International Journal of Engineering & Technology, 7 (2.8) (2018) 427-431

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

DeDuSERP: De-duplication in search engine result page

Naresh Sharma[1*], Priti Dimri[2]

[1]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Modinagar, Ghaziabad, UP, India
[2]Associate Professor and Head, Dept. of CSA, G.B. Pant Engineering College, Ghurdauri, Pauri, Uttarakhand, India
**Corresponding Author E-mail: nrssharma@gmail.com*

Abstract

Web offers a new way of service provision by arranging different resources over the web. The most critical and prominent is web searches. The purpose of this research is to identify a subtype of De-Duplication. DeDuSERP is de-duplication in search engine result page. It restricts the showcasing of urls with duplicate or similar data and hence enhances the search result experience of any client. By duplicate results we mean different links containing the same content or information. To solve this problem, we have designed a filter between Search engine result page and indexed-ranked pages which we get from the search engine in response to the query of the searcher. This filter eliminates the duplicate links idiosyncratically and displays the unique results on the SERP for the searcher. We have performed the string to string comparison of web pages and if the content is 90% similar then we adjudge them as duplicates and then check their inventiveness of these duplicate links on the basis of timestamp. By this we mean then the web page crawled earlier is original. The process of comparison and timestamp matching is done using an open source apache API Commons IO 2.4.

Keywords: *Search Engine; SERP; inventiveness; DeDuSERP (De-duplication in search engine result page).*

1. Introduction

The programs that explore the documents for some specific keywords and return a list of documents in which keywords were mentioned is known as search engines.
In fact search engine is a broad class of programs, though, the term is frequently used to specially to describe the systems like GOOGLE, ASK, BING and YAHOO search that permit users to search for documents from the Internet [1].
In general, the working of a web search engine starts by sending a query through spider to fetch maximum documents as possible. In next step, called indexing, reads downloaded documents and then creates an index depends on the keywords contained in each document.
Further a proprietary algorithm is used by search engine to generate its indices so that idyllically only significant results are returned for that particular query.
The important thing about the Internet and its most important element, the World Wide Web, is that there are millions of pages accessible, waiting to present the information on a remarkable diversity of topics. The terrible news about the Internet is that there are a huge number of pages accessible; a large portion of them contain copy or comparative information. When you have to think about a specific subject, how would you know which pages to peruse? If you're like most people you click on the most optimized SERP link.

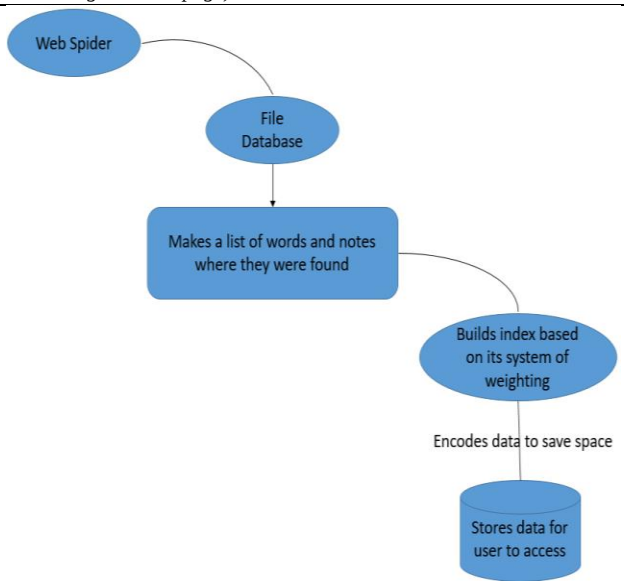

Fig. 1: Working of Search Engine

A. SERP

SERP, a web search tool comes about pages are site pages served to clients when they look for something web based utilizing a web index, for example, Google, Yahoo, Bing etc.. The client composes their pursuit question (frequently utilizing particular dialect and expressions known as keywords), whereupon the web index presents them with a SERP. Each SERP is one of a kind, not with standing for look inquiries performed on a similar web index utilizing similar keywords or pursuit questions. This is on the grounds that for all intents and purposes all web indexes redo the

experience for their clients by introducing comes about in view of an extensive variety of elements past their hunt terms, for example, the client's physical area, perusing history, and social settings. Two SERPs may seem indistinguishable, and contain huge numbers of similar outcomes, yet will frequently include unpretentious contrasts. The presence of internet searcher comes about pages is always in transition because of analyses led by Google, Bing, and other web crawler suppliers to offer their clients a more natural, responsive experience [8].

