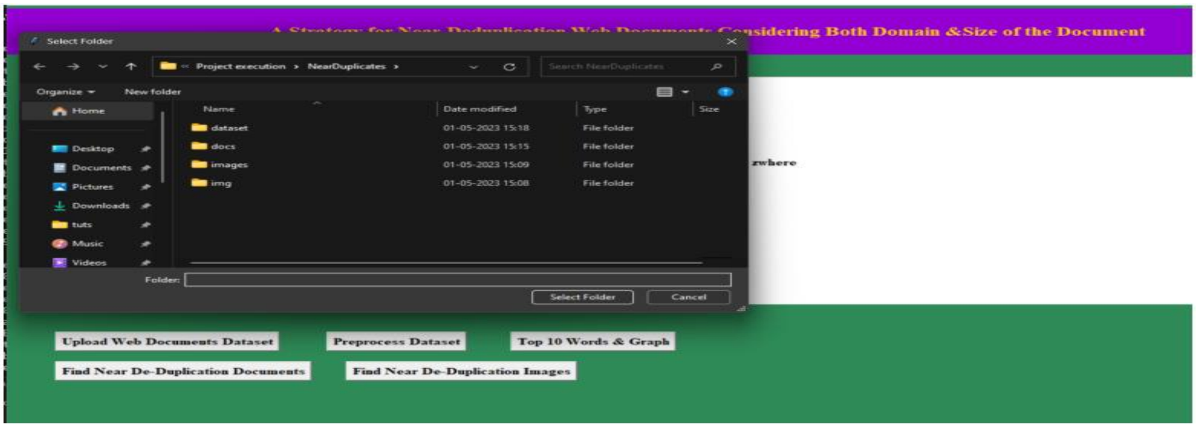

Near De-Duplication Web Document
A Strategy for Near-Deduplication Web Documents Considering Both Domain &Size of the Document
Upload Web Documents Dataset
Preprocess Dataset
Top 10 Words & Graph
Find Near De-Duplication Documents
Find Near De-Duplication Images

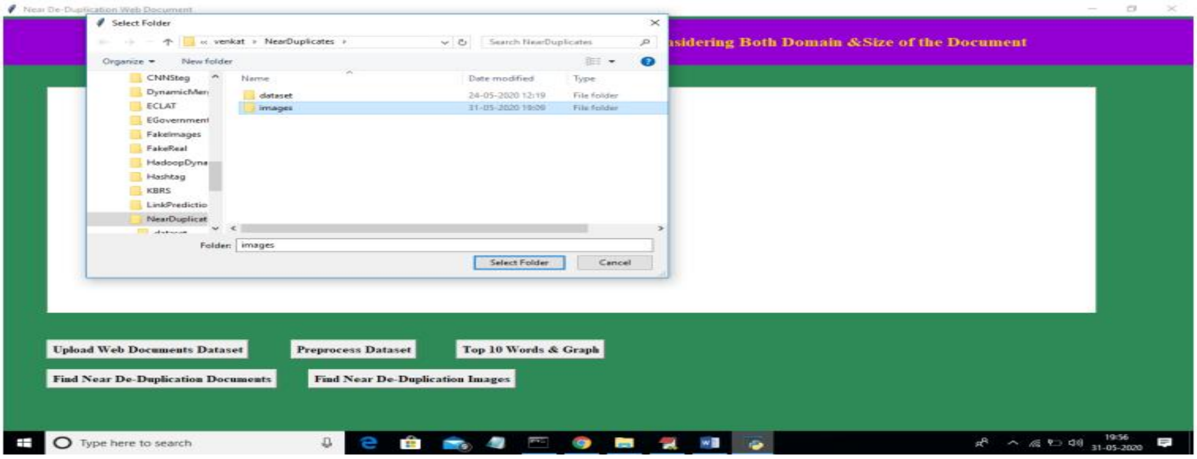
Fig.5 Near-Deduplication Interface

In above screen click on last button called 'Find Near Duplicates Images' button to allow application to find near duplicates and to get below screen


A Strategy for Near-Deduplication Web Documents Considering Both Domain &Size of the Document
Select Folder
Project execution > NearDuplicates
Search NearDuplicates
Organize
New folder
Home
Desktop
Documents
Pictures
Downloads
tuts
Music
Videos
Folder:
Select Folder
Cancel
Upload Web Documents Dataset
Preprocess Dataset
Top 10 Words & Graph
Find Near De-Duplication Documents
Find Near De-Duplication Images

Name	Date modified	Type	Size
dataset	01-05-2023 15:18	File folder	
docs	01-05-2023 15:15	File folder	
images	01-05-2023 15:09	File folder	
img	01-05-2023 15:08	File folder	

Fig.6 Dataset (documents)



