International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Role of Pre-processing Phase in Document Clustering Technique for Gurmukhi Script

Mukesh Kumar, Amandeep Verma

Abstract: ***Document clustering plays a central role in knowledge discovery and data mining by representing large data-sets into a certain number of data objects called clusters. Each cluster consists similar data objects in such a way that data objects in the same cluster are more similar and dissimilar to the data objects of other clusters. Document clustering technique for Gurmukhi script consists two phases namely: 1) Pre-processing phase 2) Processing phase. This paper concentrates pre-processing phase of document clustering technique for Gurmukhi script. The purpose of pre-processing phase is to convert unstructured text into structured text format. Various sub-phases of pre-processing phase are: segmentation, tokenization, removal of stop words, stemming, and normalization. The purpose of this paper is to present the significant role of pre-processing phase in an overall performance of document clustering technique for Gurmukhi script. The experimental results represent the significant role of pre-processing phase in terms of performance regarding assignment of data objects to the relevant clusters as well as in creation of meaningful cluster title list. .***

Keywords : Document clustering; Gurmukhi script clustering technique; Pre-processing phase; Punjabi Text document clustering; Data mining techniques; Machine learning; Unsupervised learning.

I. INTRODUCTION

Data mining is an analytical process of analyzing the potential data patterns from a large amount of raw data. To utilize large amount of raw data streams and to transform into valuable information, data mining techniques have been proved to be an important tool in many areas like medicine, education, agriculture, biology, transportation, etc. Data mining involves collection of effective and useful data streams that makes use of sophisticated algorithms to evaluate further processing. Data mining is a multi-disciplinary technique that incorporates various approaches like machine learning, statistics, artificial intelligence, visualization, data-warehouse, pattern matching, decision trees etc. Alternatively, data mining is also referred to as data knowledge recovery in databases (KDD).

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Mukesh Kumar*, Dept. of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India.
Amandeep Verma, Punjabi University Regional Centre for Information Technology & Management, Mohali, Punjab, India.

Machine learning is a process of developing a machine to learn from existing data and to adapt behavior for future incidents. It enables the machines to make data-driven decisions rather than to be programmed explicitly and learn over time. Supervised learning deals with an algorithm which learns from previously available data sets known as training data to get useful information for the further processing. Whereas, unsupervised learning deals with an algorithm which learns from itself (without any training data) based on input data sets only [1].

Clustering is an unsupervised machine learning technique that divides the data sets into number of groups (called clusters) such as there is high intra-cluster similarity and low inter-cluster similarity [4]. Clustering is different from classification as classification deals with supervised learning and clustering deals with unsupervised learning [9]. Clustering algorithms are used in many real application areas like: in searching, the search engines arrange group of similar items that can be explored by user; in commercial applications, the customer details can be grouped and summarized through the buying habits; in geography, clusters of different species can be created based upon geographical climate etc.

II. PROPOSED WORK

Domain independent document clustering for Gurmukhi script provides a solution for text clustering in Gurmukhi script. The proposed solution for text clustering in Gurmukhi script mainly deals with two phases namely: pre-processing phase and processing phase. In pre-processing phase, various steps are performed on input text in Gurmukhi script for converting unstructured text into structured text; it is prior requirement of clustering algorithm. Various sub-phases of pre-processing phase are namely: segmentation, tokenization, removal of stop words, stemming, and normalization. This paper will focus on pre-processing phase and its significant role in terms of performance for assigning data sets to relevant cluster as well as in terms of creation of a list of meaningful cluster titles.

A. Pre-processing phase

Pre-processing phase is basic requirement of document clustering technique that is applied on input text documents to convert unstructured input text into structured text format. Various sub-phases of the pre-processing phase are given in "Fig. 1".

- **Segmentation:** The process of segmentation is an initial step to be performed in pre-processing phase which deals with the identification and extraction of sentences by sentence boundary.

In Gurmukhi script, each sentence boundary is represented with dandi (i.e. "।").

