International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Retrieving and Saving Meaningful Keywords in Unstructured PDF Documents using Binary Decision Diagrams

Anuragini Sharma

Abstract*: With the growing intricacy in data engendered and processed across sundry platforms today, the desideratum for consistency has grown. Structured data is utilized for a number of purposes which is not feasible with unstructured data. The purpose of this study was to convert data from unstructured format to structured in portable document format with the help of new framework using the concept of Binary Decision Diagrams and Boolean operations. Binary decision diagrams are data structures for representing Boolean functions taking Boolean as input and generating Boolean as output and hence creating a binary diagram. This research is mainly carried out to show how we can store large number of data easily in the form of bits. The entire focus is on retrieving the meaningful information from unstructured textual data in PDF documents using Boolean operations and bag model, thus, saving the meaningful keywords in the form of binary decision trees. Later on clustering the documents based on commonalities between the documents. This research presents a way for increasing the efficiency of converting unstructured data to structured in PDF and saving huge number of data in the form of bits using this novel framework.

***Index Terms*: Unstructured data, structured data, binary decision diagram, bag model, clustering, PDF data retrieval.**

I. INTRODUCTION

The term 'Big Data' was first coined in 1990s, but it took a considerable amount of time for organizations to understand and adopt the concept for internal use.

Big data means immense volumes of high velocity, complex and changing data which require fast paced techniques to store, manage and analyze the information. Main issue is with retrieving and storing the meaningful insights from unstructured data.

For the potential of unstructured data to be realized, the data must be converted to more utilizable structured form. Text mining plays an important role in transformation which is used to discover the previously unknown as well as the interesting information from a plethora of textual data. Our research focuses only on textual data in the portable document formats.

Unstructured Data: the scale of data has changed reality from terabytes to petabytes and quickly growing. In unstructured data, the information does not fit neatly within the confines of a database. Total amount of unstructured data is growing by 62 percent every year [6]. Structured data grows predictably but unstructured data grows exponentially. Despite of tools availability, Still we have lot many problems on how we can store unstructured data, retrieve meaningful insights from it and save it for later use. This is a problem not only in management terms but also from the perspective of data retrieval and storage.

Structured Data: We can turn unstructured data into goldmine i.e. structured data which is information that can be easily processed.

II. ORGANISATION OF PAPER

In section II, we discuss about the aim of paper. In section III, we start by laying the theoretical background on which the construction of proposed tool stands. Then we briefly introduce BDD. In section IV, we discuss the methodology along with the new framework on which our tool works. Next, we present and discuss the example to illustrate the use of tool and its functionalities. In section V, we conclude and point to further extensions and improvements to the tool.

III. AIM OF THE PAPER

The aim of this paper is to propose a new framework for retrieving and saving meaningful keywords in PDF using binary decision diagrams

IV. MATHEMATICAL BACKGROUND

The tool presented in this paper is based on mathematical paradigm that is binary decision diagram. BDD is a finite directed acyclic graph with a unique initial node, where all terminal nodes are represented by 0 or 1 and all non-terminal nodes are labelled with a node names. Each non-terminal node has two edges true and false represented by a dashed line and a solid line, respectively. For more details on BDDs, we refer the reader to [1, Chapter 6]. BDDs represent Boolean functions.

Example: The BDD for f = a&b | a&c is shown in figure[1]

Manuscript published on 30 September 2019
* Correspondence Author
Anuragini Sharma*, Assistant Professor in School of Computer Application at Lovely Professional University, Punjab India.

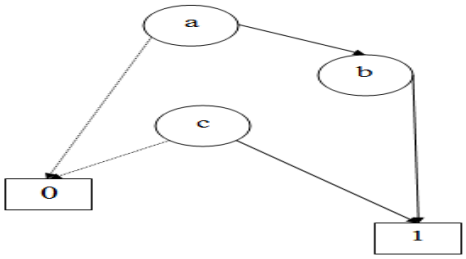

Figure 1: BDD for f = a&b | a&c

V. METHODOLOGY

A. Proposed Framework

Because the purpose of this study, the PDF dataset was first uploaded.