B. Page-Rank

A page ranking is a system from different service providers that ranks pages and their importance. Page ranks places an emphasis on inbounds links to determine the importance of pages. Page rank uses a ten points or 100 point scale. Page rank was first named after its founder, Lawrence Edward Page, who is likewise one of the founders of Google. The importance of websites is measured by the page ranking. By counting the number as well as the quality of the links to a website, it is a algorithm that was used by Google and all other search engine providers. By use of such algorithm, we will work out the ranking of a page and how popular and important the content is, before winnowing out the duplicate data.

C. De-Duplication

As the data is at its uncontrollable growth, duplicacy in data is also increasing with this comes the need for an efficient technique to manage such huge amount of data and its inventiveness [1]. De-Duplication is done by string to string comparison of web pages and if the content is 90% similar then we consider them as duplicate and then check their originality of these duplicate links on the basis of timestamp. By this we mean then the web page crawled earlier is original. The process of comparison and timestamp matching is done using an open source apache API Commons IO 2.4 [3, 4, 5].

2. Problem Statement

Due to uncontrollable growth of data, its management becomes difficult and precise search for searcher is also an issue sometimes. Tracking all the data on millions of web pages is not possible unless you stenograph a particular file with a tracking piece of code with incremental probability of every movement or Copy which is not possible, But there are programs like Antivirus which do not run without registration or specifically taking care of local computers. And you can track it like that. But if you really mean a single file, it is not possible to find how much data is duplicated over the net.

If you delimit it to an Enterprise, you can use 6 levels of Data Quality measures which we usually use in ERP DQ (enterprise data planning and data quality) measurement.

You know, there are nodes, branches in this network which leads to a million pieces of ends and many ends are eventually infinite also known broken links. One Single file can be anything, picture, text, video, GIFs but here we are dealing with textual data content only.

Search quality is a basic issue in the course of the execution of a Search Engine. Search quality issues can significantly affect an association's data framework and client encounter. Therefore, it is essential to comprehend search quality issues to guarantee achievement in executing DeDuSERP. This paper utilizes Google for instance of a web crawler and portrays an examination, which investigates search quality issues with existing frameworks, and distinguishes basic achievement factors that affect search quality [1].

3. Related Work

How about we take a gander at the 5-year picture, Internet clients have developed by 82%, or very nearly 1.7 billion individuals, since January 2012[7]. That means right around 1 million new clients every day or more than 10 new clients consistently. As the number of users is increasing day by day and the big amount of data is searched and users face the problem of duplicate data. So, several de-duplicating schemes [1] have be proposed and, much of the research has been done in the field of data de-duplication. Some de-duplication schemes have been proposed by the research community [4, 5, 6] demonstrating how de-duplication permits exceptionally engaging decreases in the utilization of storage resources.

4. Proposed Method

In this method we are trying to eliminate a sequence of words that is formed with the help of algorithm, designed on keywords as input entered by the user. To achieve this, we have made a CMM (Content matching machine).

In CMM a string is prepared with help of keywords entered by the user/client in the sequence as entered and a string to sting comparison is done to the whole content of the server.

A SERP is prepared just like a normally crawled webpage showing result of a search. The originality of any data required is decided with the help of deciding factor that could be anything like time stamp, author's validation and many more. Then refined page content is made with pure de – duplicated data is shown on the browser. The page crawled earlier and the final page on the browser screen will showcase the huge difference in the user experience.

The software used for comparison and timestamp matching is being developed and is inspired by very famous open source software – apache API Commons IO.

A. System Architecture

The whole software and its architecture are being developed on python as a core language. There are many modules which work in synchronization to achieve the overall results.