So, in Gurmukhi script, the input text can be segmented into a number of sentences on occurrence of every "।" (i.e. dandi) at the end of sentence.

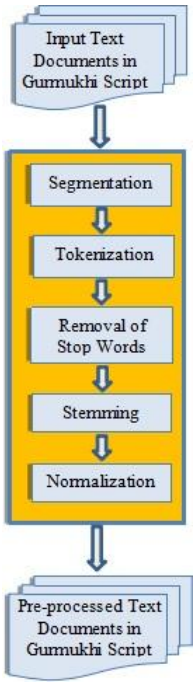

Fig.1- Steps of Pre-processing Phase

- **Tokenization:** After identifying and extracting the sentences from input text documents, it is required to break down sentences into words. The tokenization step of pre-processing phase is performed to break down sentences into words. Using the process of tokenization, the sentences in Gurmukhi script are fetched and broken down into words on occurrence of any of the space, comma, colon, hyphen, single quote and double quote.
- **Removal of stop words:** After performing tokenization, a list of words from each input text document is achieved. At this stage, each word can be recognized if it is useful and consist meaningful information for the further processing. To do so, some unimportant words which do not provide any useful information and need not to include in the list of indexing terms for further processing, are eliminated from the list so as to lessen the text weight. In Gurmukhi script, to recognize unimportant words (known as stop words) and to eliminate these words from of input text, apart from a list of 184 word in Gurmukhi [5], a self created list of 475 words is created and used. A list of self created stop words is given as below:

Table-1: list of self created stop words

ਬਾਰੇ	ਜਾਣ	ਪਾਰ	ਵਿਰੁਧ	ਖ਼ਿਲਾਫ਼
ਨੇੜੇ	ਤੇੜੇ	ਲਾਗੇ	ਆਸ	ਪਾਸ
ਜਿਉਂ	ਕਿਉਂਕੀ	ਪਰਾਂ	ਲੇਕਿਨ	ਨੇਝੇ
ਜਗਾ	ਦੂਜਾ	ਪੁੱਠਾ	ਉਲਟਾ	ਪਾਸੇ
ਪ੍ਰਤੀ	ਬਗਲ	ਤਹਿਤ	ਪਾਸਾ	ਹੇਠ
ਥੱਲੇ	ਭਿੰਨ	ਤਾਈਂ	ਉੱਪਰ	ਸਹਿਤ
ਅੰਦਰ	ਬਗੈਰ	ਦੋ	ਫੇਰਾ	ਕਿਉਂਕਿ
ਬਾਵਜੂਦ	ਦੇ	ਛੱਡ	ਵੇਖ	ਪਰੇ
ਰਾਹੀਂ	ਜਿਹੜਾ	ਬਜਾਏ	ਬਦਲੇ	ਅੱਗਾਹ
ਅਗਲਾ	ਜਿਹਨਾਂ	ਕਿੱਧਰ	ਬਾਹਰ	ਪਹਿਲਾਂ