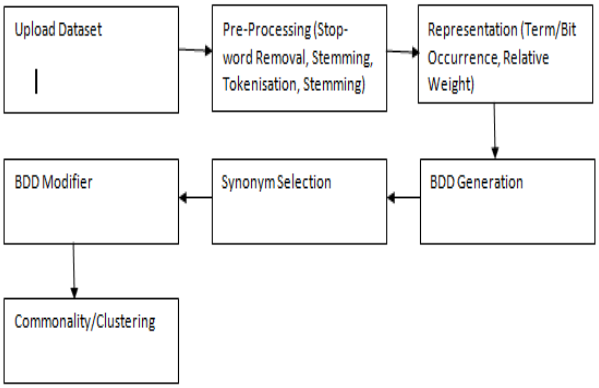

Figure 2: Framework

B. Illustrative Example

Step1: First step is to Browse and Upload the Dataset as shown in Figure [3][4].

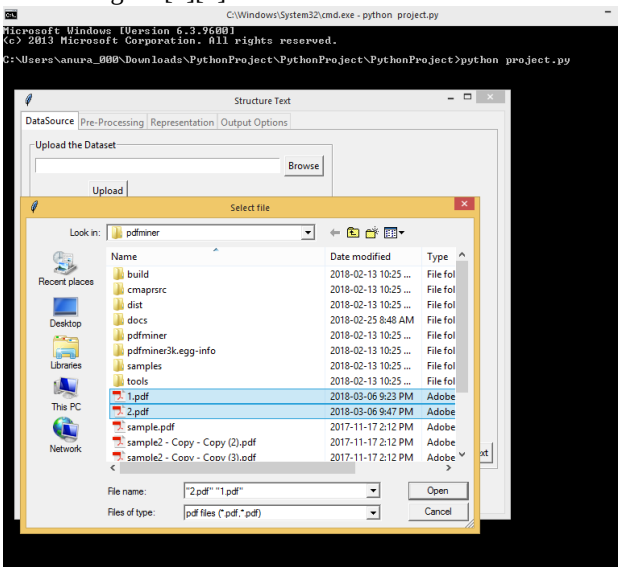

Figure 3: Selecting Dataset

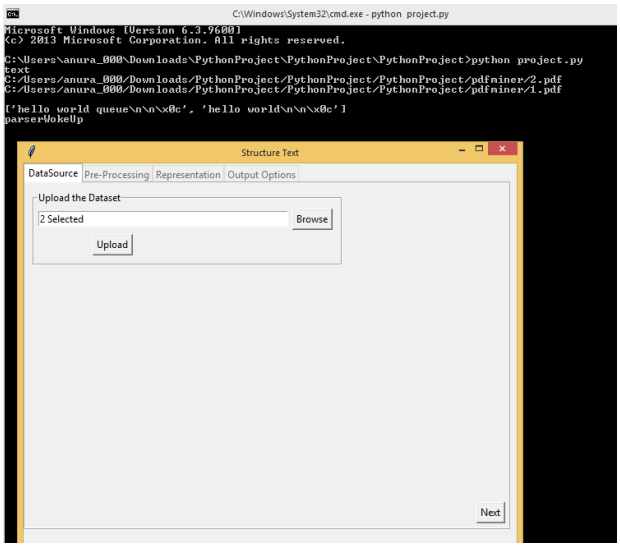

Figure 4: Uploading Dataset

Once we upload the dataset or the documents, the content on the documents is read as it can be seen in cmd prompt.

Step2: ***Pre-processing***: This technique is used to minimize the complexity of documents and make them easier to handle. Basic idea is to extract unique content bearing words from set of documents treating these words as meaningful keywords.

i. Vocabulary Reduction: It includes stopwords removal. A stop words list is a list of commonly repeated features which emerge in every document. The common features such as or, and, but, he, she, it etc. need to be removed as it does not have effect on the categorization process. So removing stop words accounts 20-30% of total word counts. Tokensiation here refers to removing any kind of characters like question marks, wild characters etc[4].

ii. Term Normalization: It includes stemming which is the process of removing affixes from words i.e. the process derived for reducing inflected words to their stem. The stem need not to be identified to the original morphological root of the word. For example: (connect, connects, connected, and connecting) from the mentioned above example, the set of words is conflated into a single word by removal of the different suffixes -s, -ed, -ing to get the single word connect. This study applied standard Porter Stemming Algorithm for finding the root words in the document. Normalize case would just normalize all the characters into lowercase. So stemming words may reduce size as much as 40-50% in the pre-processing task. This is represented in Figure[5].

