Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN: 2521-3520 (online)
Volume 28, Issue 2, 2017
DOI: http://doi.org/10.23851/mjs.v28i2.509

Research Article

Best Approximate of Vector Space Model by Using SVD

Raghad M. Hadi, Soukaena H. Hashem, Abeer T. Maolood
Departement of Computer Science, College of Science, Mustansiriyah University, IRAQ.
*Correspondent Author Email: {Raghad_alrudieny, soukaena. hassen, abeer282003}@yahoo.com

ArticleInfo

Received
10/11/2016

Accepted
21/05/2017

Abstract

A quick growth of internet technology makes it easy to assemble a huge volume of data as text document; e. g., journals, blogs, network pages, articles, email letters. In text mining application, increasing text space of datasets represent excessive task which makes it hard to pre-processing documents in efficient way to prepare it for text mining application like document clustering. The proposed system focuses on pre-processing document and reduction document space technique to prepare it for clustering technique. The mutual method for text mining problematic is vector space model (VSM), each term represent a features. Thus the proposed system create vector-space model by using pre-processing method to reduce of trivial data from dataset. While the hug dimensionality of VSM is resolved by using low-rank SVD. Experiment results show that the proposed system give better document representation results about 10% from previous approach to prepare it for document clustering.

Keywords: High Dimensional Datasets, Dimensionality reduction, SVD, Vector Space Model.

الخلاصة

ان النمو السريع لأجهزة الكمبيوتر والإنترنت يجعل من السهولة تجميع وتوريد كمية كبيرة من المعلومات على شكل النص؛ على سبيل المثال، استعراض، المدونات الالكترونية، وصفحات الويب، مقالات، رسائل البريد الإلكتروني. وغيرها ، ان زيادة نطاق قواعد البيانات النص والأبعاد العالية غير مهمة مما يجعل من الصعب تصنيف الوثائق في مختلف الفئات. لذلك فان النظام المقترح ركزفي معالجة هذه البيانات الكبيرة وتقليص من ابعادها ليتم تهيئتها الى عملية تصنيف البيانات. وذلك من خلال تجهيز تقنية VSM هو نموذج فضاء المتجه ، حيث تعتبر الكلمات هي الخصائص المهمة فيها. وهكذا يهدف النظام المقترح إلى استخدام نموذج فضاء المتجه والتي تعتمد على طريقة المعالجة المسبقة للحد من الخسائر من المعلومات التافهة. يتم حل مشكلة الأبعاد العالية مع تقنية اختيار ميزة من خلال تطبيق رتبة منخفضة لل SVD على القيم VSM. وتشير النتائج أن النظام المقترح اعطى نتائج أفضل بتمثيل البيانات الكبيرة مع تقليص ابعادها بنسبة 10% ليتم ادخالها الى عملية تصنيف البيانات.

Introduction

In the latest years, there has been a growing attention in English Language pre-processing research. English is the intuitive language of residents of more than 380 million. So the big documents datasets convert to term-document matrices which called a Vector Space Model which the term-specific weights in the document vectors are products of local and global parameters. The model is known as term frequency-inverse document frequency model of information (TF-IDF) [1]. In the VSM, a vector is castoff to characterize all piece or paper in a gathering. All section of the direction mirrors key word, or word related through the known paper. The cost given near that module replicates the status of the word in signifying the document. Typically, the cost is a role of the occurrence by the word ensues in the paper or in the paper gathering as a total. To create term-document matrix the collection of document must preprocessing first [2].

Preprocessing phase

The preprocessing holds a practice for increase the set of words to categorize meeting. The drive for pre-processing stage is to clip all words from the datasets that have cheap material that container perhaps touch an excellence of the collection reports. While Singular value decomposing skilled the commerce by loud information, the popular of open immaterial ideas would be connected to empty common relations [3].

The chief procedure of preprocessing phase exists through eliminating stop words. The stop words are common words that transmit not at all evidence then empty after castoff as a hunt relations (i. e., pronouns, prepositions, conjunctions etc) [4].

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licenses.

The additional procedure is to stem a term. Morphological variations of arguments typically need related imports. Unknown these arguments are conflated into only word, the presentation of document reclamation can be enhanced. By consuming the procedure of stemming in a method that verses are stemmed hooked on a root formula through eliminating their affixes [5].

