공학석사 학위논문

# Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation

## Bag-of-Concepts: 단어에 대한 분산표상의 군집화를 통한 해석 가능한 문서 표현법

2016년 8월

서울대학교 대학원

산업공학과 데이터마이닝 전공

김 한 결

# Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation

## Bag-of-Concepts: 단어에 대한 분산표상의 군집화를 통한 해석 가능한 문서 표현법

지도 교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함
2016 년 6 월

서울대학교 대학원
산업공학과 데이터마이닝 전공
김 한 결

김한결의 석사 학위논문을 인준함
2016 년 6 월

위 원 장 ______ 박 종 헌 ______ (인)

부위원장 ______ 조 성 준 ______ (인)

위 원 ______ 이 재 욱 ______ (인)

# Abstract

# Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation

Han Kyul Kim

Department of Industrial Engineering

The Graduate School

Seoul National University

Two document representation methods are mainly used in solving text mining problems. Known for its intuitive and simple interpretability, the bag-of-words method represents a document vector by its word frequencies. However, this method suffers from the curse of dimensionality, and fails to preserve accurate proximity information when the number of unique words increases. Furthermore, this method assumes every word to be independent, disregarding the impact of semantically similar words on preserving document proximity. On the other hand, doc2vec, a basic neural network model, creates low dimensional vectors that successfully preserve the proximity information. However, it loses the interpretability as meanings behind each feature is indescribable. This paper proposes the bag-of-concepts method as an alternative document representation method that overcomes the weaknesses of these two methods. This proposed method creates concepts through clustering word vectors generated from word2vec, and uses the frequencies of these concept clusters to represent document vectors. Through these data-driven concepts, the proposed method incorporates the impact of semantically similar words on preserving document proximity effectively. With appropriate weighting scheme such as concept frequency-inverse document frequency, the proposed method provides better document representation than previously suggested methods, and also offers intuitive interpretability behind the generated document vectors. Based on the proposed method, subsequently constructed text mining models, such as decision tree, can also provide interpretable and intuitive reasons on why certain collections of documents are different from others.

**Keywords:** bag-of-concepts; interpretable document representation; word2vec clustering
**Student Number:** 2014-22642

# Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

With growing importance of unstructured data, text data, as one of the most common forms of unstructured data, have inevitably became an important source for data analysis. To extract interesting patterns and insights from them, most fundamental and crucial step is document representation. In order to apply various machine learning and data mining techniques, raw documents need to be transformed into numerical vectors, in which each documents defining characteristics are captured. If these document vectors can preserve proper proximity between the documents and their unique characteristics, subsequent text mining algorithms can extract more accurate and valuable information hidden in the data.

Most popular document representation methods have often relied on the bag-of-words based approaches (Baeza-Yates & Ribeiro-Neto, 1999; Manning & Schtze, 1999; Sayeedunnissa, Hussain & Hameed, 2013), in which a document is fundamentally represented by the counts of word occurrences within a document. For decades, this approach has been shown to be effective in solving various text mining tasks(Huang, 2008; Kolari, Java, Finin, Oates & Joshi, 2006; Wu, Hoi & Yu, 2010). One of its major advantages is that it produces intuitively interpretable document vectors as each feature of the document vector indicates an occurrence of a specific word within a document. The bag-of-words approach, however, can be problematic when a number of documents being represented are enormous. As a number of documents increase, a number of unique words in the entire document set will also naturally increase. Consequently, not only will the generated document vectors be sparse, but their dimensions will also be huge. As the dimension and the sparsity of the document vectors increase, conventional distance metrics become ineffective in representing the proper proximity between the documents. Furthermore, the bag-of-words methods assumes that all words within the documents are independent. However, different word types such as synonyms and hypernyms usually describe similar information within the document. Thus, this word independence has adverse impact on capturing document proximity. Consequently, the text mining models constructed from the bag-of-words approach can be unsuccessful in capturing the proper differences between high dimensional and sparse document vectors. Although various dimension reduction techniques (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990; Hofmann, 1999) do exist, these techniques lose the innate interpretability of the bag-of-words approach.

To overcome such limitation of the bag-of-words approach, doc2vec model (Le & Mikolov, 2014), an extension of word2vec (Mikolov, Chen, Corrado & Dean, 2013a; Mikolov, Sutskever, Chen, Corrado & Dean, 2013) model, utilizes contextual information of each word and document to embed documents into a continuous vector space with manageable dimension. While a context of a word indicates surrounding words of a given word, a context of a document is defined as the distribution of its composing words. With this contextual information, the document vectors with similar contextual information are located close to each other in the embedded space. Consequently, its performance in document clustering and classification tasks have previously been reported to be better than those of the bag-of-words based models (Dai, Olah, Le & Corrado, 2014). However, each feature of document vectors generated from doc2vec is difficult to interpret as its value indicates the weight of the neural network used to train doc2vec.

Despite the outstanding representational performance of doc2vec in document clustering and classification tasks, having a good representation itself is not the ultimate goal of text mining. In order to apply such method in real text mining tasks, the document vectors, similar to those produced by the bag-of-words method, need to be interpretable in order to provide much deeper understanding of the dataset and the operating logic behind subsequently constructed text mining models. Furthermore, they can be utilized to acquire reasoning and summarization of the documents, potentially offering actionable business implications. However, the document vectors generated from doc2vec fail to provide any intuitive interpretability.

In order to compensate for the limitation of doc2vec, this paper suggests the bag-of-concepts approach for representing document vectors. Through clustering distributed representation of words generated from word2vec, semantically similar terms are clustered into a common concept, thereby incorporating the impact of semantically similar words on preserving document proximity. Document vectors are subsequently represented by the frequencies of these concepts. Through utilizing semantic similarity of the continuous space generated by word2vec, this proposed method captures proper proximity between the document vectors, while simultaneously providing representation interpretability and model explainability. As each feature represents an unique concept, we can perceive the documents as collections of concepts and intuitively understand the comprising components of the generated document vectors. Through model explainability, we can easily comprehend the operating logic behind the text mining model trained with the document vectors generated from the proposed method.

This paper has performed document classification on Reuter dataset to provide both quantitative and qualitative analysis of the proposed method. For quantitative analysis, we have compared the document classification accuracies of the proposed method, word2vec averaging, doc2vec, the bag-of-words and latent semantic analysis. For qualitative analysis, we have built a decision tree classifier from the document vectors generated from the proposed method. Through observing the splitting nodes of the decision tree, we have also explored the underlying reasoning behind the classifier, and successfully identified those concepts that are crucial for the classifier in classifying the documents correctly. Through these analysis, we have confirmed both the representational efficacy and the interpretability of the bag-of-concepts approach.

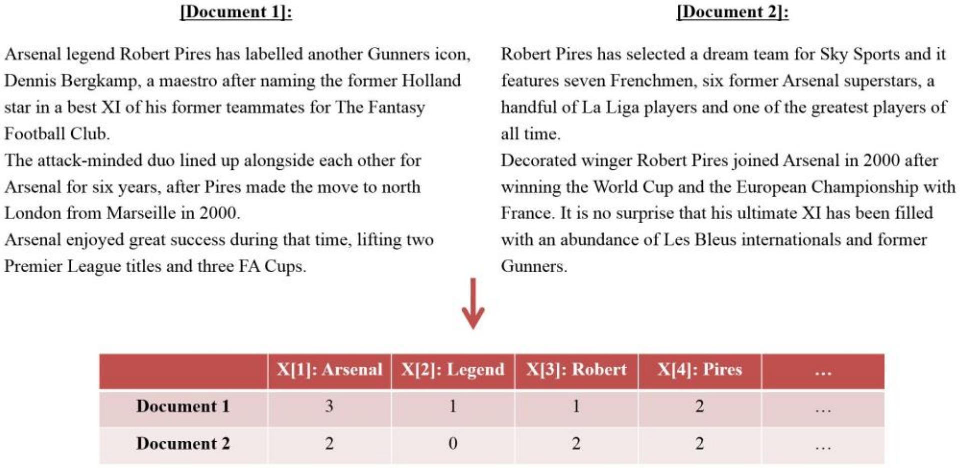
The rest of this paper is structured as follows. In Section 2, we discuss various techniques for document representation in detail. In Section 3 and 4, we propose our bag-of-concepts method and describe the dataset used throughout this paper. In Section 5, we provide experiment result of our proposed method to substantiate its representation effectiveness, interpretability and model explainability. We conclude in Section 6 with discussion and directions for future work.

# Chapter 2. Related Work

In this section, we will provide general idea and motivation behind the bag-of-words, word2vec based approach and doc2vec, and discuss their advantages and disadvantages.