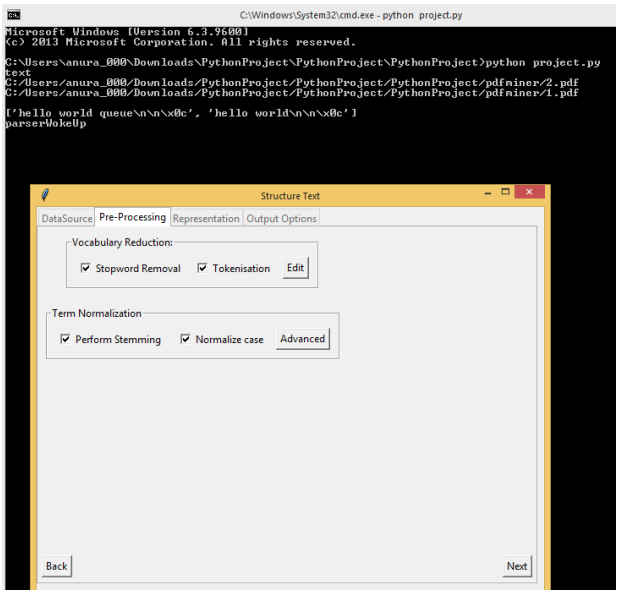

Figure 5: Data Cleaning/Pre-Processing

Step3: ***Representation***: Common representation used for processing is the TF-IDF which reduces the importance of common terms in the corpus. To keep it simple, Assume:
We uploaded 2 PDF documents.
1.pdf contains words: hello world
2.pdf contains words: hello world queue

Under representation: Vocabulary would be generated and it holds only the meaningful keywords which are "hello", "world", "queue". Then we find term occurrence of 1.pdf which is [111], i.e. the document contains all the terms present in the corpus called as vocabulary. Term occurrence of 2.pdf would be [101], which shows from the vocabulary this particular document contains only two words.

In other words, after pre-processing we have a corpus of meaningful keywords which represents vocabulary. Term occurrence is to just vectorize the terms to 0 and 1 referring to the presence(1) or absence(0) of term in the vocabulary specifying if the term has occurred in the document as compared to the vocabulary. Figure[6]-[9] shows this procedure.