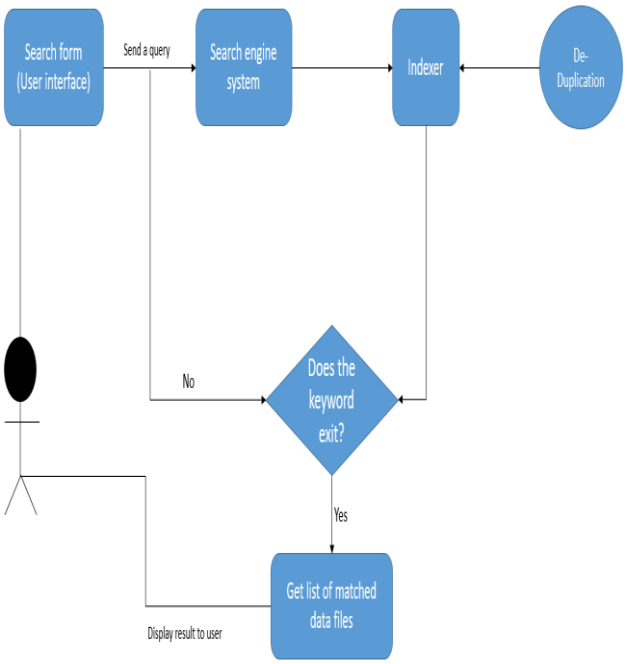

Fig. 2: Working of a Proposed System

A basic overall proposed architecture is explained with the help of figure 3.

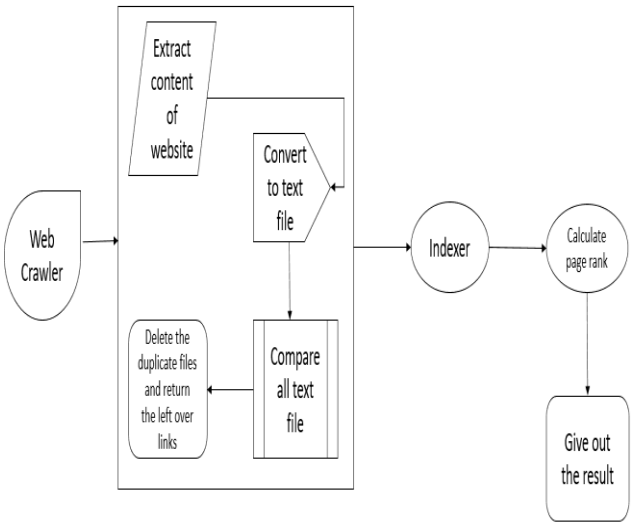

Fig. 3: Proposed Architecture of DeDuSERP

5. Implementation

DeDuSERP has been implemented in Python 3.6.3 using the pycharm IDE on a windows 10 machine. It contains separate functions for web crawling i.e. crawl(), to check the duplicates i.e. DeDuSERP, to index pages i.e. indexer() to get the URLs of the files already present on the cloud. The content matching has been done using the Sequence Matcher module from difflib(library) which checks the sequence and returns the probability of being similar content. The DeDuSERP program has been designed to shorten the number of search results we get by avoiding duplicate webpages.

6. Experiments

To test the performance of the DeDuSERP we modeled a search engine system by creating a WAMP server on the system we would use DeDuSERP system to refer it and then connecting it to client computer. The Client Computer made a request by posting the query term and the websites to be searched upon. The DeDuSERP system takes that request and crawled upon the related links via web crawler. We then opened each page and copied the content to a text document. We then searched for the similarities between those documents and removed the webpages from the crawler database whenever we found 95% of similar content in two pages. We return only the unique webpages to the database so that only the distinctive search results are obtained. The indexer then indexes the webpages and the page rank of each is decided and returned to the client.

7. Results

We get the filtered results by removing the similar content from the search results. And also, the results which are most unique to the keywords given are presented to the users. The results are well indexed and the results obtained are calculated by giving ranks to page according to the page rank algorithm. These results can also be obtained for the image, videos and other results by browsing the descriptive data they contain with them and also the keywords associated with them. This solves a critical problem of any search engine as it gives out a less search results to the user and it also save much time of the search engine as it saves the amount of duplicate pages and need not crawl those duplicate pages again.