## 2.1 Bag-of-Words

The bag-of-words approach is established upon an assumption that frequencies of words in a document can appropriately capture the similarities and differences between the documents. Consequently, the features of the document vectors generated from the bag-of-words approach represent the occurrences of each word in a document as shown in Figure 2.1.



**[Document 1]:**

Arsenal legend Robert Pires has labelled another Gunners icon, Dennis Bergkamp, a maestro after naming the former Holland star in a best XI of his former teammates for The Fantasy Football Club.
The attack-minded duo lined up alongside each other for Arsenal for six years, after Pires made the move to north London from Marseille in 2000.
Arsenal enjoyed great success during that time, lifting two Premier League titles and three FA Cups.

**[Document 2]:**

Robert Pires has selected a dream team for Sky Sports and it features seven Frenchmen, six former Arsenal superstars, a handful of La Liga players and one of the greatest players of all time.
Decorated winger Robert Pires joined Arsenal in 2000 after winning the World Cup and the European Championship with France. It is no surprise that his ultimate XI has been filled with an abundance of Les Bleus internationals and former Gunners.

| | X[1]: Arsenal | X[2]: Legend | X[3]: Robert | X[4]: Pires | … |
|---|---|---|---|---|---|
| Document 1 | 3 | 1 | 1 | 2 | … |
| Document 2 | 2 | 0 | 2 | 2 | … |

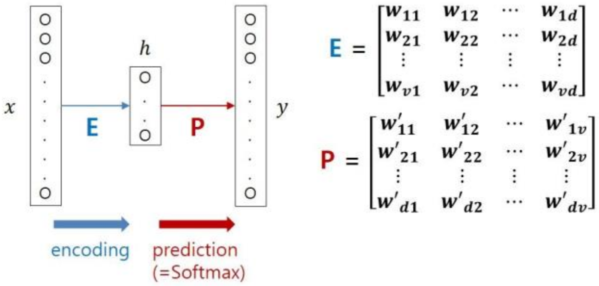
Figure 2.1: Document vectors generated via bag-of-words approach

Due to this intuitive interpretability of the generated document vectors, the bag-of-words approach has established itself as one of most influential document representation methods. However, the number of features in these vectors increases significantly as the number of documents increases in order to incorporate all word occurrences within the document set. Consequently, the dimension of the bag-of-words document vectors can become extremely large and sparse. As the dimension and the sparsity of the document vectors increase, the curse of the dimensionality occurs and conventional distance metrics such as Euclidean distance or cosine distance become meaningless. Due to such limitation, text mining models constructed from the bag-of-words based document vectors fail to capture the true proximity between the documents. Although various dimension reduction techniques such as latent semantic analysis (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990; Hofmann, 1999; Kim, Howland & Park, 2005) do exist, these techniques unfortunately lose the innate interpretability of the bag-of-words approach. Based on singular value decomposition, these techniques reduce the dimension of the document vectors into uncorrelated factors, each of which can be regarded as artificial concepts. Underlying interpretation of each concept, however, remains ambiguous.

## 2.2 Word2Vec

Although word2vec is a word representation method, it can be expanded into representing documents without much significant modification. Thus, we will first discuss word2vec prior to discussing word2vec based document representation and doc2vec.

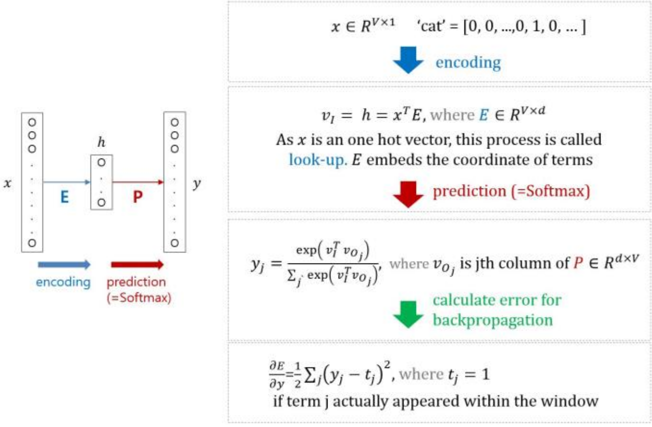
Word2vec is based on the assumption of the distributed hypothesis (Harris, 1954), which states that words that occur in similar contexts tend to have similar meanings (Turney & Pantel, 2010). Based on this assumption, word2vec uses a simple neural network to embed words into continuous vector space. Through training the weights of the network, word2vec model predicts the input word's neighboring words within certain predefined window size.


$x$
$h$
$y$
E
P
E = $\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1d} \\ w_{21} & w_{22} & \cdots & w_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ w_{v1} & w_{v2} & \cdots & w_{vd} \end{bmatrix}$
P = $\begin{bmatrix} w'_{11} & w'_{12} & \cdots & w'_{1v} \\ w'_{21} & w'_{22} & \cdots & w'_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ w'_{d1} & w'_{d2} & \cdots & w'_{dv} \end{bmatrix}$
encoding
prediction
(=Softmax)

| | |
|---|---|
| $x \in R^{V\times 1}$ | One hot representation of a word. Size of input layer is equal to the number of terms |
| $h \in R^{d\times 1}$ | Encoded context vector |
| $y \in R^{V\times 1}$ | Predicted word prob. Size of output layer is also equal to the number of terms |
| $E \in R^{V\times d}$ | Embedding Matrix |
| $w_{ij}$ | Coordinate of context vector. Weight between the ith input node and jth hidden node |
| $P \in R^{d\times V}$ | Word Matrix |
| $w'_{ij}$ | Coordinate of word vector. Weight between the ith hidden node and jth output node |

Figure 2.2: Word2vec basic architecture

As shown in Figure 2.2, the size of input layer $x$ is $V$, equivalent to the total number of unique words in a document set. Each node of the input layer represents an individual word with one-hot encoding. Through encoding matrix $E$, which essentially is an aggregate of each input nodes weight to each of hidden nodes, input word's context is embedded and represented by the hidden layer $h$. Consequently, the number of hidden nodes $d$ denotes the dimension of the word vectors and the embedding space. The embedded context vector $h$ is subsequently multiplied to the corresponding word vector from matrix $P$, which 150 is again an aggregate of each hidden nodes weight to each of the node in the output layer. Through $P$, surrounding context words of the input words are predicted with soft-max function that aims at maximizing the cross product between the embedded context vectors of the input words and the resulting word vectors. Then, this predicted probability of each word is represented by the value of each node in the output layer $y$. Identical to the input layer, the size of the output layer $y$ is once again $V$. By checking whether the predicted words actually occurred around the input words, accuracy of the prediction is evaluated. Through back-propagation, the values (weights) of the embedding vectors and the word vectors are updated. This general description for training word2vec is depicted in Figure 2.3.



$x \in R^{V\times 1}$ 'cat' = [0, 0, ...,0, 1, 0, ... ]

encoding

$v_I = h = x^T E$, where $E \in R^{V\times d}$

As $x$ is an one hot vector, this process is called look-up. $E$ embeds the coordinate of terms

prediction (=Softmax)

$y_j = \frac{\exp(v_I^T v_{O_j})}{\sum_{j'} \exp(v_I^T v_{O_j})}$, where $v_{O_j}$ is jth column of $P \in R^{d\times V}$

calculate error for backpropagation

$\frac{\partial E}{\partial y} = \frac{1}{2}\sum_j (y_j - t_j)^2$, where $t_j = 1$

if term j actually appeared within the window

Figure 2.3: Word2vec training

One of the biggest contributions of word2vec is that the words that occur in similar context - consequently with similar meaning according to the distributed hypothesis - are located close to each other in the embedded space, preserving the semantic similarities between the words. As the words are represented in a continuous embedded space, various conventional machine learning and data mining techniques can be applied in this space to resolve various text mining tasks (Bansal, Gimpel & Livescu, 2014; Ren, Kiros & Zemel, 2015; Xue, Fu & Shaobin, 2014; Cui, Shi & Chen, 2016; Cao & Wang, 2015). Figure 2.4 shows an example of such embedded space visualized by t-sne (Van der Maaten & Hinton, 2008). In this figure, we have embedded words that represent the names of baseball players, the names of soccer players and the names of countries. While the words with similar meanings are located closer to each other, the words with different meanings are located distant from each other.