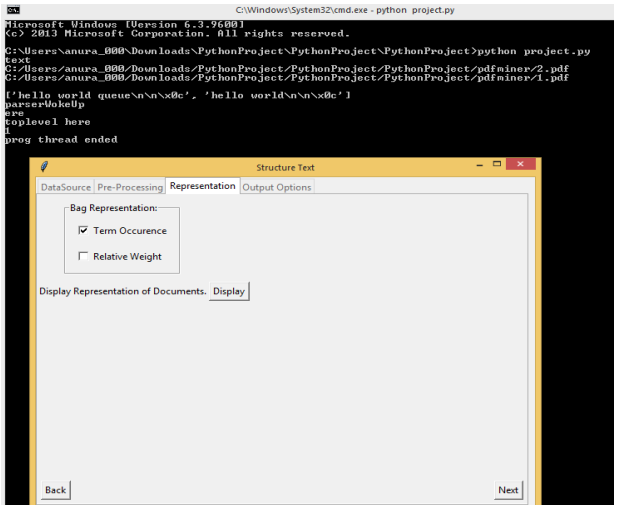

Figure 5: Vectorization

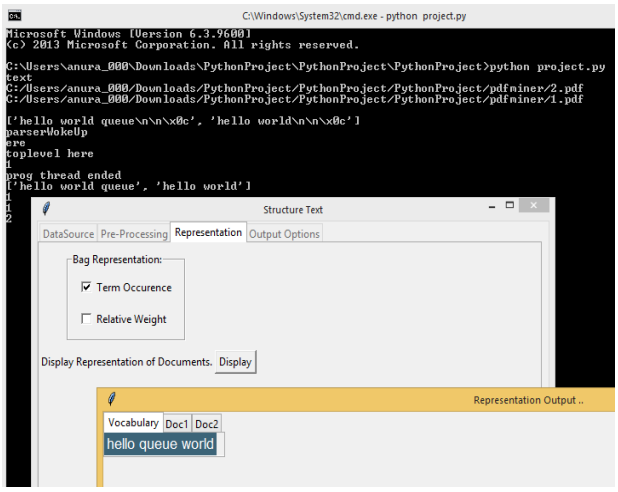

Figure 6: Representation

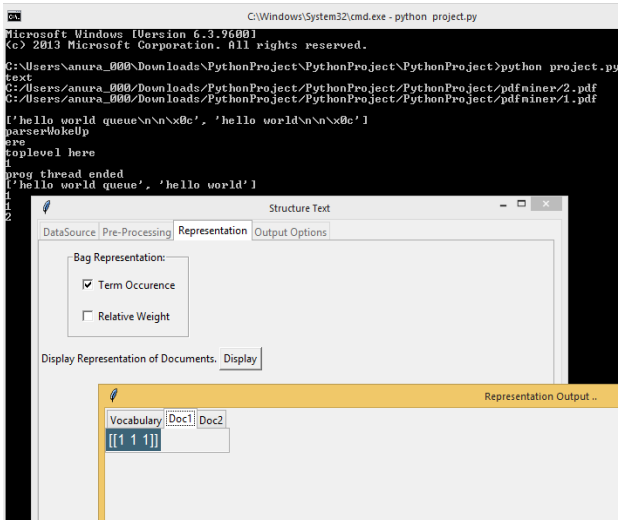

Figure 8: Term Occurence

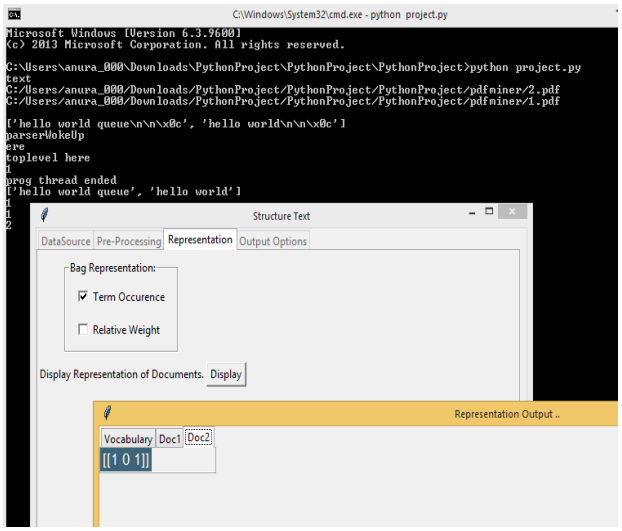

Figure 9: Term Occurence

Relative Weight: Next step is to weight every word by its inverse document frequency as shown in Figure [11][12]