Dimension Reduction techniques

- **Singular Value Decomposition (SVD)**

Is built by deduction since linear algebra which states a four-sided m-by-n matrix. A container by damaged into the creation of three array - an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V. SVD of an n x n matrix A is defined by the operation:

$A=U_*S_*V^T$

$$\begin{bmatrix} A_{1,1} & \dots. & A_{1,n} \\ A_{2,1} & \dots. & A_{2,1} \\ A_{n,1} & \dots. & A_{n,n} \end{bmatrix}$$

$$= \begin{bmatrix} U_{1,1} & \dots & U_{1,n} \\ U_{2,1} & \dots & U_{2,1} \\ U_{n,1} & \dots & U_{n,n} \end{bmatrix} . \begin{bmatrix} S_{1,1} & 0 & 0 \\ 0 & S_{2,1} & 0 \\ 0 & 0 & S_{n,n} \end{bmatrix}$$

$$. \begin{bmatrix} V_{1,1} & \dots & V_{1,n} \\ V_{2,1} & \dots & V_{2,1} \\ V_{n,1} & \dots & V_{n,n} \end{bmatrix}^T \dots \mathbf{(1)}$$

The matrix is formerly rotten via singular value decomposition into: word path matrix involving in the left singular vectors, the document vector matrix involving in the right singular vectors and the diagonal matrix involving of singular values [117].

The following steps show how SVD applied on a matrix A: [117]

$$A=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \longrightarrow A^T=\begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}$$

$$AA^T=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}\begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}=\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

$$\downarrow$$

$$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}\begin{bmatrix} x1 \\ x2 \end{bmatrix}=\lambda \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

$$11x1+ x2 = \lambda x_1$$
$$x1+ 11x2= \lambda x_2$$

$$(11-\lambda)x1 +x2 = 0$$
$$x1+ (11-\lambda)x2 = 0$$

$$\begin{vmatrix} (11-\lambda) & 1 \\ 1 & (11-\lambda) \end{vmatrix} =0$$

$$(11- \lambda)(11- \lambda)-1.1=0$$
$$(\lambda-10)\ (\lambda-12)=0$$
$$\lambda =10,\ \lambda=12$$

$$(11-10)x1+x2=0$$
$$x1=-x2$$

$$(11-12)x1+x2=0$$
$$x1=x2$$

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$U=\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$A^TA=\begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}=\begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}\begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \lambda \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix}$$

$$10x1+2x3= \lambda x1$$
$$10x2+4x3= \lambda x2$$
$$2x1+4x2+2x3= \lambda x2$$

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

$$V=\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

$$V^T=\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

$$S=\begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}$$

$$A=U*S*V^T$$

$$=\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

Related works

The approach presented in [7] is to find the damage to the use of electronic documents over databases. The solution is by text illustration which critical stage for writing pre-treating. Text (article) is a pool of arguments, in [7] Research was recognized in numerous steps. Text assembly, Format cancelling, Data pre-processing on numerous levels, with subsection serial identification. With a stretch sequence identification. By stop words subtraction and section order documentation with stop words elimination then a judgment order identification. In [8] the paper discussed about the text mining and its preprocessing techniques, discuss the three key steps of preprocessing namely, stop words removal, stemming and TF/IDF algorithms. In [9] their methodology to use an actual Preprocessing stages to protect both galaxy then while supplies through consuming developed stemming algorithm. Stemming algorithms are castoff near alter the arguments in editions into their correct origin formula. In [10] Mining text document from a preprocessed stage is calm as relate to natural languages documents. So, preprocessing phase it is a significant process in text mining application. This paper talk about shrink the dimensionally of the words space, different procedures such as cleaning (filtering) and stemming are practical. Filtering methods eliminate those words from the regular of wholly words, which do not offer related evidence; stop word filtering is a typical filtering manner.

Proposal of Preprocessing and reduction techniques

- **Module1: Preprocessing**

The proposed system selects a domain from Reuters 21578 datasets. collect whole documents from datasets by using Body based feature: All body-based features existing in the body of Reuter's document that includes: (body-keyword), (<body >), (body-java script), and etc. after these body the content of document begin, each body document in datasets was represented using the bag-of-words approach, also these representation known as Vector space model (VSM): it includes the words as column and the documents as rows in VSM matrix. The proposed system tokenize the file content into individual word as shows in the Figure 1, then removed stop words. In order reduces the dimensionality of TF-IDF matrix (VSM representation). In the stop words removal function the proposal system use the classic method which it is traditional and simple method based on removing stop words by compared the words of the text with in the words store in list so if there are any match the word is remove from text, then the proposed system apply porter stemming algorithm with enhancement on its rules, at each step, a certain suffix is deleted by uses of set rules to decrease amount of verses, to must accurately similar stems, and to protect recollection space and period. The proposed system used Porters algorithm and table look up approach by having two dictionaries, one for various irregular English words, and another for various suffixes. To applied the following:

Root = past simple or past participle.