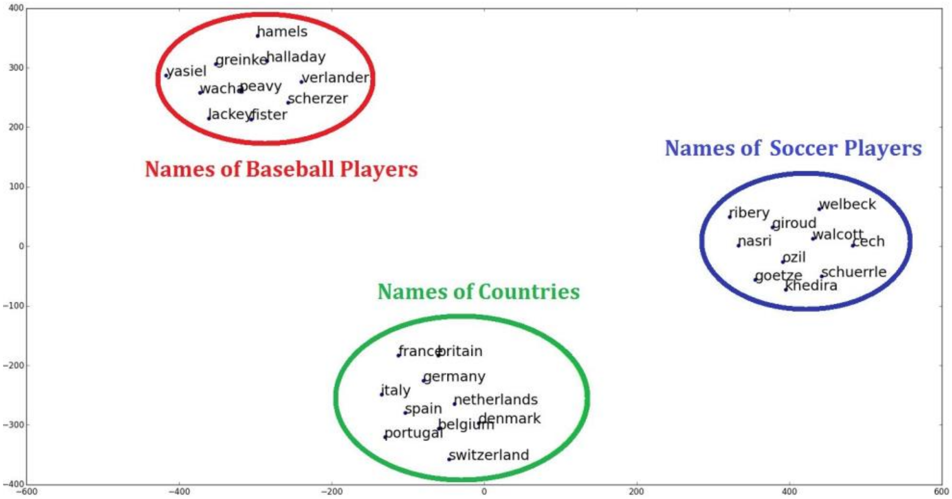
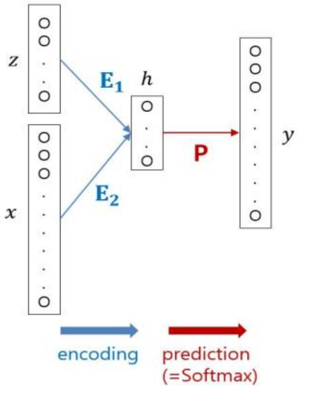


Figure 2.4: Embedded space using t-sne

Compared to the bag-of-words approach, in which dimension and sparsity of a document vector can increase significantly, word2vec model can be utilized to construct dense document vectors with reasonable dimension. One of the simplest approach for representing a document using word2vec is averaging the word vectors of the words that occurred in the document (Xing, Wang, Zhang & Liu, 2014). Despite its simplicity, its representation effectiveness, as validated by the document classification task, has been shown to be quite promising.

## 2.3 Doc2Vec

Instead of averaging the embedded word2vec vectors to represent a document vector, doc2vec directly embeds documents along with their words as shown in Figure 2.5.



$$\mathbf{E} = \begin{bmatrix} \mathbf{E_1} \\ - \\ \mathbf{E_2} \end{bmatrix} = \begin{bmatrix} w^z_{11} & w^z_{12} & \cdots & w^z_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ w^z_{m1} & w^z_{m2} & \cdots & w^z_{md} \\ w_{11} & w_{12} & \cdots & w_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ w_{v1} & w_{v2} & \cdots & w_{vd} \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} w'_{11} & w'_{12} & \cdots & w'_{1v} \\ w'_{21} & w'_{22} & \cdots & w'_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ w'_{d1} & w'_{d2} & \cdots & w'_{dv} \end{bmatrix}$$

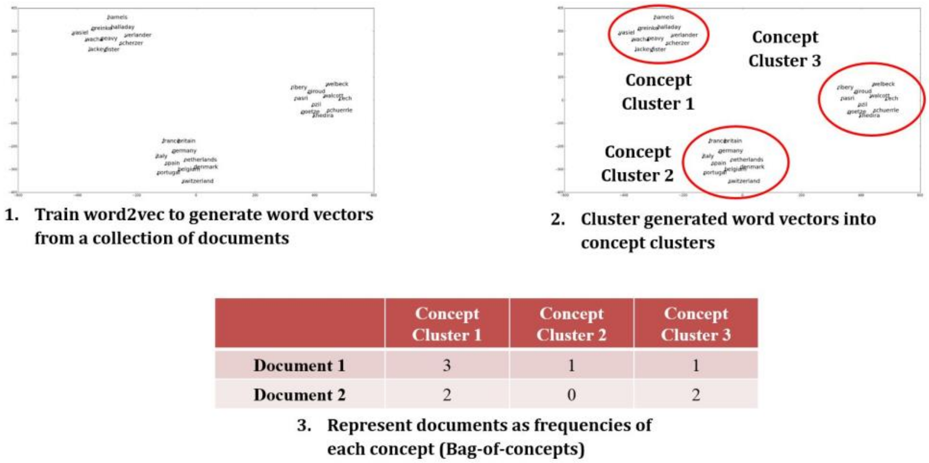
| | |
|---|---|
| $z \in R^{m\times 1}$ | One hot representation of a document. Size of input layer is equal to the number of documents |
| $w^z_{ij}$ | Coordinate of document embedding vector. Weight between the ith input node of z and jth hidden node |

Figure 2.5: Doc2Vec architecture

The architecture and the training of the neural network in doc2vec are essentially identical to those of word2vec. The only difference lies in the fact that documents are also incorporated into the network. Similar to the words in word2vec model, documents are represented by one hot encoding and embedded into a continuous space through an embedding matrix. As shown in Figure 2.5, $E_1$ represents an embedding matrix for the documents, while $E_2$ indicates an embedding matrix for the words. Their coordinates, the values of the weight towards the hidden nodes, are similarly updated by back-propagation.

The representation power of doc2vec has been shown to be effective in document clustering and classification tasks, outperforming word2vec averaging method (Dai, Olah, Le & Corrado, 2014). Although the dimensions of document vectors generated from doc2vec are generally smaller than that of the bag-of-words approach, these features sufficiently incorporate the contextual information of the words and the documents, consequently outperforming the bag-of-words based models. Despite its effective representation power, doc2vec model fails to provide intuitive interpretation behind its generated document vectors. Since each document vector is trained through a neural network, each value of the vector represents only the strength of the connection between the nodes. Consequently, it is hard to comprehend what each feature of a document vector represents in terms of the contents of the document. Therefore, if a text mining model such as a document classifier is trained from these document vectors generated from doc2vec, it fails to provide any intuitive explanation for the operating logic behind the model. Having a good representation of a document itself is not be the ultimate goal of text mining. In order for these representation methods to have meaningful impact and implication in real business environment, it is essential that document representation should be able to provide clear understanding and intuition behind the representation and its subsequently constructed text mining model.

# Chapter 3. Proposed Method



1. Train word2vec to generate word vectors from a collection of documents

2. Cluster generated word vectors into concept clusters

| | Concept Cluster 1 | Concept Cluster 2 | Concept Cluster 3 |
|---|---|---|---|
| Document 1 | 3 | 1 | 1 |
| Document 2 | 2 | 0 | 2 |

3. Represent documents as frequencies of each concept (Bag-of-concepts)

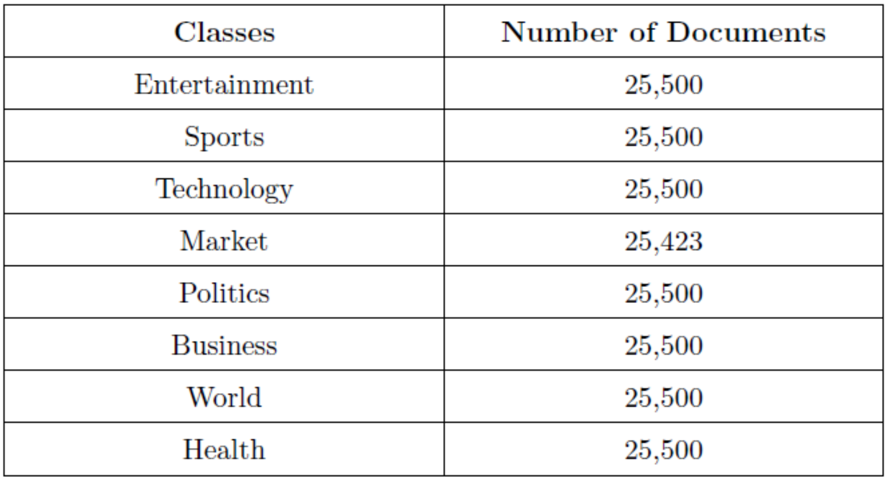
Figure 3.1: Bag-of-Concepts

This paper suggests the bag-of-concepts method as an alternative method for document representation (Figure 3.1). In this proposed method, word vectors of the documents are trained via word2vec. As word2vec embeds semantically similar words into neighboring area, the proposed method clusters neighboring words into one common concept cluster. Similar to the bag-of-words method, each document vector will then be represented by the counts of each concept clusters in the document. As each concept cluster will contain words with similar meaning or common hypernym, the features of the document vectors generated from the proposed method will be interpretable and intuitive. Furthermore, the bag-of-concepts method can be understood as a non-linear dimension reduction technique for transforming a word space into a concept space based on semantic similarity. As the proposed method represents a document with concept frequencies instead of word frequencies, it incorporates both the interpretability of the bag-of-words method, and the representational superiority of the distributed representation method, while overcoming their limitations.