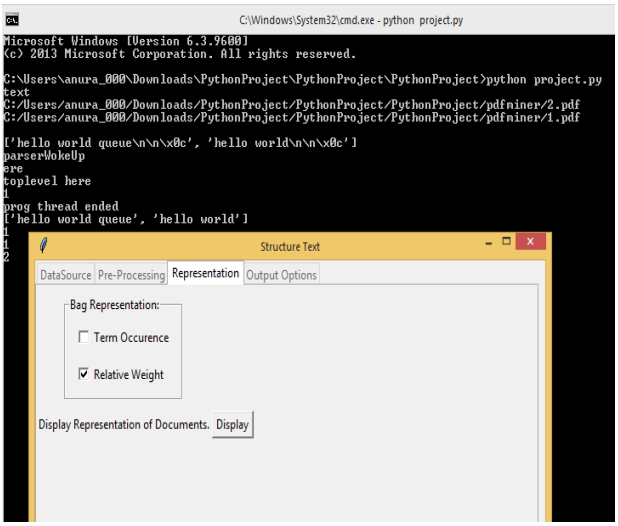

Figure 7

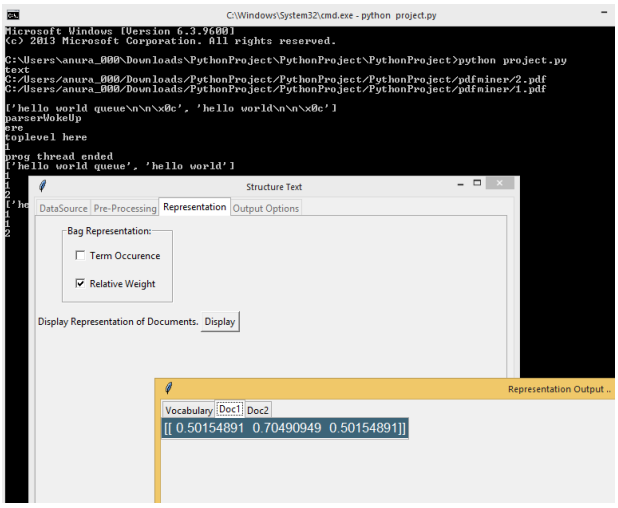

Figure 8: Relative Weight

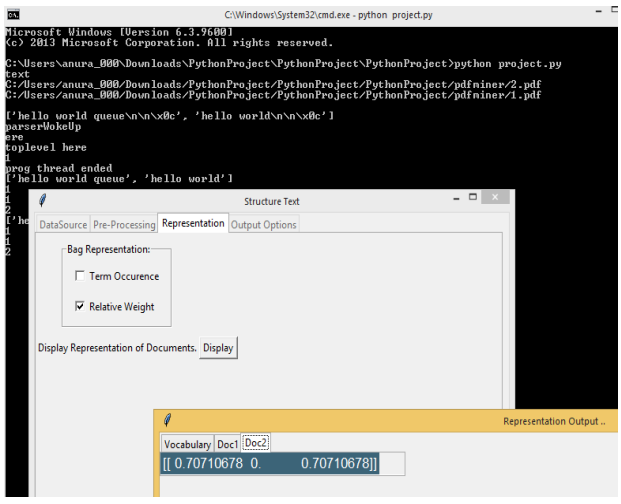

Figure 9

Step4: Creating the binary decision diagram from the vocabulary (Refers to Dictionary in below figure, Dictionary and Vocabulary are used interchangeably). The concept is to initialize all the bits to zero, i.e. all the terms in the vocabulary are traversed down to zero and a corresponding BDD called as "Main BDD" is generated. After that we can compare each and every document with the main BDD and set the bit to 1 for the term which is present in a particular document as compared to main vocabulary. Hence for each document a corresponding BDD is generated.

Example: Assume our

vocabulary (Dictionary) includes terms "hello" "queue" "word" "bank" "home".

Document 1 includes terms "queue" "word"

Main BDD would have all the aforesaid five terms set to 0. After that we have to apply logic and generate expression in such a way that a new BDD for document 1 is generated from the main BDD where "queue" and "word" are now set to 1 and all other terms are still zero. The expression for the same has been generated in the cmd and shown in Figure[13]

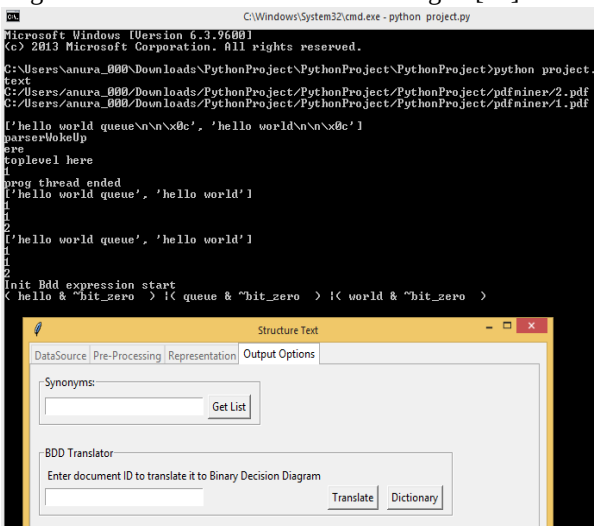

Figure 10

After clicking on Dictionary, Main BDD in the form of PDF would be generated as Figure[14]