Suffixed = root + suffix.

As the result of preprocessing phase which it produce the TF-IDF matrix with hug dimension, so the second phase is dimensionality reduction techniques by using SVD.

- **Module2 :Dimensionality Reduction Techniques using SVD**

The proposed system decomposition the Term Frequency-Inverse Document Frequency matrix (TF-IDF) matrix by using Singular Value Decomposition (SVD) in to three matrixes $\mathbf{USV}^{T}$, then find k greatest chief scopes (through the top singular values in S matrix) is nominated.completely additional features stay absent. The summary matrix perfectly denotes the significant and dependable patterns underlying the data in TF-IDF matrix. The proposed system dropping the rank of the TF-IDF matrix is incomes of eliminating unimportant info or clatter from the datasets it embodies.

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licenses.

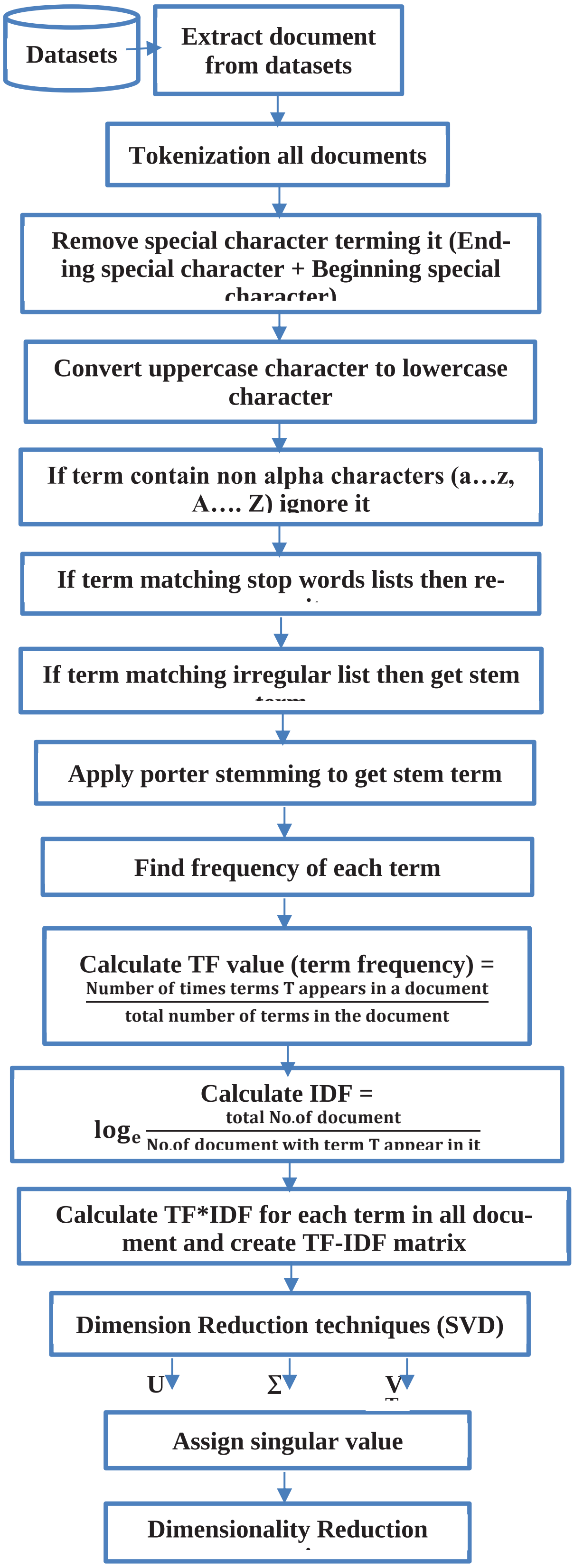

Datasets
Extract document from datasets
Tokenization all documents
Remove special character terming it (Ending special character + Beginning special character)
Convert uppercase character to lowercase character
If term contain non alpha characters (a…z, A…. Z) ignore it
If term matching stop words lists then re-
If term matching irregular list then get stem
Apply porter stemming to get stem term
Find frequency of each term
Calculate TF value (term frequency) = Number of times terms T appears in a document / total number of terms in the document
Calculate IDF = log_e total No.of document / No.of document with term T appear in it
Calculate TF*IDF for each term in all document and create TF-IDF matrix
Dimension Reduction techniques (SVD)
U
Σ
V
Assign singular value
Dimensionality Reduction

Figure 1: The system Architecture

The proposed system is summarized using the following algorithm:

Algorithm 1: Best approximate vector space model algorithm
Input: The documents datasets

Step1: Input datasets, a document by term matrix is created using only term counts. The proposed system set the following:

- The stop words list and the ignore characters are specified.
- The short words are also declared (the proposed system propose to exclude words less than 4 characters).
- Apply porter stemming enhancement and finally, create TF*IDF matrix.

