Variance-based Features for Keyword Extraction in Persian and English Text Documents

Hadi Veisi*
Faculty of New Sciences and Technologies (FNST), University of Tehran, Tehran, Iran
h.veisi@ut.ac.ir
orcid.org/0000-0003-2372-7969

Niloofar Aflaki
Kish International Campus, University of Tehran, Kish, Iran
n.aflaki@ut.ac.ir
https://orcid.org/0000-0002-4643-7692

Pouyan Parsafard
Kish International Campus, University of Tehran, Kish, Iran
poovanparsafard@ut.ac.ir

Abstract
This paper addresses automatic keyword extraction in Persian and English text documents. Generally, for keyword extraction in a text, a weight is assigned to each token and words having higher weights are selected as the keywords. We have proposed four methods for weighting the words and have compared these methods with five previous weighting techniques. The previous methods used in this paper are term frequency (TF), term frequency inverse document frequency (TF-IDF), variance, discriminative feature selection (DFS), and document length normalization based on unit words (LNU). The proposed weighting methods are based on using variance features and include variance to TF-IDF ratio, variance to TF ratio, the intersection of TF and variance, and the intersection of variance and IDF.

For evaluation, the documents are clustered using the extracted keywords as feature vectors, and K-means, expectation maximization (EM), and Ward hierarchical clustering methods. The entropy of the clusters and pre-defined classes of the documents are used as the evaluation metric. For the evaluations, we have collected and labeled Persian documents. Results show that our proposed weighting method, variance to TF ratio, has the best performance for Persian. Also, the best entropy is resulted by variance to TD-IDF ratio for English.

Keywords: Keyword Extraction; Term Frequency; Variance; Clustering; Persian Text Processing.

1- Introduction
Text mining as a subfield of data mining focuses on the extraction useful data and knowledge from textual data. A means of getting the knowledge about a document is the accessing of its keywords. Keywords are important elements in searching for and getting information. They can be considered as a collection of words that describe the document during the search and information retrieval operations. In other words, each important word that describes the content

---

1 Present Address: Geoinformatics Collaboratory and School of Natural and Computational Sciences, Massey University, Auckland, New Zealand (n.aflaki@massey.ac.nz)
of a document is called a keyword. These words are often used to define and display the information retrieval systems, because they are short and likely to remain in the mind. Consequently, keywords of a text document which are the most relevant words concerning the substance of the documents, can be great candidates to be chosen as features in document processing tasks such as classification and clustering [1,2].

Keywords can be assigned to or extracted from a document [3]. In keyword assignment, a set of conceivable keywords are chosen from a controlled vocabulary of words, while keyword extraction distinguishes the pertinent words accessible in the examined document [4]. Also, keywords can be categorised into two groups: functional and informative [5]. Informative keywords have a strong connection with the contents of the text and introduce its main content. For example, in case of sport news about the Barcelona football team, football can be defined as an informative keyword in this news item. On the other hand, functional keywords such as prefixes and conjunctions have a less degree of connection with the text [6].

Although there are several methods for keyword extraction [4, 7-12], and also methods are used for various languages such as English [13] and Persian [14], finding the appropriate keywords of document is still a challenging problem. The main focus of the methods is on the assigning reliable weights to the keyword candidates and then selecting the best ones. In this paper, we have proposed new methods for weighting the words and evaluated out methods in Persian and English languages.

In this paper, Section 2 provides a summary of the related works. The third section refers to a keyword extraction framework that includes the known weighting methods like variance, Discriminative Feature Selection (DFS), Document Length Normalization based on Unit words (LNU), and clustering methods. The fourth section presents our proposed weighting methods, and the fifth contains experiments and evaluations. The discussions and comparative results are also given in this section. At the end, the conclusion and future works will feature in Section 6.

### 2- Related Works

There are several methods for keyword extraction from a text which can be classified into five general approaches:

**2-1 Statistical methods based on term frequency analysis**

The first commonly used statistical method is Term Frequency (TF), which calculates the occurrence of a word in a document. Another common method is the term frequency inverse document frequency (TF-IDF) that measures the occurrence of a word in a document and all the documents. These approaches, which have been used in this paper, are reviewed in Section 3-1. Statistical methods are well-known and are reliable for keyword extraction because when a word occurs in a document several times, we can consider it to be a keyword candidate. The statistical keyword extraction techniques can be domain-independent and do not require training data [4].

In [8], TF, TF-IDF and variance are used as their weighting methods in document clustering using K-means. They calculate the entropy value to compare the quality of
clustered documents with their prior classes. In addition to TF and TF-IDF, there are other statistical methods such as variance and word co-occurrence [13].

2-2 Linguistic methods based on language parsing
The linguistic features of words are used in linguistic methods. Synthetic analysis [9] and lexical analysis [10] are examples of these methods that express the semantic content of a part of the text. Lexical chains, too, are used in text summarization. The recall of synthetic analysis is 66% and precision 64% for lexical chains.

2-3 Machine learning methods such as supervised learning
Keyword extraction is done through a supervised action in this method. In this case, models are trained and keywords extracted using these trained models. Examples of this approach are the Naïve Bayes [15] and SVM methods [11]. An ongoing report on keyword extraction presented a model-based on fractal patterns [16]. The outcomes demonstrate that the most relevant terms about the topic of the text document have fractal dimensions not quite the same as one, though insignificant terms have a fractal dimension value of one.