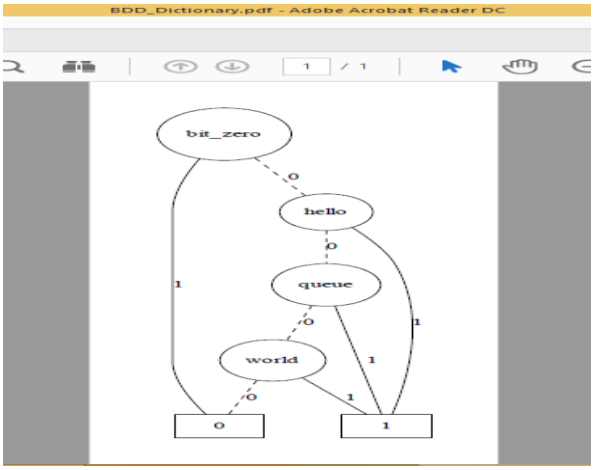

Figure 11: Main BDD

Above BDD is the main BDD. Next when we click on translate button(it is called as modifier, which modifies the main BDD) along with the ID (ID refers to serial; number of documents, i.e. if i upload 100 documents, ID would be from 0-99.).

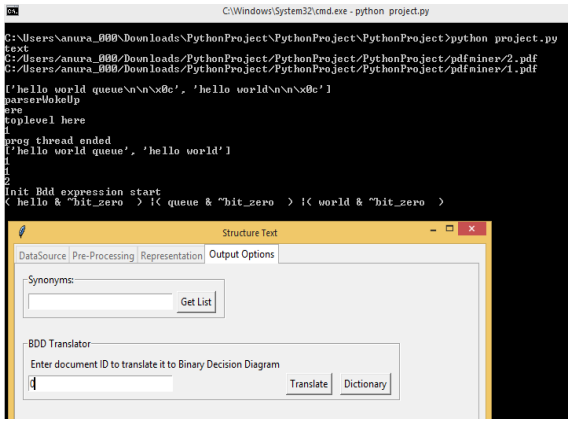

Figure 12

The following BDD is generated for first document in the dataset:

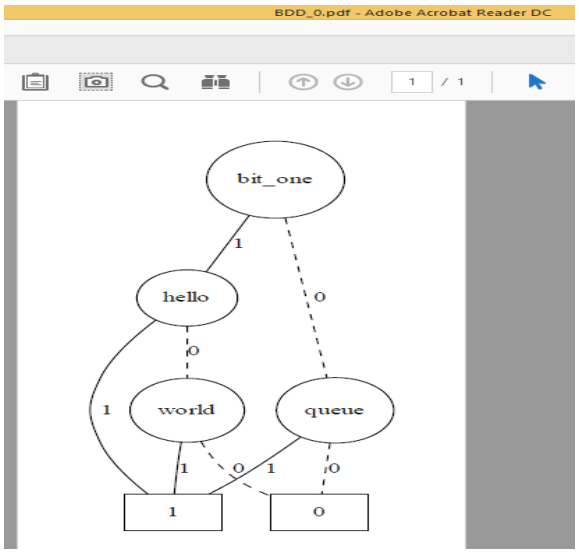

Figure 13: Generated BDD

It can be seen in Figure[15], we entered 0 as ID for first document i.e. 1.pdf which has words hello and world. The generated BDD's are the reduced ordered binary decision diagrams with expression.

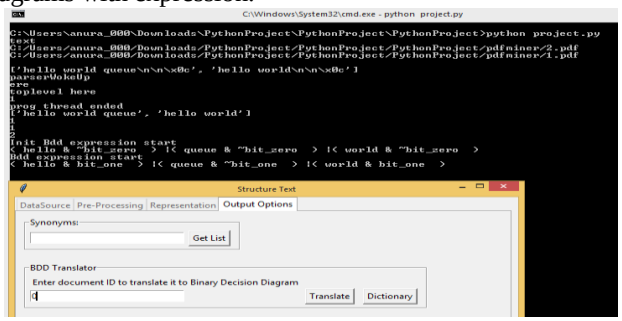

Figure 14

As it can be seen in Figure[17], hello and world are turned to bit1, but queue is still 0.

Step5: Synonyms: Next step is to find the synonyms of the terms and modifying the BDD with again clicking on the translate button in the BDD module with the updated synonym terms. Figures [18] illustrate the process.