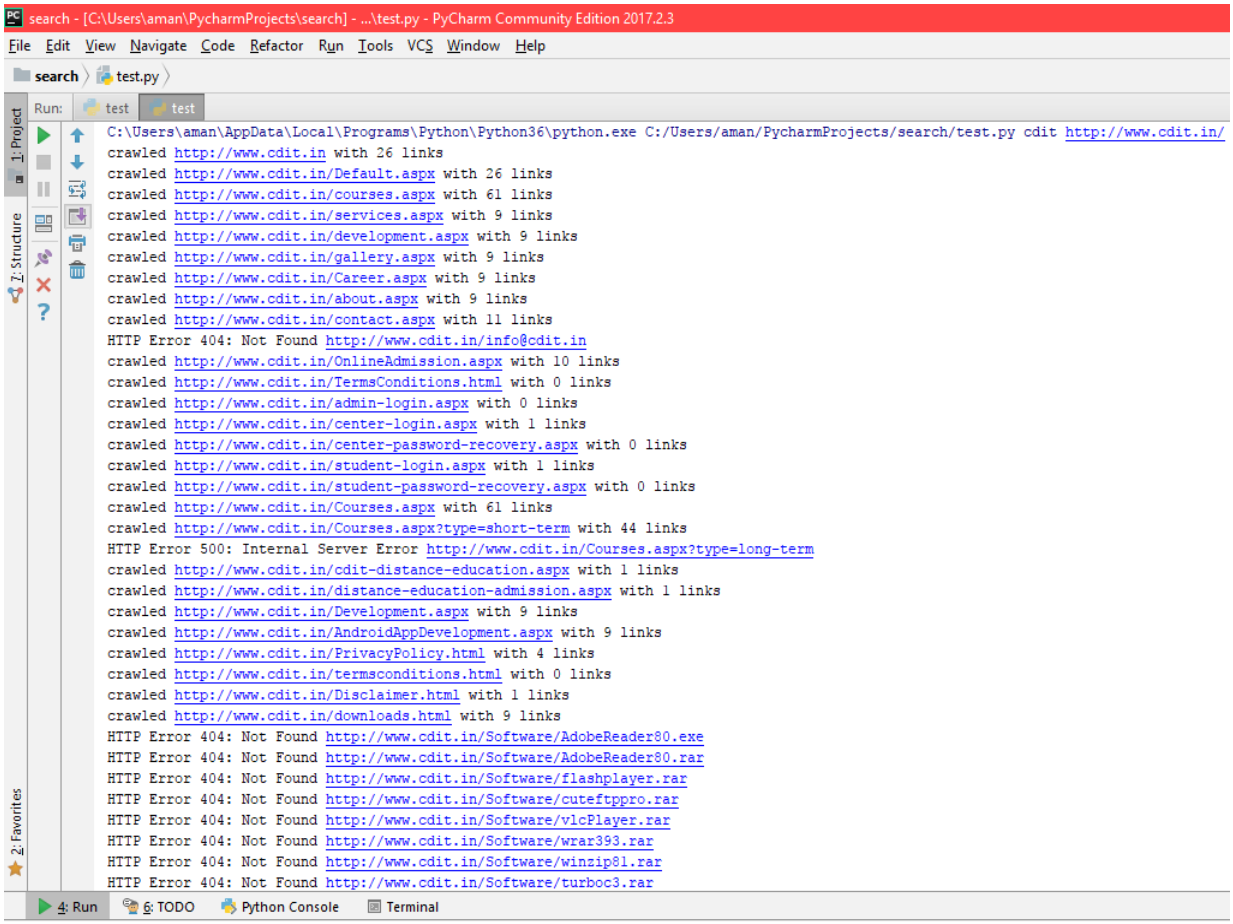

Fig. 4: Crawling Results of a Particular Query

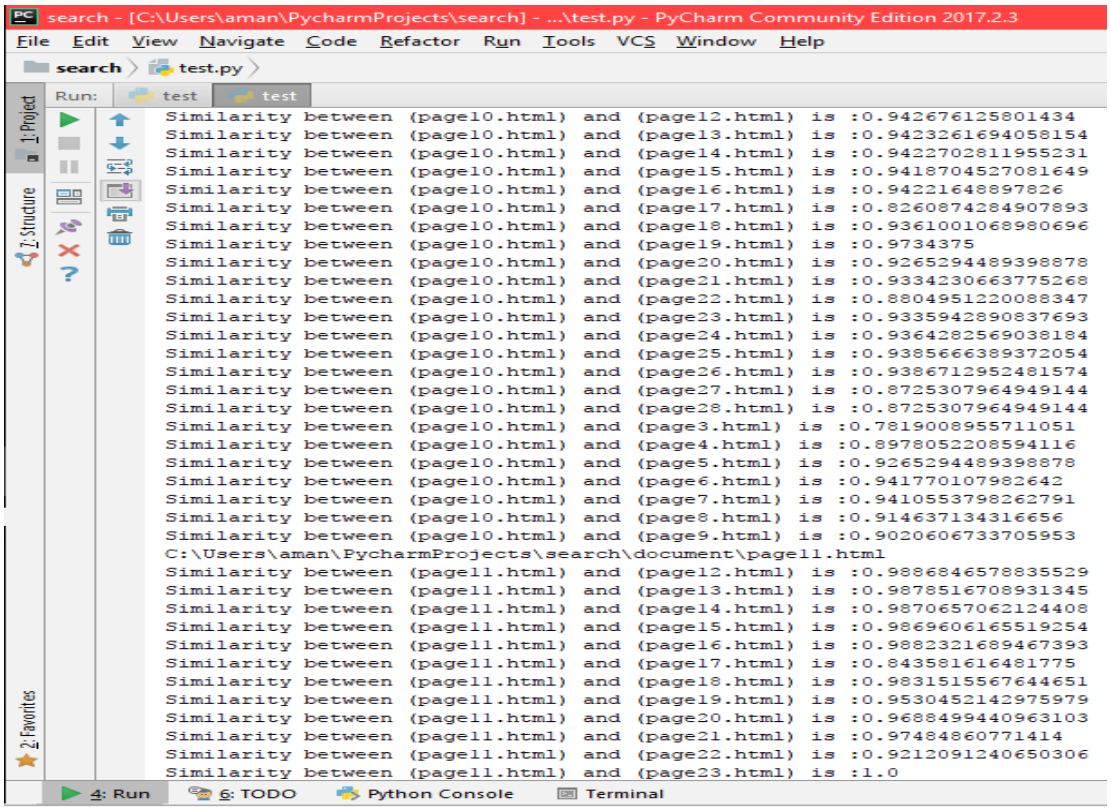

Fig. 5: Similarity between Duplicate Pages

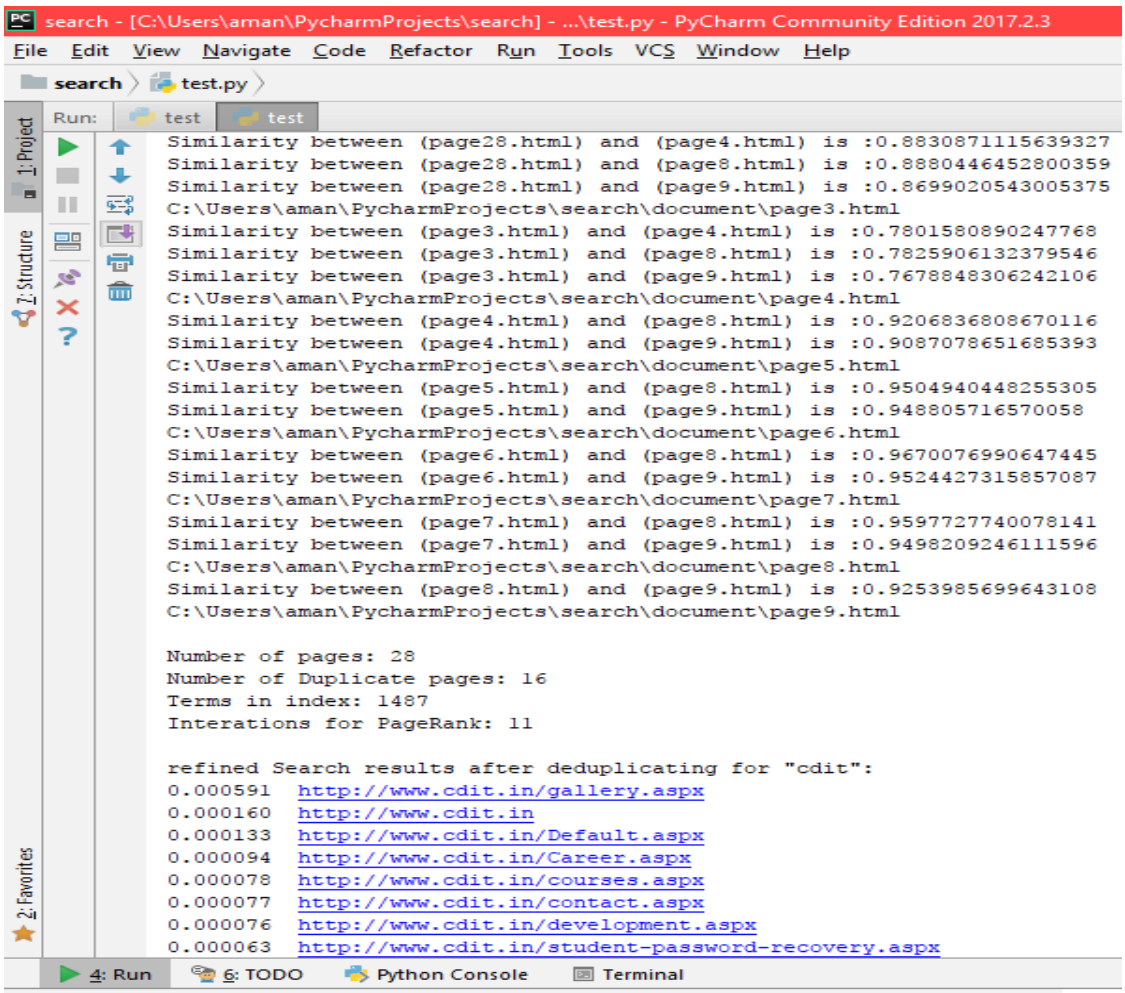

Fig. 6: Refined Search Results after De-duplication (Part – A)

8. Conclusion

Now-a-days an average man on computer with internet encounters a huge quantity of information on a single day. A century ago, an equivalent amount of information was accessible to an individual over the period of a year. The advent and subsequent boom of the internet had led to this tremendous amount data duplicacy ruining the authenticity of data.

The World Wide Web is treasure trove of data. Accessibility of goliath information on the web prompts the need to avoid irrelevant data and acclimatize related data and DeDuSERP plays crucial role in separating data (BIG DATA). This research takes care of a basic issue of any search engine as it gives out a less indexed lists to the client and it additionally spare much time of the web index as it spare the measure of copy pages and need not creep those copy pages once more.

Google which owns 65% of the market, now processes over 40,000 search queries every second on average (visualize them here), which translates to over 3.5 billion searches per day and 1.2 trillion searches per year and a single google query uses 1000 computers in 0.2 seconds to retrieve a SERP by travelling an average of 1500 miles to a data centre and back. And our page ranking algorithm and de-duplicating algorithm will refine the facts stated above.