2-4 Conceptual methods based on the use of knowledge data base to interpret the meaning and concept
This method uses semantic analysis and the dictionary [17]. In semantic analysis and dictionary that is a method of keyword extraction, documents are divided into a set of sentences. Later, a model finds the best concept of each term. Then, Lesk algorithm calculates the word-to-word similarity for each pair. So, it computes the best concept of each two words in a pair with their similarity score. Each pair will have clusters based on its similarity score. This is followed by the calculation of the average similarity score for every cluster. In the next step, inverse similarity score is computed in order to evaluate the importance of a word in a cluster’s similarity score. The average of term similarity weights is calculated after removing a term. This step is repeated for all clusters to determine clusters coherence.

2-5 Combination of the above approaches
These techniques are the combination of some or all of the approaches to create a heuristic method, for example, by using html tags [18]. The heuristic method uses phrase rates, which are an interactive aid for keyword extraction for human classifiers in Informine projects (http://infomine.ucr.edu). This method is a heuristic key phrase extraction for web pages that require no training. Instead, it is based on the hypothesis that most of the well-written web pages offer key phrases based on their inner structure.
3- Keyword Extraction Framework
To extract keywords, after the stop word removal to do away with all unimportant and meaningless words, the weights are calculated for other terms, using the known weighting methods mentioned in Section 3-1 and the proposed methods in Section 4. The terms with higher weights are selected after the extraction of the document features. Then, the quality of the selected terms is evaluated, using clustering techniques given in Section 3-2. To do the clustering, document-feature matrices are formed for the selected terms. Three types of such matrix are used in this paper.

1- Frequency matrix: A two-dimensional matrix in which documents are rows and terms are columns and the frequency of every term in every document forms the values of the related cell.

2- Normal matrix: This matrix is similar to the frequency matrix in which the values of each cell are the normalized frequency of the related term. The normalization is done by dividing the term frequency in each document into the maximum term frequency in that document.

3- Boolean matrix: As the name suggests, the values of each cell in this case is 1 or 0, denoting the presence or absence of that term in the document.
The matrices are formed for a number of features (i.e., terms) and are given to Weka tool [19] for clustering. For clustering, three methods such as K-means, EM, and Ward, which are reviewed in Section 3-2, are used. After clustering, the entropy matrix is prepared to evaluate the results. The entropy value is between 0 and 1, with lower values (i.e., nearer to 0) being appropriate. The flowchart of the keyword extraction steps is shown in Fig1; each of these steps is described in the following sections.

3-1 Weighting Methods
Term extraction is the first step to extract keywords from a text. The next step is the selection of features for assigning weights. Weighting of features can be done in various ways, with different weighting approaches yielding different values. In this section, we describe the current weighting methods such as Variance, LNU, and DFS.

3-1-1 Term Frequency (TF) and TF-Inverse Document Frequency (TF-IDF)
The first statistical method to extract keywords from a document is TF, which represents the occurrence of a term in a document. If a term occurs more than other terms, it is most likely to be a keyword. Another recent method for keyword extraction is TF-IDF. In this approach, the frequency of usage of a term is measured in the document concerned and all other documents. The term that occurs in most of the documents has a low TF-IDF value because of its lesser power of discrimination. Therefore, if a term occurs in fewer documents, its TF-IDF value would be high, indicating it can identify documents. TF and TF-IDF formulas are shown in Equations 1, 2 and 3, respectively.
\( TF(Term_i, Doc_j) = \frac{\text{Number of times } Term_i \text{ appears in document } Doc_j}{\text{Total number of terms in document } Doc_j} \)  

(1)

\[ IDF(Term_i) = \log \left( \frac{\text{Total number of documents}}{\text{Number of documents with } Term_i \text{ in it}} \right) \]

(2)

\[ TF-IDF(Term_i, Doc_j) = TF(Term_i, Doc_j) \times IDF(Term_i) \]

(3)

### 3-1-2 Variance

This method calculates the usage variance of each word in a document. Variance is equal to the mean square difference from the average and is calculated by equation 4:

\[ Variance(Term_j) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \]

(4)

In which \( N \) is the total number of documents containing \( Term_j \), \( \mu \) is the term’s frequency average in all documents, and \( x_i \) is the occurrence of \( Term_i \) in the document.

### 3-1-3 Document Length Normalization based on Unit words (LNU)

The LNU weighting method is based on the frequency factor (known as the L factor) and the normalization unit (U factor), and word weighting is based on unique words in documents. This relation is expressed in the following equation (5) [20].

\[ LNU(Term_i, Doc_j) = \frac{1 + \log(TF(Term_i, Doc_j))}{1 + \log(\text{average}(TF(:, Doc_j))))} \times \frac{1}{(1 - \text{slope}) \times \text{pivot} + (\text{slope} \times \text{num unique terms})} \]

(5)

In which \( \text{average}(TF(:, Doc_j)) \) is the average of term frequency for all words in document \( Doc_j \), \( \text{slope} \) is the experimental slope of the curve and often considered as a constant value 0.25. \( \text{Pivot} \) is the ratio of the total number of unique words in all documents to the total number of documents, and \( \text{num unique terms} \) is the number of unique terms in document \( Doc_j \).