ਕਾਰਨ	ਜਿੰਨੀ	ਵੇਖ	ਸਕੇ	ਤੋਂ
ਇਲਾਵਾ	ਕਰੀਬ	ਸਾਹਮਣੇ	ਹੋਇਆਂ	ਓਹ
ਦੇਖ	ਨਾਲੋਂ	ਉਚੇਰਾ	ਵਾਧੂ	ਕੋਲ
ਵਿਖੇ	ਏ	ਓ	ਵੇ	ਨੀ
ਦੁਆਰਾ	ਚ	ਤੱਕ	ਹਿਤ	ਵਾਸਤੇ
ਆਲ਼ੇ	ਦੁਆਲ਼ੇ	ਬਿਨਾਂ	ਅਸੀਂ	ਅਸਾਂ
ਤੁਸਾਂ	ਤਾ	ਦੁਰਾਡੇ	ਸਾਨੂੰ	ਮੈਥੋਂ
ਸਾਥੋਂ	ਤੈਥੋਂ	ਤੁਹਾਥੋਂ	ਸਾਡਾ	ਤੁਹਾਡਾ
ਉੱਤੇਰਾ	ਇਹਨਾਂ	ਇੰਨ੍ਹਾਂ	ਦੇਂਦੀ	ਉੱਤੇਰਾ
ਓਦੋਂ	ਕਿੱਥੇ	ਓਥੇ	ਕਿੱਥੋਂ	ਜਿੱਥੋਂ
ਏਥੋਂ	ਇੱਥੋਂ	ਓਥੋਂ	ਉੱਥੋਂ	ਉਥੇ
ਕਰਦੇ	ਏਧਰ	ਓਧਰ	ਕਿੱਦਾਂ	ਜਿੱਦਾਂ
ਏਦਾਂ	ਓਦਾਂ	ਕਿਵੇਂ	ਇਵੇਂ	ਓਵੇਂ
ਕਿਹੋ	ਇਹੋ	ਵਾਲਾ	ਦੀਆਂ	ਇਨ੍ਹਾਂ
ਕਿੱਡਾ	ਦਾ	ਦੀ	ਦੇਂਦਾ	ਉਸ
ਉਸਨੂੰ	ਵਿਚ	ਵਿੱਚ	'ਚ	ਕੀਤੇ
ਉਤੇ	'ਤੇ	ਤੇ	ਤੋ	ਤੋਂ
ਨਾਲ	ਨੇ	ਇਸਨੂੰ	ਇਸਨੇ	ਉਹਨਾ
ਵਾਲਿਆ	ਵਾਲਿਆਂ	ਤਦੋਂ	ਹੋਰਨਾ	ਹੋਰਨਾਂ
ਹੋਰ	ਹੋਰਾਂ	ਜਾਂਦਿਆਂ	ਜਾਂਦੇਂ	ਜਾਂਦੀਂ
ਅਤੇ	ਨਾਹ	ਕਰਵਾਈ	ਅਤੇ	ਨਾਹ
ਹੁੰਦਿਆ	ਹੁੰਦਿਆਂ	ਤੇਰੀ	ਆ	ਆਇਆ
ਇਸ	ਵਰ੍ਹੇ	ਜਿੰਨਾ	ਮਗਰ	ਕਰਵਾਇਆ
ਆਏ	ਆਈਆਂ	ਬਿਨਾ	ਲੈਂਦਾ	ਲੈਂਦੀ
ਲੈਂਦੇ	ਮੈਨੂੰ	ਤੁਹਾਨੂੰ	ਦੌਰਾਨ	ਲਾਇਆ
ਹਨ	ਦਿੰਦਾ	ਦਿੰਦੀ	ਦਿੰਦੇ	ਕਿਰਪਾ
ਦੇਣ	ਦੇਣੇ	ਰਹਿੰਦੇ	ਰਹਿੰਦੀ	ਰਹਿੰਦਾ
ਹਰੇਕ	ਹਰ	ਬਿਲਕੁਲ	ਕਰਦੀ	ਸਕਦਾ
ਸਕਦੀ	ਲੱਗ	ਆਪਣੀ	ਆਪਣੇ	ਜਦੋਂ
ਪੈਂਦਾ	ਪੈਂਦੀ	ਪੈਂਦੇ	ਵਾਲੀ	ਬਣਾਇਆ
ਵਾਸਤੇ	ਇਹਨਾ	ਕਦਾਈ	ਮਨਾ	ਹੋਇ
ਉਹੀ	ਵਿਚੋਂ	ਵਿੱਚੋਂ	ਪਾਏ	ਪਾਈ
ਪਾਇਆ	ਅੰਦਰਲਾ	ਕੀਤੀਆਂ	ਪਾ	ਸਾਰੀ
ਕਈਆਂ	ਲਿਆਏ	ਲਿਆਈ	ਲਈ	ਦਿੱਤੇ
ਤੌਰ	ਆਪੇ	ਉੱਤੇ	ਕਰ	ਵਾਹ
ਮੇਰੇ	ਨਾਹ	ਇਉ	ਉਸੇ	ਸੈ
ਇਹੁ	ਵਲ	ਵੱਲ	ਜਿੱਡਾ	ਏਡਾ
ਓਡਾ	ਤਕ	ਵੀ	ਨਹੀ	ਲਗਾਇਆ
ਇਹ	ਜਦੋ	ਕਈ	ਤੱਦ	ਲਗਾਉਂਦਾ
ਹੋਵੇ	ਜੇਕਰ	ਬਾਅਦ	ਬਾਦ	ਸਾਰਾ
ਚੋਂ	ਕਦੀ	ਨਾ	ਹੁਣ	ਲਗਾਉਂਦੇ
ਕੇ	ਪਿਛੇ	ਪਿਛੋਂ	ਅਜਿਹੇ	ਹੀ
ਹਾਂ	ਬਹੁਤ	ਕਾਫ਼ੀ	ਹੁਣੇ	ਲਿਆਇਆ
ਹਾ	ਭੀ	ਵਲੋ	ਵਲੋਂ	ਵੱਲੋਂ
ਇੱਥੇ	ਜਿਨੂੰ	ਜਦ	ਵਾਂਗ	ਦਰਮਿਆਨ
ਹੋਣਗੇ	ਹੋਏਗੀ	ਹੋਏਗਾ	ਜਿਹਾ	ਮੁਤਾਬਕ
ਕਦੋਂ	ਕਦੇ	ਹੋਏ	ਰਹੇ	ਅਨੁਸਾਰ
ਦਿੱਤੀ	ਦਿੱਤਾ	ਦਿੱਤੀਆਂ	ਤੇਰੇ	ਬਾਹਰਲਾ
ਐਹੋ	ਕੌਣ	ਫਿਰ	ਫੇਰ	ਅੰਤਰਗਤ
ਜਿੱਥੇ	ਕੋਈ	ਕੀ	ਜੀ	ਅਤਿਰਿਕਤ
ਕਰ	ਪੈਣ	ਕਹਿ	ਦਿਨੀਂ	ਦਿਖਾਇਆ
ਕਿਸੇ	ਕਿਸ	ਤਰਾ	ਤਰਾਂ	ਮਗਰੋਂ
ਤਰ੍ਹਾਂ	ਵੇਲੇ	ਉੱਥੇ	ਕਿਤੇ	ਜੋ
ਪੂਰਾ	ਨਾਲੇ	ਹੋਈਆਂ	ਹੋਈ	ਹੋਣਾ
ਹੋਣੀ	ਜਿੰਨਾਂ	ਕੁਝ	ਸਦਾ	ਏਥੇ
ਬਾਰੇ	ਕਦ	ਕਦੋ	ਦੇਣਾ	ਹੋਇਆ
ਗਈ	ਗਈਆਂ	ਗਏ	ਗਿਆ	ਹੁੰਦਾ
ਜਾਣ	ਜਾਵੇ	ਜਾਵੇਂ	ਜਾਵਾਂ	ਕਰਕੇ
ਕਾਰਨ	ਕਿੰਨਾ	ਕਿੰਨਾਂ	ਕਿੰਨੇ	ਜਿਵੇ
ਜਿਵੇਂ	ਹੇਠਾ	ਹੇਠਾਂ	ਸਾਰੇ	ਚੱਲਾ
ਚੱਲਾਂ	ਚੱਲੇ	ਚੱਲਦਾ	ਚੱਲਦੇ	ਬਣ
ਚੁੱਕਾ	ਬਣੇ	ਏਸ	ਬਣਾਏ	ਚੁੱਕੇ
ਕੀਤਾ	ਗੱਲ	ਆਦਿ	ਲਿਆ	ਚੁੱਕਾ