Name	Date modified	Type
dataset	24-05-2020 12:19	File folder
images	31-05-2020 19:09	File folder

Fig.7 Dataset (images)

In above screen I am selecting and uploading 'images' folder and then click on 'Select Folder' button to get below output

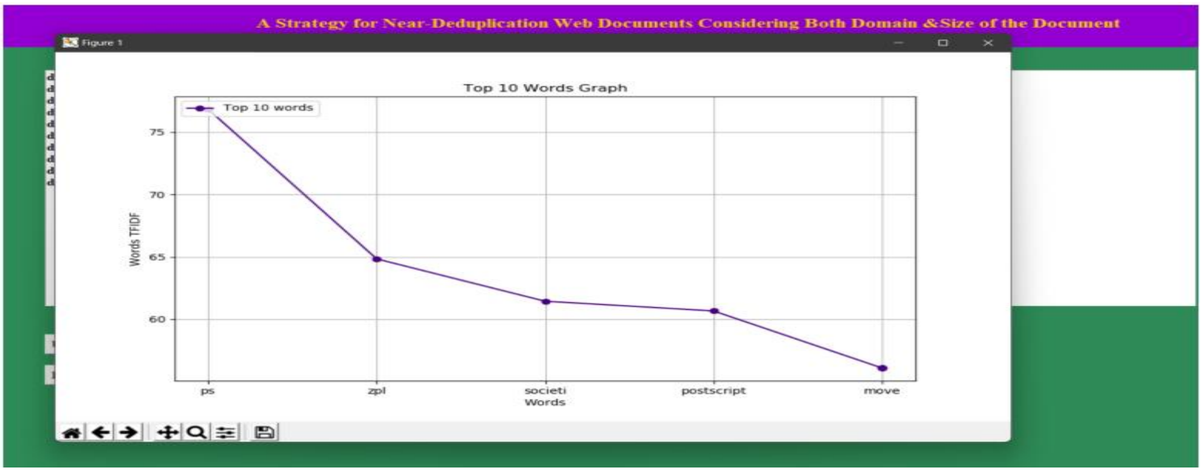
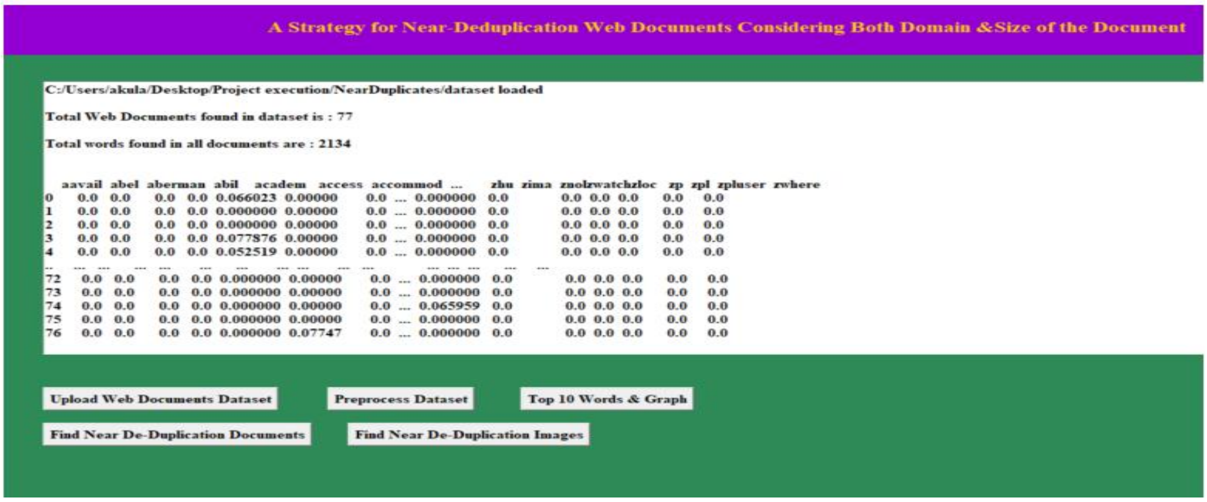
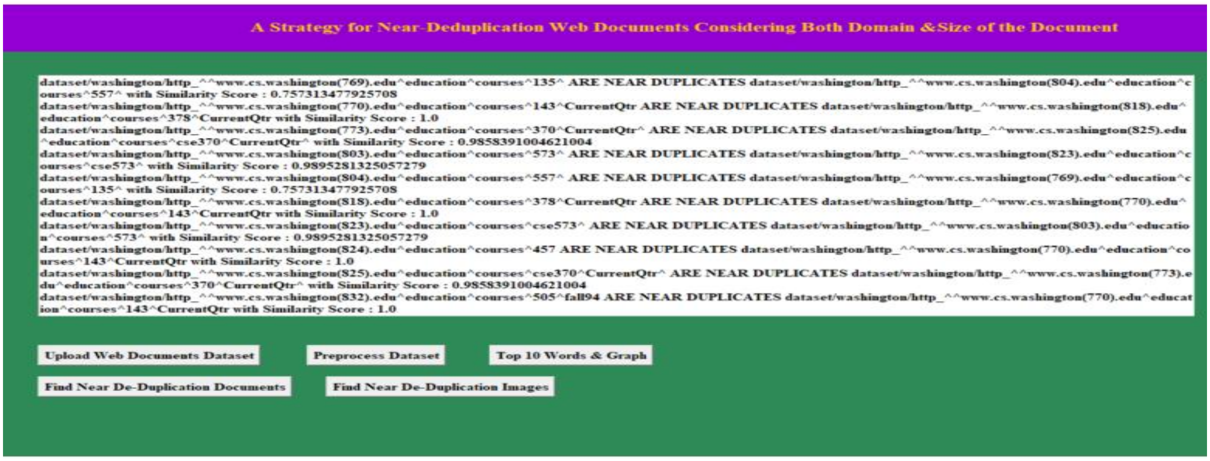