### 3-1-4 Discriminative Feature Selection (DFS)

Discriminative Feature Selection (DFS) is another feature extraction method proposed for document classification [21]. Here, discriminative features are those that have a higher weight in their categories in comparison to the others. For extracting these features, the parameters of Table 1 are first calculated.
The DFS will select features having the highest average term frequency in the mentioned category, picking up those with the highest occurrence rate in most of the documents in the mentioned category. The DFS also does not put features that occur in most of the documents belonging to both $c_j$ and $\overline{c_j}$, under consideration. With these descriptions, equation 6 is presented for the estimation of DFS features.

$$\text{DFS} (t_i, c_j) = \frac{TF(t_i, c_j)}{df(t_i, c_j)} \times \frac{a_{ij}}{(a_{ij} + b_{ij})} \times \frac{a_{ij}}{(a_{ij} + c_{ij})} \times \frac{a_{ij}}{(a_{ij} + d_{ij})}$$

That $TF(t_i, c_j)$ and $TF(t_i, \overline{c_j})$ show term frequency of feature $t_i$ in category $c_j$ and $\overline{c_j}$. $df(t_i, c_j)$ and $df(t_i, \overline{c_j})$ show the number of documents that consisting feature $t_i$ in category $c_j$ and $\overline{c_j}$. Then feature $t_i$ has the DFS value in each category and final DFS score value for $t_i$ calculated as the following equation (7) which means the features with the highest values.

$$\text{DFS} (t_i) = \max_{t_i, c} \{ \text{DFS} (t_i, c_j) \}$$

### 3-2 Clustering Methods

To evaluate the quality of the extracted keywords, we cluster the documents using the selected terms by the weighting techniques. After that, the clusters are compared with the predefined classes, using entropy of the clusters. The clustering methods categorize documents in an unsupervised manner. Here, we use three clustering approaches including centroid-based (i.e., K-means) [22,23], distribution-based (i.e., EM: Expectation Maximization) [24], and hierarchical (i.e., Ward) [25]. These methods are selected among a variety of the clustering methods due their popularity in text processing [22,25]. In the following sections, we review these methods briefly.

#### 3-2-1 K-means

K-means is a classic and well-known unsupervised clustering algorithm [22,26]. This method is an easy way to categorize information by certain number of clusters, i.e., K clusters. The main idea is that K centres are determined for the clusters. The centres must be chosen carefully, because each centre will produce different results. So, it is better to put them as far apart as possible. To begin with, K-cluster centres are chosen randomly, and, in the next step, all points are assigned to the nearest centre. Then, all the centres are re-calculated as the mean value of the assigned data to each cluster. The process of assigning the points to the clusters and updating the cluster centres are repeated interactively until the centres show no change. This algorithm aims to minimize the objective function as a square function of error, shown in the following equation (9) [27, 28].
\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i - \mu_j \| \]  

That \( \| x_i - \mu_j \| \) is the Euclidean distance between data point \( x_i \) and cluster centre \( \mu \), \( n \) is the number of points in cluster \( i \) and \( k \) is the number of cluster centres.

### 3-2-2 Expectation Maximization (EM)

In situations without a specific number of clusters, one of the clustering algorithms used is expectation maximization [24]. It is a computational method for estimating data, particularly hidden data. This algorithm would be suitable for lost data; it can also be an efficient method to calculate the maximum likelihood estimates in repeated computations. The algorithm is related to specific methods of hidden data approximation, which, in these approach parameters, re-estimate and continue the process until they converge on a particular value. The name is chosen because, in each repetitive step of the algorithm, there is a phase of expectation and maximization [29].

The probability distribution used in the algorithm is mostly normal distribution as shown in Equation 9, because the assumption is that the data could be transformed as a linear sequence from the multivariate normal distribution. EM is an iterative method to calculate parameters.

\[ p(x) = \sum_{z} p(z) p(X|z) = \sum_{k=1}^{K} \pi_k N(X | \mu_k, \Sigma_k) \]  

The connected distribution \( p(x, z) \) in terms of a marginal distribution \( p(z) \) and an eventual distribution \( p(x|z) \), the marginal distribution over \( z \) is specified in terms of the mixing factor \( \pi_k \), that \( p(z_k = 1) = \pi_k \cdot \Sigma_k \) is the covariance and \( p(x) = \sum_{z} p(x,z) \) shows that for every data point \( X_n \) there is a comparable hidden variable \( z_n \).

EM is an iterative method to calculate parameters in two steps: Expectation and Maximization. In the first step, primary values are given to parameters. Then, in the Expectation step, weighting factor for data point \( X_n \) is noted by the posterior probability of \( \gamma(z_{nk}) \) that parameter \( k \) generates \( X_n \) and it is shown in equation 10. In this equation \( z_{nk} \) represents value of responsibilities, \( X_n \) is data point \( n \), \( \mu_k \) is the mean. Following formulas represent EM Gaussian mixture model.

\[ \gamma(z_{nk}) = \frac{\pi_k N(X_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j N(X_n | \mu_j, \Sigma_j)} \]  

The second step is Maximization, which calculates the values of the parameters based on the estimated value for \( \gamma(z_{nk}) \) as shown in Equation 11 that \( N_k \) is the effective number of points assigned to cluster \( k \).
The two Expectation and Maximization steps are repeated until the optimal parameter values of converge (equation 12,13). As EM uses the maximum likelihood estimation in each iteration, the following likelihood (equation 14) is increased [30].

\[
\ln p(X | \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k N \left( X_n | \mu_k, \Sigma_k \right) \right)
\]

For simplicity, consider a Gaussian mixture, whose components have covariance matrices given by \( \Sigma_k = \sigma_k^2 I \), where I is the unit matrix, although the conclusions will hold for general covariance matrices. Other distributions such as Poisson and log-normal will also be used to improve results. K-means clustering is a special case of expectation maximization clustering as well [31].