As word2vec maximizes the cross product between the embedding vectors and the context vectors, cosine distance metric is used for clustering the word vectors in the embedding space. Consequently, spherical k-means algorithm (Zhong, 2005) is used to cluster word vectors into concept clusters. For predetermined value of k, spherical k-means clustering, similar to k-means clustering, iteratively assigns each data point to one of k centroids, and updates each centroid given the membership of the data points. However, spherical k-means clustering, instead of Euclidean distance, uses cosine similarity as a distance metric.

# Chapter 4. Data Set Description

Table 4.1: Reuter dataset



| Classes | Number of Documents |
|---|---|
| Entertainment | 25,500 |
| Sports | 25,500 |
| Technology | 25,500 |
| Market | 25,423 |
| Politics | 25,500 |
| Business | 25,500 |
| World | 25,500 |
| Health | 25,500 |

In order to show the representational performance of the proposed method and its applicability, document classification task has been carried out using the document vectors generated from the proposed method. Document classification task aims at differentiating the documents according to their classes. If the proposed method can truly capture the semantic differences between the documents, it should perform well in this task. In this paper, Reuter dataset has been used. To avoid class imbalance problem, Reuter dataset consists of 203,923 randomly selected articles from Reuter website, published between September 1st, 2006 and June 6th, 2015. These articles are labeled by Reuter website into 8 different classes as shown in Table 4.1. The total number of sentences amounts to 3,076,016, while the total number of tokens is equivalent to 89,146,031. For faster word2vec training, we have ignored those words that occurred less than 20 times in the entire dataset, leaving total of 65,159 unique words for training.

# Chapter 5. Experiment Result

Biggest contribution of the proposed bag-of-concepts method is that it incorporates the advantages of the bag-of-words method and doc2vec model. Similar to doc2vec model, the proposed method offers superior representational performance derived from utilizing contextual information. Furthermore, it creates document vectors with reasonable number of dimensions. Most importantly, the proposed method provides explicitly explanatory features for the document vectors, providing interpretability for the vectors themselves and explainability for the text mining models built from these vectors. These three aspects of the proposed method (representational performance, representation interpretability, and model explainability) are established through performing document classification task on Reuter dataset.

## 5.1 Representation Effectiveness

In order to analyze the representation effectiveness of the proposed method, document classification task has been carried out on the document vectors generated from the proposed method. Classification performance is subsequently compared to those calculated from the document vectors generated from the bag-of-words method, latent semantic analysis, word2vec averaging method and doc2vec method as shown in Figure 5.1.

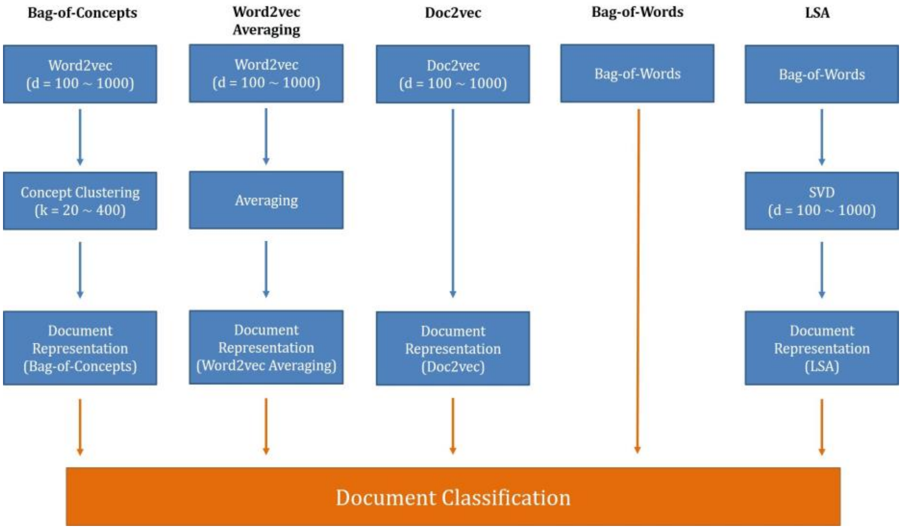

Bag-of-Concepts
Word2vec Averaging
Doc2vec
Bag-of-Words
LSA
Word2vec (d = 100 ~ 1000)
Word2vec (d = 100 ~ 1000)
Doc2vec (d = 100 ~ 1000)
Bag-of-Words
Bag-of-Words
Concept Clustering (k = 20 ~ 400)
Averaging
SVD (d = 100 ~ 1000)
Document Representation (Bag-of-Concepts)
Document Representation (Word2vec Averaging)
Document Representation (Doc2vec)
Document Representation (LSA)
Document Classification

Figure 5.1: Document classification experiment design

Document classification task similar to that of Dai et al. (Dai, Olah, Le & Corrado, 2014) has been carried out. In this document classification task, triplets of documents have been constructed, in which two documents are chosen from the same class, while the remaining document is selected from a different class. We then have computed all pair-wise cosine distance between all of the documents within the triplet. If the document calculated to be most distant is indeed from the different class, classification result is regarded as correct, implying that the representations of these documents are indeed effective in capturing their characteristics and differences. As our dataset contains 8 different classes, 56 unique combinations of the triplets exist. Randomly creating 5,000 triplets for each unique combination, we have performed the document classification task on 280,000 set of triplets.

Numerous hyperparameters are involved in training effective word2vec and doc2vec models. In order to minimize the impact of the hyperparameters in the overall performance, the proposed method, word2vec averaging method and doc2vec method are designed to share same window size of 9 and training epoch of 3. All word2vec and doc2vec training have been carried out by using Gensim library[1] in Python. Furthermore, various embedding dimensions of the document and word vectors have been tested. Starting with the dimension of 100, the dimension is increased by 100 until 1000. The proposed method is additionally influenced by an extra hyperparameter k, the number of concept clusters. In order to observe its impact on the representation performance, several values for the number of concept clusters have also been tested. Starting with 20, the value of k is increased by 10 until 400.

Furthermore, term frequency-inverse document frequency(TF-IDF) (Salton & Buckley, 1988) is commonly applied in the bag-of-words document representation for improved performance. It is a weighting scheme that readjusts the count of a word based on its frequency in the entire corpus. If a certain word occurs in every document in the corpus, it will be regarded as relatively unimportant, thus reducing its frequency. The word occurrences of the bag-of-words model with TF-IDF weighting scheme are calculated by Equation 1.

$$TF-IDF(t_i,d_j,D)=TF(t_i,d_j)\times log\frac{|D|}{|d\in D;t_i\in d|} \qquad (1)$$

where,

[1] https://radimrehurek.com/gensim/

$$(t_i, d_j, D) = (Term_i, Document_j, Corpus)$$

$$|D| = \text{Number of documents in Corpus}$$

$$|d \in D; t_i \in d| = \text{Number of documents in Corpus with } Term_i$$

As certain concepts can also occur frequently in the corpus, similar weighting scheme can also be applied to the Bag-of-Concepts method. Along with the Bag-of-Concepts method that calculates the raw frequencies of the concepts, we have also applied concept frequency-inverse document frequency(CF-IDF). Similar to TF-IDF, CF-IDF weighting scheme is calculated by Equation 2.