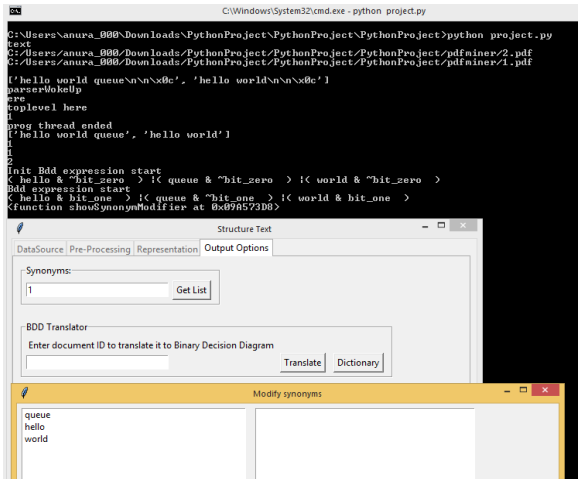

Figure 15

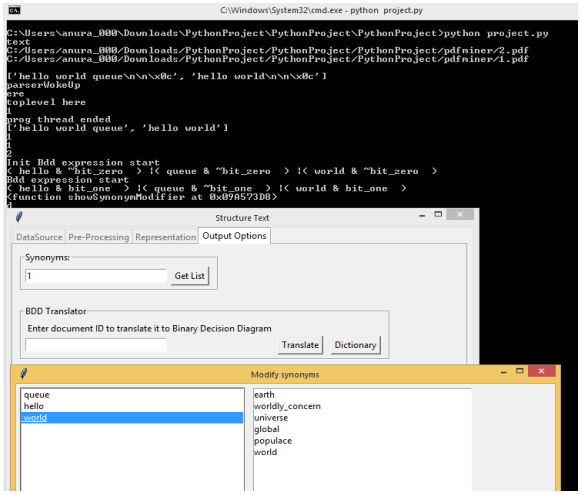

Figure 16

Figure[19] shows the ID=1, i.e the second document in the dataset, here it refers to 2.PDF which holds words hello, queue and world. We click on "world" to check the synonyms. In Figure[20] we choose Universe and now again we will update the BDD with new word "Universe" instead of "world" by clicking on the translate button as shown in Figure[22].