### 3-3-3 Hierarchical Clustering

In the hierarchical clustering method [32], the tree structure is assigned to the final clusters in accordance with their popularity. This hierarchical tree is called a ‘dendrogram’. Hierarchical trees are usually divided into two categories [33,25].

1. Top-down or divisive: In this approach, initially, all data are considered a cluster. Then, in every step of a repetition process, the data that are less similar to each other are put in separate clusters. This step continues until the clusters have only one member. Examples of this clustering are bisecting K-means [33].
2. Bottom-up or agglomerative: In this method, each data is considered a separated cluster, and in every step of a repetition process, data that are similar to each other are combined to produce a cluster or a certain number of clusters. The examples of this type of clustering are average link, complete link, and single link [33].

In order to decide which clusters must be merged (for agglomerative), or where a cluster should split (for divisive), an evaluation of dissimilarity between sets is needed. In most hierarchical clustering methods, this is done by the use of an appropriate metric (a measure of distance between pairs), and a linkage criterion that determines dissimilarity of sets as a function of the pair wise distances of observations in the sets. The choice of a proper evaluation metric has a
direct impact on the final result. Based on different metrics such as Manhattan, Euclidean, etc., points have different distances from each other, resulting in different forms of clusters and different results of clustering [34,35]. Linkage metrics are the same as those that denote distances between sets of points. In hierarchical clustering, there are some linkage metrics like complete-linkage, minimum-linkage, average-linkage, Ward, etc.

Ward clustering is a method to reduce the loss of remote data [36]. This method uses new criteria to calculate dissimilarity between clusters. In this process, difference of square’s summation between each data from a cluster to the cluster’s mean vector is calculated with the aim of evaluating the cluster. The following algorithm could be considered for Ward [37]:

1- Each data is considered a cluster.
2- For all pairs from a set of clusters, those two clusters whose sum of the squares of differences between the clusters’ data to the obtained mean vector is less than the others, are going to be selected.
3- Two selected clusters are combined and the new cluster centre is calculated.
4- As long as the number of clusters is not the target number, Steps 2 and 3 are repeated.

In Ward’s method, the distance between clusters A and B is shown by Equation 15.

\[
\Delta(A, B) = \sum_{i \in A} \| x_i - \mu_{A \cup B} \|^2 - \sum_{i \in A} \| x_i - \mu_A \|^2 - \sum_{i \in B} \| x_i - \mu_B \|^2 = \frac{n_A n_B}{n_A + n_B} \| \mu_A - \mu_B \|^2
\]

(15)

That \( m_j \) is the centre of cluster \( j \), \( n_j \) is the number of points that exist in cluster \( j \), and \( \Delta \) is the merging cost of two clusters \( A \) and \( B \).

4- Proposed Weighting Methods

In this section, we describe our proposed weighting methods that are a combination of three known weightings, TF, TF-IDF and Variance. The proposed methods are the variance to TF-IDF ratio, the intersection of TF-IDF and variance, and the intersection of TF and variance.

4-1 The Ratio of Variance to TF-IDF (Var2TF-IDF)

In this method, the variance of term frequency to TF-IDF ratio of a term is computed. For normalization, the numerator multiplied in \( 10^{-5} \) and denominator multiplied in \( 10^3 \). This operation is done for all of the terms, and the terms with the highest values are selected as the keywords. This weighting method is calculated using Equation 16 in which \( t_i \) denotes the \( i \)th term.

\[
Var2TFIDF(Term_i) = \frac{var(TF(Term_i)) \times 10^{-5}}{TFIDF(Term_i) \times 10^3}
\]

(16)

Fig 2 represents the entropy of document clustering (5 clusters) using variance, TFIDF, and variance to TFIDF ratio feature selection methods. It shows the effectiveness of the proposed method in comparison with the reference methods. Detailed results are given in Section 5.
4-2 The Ratio of Variance to TF (Var2TF)
In the second proposed method of this paper, after term extraction, the values of variance of term frequency and TF of a word is calculated in addition to other terms. Then, the ratio of variance to TF is computed. For normalization, the numerator multiplied by $10^{-5}$ and the denominator by $10^3$. This is done for all the terms, and the terms with the highest values are selected as the keywords. This weighting method is calculated as Equation 17, in which $t_i$ denotes the $i^{\text{th}}$ term.

$$Var2TF(Term_i) = \frac{\text{var}(TF(Term_i)) \times 10^{-5}}{TF(Term_i) \times 10^3}$$ (17)

Fig 3 represents the entropy of document clustering (5 Clusters) using variance, TF, and variance to TF ratio feature selection methods. It shows the effectiveness of the proposed method in comparison with the reference methods. The detailed results are given in Section 5.

4-3 The Intersection of Variance and TF-IDF (Var∩TF-IDF)
In the third proposed method of this paper, the values of features are computed by using the variance weighting method and TF-IDF method, as an intersection is made between 150 terms of both methods. So, for common words, the matrix will form.
Fig 4 represents the difference between variance and TF-IDF values to variance $\cap$ TF-IDF with Ward clustering (5 clusters). The effectiveness of the proposed method is shown in comparison to the reference methods. More detailed results are given in Section 5.