ਪਰ	ਬਣੇ	ਲੈਣਾ	ਕਰਨ	ਕਰਨਾ
ਕਰੀ	ਕਰੀਂ	ਤਾ	ਜਿਨ੍ਹਾਂ	ਕਿਹਾ
ਦੀਆ	ਤਦ	ਕਹਿੰਦੇ	ਤੈਨੂੰ	ਰਹਿ
ਰਿਹਾ	ਉਹ	ਸਾਂ	ਸਭ	ਸੱਭ
ਹੋ	ਹੈ	ਹੈ-	ਹੈਂ	ਕਹਿੰਦਾ
ਆਪਣਾ	ਅਪਣਾ	ਜੇ	ਜਾ	ਜਾਂ
ਕੁੱਲ	ਕੁਲ	ਵਗ਼ੈਰਾ	ਵਗੈਰਾ	ਰੱਖ
ਲੱਗ	ਪਿਛਲਾ	ਪੱਧਰ	ਰਹੀ	ਉਸਨੇ
ਉਹਨੇ	ਉਸ	ਉਸਦਾ	ਉਸਦੇ	ਉਸਦੀ
ਤੁਸੀਂ	ਮੇਰਾ	ਮੇਰੀ	ਪਿਛਲੀ	ਆਪ
ਆਪਾ	ਆਪਾਂ	ਸਨ	ਸੀ	ਮੈਂ
ਮੈ	ਤੁਸੀ	ਤੂੰ	ਤੂੰ	ਅਸੀ
ਪਿਛਲੇ	ਤਾਂ	ਭਾਵੇ	ਭਾਵੇਂ	ਅਗਲੀ
ਬੀਤੇ	ਆਮ	ਕਿਉਂਕਿ	ਕਿੰਨੇ	ਭਾਗ
ਭਾਗਾ	ਤੋੜ	ਮਰੋੜ	ਕਿ	ਦੇ
ਜਿਹੜੀ	ਅਗਲੇ	ਜਾਰੀ	ਦੂਜੇ	ਅੱਗੇ
ਦੱਸਿਆ	ਹੋਣ	ਨਹੀਂ	ਬਣੀ	ਨਾਲੇ
ਜਿਹੜੇ	ਥੋੜ੍ਹੀ	ਥੋੜ੍ਹੇ	ਜਾਂਦੀ	ਜਾਂਦਾ
ਜਾਂਦੇ	ਪਏ	ਪਈ	ਪਿਆ	ਪਿੱਛੇ
ਉਨ੍ਹਾਂ	ਵੱਡੀ	ਸਿਰ	ਉਨਾ	ਸਮੇ
ਵੱਡੇ	ਵੱਡਾ	ਹਿੱਸਾ	ਹਿਸਾ	ਸਮੇਂ
ਸਕਦੇ	ਸ.	ਸ੍ਰੀ	ਮਾਨ	ਸ਼੍ਰੀਮਤੀ