$$CF - IDF(c_i, d_j, D) = CF(c_i, d_j) \times log \frac{|D|}{|d \in D; c_i \in d|} \qquad (2)$$

where,

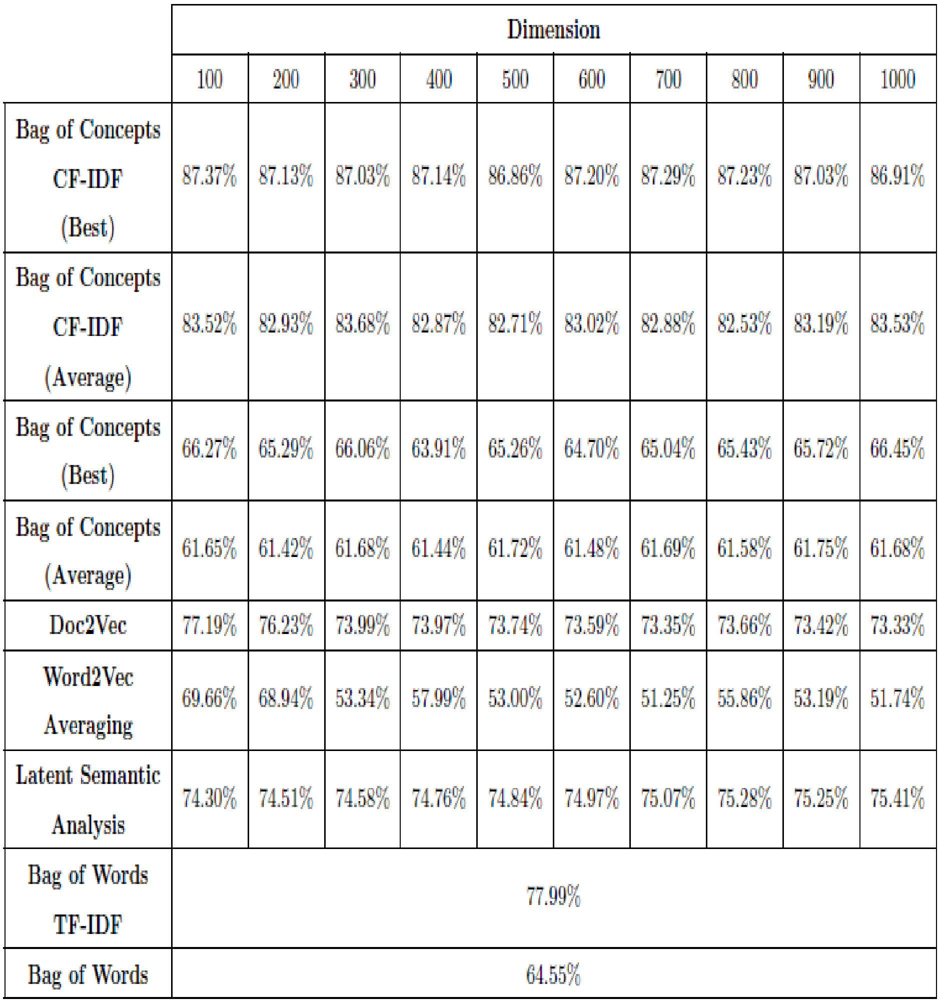
$$(c_i, d_j, D) = (Concept_i, Document_j, Corpus)|$$

$$|D| = \text{Number of documents in Corpus}$$

$$|d \in D; t_i \in d| = \text{Number of documents in Corpus with } Concept_i$$

Figure 5.2 and Table 5.1 show the classification accuracy of different document representation methods with respect to varying dimensions of the generated document vectors. For a fixed dimension, "Bag-of-Concepts CF-IDF (Best)" and "Bag-of-Concepts (Best)" indicate models with the highest accuracy amongst the models trained with different number of concept clusters, k, between 20 ~ 400. On the other hand, "Bag-of-Concepts CF-IDF (Average)" and "Bag-of-Concepts (Average)" represent average accuracies from all of the proposed models with different values of k for a fixed dimension.

Table 5.1: Accuracy of document classification task



| | Dimension | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| Bag of Concepts CF-IDF (Best) | 87.37% | 87.13% | 87.03% | 87.14% | 86.86% | 87.20% | 87.29% | 87.23% | 87.03% | 86.91% |
| Bag of Concepts CF-IDF (Average) | 83.52% | 82.93% | 83.68% | 82.87% | 82.71% | 83.02% | 82.88% | 82.53% | 83.19% | 83.53% |
| Bag of Concepts (Best) | 66.27% | 65.29% | 66.06% | 63.91% | 65.26% | 64.70% | 65.04% | 65.43% | 65.72% | 66.45% |
| Bag of Concepts (Average) | 61.65% | 61.42% | 61.68% | 61.44% | 61.72% | 61.48% | 61.69% | 61.58% | 61.75% | 61.68% |
| Doc2Vec | 77.19% | 76.23% | 73.99% | 73.97% | 73.74% | 73.59% | 73.35% | 73.66% | 73.42% | 73.33% |
| Word2Vec Averaging | 69.66% | 68.94% | 53.34% | 57.99% | 53.00% | 52.60% | 51.25% | 55.86% | 53.19% | 51.74% |
| Latent Semantic Analysis | 74.30% | 74.51% | 74.58% | 74.76% | 74.84% | 74.97% | 75.07% | 75.28% | 75.25% | 75.41% |
| Bag of Words TF-IDF | 77.99% | | | | | | | | | |
| Bag of Words | 64.55% | | | | | | | | | |

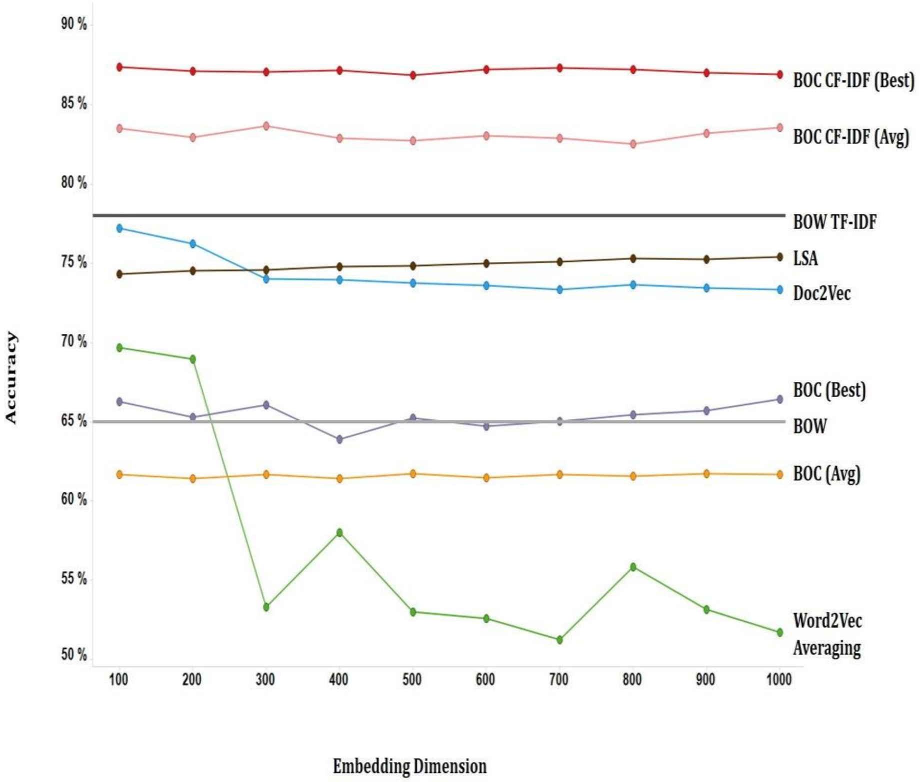


Figure 5.2: Accuracy of document classification task

As shown by Figure 8, the bag-of-concepts with CF-IDF weighting outperforms all other document representation methods. Furthermore, its performance is robust to varying embedding dimensions. The reason behind its outstanding performance lies in the fact that the bag-of-concepts with appropriate weighting scheme captures truly defining characteristics of the documents. Doc2vec, word2vec averaging, latent semantic analysis and the bag-of-words methods treat words as basic units for representing the document vectors. However, the bag-of-concepts method exploits the semantic similarities between the words to disregard redundancy between similar words, and emphasizes only the essential defining concepts within the documents. Therefore, it is capable of effectively preserving true proximity between the documents. Furthermore, previously proposed concept-based document representation methods (Bing, Jiang, Lam, Zhang & Jameel, 2015; Jiang, Bing, Sun, Zhang & Lam, 2011; Sedding & Kazakov, 2004) relied on ontologies or entity information from external sources. However, the proposed method creates data-driven concepts without relying on external ontologies.