4-4 The Intersection Variance and TF (Var∩TF)
In the fourth proposed method of this paper, the values of features are calculated by using the variance weighting method and TF. Subsequently, an intersection is made between 150 terms of both methods. For common words, the matrix would form.
Fig. 5 represents the difference between variance and TF values to Var∩TF with Ward clustering (5 Clusters). The effectiveness of the proposed method is shown in comparison to the reference methods. More detailed results are given in Section 5.

5- Experiments and Evaluations
In this section, first, the data are explained. Then, the evaluation methods and the results are given in entirety. The comparison between the results is shown at the end of the section. All the evaluations and results are calculated for both English and Persian text documents.

5-1 Data
We have evaluated the keyword extraction methods for both Persian and English languages. For Persian, evaluations are done on 500 documents collected for this research. The Persian data includes news (collected from the ISNA website - https://www.isna.ir/) and scientific articles.
(articles of the Iran Computer Association 2014 - http://csicc2014.sbu.ac.ir/). This dataset has five classes including cultural, medical, sport, information technology and political categories. The data of classes are balanced and each class consists of 100 documents. The English data consist of 500 patent documents [8]. This dataset also has five classes including five patent categories in subjects of gasification, genetically engineered organisms, solar cells, passive space heating, and wind. Each class has 100 documents. The documents of these datasets did not have any pre-defined keywords. They were extracted by weighting methods; then, document clustering, based on the keywords, was done to evaluate the weighting and clustering methods.

5-2 Evaluation Method
As each document in our datasets has a label of its class, the evaluation metric used here is cluster entropy. To calculate entropy, a cluster matrix is calculated in which rows represent the classes and columns represent the clusters. Fig. 6 shows the matrix in which \( X_{ij} \) is the number of elements in class \( i \) and cluster \( j \), \( N \) is number of classes, \( M \) is number of clusters. The purpose of forming this matrix is to calculate entropy. Scattering is reduced as the amount of entropy moves close to zero, indicating an improvement in the chosen words and clustering. In contrast, if the entropy value is close to 1, the method will be less accurate. Using the mentioned matrix, the entropy is calculated using Equations 18 and 19.

\[
e(c_j) = \sum_i^N \left( -\frac{X_{ij}}{N} \log \frac{X_{ij}}{N} \right) (i = 1, 2, ..., N, j = 1, 2, ..., M)
\]

\[
e = \sum_j^M \left( \frac{1}{n} \sum_i^N X_{ij} \right) e(c_j)
\]

In these equations, \( n \) is the total number of documents, \( e(c_j) \) is entropy of cluster \( j \), \( w(c_j) \) is the weight of cluster \( j \) and \( e \) is total entropy.

5-3 Keyword Extraction for English
5-3-1 Evaluation of K-means Clustering
In this section, the results of all the weighting methods by K-means clustering for English documents are presented (5 Clusters). After extracting the keywords by the weighting methods mentioned in Sections 3-1 and 4, the most important terms (i.e., keywords with higher weights) are selected. Our evaluations are performed for different number of terms—30, 70 and 150. For each of these numbers of keywords, three mentioned features i.e., frequency, normalized frequency, and Boolean are used for clustering. After performing K-means clustering on each feature, the entropy value is calculated to compare the obtained clusters with predefined classes. Fig 7 represents the results.
It shows that, for 30 features, the best response is the one which relates to the Var2TF-IDF with the Boolean matrix. The worst response is for TF-IDF with frequency and normal matrices. For 70 features, the best response is the same as in 30 features, and the worst response is for the TF-IDF method with Boolean and frequency matrices. In 150 features, the best response is for the proposed method, Var∩TF, and the worst is for TF-IDF with normal and frequency matrices. The experiments in this section show that the proposed methods, Var2TF-IDF and Var∩TF, are the better methods for keyword extraction than the reference methods.

5-3-2 Evaluation of EM Clustering
This section represents the results of all weighting methods evaluated by the expectation maximization clustering algorithm of English documents (5 Clusters). As shown in Section 5-3-1, in this experiment 30, 70, and 150 features are extracted and EM clustering is applied to them. The results are shown in Fig. 8. It shows, for 30 features, the best response is the one that relates to TF with all three matrices and the worst response is for TF-IDF with normal and frequency matrices. In 70 and 150 features, the best answers are achieved by the Var2TF-IDF proposed method with the Boolean and normal matrices. In contrast, the worst response in 70 features is for the TF-IDF method with Boolean and frequency matrices and in 150 features is the one relating to the TF-IDF with all three matrices. As a result, the recommended keyword extraction methods are TF and Var2TF-IDF.

5-3-3 Evaluation of Ward Clustering
The results of all weighting methods evaluated by the hierarchical Ward clustering method for English documents are given in Fig. 9 (5 Clusters). Similar to Sections 5-3-1 and 5-3-2, for this method, too, 30, 70 and 150 features are extracted and Ward clustering applied to the features. For this number of features, the best response is the one which relates to Var2TF-IDF with frequency and Boolean matrixes. In 30 features, the worst performance is the one which relates to DFS with a frequency matrix. In 70 and 150 features, the worst response is the one which relates to DFS and TFIDF with frequency and normal matrices. The experiments of this section define the proposed Var2TF-IDF method for keyword extraction.