- **Stemming:** Stemming is the process of reducing a word to its root word known as ‘stem’. In clustering of text document in Gurmukhi script, stemming has its own significant role in defining the list of cluster titles and assigning key terms under the relevant cluster title; results to enhancement of an overall performance of clustering process. In Gurmukhi script, there are various grammatical forms of a single root word for instance, a root word ‘ਦੌੜ’, has grammatical word forms ‘ਦੌੜੀ’, ‘ਦੌੜਿਆ’, ‘ਦੌੜਦੀ’, ‘ਦੌੜਦਾ’, ‘ਦੌੜਦੇ’, ‘ਦੌੜੇ’, ‘ਦੌੜੇਗੀ’, ‘ਦੌੜੇਗਾ’, ‘ਦੌੜਨਗੇ’, ‘ਦੌੜਨ’. Using the process of stemming, various grammatical forms of a stem word are reduced to a single stem word and will be considered as similar key terms to represent similarity among the text documents. In the proposed work, an automated stemming is performed [6].
- **Normalization:** In Gurmukhi script, spellings of some words are not standardized and are not represented uniformly. Most o the words in Gurmukhi are written in different manners which cause difficulty to recognize uniformly for e.g. the words “ਵਿੱਤੀ”, “ਫੁੱਟਬਾਲ”, “ਮੁਸ਼ਕਿਲ”, “ਅਰਜ਼ੀ”, “ਦੇਸ਼”, “ਲਾਪ੍ਰਵਾਹੀ” can also be written as “ਵਿਤੀ”, “ਫੁਟਬਾਲ”, “ਮੁਸ਼ਕਲ”, “ਅਰਜੀ”, “ਦੇਸ”, “ਲਾਪਰਵਾਹੀ” respectively. Thus, the text documents consisting same words with variation in spellings do not represent similarity and uniformity of words. In such situation, the text documents consisting same words with non-uniform spellings will create difficulty in placing these text documents in the same cluster. To overcome the problem of non-uniformity of word spellings in Gurmukhi script, the process of normalization is performed [7]. The process of normalization allows representing similarity among the text documents which consist same words with non-uniform of spellings also; and placed such text documents in the same cluster. It has its own significant role in providing uniformity of word spellings which results in high accuracy of clustering process.