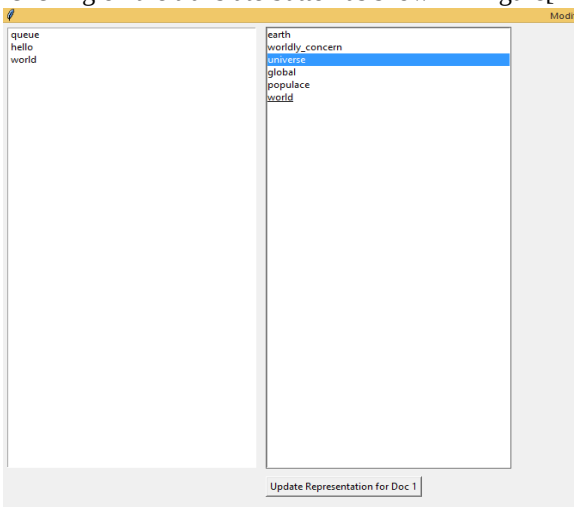

Figure 17

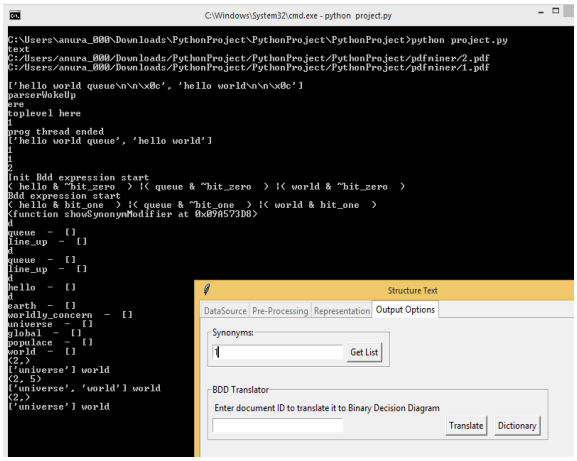

Figure 21

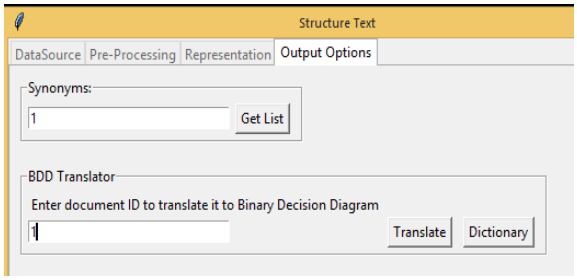

Figure 22

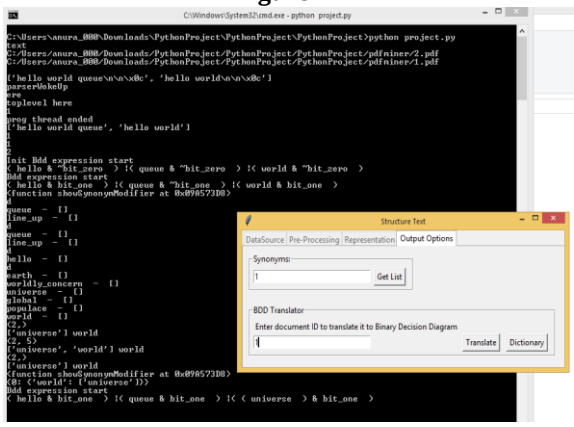

Figure 23

Step6: Clustering: Last step is to perform clustering. Like group together the documents that have same terms. For this purpose we need to use AND and OR Boolean operators. We can apply "AND" operation to find out the commonalities between the documents. AND operation takes the multiple BDD's and generates a new BDD with the exactly same terms. OR operation meets atleast one criteria or another criteria. These operations corresponds to Intersection and Union respectively in Mathematical Terminology.

VI. CONCLUSION

In this study, we extracted the meaningful keywords from unstructured PDF documents by first cleaning the data. Then we represented the bit occurrences of the terms and find out commonalities between the documents with the help of AND, OR operators. Hence representing, how the data can be cleansed to retrieve the useful keywords using Binary Decision Diagrams.

REFERENCES

1. M. R. A. Huth and M. D. Ryan, Logic in Computer Science: Modelling and reasoning about Systems. Cambridge University Press, 2000.
2. S. B. Akers, “Binary decision diagrams,” IEEE Trans. Computer, vol. 27, no. 6, pp. 509–516, 1978.
3. R. E. Bryant, “Graph-based algorithms for Boolean function manipulation,” IEEE Transactions on Computers, vol. C-35, pp. 677–691, 1986
4. Maluf D.A. Tran, P .B “Managing unstructured data with structured legacy systems” , Aerospace conference 2008, IEEE.
5. Seth Grimes. “is unstructured data merely modeled” published in Intelligent Information week journal. 2005.
6. Robert Malone. “Structuring unstructured data” published in Forbes magazine, USA. 04-may-2007.
7. Ramesh Nair, Andy Narayanan, “Benfitting from Big data Leveraging Unstructured data Capabilities for Competitive Advantage”, Booz & Company Inc. 2012.
8. Clinton Wills Smullen, “A Benchmark Suite for Unstructured Data Processing.
9. Shin-ichi Minato, A fast algorithm for cofactor implication checking and its applications fo knowledge discovery, IEEE.
10. Wassim Ayadi, Khedija Arour, A Novel Parallel Boolean approach for discovering frequent item sets, IEEE.
11. Mita K. Dalal, Mukesh A. Zaveri, Automatic classification of unstructured blog text.
12. "Symbolic Boolean manipulation with ordered binary decision diagrams, "Carnegie Mellon University, Pittsburgh", PA, US, Tech, Rep., 1992.
13. https://www.peopledatalabs.com/blog/searching-unstructured-data

AUTHORS PROFILE

Mrs. Anuragini Sharma received M.Tech(CSE) in Department of Computer Science and Engineering, from Guru Nanak Dev University, Punjab in 2010. She served as an Assistant Professor in School of Computer Application at Lovely Professional University, Punjab for five years (2010-2015). She is currently pursuing MSc(Computer Science) from Mcmaster University, Canada, but is on term off with good standing. She is also preparing herself for Ph.D entrance exams. Her key area of research focuses on Data mining.