Step 2: Execute the Singular value decomposing on the TF-IDF matrix to get matrix Σ, the right matrix V^T, and matrix U.

Step 3: Select the K-th rank approximation through save the chief k singular values from matrix Σ and set the residual singular values to null, The proposed system propose to investigate different k values around k>30.

Step 4: Choice the top k'th from matrix VT: set $\mathbf{V^T}_{\mathbf{k}}$ equal to the first k rows of $\mathbf{V^T}$ ($\boldsymbol{TF - IDF_{k+n}}$) matrix.

Step 5: Select only the top k singular values: set Σ_k equal to the first k rows and columns of Σ, ($\mathrm{TF - IDF}_{k+k}$) matrix, corresponding to the selected k singular values of TF-IDF matrix.

Step 6: Select only the top k left singular vectors: set $\mathbf{U_k}$ equal to the first k columns of U ($\boldsymbol{TF - IDF_{m+k}}$) matrix.

Step 7: proposed system computed low- rank approximation is then $\mathrm{TF - IDF}_k = \mathrm{U}_k * \Sigma_k * \mathrm{V}_k{}^T$

Step 8: If k ranges the required number then end the process; else, addition k value by one, go to Step 4.
End.

The Experimental Results

The proposed system presents the experimental results. The stop words list is specified as: {a, an, the, all, am, able, and, after, also, anybody, among, yet, yes….. ect}. The ignore non alpha characters list is: {~, `, !, @,, $, %, , &, *, (,), -, _, +, =, ,, {, }, [,], |, \, /, :, ;, 0,1,2,3,4,5,6,7,8,9}. All words that are less than 4 characters in lengths are ignored. The results are shown by applied tokenization with vectors (with pre-processing) on given input documents. The results created by testing the proposed system on

925 input documents, and whole number of tokens made in all effort documents after treating are (13195). Lacking tokenization treating to huge number of tokens, which is hard to supply, and time expended in complete tokenization procedure is right relative to show degree of an information retrieval system, as it acutely moves the indexing and storing features. The proposed system extraction the documents from Reuters 21578 datasets as shown Table 1, and finally the proposal system calculates TF-IDF value for each term in datasets, a small example from huge TF-IDF matrix shown in Table 2.

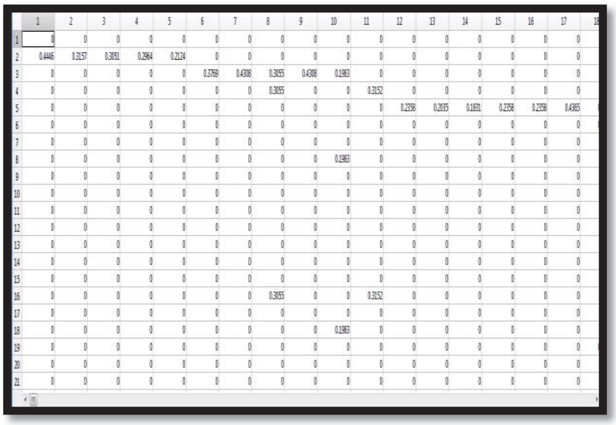
Table 1: Datasets Extraction

Do. Id	Document contents
1	Showers continued throughout the week in behin coca zone alleviating drought since early January improving prospects coming tempora
2	Standard oil co and bp north America said they plan form venture manage borrowing investment activities both companies north
3	Texa commerce Bancshares incs texa commerce bank Houston said filed application with comptroller currency
4	Bankamerice corp is not under pressure quickly proposed equity offering would well delay because stocks recent poor
5	The u. s. agriculture department reported farmer-owned reserve national five day average price through February follows dlrs/ bu
6	Argentine grain board figures show crop registrations grains oilseeds their products February thousands tonnes showing
7	Lion inns limited partnership said filed registraction statement with securities exchange commission covering proposed

The role of the preprocessing is to prepare the datasets as shown in table1 for next proposed system stage. This is basically to reduce the noise from the dataset and keep only the desired information represent by document body.