B. Processing Phase

In the processing phase, in first step, noun entities from the structured text documents are extracted that will be used further to create a list of cluster titles. In the next step of processing phase, the cluster titles are created by calculating fuzzy term frequency (TF) of the key terms in text documents. In the third step, clusters of similar text documents are created by placing the text documents in the same cluster which consist same terms as cluster titles. In this paper, the processing phase is not elaborated as more concentration is given to the role and importance of pre-processing phase only.

III. EXPERIMENTAL ANALYSIS

To evaluate the role and importance of pre-processing phase of the clustering technique, the proposed algorithm have been implemented using Python 3.4.

A. Data Set

As the domain independent document clustering for Gurmukhi script is first ever work, no standard data-set is available for this work in Gurmukhi script. To analyze the role and importance of pre-processing phase of clustering technique, text documents in Gurmukhi script have been collected from various e-resources and taken as an input to the system.

B. Results and Discussions

To execute the proposed algorithm of document clustering technique for Gurmukhi script and to analyse the role of pre-processing phase, the clustering technique for Gurmukhi script is executed in two rounds. In first round, the clustering technique for Gurmukhi script is executed using all the steps of pre-processing phase. Whereas, in second round, the same clustering technique for Gurmukhi script is executed without using the pre-processing steps. The screen shots of system are shown in Fig. 2. The resultant clusters of clustering technique without using pre-processing steps are shown in Table-2. Whereas, the resultant clusters of clustering technique using pre-processing steps are shown in Table-3.

Table-2: resultant Clusters for Document Clustering Technique without using removal of stop words, stemming and normalization steps of pre-processing phase.

Sr. No.	Cluster Name	Result
1.	ਵਿਦਿਆਰਥੀ (Vidiarthi)	Expected
2.	ਵਿਦਿਆਰਥੀਆਂ (Vidiarthia)	Unexpected
3.	ਲੋਕਤੰਤਰ (Loktantar)	Expected
4.	ਕਿਸਾਨ (Kisan)	Expected
5.	ਕਿਸਾਨਾ (Kisana)	Unexpected
6.	ਦੇਸ਼ (Desh)	Expected
7.	ਦੇਸ਼ਾਂ (Desha)	Unexpected
8.	ਸਕੂਲ (School)	Expected
9.	ਫੁੱਟਬਾਲ (Football)	Expected
10.	ਟੀਮ (Team)	Expected

11.	ਖੇਡ (Khed)	Expected
12.	ਖਿਡਾਰੀ (Khidari)	Expected
13.	ਹਾਕੀ (Hockey)	Expected
14.	ਇੰਗਲੈਡ (England)	Expected
15.	ਸਰਕਾਰ (Sarkar)	Expected
16.	ਸਰਕਾ**ਰਾਂ** (Sarkara)	Unexpected
17.	ਬੈਂਕ (Bank)	Expected
18.	ਬੈਂਕਾਂ (Banka)	Unexpected

Table-3: Clusters for Document Clustering Technique for Gurmukhi Script using removal of stop words, stemming and normalization steps of pre-processing phase.