5-4 Keyword Extraction for Persian
5-4-1 Evaluation of K-means Clustering
In the current section, the results of all weighting methods evaluated by K-means for Persian documents are presented in Fig. 10 (5 Clusters). Similar to the English experiments, the matrices are formed for 30, 70 and 150 features after extracting keywords by all the weighting methods. Afterwards, the K-means clustering is performed on the matrixes to compare the new clusters with predefined classes. This process for 30 features has the best performance with Var2TF with the Boolean matrix and the worst performance with the TF-IDF by all three matrices. For 70 features, the best response is, again, for the proposed Var2TF method with the Boolean matrix and the worst response is for the TF-IDF method with all three matrixes. In 150 features, the best response is for Var2TF with the Boolean matrix and the worst response is for the TF-IDF with
normal and frequency matrices. It means the proposed Var2TF is the best method for keyword extraction.

5-4-2 Evaluation of EM Clustering
In this section, the results of the weighting methods for Persian documents are evaluated by the expectation maximization method (5 Clusters). The results are given in Fig. 11. As in Section 5-4-1, the 30, 70 and 150 features are extracted and EM clustering is applied to them. For 30 and 70 features, the best response is for Var2TF with normal and frequency matrices. For 150 features, the best response is the one for the TF method with the Boolean matrix. The worst response for all features is for the TF-IDF method with normal and frequency matrices. The results of this section show that the best keyword extraction methods are TF and Var2TF.

5-4-3 Evaluation of Ward Clustering
Now, the results of the weighting methods are evaluated by the hierarchical Ward for Persian documents and the results are shown in Fig. 12 (5 Clusters). Similar to Sections 5-4-1 and 5-4-2, in this section, too, the 30, 70 and 150 features are extracted before applying the Ward clustering to them. In 30 and 70, the best responses in both are for the proposed Var2TF method with the Boolean matrix. In addition, for 150 features, the best response is the one which relates to Var2TF with normal and Boolean matrices. In this experiment, the worst response for all features is the one which relates to the TF-IDF method with normal and frequency matrices. The results of this section also show the effectiveness of the Var2TF for keyword extraction in Persian in comparison with the other methods.

5-5 Comparative Results
This section presents a comparative study of the keyword extraction methods, clustering methods, and the numbers of features. In Fig. 13 average entropy values for the proposed and the reference keyword extraction methods (i.e., weighting) are shown for English and Persian text documents. This figure presents the evaluation results of three clustering methods. The entropy values of this figure are the average of the entropy values of three sets of features (i.e., 30, 70 and 150) and three feature types (i.e., Boolean, normal, and term frequency matrices).

As shown by Fig. 13, for keyword extraction in English, the proposed Var2TF-IDF method in Ward clustering has the best entropy value (i.e., the lowest entropy value). Besides, in Persian keyword extraction, the proposed weighting, Var2TF, has the best entropy value in Ward clustering.

For other clustering methods, the proposed keyword extraction methods have good entropy values as well. For English, the Var2TF-IDF and Var∩TF are the best methods for the K-means and EM clustering methods, respectively. In Persian, the Var2TF is the best method for other clustering methods, too.

Moreover, it is notable that the average of the entropy values for Persian are higher than those for English. This is probably due to a higher overlap of the documents belonging to different classes for Persian than English.

The average of entropy values obtained from the three clustering methods is show in Fig. 14. (8 Clusters). In this figure, the average is calculated with the entropy values of nine weighting
methods, three numbers of keywords and also all three feature extraction techniques. The results show, the Ward method produced lower entropy than others and K-means achieved the highest value. This demonstrates the higher clustering power of Ward than the other methods.

The results shown in Fig. 15 are obtained by averaging the outcome of all nine weighting methods, three clustering techniques, and three numbers of keywords to show the entropy values for three types of features used in clustering (i.e., matrices). It is clear the Boolean matrix in both English and Persian text documents has the best result (i.e., lower entropy values). It is also evident that average entropy values for Persian are higher than English.

The averaging result of all nine weighting methods, three clustering techniques, and three features extraction methods to show the entropy values for numbers of keywords is shown in Fig 16. As expected, an increase in the number of keywords resulted in better clustering (i.e., lower entropy values). Moreover, it can be seen that the average entropy values for Persian are higher than those for English.

6- Summary and Conclusions
Keyword extraction is often done to realize the overall concept of documents and giving the reader an overall view. In this paper, five statistical methods of keyword extraction from a text document were presented and four others were proposed. In a keyword extraction method, first, all terms are weighted by using the reference and proposed methods, and the terms with the highest weights are select as the keywords. The proposed weighting methods were variance to TF-IDF ratio, variance to TF ratio, the intersection of TF and variance, and the intersection of variance and TF-IDF. To evaluate our methods, we used documents with predefined classes. All documents were clustered using three clustering methods, K-means, EM, and hierarchical Ward. The clusters were then compared with predefined classes using entropy value as the evaluation metric.

We did the evaluations for both English and Persian documents for three sets of keywords. For English documents, the best keyword extraction method was the proposed variance to TF-IDF ratio method and, for Persian documents, the best method was the variance to TF ratio method. Moreover, we evaluated the methods using three different feature extraction techniques during the clustering, with results showing the effectiveness of the Boolean method in comparison with the other methods.