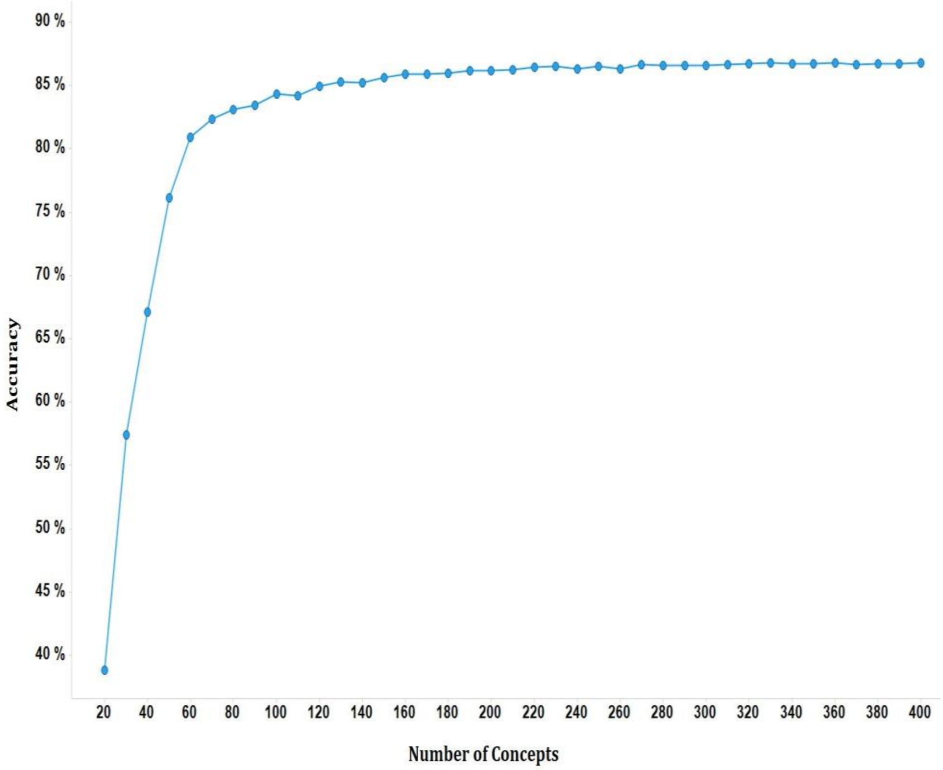
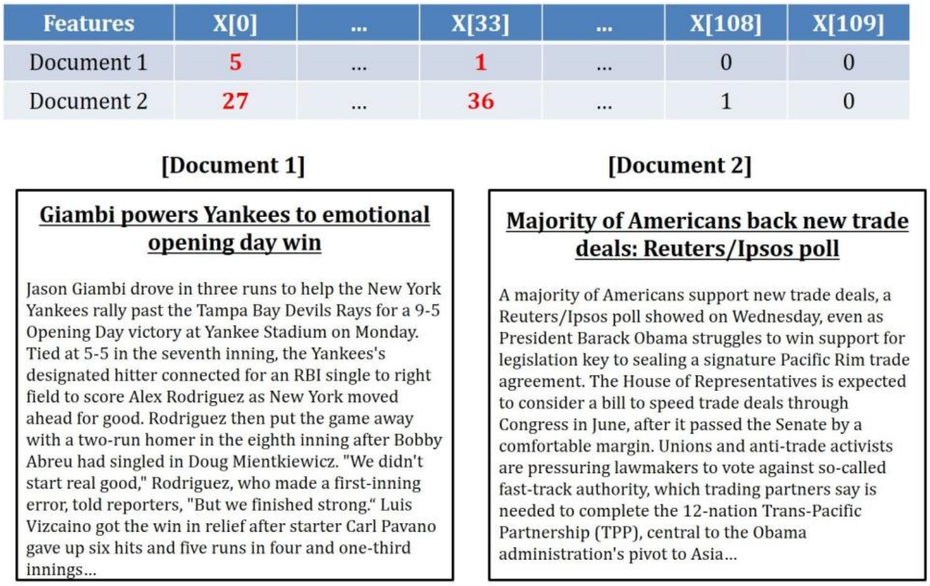


Figure 5.3: Classification accuracy with respect to the number of concept cluster

Figure 9 shows the impact of k, the number of concept clusters, on the classification accuracy. Although the classification accuracy is low when k is initially set at 20, it starts to improve significantly when k is increased up to 100, after which the classification accuracy becomes stable. Therefore, selecting appropriate value of concept clusters, along with the appropriate weighting scheme, is crucial for generating high quality document representation from the bag-of-concepts method.

## 5.2 Representation Interpretability

Unlike doc2vec, the proposed method, while offering superior document representation, is capable of providing intuitive interpretation for the generated document vectors. In order to show this representation interpretability, we will use a model, in which the words are embedded into continuous space of 500 dimensions, and are clustered into 110 concepts (k = 110). Consequently, each document vector in this model is represented as a vector with 110 features. As shown in Figure 5.4, we have selected two clearly different documents as examples.



| Features | X[0] | ... | X[33] | ... | X[108] | X[109] |
|---|---|---|---|---|---|---|
| Document 1 | 5 | ... | 1 | ... | 0 | 0 |
| Document 2 | 27 | ... | 36 | ... | 1 | 0 |

[Document 1]

**Giambi powers Yankees to emotional opening day win**

Jason Giambi drove in three runs to help the New York Yankees rally past the Tampa Bay Devils Rays for a 9-5 Opening Day victory at Yankee Stadium on Monday. Tied at 5-5 in the seventh inning, the Yankees's designated hitter connected for an RBI single to right field to score Alex Rodriguez as New York moved ahead for good. Rodriguez then put the game away with a two-run homer in the eighth inning after Bobby Abreu had singled in Doug Mientkiewicz. "We didn't start real good," Rodriguez, who made a first-inning error, told reporters, "But we finished strong." Luis Vizcaino got the win in relief after starter Carl Pavano gave up six hits and five runs in four and one-third innings...

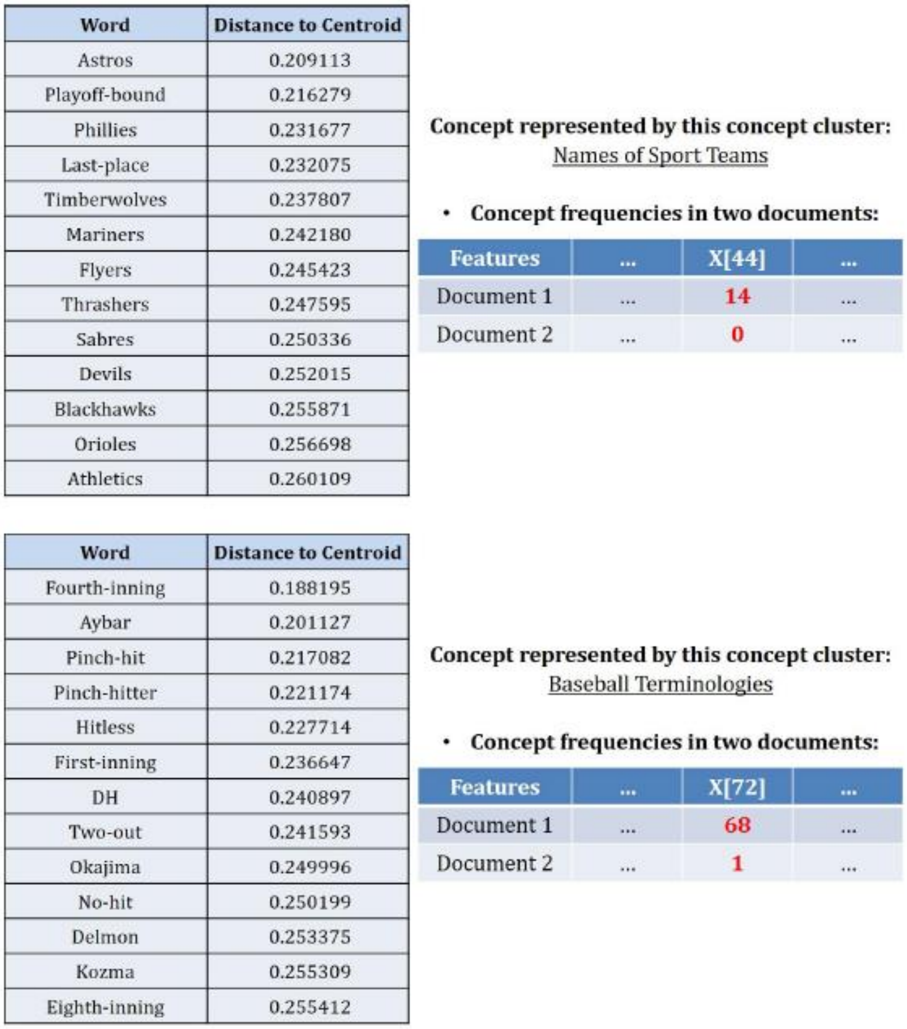
[Document 2]

**Majority of Americans back new trade deals: Reuters/Ipsos poll**

A majority of Americans support new trade deals, a Reuters/Ipsos poll showed on Wednesday, even as President Barack Obama struggles to win support for legislation key to sealing a signature Pacific Rim trade agreement. The House of Representatives is expected to consider a bill to speed trade deals through Congress in June, after it passed the Senate by a comfortable margin. Unions and anti-trade activists are pressuring lawmakers to vote against so-called fast-track authority, which trading partners say is needed to complete the 12-nation Trans-Pacific Partnership (TPP), central to the Obama administration's pivot to Asia...