Sr. No.	Cluster Name	Result
1	ਵਿਦਿਆਰਥੀ (Vidiarthi)	Expected
2	ਲੋਕਤੰਤਰ (Loktantar)	Expected
3	ਕਿਸਾਨ (Kisan)	Expected
4	ਦੇਸ਼ (Desh)	Expected
5	ਸਕੂਲ (School)	Expected
6	ਫੁੱਟਬਾਲ (Football)	Expected
7	ਟੀਮ (Team)	Expected
8	ਖੇਡ (Khed)	Expected
9	ਖਿਡਾਰੀ (Khidari)	Expected
10	ਹਾਕੀ (Hockey)	Expected
11	ਇੰਗਲੈਡ (England)	Expected
12	ਸਰਕਾਰ (Sarkar)	Expected
13	ਬੈਂਕ (Bank)	Expected

As shown in Table-2, the clustering technique without using removal of stop words, stemming and normalization steps of pre-processing phase resulted in creation of some unexpected clusters like ਵਿਦਿਆਰਥੀਆਂ, ਕਿਸਾਨਾ, ਦੇਸ਼ਾਂ, ਸਰਕਾਰਾਂ and ਬੈਂਕਾਂ. Whereas, in Table-3, the clustering technique using removal of stop words, stemming and normalization steps of pre-processing phase resulted in handling the creation of unexpected clusters. Moreover, the list of resultant clusters represent that there is significant role of pre-processing phase in creation of expected clusters. The pre-processing step 'Removal of stop words' reduces the execution time of proposed clustering technique as this step eliminates a huge number of unimportant words from the input text documents. The next pre-processing steps 'Stemming' and 'Normalization' allow the proposed clustering technique to create a list of meaningful cluster titles as the process of 'Stemming' leads the creation of cluster titles using stem words only and the process of 'Normalization' provides high accuracy to place the text documents in most relevant cluster.

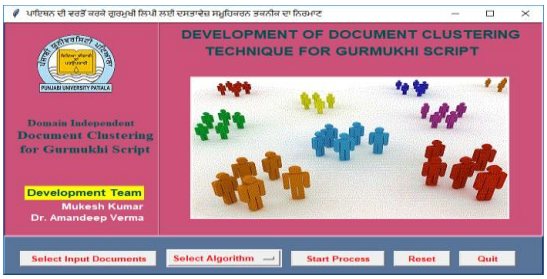

Fig.2(a)- Main Screen of Proposed Clustering Technique Implemented using Python 3.4