In the continuation of this research, the use of a semantic approach to extract and select features will be pursued. It means, after the extraction and the counting the features, features that are semantically similar but lack identical written forms are considered as a single word (feature). The similarity of the words can be discerned by using a semantic dictionary like WordNet in English and FarsNet in Persian [38]. By considering every sense of each word, similar words are counted as a single word and the rank of word counting would increase in TF, TF-IDF, Variance, DFS, the ratio of variance to TF, and the ratio of Variance to TF-IDF methods.
References

1. Liu, J., and Wang, J. Keyword extraction using language network. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, pp. 129-134 (2007).
2. Rossi, R.G., Maracini, R.M., and Rezende, S.O. “Analysis of domain independent statistical keyword extraction methods for incremental clustering”. Learning and Nonlinear Models, 12(1), 17-37 (2014).
3. Siddiqi, S., and Sharan, A. “Keyword and keyphrase extraction techniques: a literature review”. International Journal of Computer Applications, 109(2), 18-23 (2015).
4. Beliga, S., Mestrovic, A., and Martinic-Ipsic, S. “An overview of graph-based keyword extraction methods and approaches”. Journal of Information and Organizational Sciences, 39(1), 1-20 (2015).
5. Taeho, C. J. “Text categorization with the concept of fuzzy set of informative keywords”. Fuzzy Systems Conference Proceedings. FUZZ-IEEE'99. 1999 IEEE International. Vol. 2. IEEE (1999).
6. Mohammadi, M., and Analouyi, M. “Keyword Extraction in Persian Documents”. 13th conference of Iran computer Association, Kish, Iran (2007).
7. Biswas, S. K., Bordoloi, M., and Shreya, J. “A graph based keyword extraction model using collective node weight”. Expert Systems with Applications, 97, 51-59 (2018).
8. Noh, H., Joe, Y., and Lee, S. “Keyword selection and processing strategy for applying text mining to patent analysis”. Expert Systems with Applications Elsevier (2015).
9. Haulth, A. “Improved Automatic Keyword Extraction Given More Linguistic Knowledge”. In proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, Sapporo, Japan: 216-223 (2003).
10. Ercan, G., and Cicicli, I. “Using Lexical Chains for Keyword Extraction”. Information Processing and Management: 1705-1714 (2007).
11. Zhang, K., Xu, H., Tang, J., et al. “Keyword Extraction Using Support Vector Machine”. In Proceeding of the 7th International Conference on Web-Age Information Management. Hong Kong, China: 85-96 (2006).
12. Onan, A., Korukoğlu, and S., Bulut. “Ensemble of keyword extraction methods and classifiers in text classification”. Expert Systems with Applications (2016).
13. Wartena, C., Brussee, R., and Slakhorst, W. “Keyword Extraction Using Word Co-occurrence”. IEEE 2010 Workshops on Database and Expert Systems Applications: 54-58 (2010).
14. Arabi, S., Vahidi, M., and Minaei, B. “Keyword Extraction for Persian Text Categorization”. First Iranian Data Mining conference, Amirkabir University of Technology (2007).
15. Frank, E., and Paynter, I. H. “Domain-Specific Keyphrase Extraction”. In proceedings of the 16th International Joint Conference on Artificial Intelligence. Stockholm, Sweden. Morgan Kaufman: 668-673 (1999).
16. Najafi, E., and Darooneh, A.H. “The fractal patterns of words in a text: a method for automatic keyword extraction”. Plos One, 10(6), e0130617(2015).
17. Haggag, M. “Keyword Extraction using Semantic Analysis”. International Journal of Computer Applications: 1128-1132 (2013).
18. Keith, J. B. “Phraserate: An Html Keyphrase Extractor”. Technical Report, University of California, Riverside: 1-16 (2002).
19. Sharma, N., Bajpai, A., and Litoriya, R. “Comparison the various clustering algorithms of Weka tool”. International Journal of Emerging Technology and Advanced Engineering: 73-80 (2012).
20. Aleahmad, A., Hakimian, P., Mahdikhani, F., et al. “N-gram and Local Context Analysis for Persian Text Retrieval”. Signal Processing and Its Applications (2007).
21. Zong, W., Wu, F., Chu, L., et al. “A discriminative and semantic feature selection method for text categorization”. Int. J. Production Economics Elsevier (2015).
22. Kaur, M., and Kaur, N. “Web Document Clustering Approaches Using K-means Algorithm”. International Journal of Advanced Research in Computer Science and Software Engineering: 861-864 (2013).
23. Zarandi, M. F., Faraji, M. R., and Karbasian, M. “An exponential cluster validity index for fuzzy clustering with crisp and fuzzy data”. Scientia Iranica. Transaction E, Industrial Engineering, 17(2), 95 (2010).
24. McLaclan, G., and Krishnan, T. “The EM Algorithm and Extensions”. Wiley 2nd Edition. New Jersey (2008).
25. Zhao, Y., and Karypis, G. “Hierarchical Clustering Algorithms for Document Datasets”. Journal of data mining and knowledge discovery Elsevier: 141-168 (2005).
26. Montazeri-Gh, M.,and Fotouhi, A. “Traffic condition recognition using the k-means clustering method”. Scientia Iranica, Transactions B: Mechanical Engineering, 18, 930–937 (2011).
27. Hartigan, J.A, and Wong, M.A. “Applied Statistics”. Journal of the Royal Statistical Society: 100-108 (1979).
28. Jain, A.K. “Data Clustering: 50 years beyond K-means”. International Conference on Pattern Recognition (ICPR) Elsevier: 651-666 (2010).
29. Jeff, Wu. C.J. “On the Convergence Properties of the EM Algorithm”. The Annals of statistics: 95-103 (1983).
30. Bock, R.D., and Aitkin, M. “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm”. Psychometrika 46(4): 443-459 (1981).
31. Bottou, L., and Bengio, Y. “Convergence properties of the k-means algorithms”. Advances in neural information processing systems: 585-592 (1995).
32. Johnson S.C. “Hierarchical Clustering Schemes”. Psychometrika 32:3: 241-254 (1967).
33. Web A.R. “Statistical Pattern Recognition”. John Wiley & Sons. 2nd Edition (2002).
34. Olson, C.F.” Parallel Algorithms for Hierarchical Clustering”. International journal of parallel computing Elsevier: 1313-1325 (1995).
35. Mangiameli, P., Chen, S.K., and West, D. “Comparison of SOM Neural Network and Hierarchical Clustering Method”. European Journal of operational Research: 402-417 (1976).
36. He, Q. “A Review of Clustering Algorithms as Applied in IR”. Graduate School of Library and Information Science University of Illinois at Urbana-Champaign (1999).
37. Batagelj, V. “Generalized Ward and Related Clustering Problem. Classification and Related Methods of Data Analysis”. H.H. Bock (editor). NorthHolland, Amsterdam: 67-74 (1986).
38. Shamsfard, M., Hesabi, A., Fadaei, H., et al. “Semi-Automatic Development of FarsNet; The Persian WordNet”. Proceedings of 5th Global WordNet Conference. Mumbai, India, vol. 29 (2010).
Figures’ list