Figure 5.4: Examples of interpretable document vectors

As shown in Figure 5.4, Document 1 belongs to Sports class as it discusses about an opening day win for New York Yankees, a baseball team. Document 2, on the other hand, belongs to Politics class as it discusses a recent survey regarding the Trans-Pacific Partnership agreement, a economic trade agreement between twelve countries around Pacific Rim. Both doc2vec and the proposed method can successfully classify Document 1 as a member of Sports class and Document 2 as a member of Politics class. Observing the document vectors generated from the proposed method, however, provides more insightful and profound understanding behind the result. The features of the document vectors generated from doc2vec represent the coordinates of the vectors in 500 dimensional space, but it fails to provide any clear intuitive understanding behind the meaning of each axis. The proposed method, however, successfully offers clear interpretation of the meaning behind each features.



| Word | Distance to Centroid |
|---|---|
| Astros | 0.209113 |
| Playoff-bound | 0.216279 |
| Phillies | 0.231677 |
| Last-place | 0.232075 |
| Timberwolves | 0.237807 |
| Mariners | 0.242180 |
| Flyers | 0.245423 |
| Thrashers | 0.247595 |
| Sabres | 0.250336 |
| Devils | 0.252015 |
| Blackhawks | 0.255871 |
| Orioles | 0.256698 |
| Athletics | 0.260109 |

**Concept represented by this concept cluster:**
Names of Sport Teams

- **Concept frequencies in two documents:**

| Features | ... | X[44] | ... |
|---|---|---|---|
| Document 1 | ... | 14 | ... |
| Document 2 | ... | 0 | ... |

| Word | Distance to Centroid |
|---|---|
| Fourth-inning | 0.188195 |
| Aybar | 0.201127 |
| Pinch-hit | 0.217082 |
| Pinch-hitter | 0.221174 |
| Hitless | 0.227714 |
| First-inning | 0.236647 |
| DH | 0.240897 |
| Two-out | 0.241593 |
| Okajima | 0.249996 |
| No-hit | 0.250199 |
| Delmon | 0.253375 |
| Kozma | 0.255309 |
| Eighth-inning | 0.255412 |

**Concept represented by this concept cluster:**
Baseball Terminologies

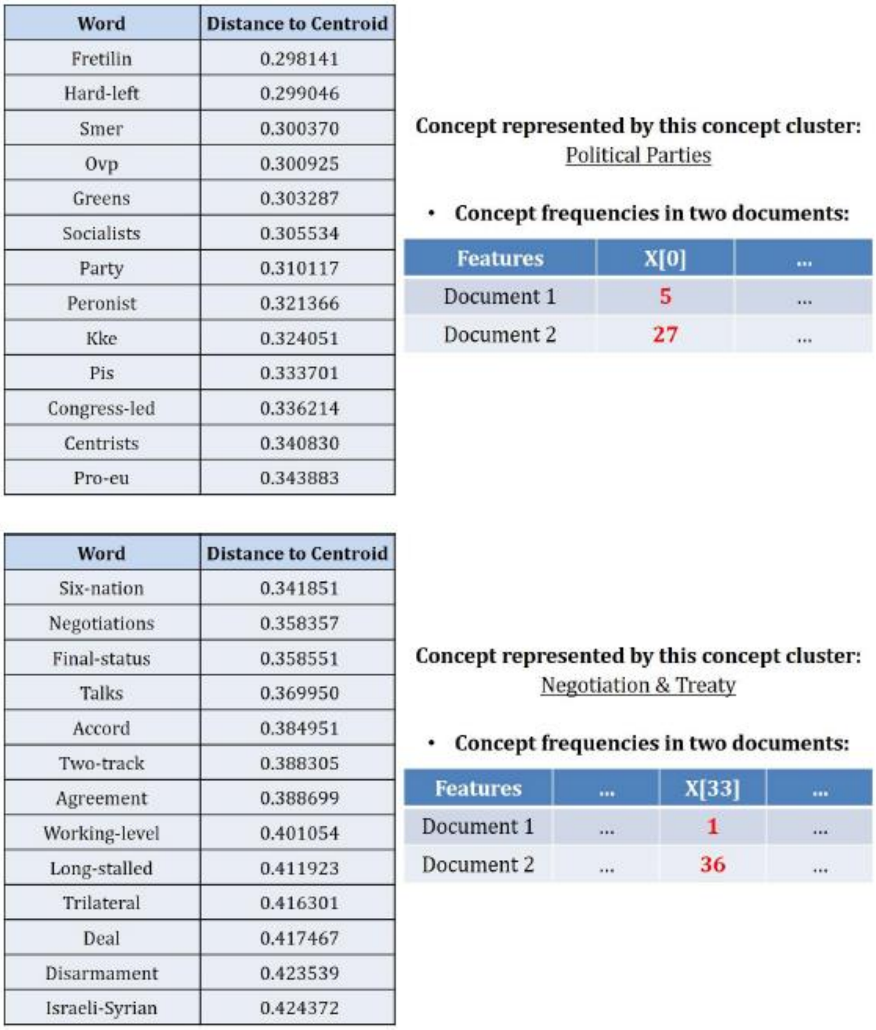
- **Concept frequencies in two documents:**

| Features | ... | X[72] | ... |
|---|---|---|---|
| Document 1 | ... | 68 | ... |
| Document 2 | ... | 1 | ... |

Figure 5.5: Concept clusters that are strongly related to Document 1

Figure 5.5 ~ 5.8 list some examples of contrasting features between two document vectors generated from the proposed method that can provide additional intuition. Looking at the words in the concept clusters depicted in Figure 5.5, we can understand that these two concept clusters contain words that are related to the names of sports teams, and to baseball terminologies, respectively. In Document 1, words belonging to the concept cluster related to the names of sports teams occurred 14 times compared to none in Document 2. Similarly, the concept cluster related to baseball terminologies occurred 68 times in Document 1, while once in Document 2. Consequently, we can understand that Document 1 contains more words related to the names of sports teams and to baseball terminologies. As Document 1 is indeed an article about a baseball game, it seems inevitable for Document 1 to have high occurrences in these two concept clusters. As these concepts are more likely to be used in a sports section of a newspaper than a politics section, Document 1, therefore, can be naturally be understood as a member of Sports class.

Looking at the words in the concept clusters depicted in Figure 12, we can understand that these two concept clusters contain words that are related to the names of political parties, and to the words that describe negotiations, respectively. In Document 2, words belonging to the concept cluster related to the names of political parties occurred 27 times compared to 5 times in Document 1. Similarly, the concept cluster related to the negotiation terms occurred 36 times in Document 2, while once in Document 1. Consequently, we can understand that Document 2 contains more words related to the names of political parties and to the concept of negotiation. As these concepts are more likely to be used in a political section of a newspaper than a sports section, Document 2, therefore, therefore, can be naturally be understood as a member of Politics class.



| Word | Distance to Centroid |
|---|---|
| Fretilin | 0.298141 |
| Hard-left | 0.299046 |
| Smer | 0.300370 |
| Ovp | 0.300925 |
| Greens | 0.303287 |
| Socialists | 0.305534 |
| Party | 0.310117 |
| Peronist | 0.321366 |
| Kke | 0.324051 |
| Pis | 0.333701 |
| Congress-led | 0.336214 |
| Centrists | 0.340830 |
| Pro-eu | 0.343883 |

**Concept represented by this concept cluster:**
Political Parties

- **Concept frequencies in two documents:**

| Features | X[0] | ... |
|---|---|---|
| Document 1 | 5 | ... |
| Document 2 | 27 | ... |

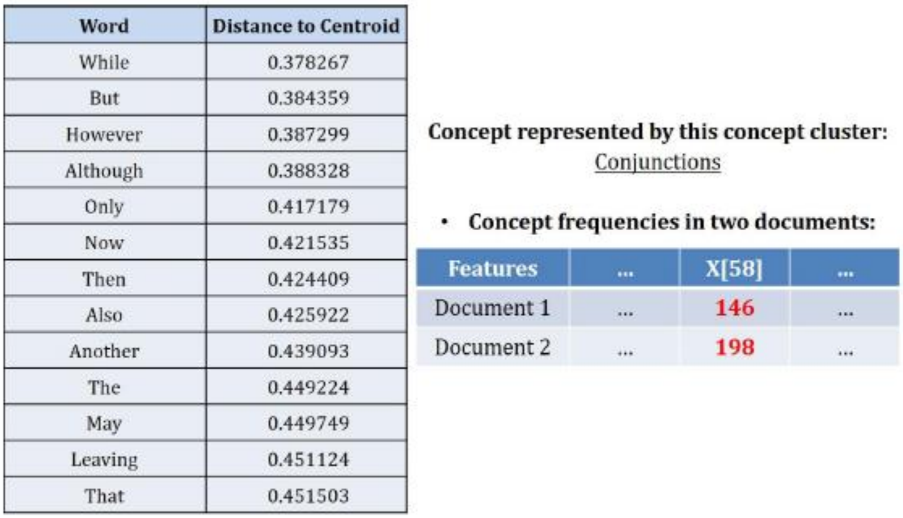
| Word | Distance to Centroid |
|---|---|
| Six-nation | 0.341851 |
| Negotiations | 0.358357 |
| Final-status | 0.358551 |
| Talks | 0.369950 |
| Accord | 0.384951 |
| Two-track | 0.388305 |
| Agreement | 0.388699 |
| Working-level | 0.401054 |
| Long-stalled | 0.411923 |
| Trilateral | 0.416301 |
| Deal | 0.417467 |
| Disarmament | 0.423539 |
| Israeli-Syrian | 0.424372 |