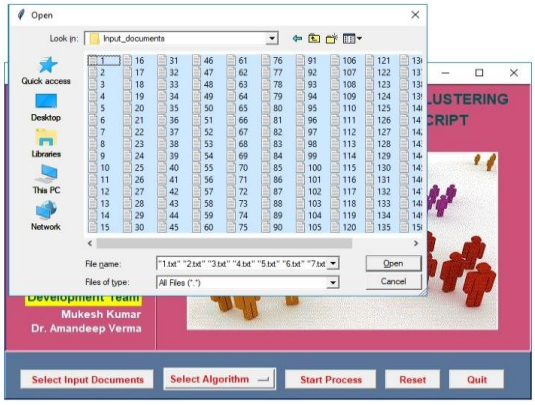

Fig.2(b)- Selection of Input Text Documents in Gurmukhi Script

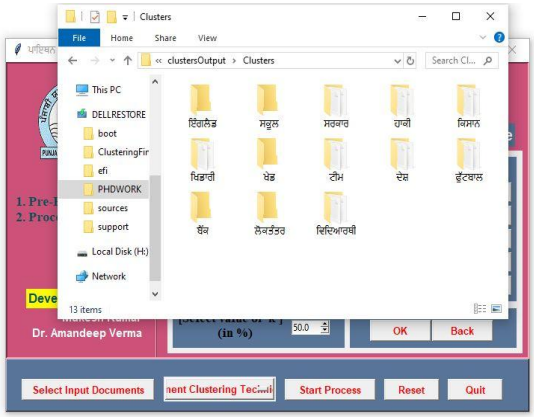

Fig.2(c)- Clusters as output without using pre-processing steps

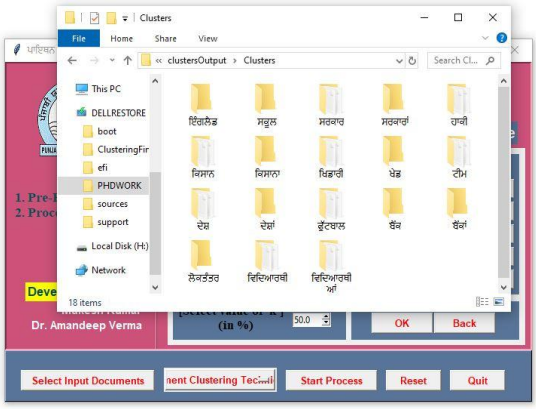

Fig.2(d)- Clusters as output using pre-processing steps

IV. CONCLUSIONS AND FUTURE SCOPE

The proposed paper represents significant role of pre-processing steps namely - 'Removal of stop words', 'Stemming', and 'Normalization'. The experimental results reveal that with the use of pre-processing steps; the clustering technique for Gurmukhi script performs better in terms of creation of meaningful cluster titles as well as in terms of placing text documents in relevant cluster. Development of more efficient pre-processing techniques may lead to further research; as these techniques play significant role in overall performance of clustering technique.

REFERENCES

1. Kishore C, Srinivasulu Asadi and Anusha G, Comparative study of software module clustering algorithms: hill-climbing, IJARCET, 1(3), (2012), 2278–1323.
2. A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Surveys, 31(3), (1999) 1-60
3. N. Dangre, A. Bodke, A. Date, S. Rungra, S.S.Pathak (2016) "System for Marathi News Clustering" ICCC-2016, Elsevier Science Ltd., Procedia Computer Science, Vol. 92, PP.18 – 22.
4. Rasool Azimi, Mohadeseh Ghayekhloo, Mahmoud Ghofrani, Hedieh Sajedi, (2017), " A novel clustering algorithm based on data transformation approaches ", Expert Systems With Applications, Elsevier Science Ltd., Vol. 76, PP. 59–70.
5. Kaur , Kumar, Saini, Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle, ACM (2016).
6. Vishal Gupta , Automatic stemming of words for Punjabi language, Advances in Intelligent Systems and Computing, Springer International Publishing Switzerland, (2014), 73-84.
7. Vishal Gupta , Automatic normalization of Punjabi words, International Journal of Engineering Trends and Technology (IJETT) – 6(7), (2013), 353-357.
8. A.F.Gómez-Skarmeta, M.Delgado, M.A.Vilab, About the use of fuzzy clustering techniques for fuzzy model identification, Vol. 106,(2), (1999), 179-188.
9. G.X. Xu, W. Sun, X.P. Peng (2015), "Clustering Research across Tibetan and Chinese Texts", Vol. 13, No. 3, PP. 162-168.

AUTHORS PROFILE

Mukesh Kumar is Ph.D. scholar and working in Dept. of Comp. Sci., Mata Gujri College (an Autonomous Institution), Fatehgarh Sahib, Punjab, INDIA. He has published a number of books with Kalyani Publishers He has 17 years of teaching experience.

Amandeep Verma is working in Punjabi University Regional Centre for Information Technology and Management, Mohali Punjab, INDIA. He completed Doctorate degree from Punjabi University, Patiala. He has number of research papers published in reputed National and International journals.