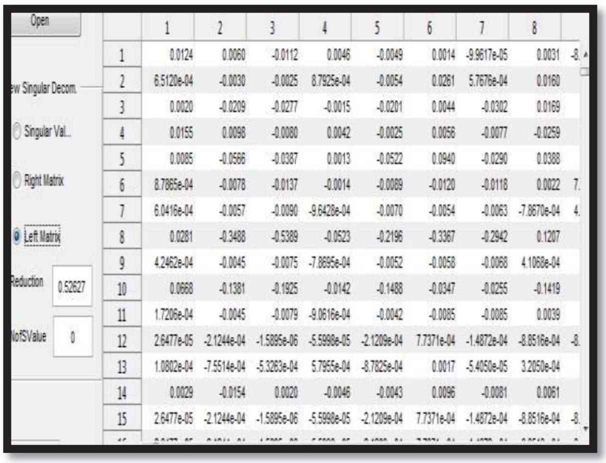
Table 2: Sample of TF-IDF value calculation

Term	TF value	IDF value	TF-IDF value
week	0.0108	4.3027	0.0464
behia	0.0144	8.0163	0.1153
cocoa	0.0216	7.3232	0.1581
come	0.0072	6.9177	0.0498
tempora	0.0072	8.0163	0.0577
have	0.0072	3.7536	0.0270
commissari	0.0180	8.0163	0.1442
said	0.0180	1.7174	0.0309
Period	0.0072	5.9369	0.0427
year	0.0072	2.9226	0.0210
arrive	0.0072	8.0163	0.0577
februari	0.0108	4.8383	0.0522
bag	0.0180	6.9177	0.1244
kilo	0.0072	6.9177	0.0498
total	0.0108	4.7582	0.0513
against	0.0108	5.1831	0.0559
consign	0.0072	8.0163	0.0577
still	0.0108	6.4069	0.0691
crop	0.0180	6.6300	0.1192
export	0.0072	4.3528	0.0313
dlr	0.0504	2.5191	0.1269
port	0.0108	6.4069	0.0691
open	0.0072	6.2246	0.0448
north	0.0476	6.6300	0.3157



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0.4446	0.3157	0.3051	0.2964	0.2124	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0.3769	0.4308	0.3055	0.4308	0.1983	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0.3055	0	0	0.3152	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0.2358	0.2035	0.1831	0.2358	0.2358	0.4365
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0.1983	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0.3055	0	0	0.3152	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0.1983	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: sample of TF-IDF matrix



○ Singular Val...
○ Right Matrix
● Left Matrix
Reduction 0.52627
NofSValue 0

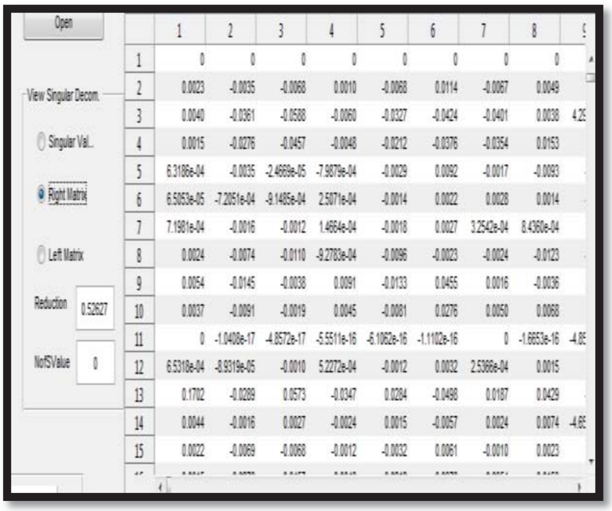
	1	2	3	4	5	6	7	8
1	0.0124	0.0060	-0.0112	0.0046	-0.0049	0.0014	-9.9617e-05	0.0031
2	6.5120e-04	-0.0030	-0.0025	8.7925e-04	-0.0054	0.0261	5.7676e-04	0.0160
3	0.0020	-0.0209	-0.0277	-0.0015	-0.0201	0.0044	-0.0302	0.0169
4	0.0155	0.0098	-0.0080	0.0042	-0.0025	0.0056	-0.0077	-0.0259
5	0.0085	-0.0566	-0.0387	0.0013	-0.0522	0.0940	-0.0290	0.0388
6	8.7865e-04	-0.0078	-0.0137	-0.0014	-0.0089	-0.0120	-0.0118	0.0022
7	6.0416e-04	-0.0057	-0.0090	-9.6428e-04	-0.0070	-0.0054	-0.0063	-7.8670e-04
8	0.0281	-0.3488	-0.5389	-0.0523	-0.2196	-0.3367	-0.2942	0.1207
9	4.2462e-04	-0.0045	-0.0075	-7.8695e-04	-0.0052	-0.0058	-0.0068	4.1068e-04
10	0.0668	-0.1381	-0.1925	-0.0142	-0.1488	-0.0347	-0.0255	-0.1419
11	1.7206e-04	-0.0045	-0.0079	-9.0616e-04	-0.0042	-0.0085	-0.0085	0.0039
12	2.6477e-05	-2.1244e-04	-1.5895e-06	-5.5998e-05	-2.1209e-04	7.7371e-04	-1.4872e-04	-8.8516e-04
13	1.0802e-04	-7.5514e-04	-5.3263e-04	5.7955e-04	-8.7825e-04	0.0017	-5.4050e-05	3.2050e-04
14	0.0029	-0.0154	0.0020	-0.0046	-0.0043	0.0096	-0.0081	0.0061
15	2.6477e-05	-2.1244e-04	-1.5895e-06	-5.5998e-05	-2.1209e-04	7.7371e-04	-1.4872e-04	-8.8516e-04