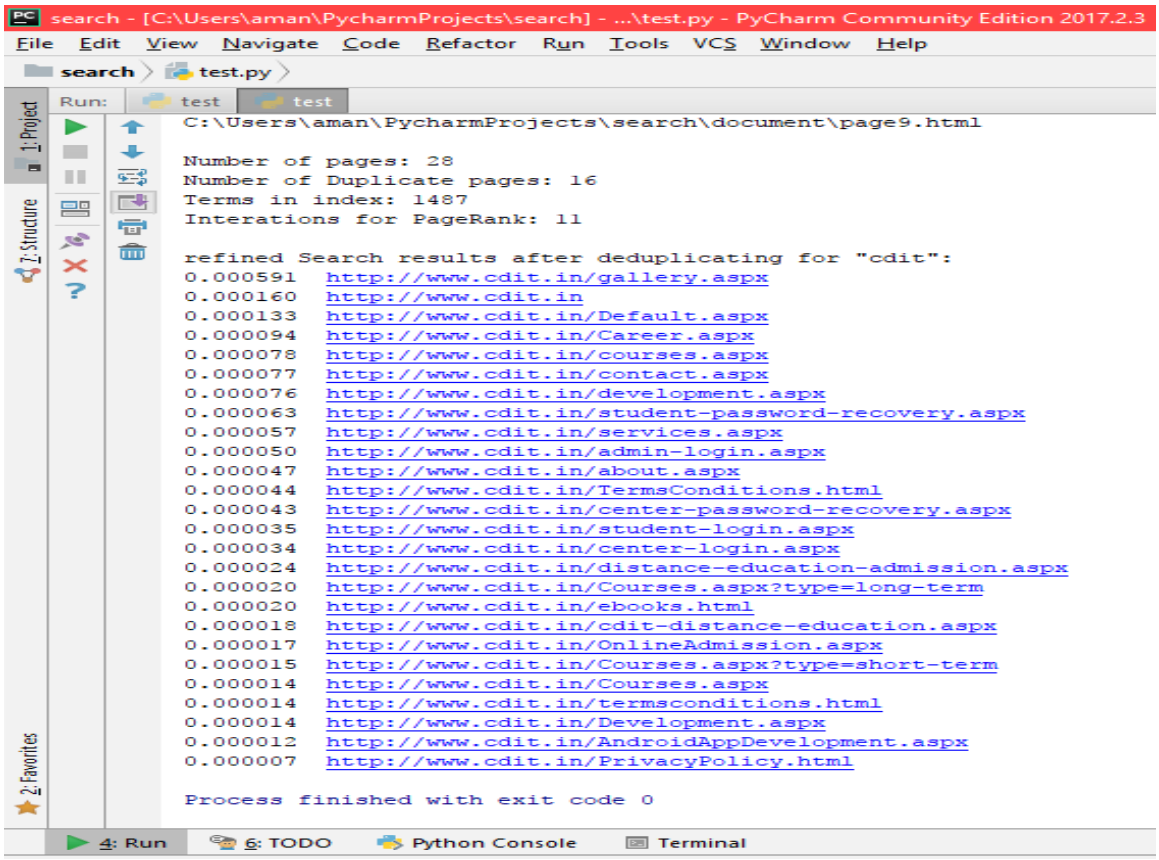

Fig. 7: Refined Search Results after De-duplication (Part – B)

References

[1] De-duplication in Search Results US20150161267A1. https://www.google.ch/patents/US20150161267

[2] S. Brin, L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems. 30. . 10.1016/j.comnet.2012.10.007. http://infolab.stanford.edu/~backrub/google.html

[3] Jin Li, Yan Kit Li, Xiaofeng Chen, Lee, P.P.C., W. Lou, "Aybrid Cloud Approach for Secure Authorized Deduplication," in Parallel and Distributed Systems, IEEE Transactions on , vol.26, no.5, pp.1206- 1216, May 1 2015

[4] L. Aronovich, R. Asher, E. Bachmat, H. Bitner, M. Hirsch, and S. T. Klein. "The design of a similarity based deduplication system" Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference (SYSTOR '09). ACM, New York, NY, USA, Article 6, 14 pages.

[5] Stanek, Jan, et al. "A secure data deduplication scheme for cloud storage." Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2014. 99-118

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The Page Rank Citation ranking: Bringing order to the web. Technical report, Stanford digital library Technologies Project, Stanford University, Stanford, CA, USA, 1998.

[7] Incredible growth of internet users. https://thenextweb.com/insider/2017/03/06/the- incredible-growth-of-the-internet-over-the-past-five- years-explained-in-detail.

[8] N. Sharma, A. Mishra, P. Garg. Nap: Improving The Quality Of Search By Deduplicating The Search Results. International Journal of Engineering Applied Sciences and Technology, 2016 Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 141-144 Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

[9] A. Murgai, N. Sharma,"Page Rank Algorithm Expressed in Terms of Link Distance and a Modified Procedure of Page Rank Calculation",3rd International Conference on Computer Modeling and Simulation (ICCMS 2011), 2011, ISBN: 978-1-4244-9243-5, pp. 586- 588

[10] T.Padmapriya and V.Saminadan, "Utility based Vertical Handoff Decision Model for LTE-A networks", International Journal of Computer Science and Information Security, ISSN 1947-5500, vol.14, no.11, November 2016.

[11] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on 26-27 Feb. 2015,Publisher: IEEE,DOI: 10.1109/ECS.2015.7124833.