Fig.8 Graph


A Strategy for Near-Deduplication Web Documents Considering Both Domain &Size of the Document
Upload Web Documents Dataset
Preprocess Dataset
Top 10 Words & Graph
Find Near De-Duplication Documents
Find Near De-Duplication Images

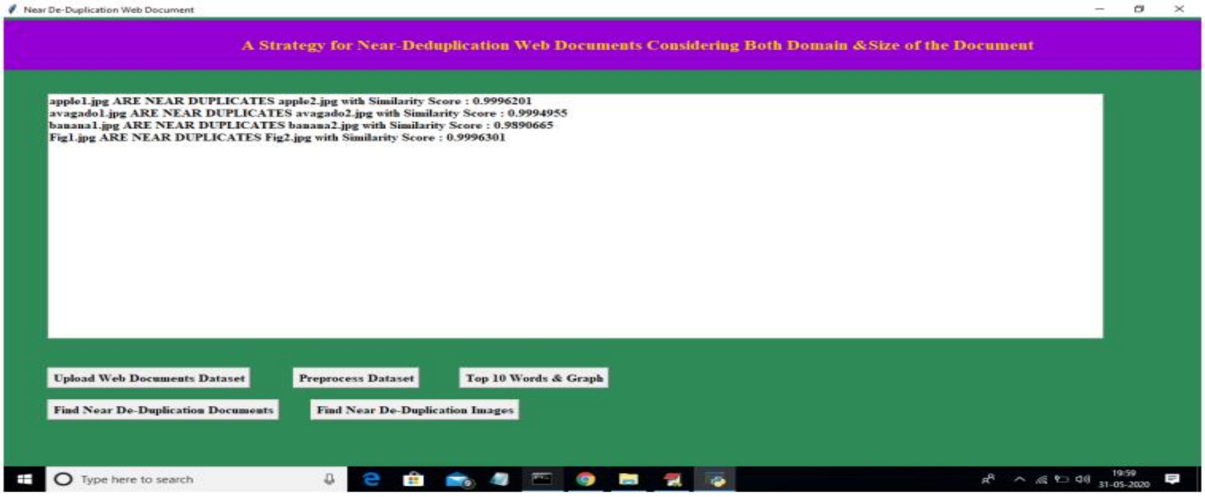
```
C:/Users/akula/Desktop/Project execution/NearDuplicates/dataset loaded

Total Web Documents found in dataset is : 77

Total words found in all documents are : 2134

  aavail abel aberman abil   academ   access accommod ...   zhu zima znolzwatchzloc  zp  zpl zpluser zwhere
0    0.0  0.0     0.0  0.0 0.066023 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
1    0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
2    0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
3    0.0  0.0     0.0  0.0 0.077876 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
4    0.0  0.0     0.0  0.0 0.052519 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
..   ...  ...     ...  ...      ...     ...      ... ...      ...  ...       ... ... ...    ...    ...
72   0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
73   0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
74   0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.065959  0.0       0.0 0.0 0.0    0.0    0.0
75   0.0  0.0     0.0  0.0 0.000000 0.00000      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
76   0.0  0.0     0.0  0.0 0.000000 0.07747      0.0 ... 0.000000  0.0       0.0 0.0 0.0    0.0    0.0
```