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licenses.

Figure 3: sample of left matrix U after applying SVD (TF-IDF matrix)

The first sample of calculate TF-IDF value, these coefficients give the best energy to reconstruct the original TF-IDF value as shows in table 2. In the proposed system, by testing the TF-IDF value it is found that the best value it have large number of TF-IDF coefficients to represent the subword as feature vector. In order to get these features and passed it to next stage in proposed system.

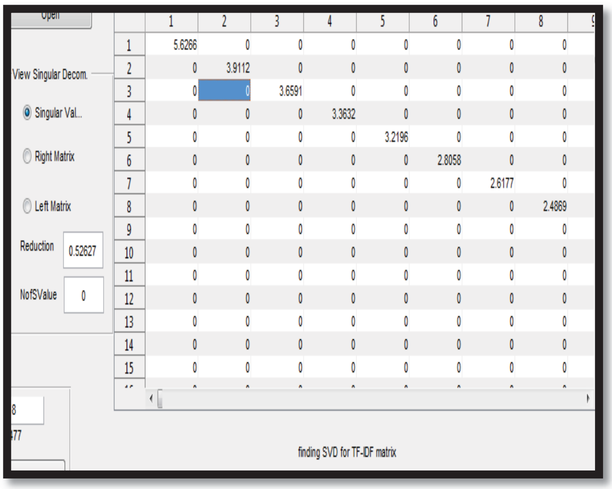
Figures 2 and 3 show the sample matrix of TF-IDF matrix in the proposed system and Figure 4,5,6,7 explain the procedure of the proposed system through applying r-rank SVD.



Reduction: 0.52627
NofSValue: 0

	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0.0023	-0.0035	-0.0068	0.0010	-0.0068	0.0114	-0.0067	0.0049
3	0.0040	-0.0361	-0.0588	-0.0060	-0.0327	-0.0424	-0.0401	0.0038
4	0.0015	-0.0276	-0.0457	-0.0048	-0.0212	-0.0376	-0.0354	0.0153
5	6.3186e-04	-0.0035	-2.4669e-05	-7.9879e-04	-0.0029	0.0092	-0.0017	-0.0093
6	6.5053e-05	-7.2051e-04	-9.1485e-04	2.5071e-04	-0.0014	0.0022	0.0028	0.0014
7	7.1981e-04	-0.0016	-0.0012	1.4664e-04	-0.0018	0.0027	3.2542e-04	8.4360e-04
8	0.0024	-0.0074	-0.0110	-9.2783e-04	-0.0096	-0.0023	-0.0024	-0.0123
9	0.0054	-0.0145	-0.0038	0.0091	-0.0133	0.0455	0.0016	-0.0036
10	0.0037	-0.0091	-0.0019	0.0045	-0.0081	0.0276	0.0050	0.0068
11	0	-1.0408e-17	-4.8572e-17	-5.5511e-16	-6.1062e-16	-1.1102e-16	0	-1.6653e-16
12	6.5318e-04	-8.9319e-05	-0.0010	5.2272e-04	-0.0012	0.0032	2.5366e-04	0.0015
13	0.1702	-0.0289	0.0573	-0.0347	0.0284	-0.0498	0.0187	0.0429
14	0.0044	-0.0016	0.0027	-0.0024	0.0015	-0.0057	0.0024	0.0074
15	0.0022	-0.0069	-0.0068	-0.0012	-0.0032	0.0061	-0.0010	0.0023

Figure 4: Sample of Right matrix $\mathbf{V}^{\mathbf{T}}$ after applying SVD (TF-IDF matrix)