Fig 1. Keyword extraction and evaluation steps
Fig 2. Entropy results of TFIDF, Variance and proposed Variance/TFIDF methods for Ward clustering (5 clusters)
Fig 3. Entropy results of TF, Variance and proposed Variance/TF methods for Ward clustering (5 Clusters)
Fig 4. Entropy results of TF-IDF and variance to proposed variance ∩ TF-IDF method with Ward clustering (5 Clusters)
Fig 5. Entropy results of TF and variance to proposed variance ∩ TF method with Ward clustering (5 Clusters)
Fig 6. Entropy matrix for evaluation
Fig 7. K-means results for different numbers of extracted keywords on English text documents using different methods (5 Clusters)
Fig 8. EM results for different numbers of extracted keywords on English text documents using different methods (5 Clusters)
Fig 9. Ward results for different numbers of extracted keywords on English text documents using different methods (5 Clusters)
Fig 10. K-means results for different numbers of extracted keywords on Persian text documents using different methods (5 Clusters)
Fig 11. EM results for different numbers of extracted keywords on Persian text documents using different methods (5 Clusters)
Fig 12. Ward results for different numbers of extracted keywords on Persian text documents using different methods (5 Clusters)
Fig 13. Average of entropy values for all weighting methods in a) English and b) Persian documents for different clustering methods (5 Clusters)
Fig 14. Average of entropy values for clustering methods in a) English and b) Persian documents
Fig 15. Average of entropy values for different features of clustering in a) English and b) Persian documents
Fig 16. Average of entropy values for different numbers of keywords in a) English and b) Persian documents
Tables' List

Table 1. Occurrence rate of feature ti and category cj (Zong et al, 2015)
Figures and Tables

Documents with predefined Classes

Pre-processing: Normalization and stop-word removal

Calculating weight for each term by using weighting methods

Selecting the highest values of each term as a feature and forming the feature-document matrix

Clustering each matrix using clustering methods

Calculating entropy values for each clustered documents to compare the clusters to predefined classes

Fig 1
Fig 2

Fig 3
Fig 4

Normal Matrix

Fig 5

Normal Matrix

Fig 6

\[
\begin{array}{cccc}
1 & 2 & \cdots & M \\
X_{11} & X_{12} & \cdots & X_{1M} \\
X_{21} & X_{22} & \cdots & X_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
N & X_{N1} & \cdots & X_{NM}
\end{array}
\]
Fig 7

A) 30 KEYWORDS

B) 70 KEYWORDS

C) 150 KEYWORDS

Fig 8
Fig 9

A) 30 KEYWORDS

B) 70 KEYWORDS

C) 150 KEYWORDS
Fig 11
Fig 12

A) 30 KEYWORDS

B) 70 KEYWORDS

C) 150 KEYWORDS
Fig 13

a) English text documents

b) Persian text documents
a) English text documents

b) Persian text documents

Fig 14
a) English text documents

b) Persian text documents

Fig 15
Table 1

| document is in category $c_j(c_j)$ | $a_{ij}$ | $b_{ij}$ |
|-----------------------------------|---------|---------|
| document is not in category $c_j(c_j)$ | $c_{ij}$ | $d_{ij}$ |
Biography

Hadi Veisi received his Ph.D. in Artificial Intelligence from Sharif University of Technology in 2011. He joined University of Tehran, Faculty of New Sciences and Technologies (FNST) in 2012 and established Data and Signal Processing (DSP) lab. The main research interests of Hadi are artificial neural network and deep learning, natural language processing and speech processing.

Niloofar Aflaki obtained her BS and MS in software engineering from Islamic Azad University Central Tehran Branch and University of Tehran, Kish International Campus in 2013 and 2016, respectively. She is currently a Ph.D. candidate and a tutor at Massey University, Auckland. Her recent research focuses on the interpretation of geospatial language.

Pouyan Parsafard in 2013 by entering the field of Computer Science at KIAU Azad University of Karaj for his bachelor’s degree and after learning various programming languages in the first year, became acquainted with Machine Learning and Data Mining. He finished his master degree in Software Engineering from University of Tehran, Kish International Campus in 2016. He carried out his research in Master’s degree at University of Tehran, Data and Signal Processing lab (DSP) under supervision of Dr. Hadi Veisi. There, he acquired strong interest in Machine Learning, Fuzzy Systems and Data Mining.