Fig.9 Pre-processed data


A Strategy for Near-Deduplication Web Documents Considering Both Domain &Size of the Document
Upload Web Documents Dataset
Preprocess Dataset
Top 10 Words & Graph
Find Near De-Duplication Documents
Find Near De-Duplication Images

```
dataset/washington/http_^^www.cs.washington(769).edu^education^courses^135^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(804).edu^education^c
ourses^557^ with Similarity Score : 0.757313477925708
dataset/washington/http_^^www.cs.washington(770).edu^education^courses^143^CurrentQtr ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(818).edu^
education^courses^378^CurrentQtr with Similarity Score : 1.0
dataset/washington/http_^^www.cs.washington(773).edu^education^courses^370^CurrentQtr^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(825).edu
^education^courses^cse370^CurrentQtr^ with Similarity Score : 0.9858391004621004
dataset/washington/http_^^www.cs.washington(803).edu^education^courses^573^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(823).edu^education^c
ourses^cse573^ with Similarity Score : 0.9895281325057279
dataset/washington/http_^^www.cs.washington(804).edu^education^courses^557^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(769).edu^education^c
ourses^135^ with Similarity Score : 0.757313477925708
dataset/washington/http_^^www.cs.washington(818).edu^education^courses^378^CurrentQtr ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(770).edu^
education^courses^143^CurrentQtr with Similarity Score : 1.0
dataset/washington/http_^^www.cs.washington(823).edu^education^courses^cse573^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(803).edu^educatio
n^courses^573^ with Similarity Score : 0.9895281325057279
dataset/washington/http_^^www.cs.washington(824).edu^education^courses^457 ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(770).edu^education^co
urses^143^CurrentQtr with Similarity Score : 1.0
dataset/washington/http_^^www.cs.washington(825).edu^education^courses^cse370^CurrentQtr^ ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(773).e
du^education^courses^370^CurrentQtr^ with Similarity Score : 0.9858391004621004
dataset/washington/http_^^www.cs.washington(832).edu^education^courses^505^fall94 ARE NEAR DUPLICATES dataset/washington/http_^^www.cs.washington(770).edu^educat
ion^courses^143^CurrentQtr with Similarity Score : 1.0
```

Fig.10 Result (a)



Near De-Duplication Web Document
A Strategy for Near-Deduplication Web Documents Considering Both Domain &Size of the Document
Upload Web Documents Dataset
Preprocess Dataset
Top 10 Words & Graph
Find Near De-Duplication Documents
Find Near De-Duplication Images
Type here to search
19:59
31-05-2020

```
apple1.jpg ARE NEAR DUPLICATES apple2.jpg with Similarity Score : 0.9996201
avagado1.jpg ARE NEAR DUPLICATES avagado2.jpg with Similarity Score : 0.9994955
banana1.jpg ARE NEAR DUPLICATES banana2.jpg with Similarity Score : 0.9890665
Fig1.jpg ARE NEAR DUPLICATES Fig2.jpg with Similarity Score : 0.9996301
```

Fig.11 Result (b)

V. CONCLUSION

Near-duplicate web documents will produce a main problem to the web crawling community and have become a significant task for the search engines. Near-duplicates raise the cost of serving answers, provoke a gigantic amount of space to store the indexes and ultimately slow down the results, hence affecting both the time complexity and space complexity. Near-duplicate documents are also resulting in irrelevant answers to the users. The near de-duplication of web documents, the search engine result in getting relevant answers and hence reducing search space.

REFERENCES

1. Chuan Xiao, Wei Wang, Xuemin Lin, “Efficient Similarity Joins for Near-Duplicate Detection”, Proceeding of the 17th international conference on World Wide Web, pp. 131 – 140. April 2008.

2. Uday Chandrakant Patkar, Sushas Haribabu Patil and Prasad Peddi, "Translation of English to Ahirani Language", *International Research Journal of Engineering and Technology(IRJET)*, vol. 07, no. 06, June 2020.

3. Spetka, Scott. "The TkWWW Robot: Beyond Browsing". NCSA. Archived from the original on 3 September 2004. Retrieved 21 November 2010.

4. M. Charikar. “Similarity estimation techniques from rounding algorithms”. In Proc. 34th Annual Symposium on Theory of Computing (STOC 2002), pp: 380-388, 2002.

5. Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, "Detecting near-duplicates for web crawling," Proceedings of the 16th international conference on World Wide Web, pp: 141 - 150, 2007.

6. V.A.Narayana, P. Premchand, and A. Govardhan, "A Novel and Efficient Approach For Near Duplicate Page Detection in Web Crawling." IEEE International Advance Computing Conference, March 6-7, 2009.

7. Thomas Dean, Mykyta Synytskyy, "Agile Parsing Techniques for Web Applications".

8. Prasadu Peddi (2016), Comparative study on cloud optimized resource and prediction using machine learning algorithm, ISSN: 2455-6300, volume 1, issue 3, pp: 88-94.