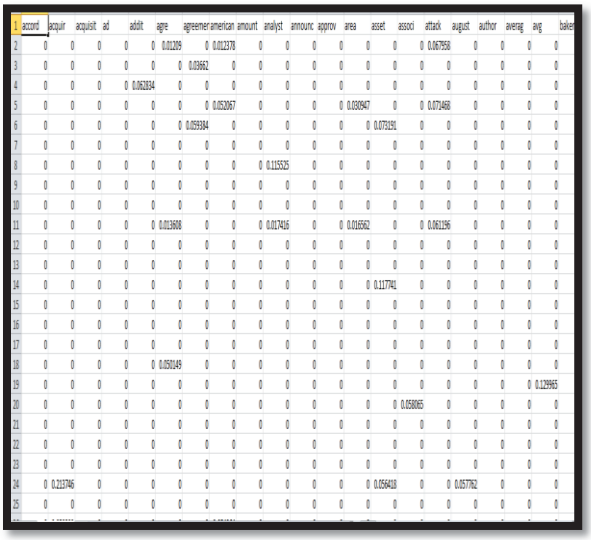


Reduction: 0.52627
NofSValue: 0

	1	2	3	4	5	6	7	8
1	5.6266	0	0	0	0	0	0	0
2	0	3.9112	0	0	0	0	0	0
3	0	0	3.6591	0	0	0	0	0
4	0	0	0	3.3632	0	0	0	0
5	0	0	0	0	3.2196	0	0	0
6	0	0	0	0	0	2.8058	0	0
7	0	0	0	0	0	0	2.6177	0
8	0	0	0	0	0	0	0	2.4869
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0

finding SVD for TF-IDF matrix

Figure 5: Sample of singular value matrix Σ after applying SVD (TF-IDF matrix)



1	accord	acquir	acquisit	ad	addit	agre	agreemer	american	amount	analyst	announc	approv	area	asset	associ	attack	august	author	averag	avg	baker
2	0	0	0	0	0	0.01209	0	0.012378	0	0	0	0	0	0	0	0.067958	0	0	0	0	
3	0	0	0	0	0	0	0.03662	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0.062834	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0.052067	0	0	0	0	0.030947	0	0	0.071468	0	0	0	0	
6	0	0	0	0	0	0	0.059384	0	0	0	0	0	0	0.073191	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	0.115525	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0.013608	0	0	0	0.017416	0	0	0.016562	0	0	0.061196	0	0	0	0	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0.117741	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0.050149	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.129965	
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.058065	0	0	0	0	0	
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	0	0.213746	0	0	0	0	0	0	0	0	0	0	0	0.056418	0	0	0.057762	0	0	0	
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

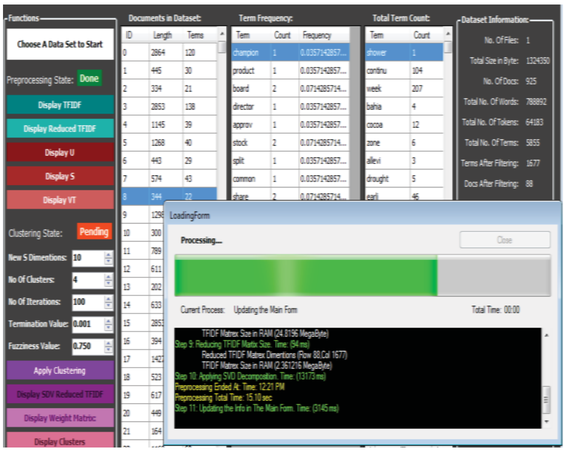
Figure 6: Final dimension reduction matrix ($\mathbf{U_K}\ \boldsymbol{\Sigma_K}\ \boldsymbol{V_K}^{\boldsymbol{T}}$)after input threshold value =0.52627

Results Discussion

The results obtained from applying the proposed algorithms and the effect of the proposed algorithm of the system are presented in tables. The test setup and the experimental results obtained for the preprocessing the datasets to obtain features represented by TF-IDF value. The proposed system is implemented in Visual Studio 2013 programming languages.

The experiments were performed on an Intel Core i7, 64 bit Operating System, 2.50 GHz processor and 6GB RAM. In order to evaluate the proposed system, number of metrics is obtained by applying singular value decomposition.

The results of applying the proposed system are shown in Figure 7. The proposed preprocessing algorithm is applied for document sets from the Reuters 21578 datasets.



Preprocessing State: Done

Clustering State: Pending

New S Dimentions: 10

No Of Clusters: 4

No Of Iterations: 100

Termination Value: 0.001

Fuzziness Value: 0.750

Documents in Dataset:

ID	Length	Terms
0	2864	120
1	445	30
2	334	21
3	2853	138
4	1145	39
5	1268	40
6	443	29
7	574	43
8	344	22
9	1298	
10	300	
11	789	
12	611	
13	202	
14	633	
15	2853	
16	394	
17	1427	
18	523	
19	617	
20	449	
21	164	

Term Frequency:

Term	Count	Frequency
champion	1	0.0357142857...
product	1	0.0357142857...
board	2	0.0714285714...
director	1	0.0357142857...
approv	1	0.0357142857...
stock	2	0.0714285714...
split	1	0.0357142857...
common	1	0.0357142857...
share	2	0.0714285714...

Total Term Count:

Term	Count
shower	1
continu	104
week	207
bahia	4
cocoa	12
zone	6
allevi	3
drought	5
earli	46

Dataset Information:

No. Of Files: 1
Total Size in Byte: 1324350
No. Of Docs: 925
Total No. Of Words: 788892
Total No. Of Tokens: 64183
Total No. Of Terms: 5855
Terms After Filtering: 1677
Docs After Filtering: 88

Current Process: Updating the Main Form
Total Time: 00:00

```
          TFIDF Matrex Size in RAM (24.8196 MegaByte)
Step 9: Reducing TFIDF Matix Size. Time: (94 ms)
          Reduced TFIDF Matrex Dimentions (Row 88,Col 1677)
          TFIDF Matrex Size in RAM (2.361216 MegaByte)
Step 10: Applying SVD Decomposition. Time: (13173 ms)
Preprocessing Ended At: Time: 12:21 PM
Preprocessing Total Time: 15.10 sec
Step 11: Updating the Info in The Main Form. Time: (3145 ms)
```

Figure 7: The proposed system

Conclusions

The proposed system introduces an enhancement to the pre-processing information retrieval system by using an existing open source Reuters 21578 datasets; this step affects the outcomes of any IR system. The lack of standard porter stemming algorithm and preprocessing steps such as, stop-word removal and stemming also motivates us to bring out these instruments.

The proposed system GUI has many options including reading dataset files, display output in tables, and produce statistics about preprocessing steps. And it is careful as a chief step through a Standard English language preprocessing systems and then applies low-rank SVD, Singular Value Decomposition (SVD) is a dimensionality technique that can be used to create lower-dimensional embedding from a full term-document matrix. The compared of porter algorithm enhancement in the proposed system with the popular porter stemming algorithms that failed with 22 English words and we understood that the proposed enhanced stemming algorithm output the best result that consider English words, this also should be regarded as a standard feature for any upcoming English stemming algorithm.

References

[1] H. Froud, A. Lachkar and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering," *Journal of university sidi mohamed ben abdellah, Morocco*, 2012.

[2] N. S. Pathak, P. P. Rajurkar and A. G. Bhor, "effective approach towards exporter IR system through comparision of various preprocessing techniques," *International conference on advances in engineering science and management*, vol.8, 2015.

[3] N. A. Samat, M. A. Azmi and M. T. Abdullah, "Malay documents clustering algorithm based on singular value decomposition," *Faculty of computer science and information technology, university of Putra Malaysia*, vol.3, 2016.

[4] M. W. Berry, Z. Drma and E. R. Jessuo, "Matrices vector spaces and information retrieval," *website www. amazon.com*, 2012.

[5] S. Lappin and C. Fox, "Vector space models of lexical meaning," *Stephen clark university of cambridge computer laboratory*, vol.25th, 2014.

[6] S. Shama and L. Padmalatha, "Performance comarison of image fusion using singular value decomposition," *International journal of innovative research in science, Engineering and technology*, vol.4, no.9, 2015.

[7] D. Munkova, M. Munk and M. Vozar, "Data pre processing evalution for text mining: Transaction/Sequence Model," *international conference on computational Science*, 2013.

[8] S. Vijayarani and J. Ilamathi, "Preprocessing Techniques for text mining an overview," *International journal of computer science and communication networks*, vol.5, 2015.

[9] C. Ramasubramanian, R. Ramya and V. Tamilnadu, "Effective preprocessing activities in text mining using improved porters stemming algorithm," *international journal of adanced research in computer and communication engineering*, vol.2, no.12, 2013.

[10] N. P. Katariya, S. Chaudhari and N. P. Katariya, "Text preprocessing for text mining using side information," *international journal of computer science and mobile application*, vol.3, no.1, 2015.

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licenses.