**Concept represented by this concept cluster:**
Negotiation & Treaty

- **Concept frequencies in two documents:**

| Features | ... | X[33] | ... |
|---|---|---|---|
| Document 1 | ... | 1 | ... |
| Document 2 | ... | 36 | ... |

Figure 5.6: Concept clusters that are strongly related to Document 2



| Word | Distance to Centroid |
|---|---|
| While | 0.378267 |
| But | 0.384359 |
| However | 0.387299 |
| Although | 0.388328 |
| Only | 0.417179 |
| Now | 0.421535 |
| Then | 0.424409 |
| Also | 0.425922 |
| Another | 0.439093 |
| The | 0.449224 |
| May | 0.449749 |
| Leaving | 0.451124 |
| That | 0.451503 |

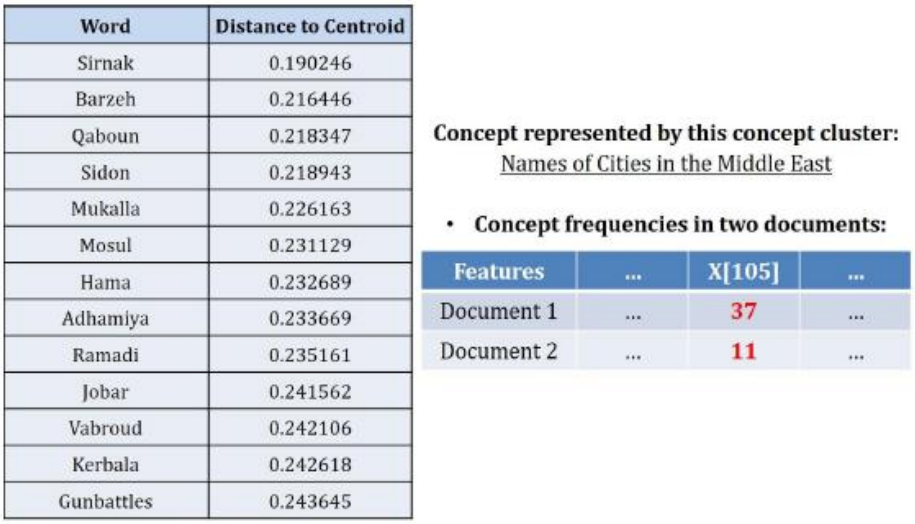
**Concept represented by this concept cluster:**
Conjunctions

- **Concept frequencies in two documents:**

| Features | ... | X[58] | ... |
|---|---|---|---|
| Document 1 | ... | 146 | ... |
| Document 2 | ... | 198 | ... |

Figure 5.7: Concept clusters that are strongly related to both documents

First four concept clusters in Figure 5.5 ~ 5.6 successfully capture the contents of documents, and provide reasons behind each document's class membership. However, not every concept clusters are effective in providing intuition behind the representation. Figure 5.7 shows the concept cluster that has occurred most frequently in both Document 1 and 2. Looking at some of the words within this concept, it becomes obvious that conjunctions are clustered into this concept cluster. As conjunctions can be common in any articles, the occurrences of this concept cluster in both documents are relatively higher than the occurrences of other concept clusters. Thus, this concept, despite its high occurrence, is irrelevant in capturing meaningful differences between these two documents. Its impact, however, can be adjusted by applying appropriate weighting scheme such as CF-IDF.



| Word | Distance to Centroid |
|---|---|
| Sirnak | 0.190246 |
| Barzeh | 0.216446 |
| Qaboun | 0.218347 |
| Sidon | 0.218943 |
| Mukalla | 0.226163 |
| Mosul | 0.231129 |
| Hama | 0.232689 |
| Adhamiya | 0.233669 |
| Ramadi | 0.235161 |
| Jobar | 0.241562 |
| Vabroud | 0.242106 |
| Kerbala | 0.242618 |
| Gunbattles | 0.243645 |

**Concept represented by this concept cluster:**
Names of Cities in the Middle East

- **Concept frequencies in two documents:**

| Features | ... | X[105] | ... |
|---|---|---|---|
| Document 1 | ... | 37 | ... |
| Document 2 | ... | 11 | ... |

Figure 5.8: Misallocated concept clusters

The concept cluster in Figure 14 represents the names of Middle Eastern cities. Although both Document 1 and 2 seem irrelevant to the cities in the Middle East, the occurrences of this concept cluster in these two documents are quite significant. Through careful observation of the words in this concept cluster, it can be discovered that such high frequency of this irrelevant concept cluster is attributed to misallocation of some irrelevant terms into this concept cluster. For example, some common words such as near and cities have been clustered into this concept cluster. Consequently, occurrence of such irrelevant yet common words in the documents has increased the frequency of the corresponding concept cluster in these document vectors without revealing their contrasting contents.

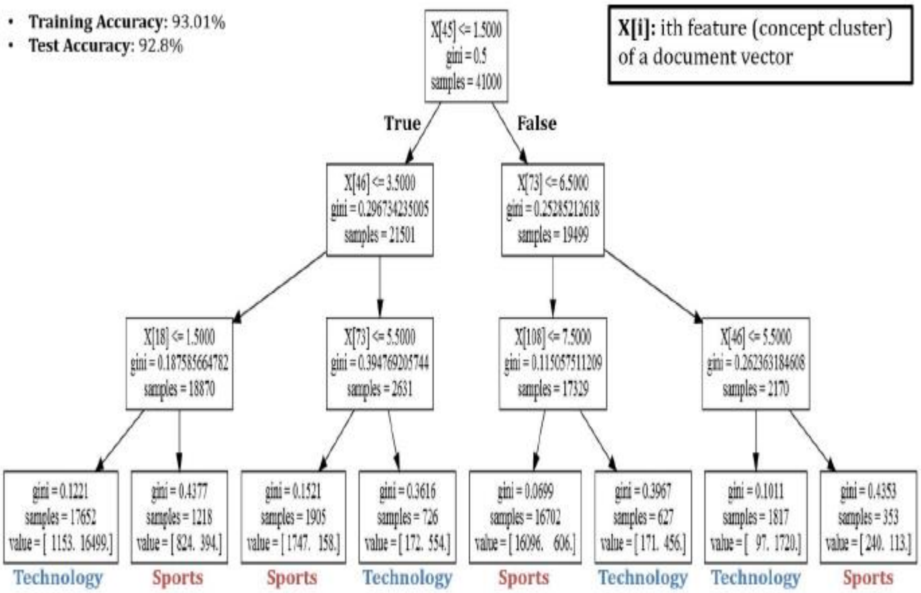
Although some of the concept clusters with high frequencies are not so intuitive in distinguishing these two document vectors, the proposed method, unlike doc2vec, is capable of providing clear interpretation behind the features of the generated document vectors. Through this representation interpretability of the generated vectors, it is now possible to understand the comprising contents of the documents, and to comprehend the similarities and the differences between the vectors.

## 5.3 Model Explainability

The proposed method can additionally provide explanatory power for a text mining model built from the generated document vectors. In order to show such model explainability, a document classifier using decision tree algorithm has been constructed to classify articles in Sports class from those in Technology class. For this decision tree, document vectors are once again represented by 110 concepts that have been constructed from the word embedding space of 500 dimension. Amongst 110 concept clusters, this decision tree seeks to identify important concept clusters that can distinguish Sports class from Technology class. Amongst 25,500 articles from each class, 20,500 articles from each class (total of 51,000) have been used to build a decision tree, while remaining 5,000 articles from each class (total of 10,000) have been used as a test set (Table 5.2). The constructed decision tree and its training and test accuracy are shown in Figure 5.9

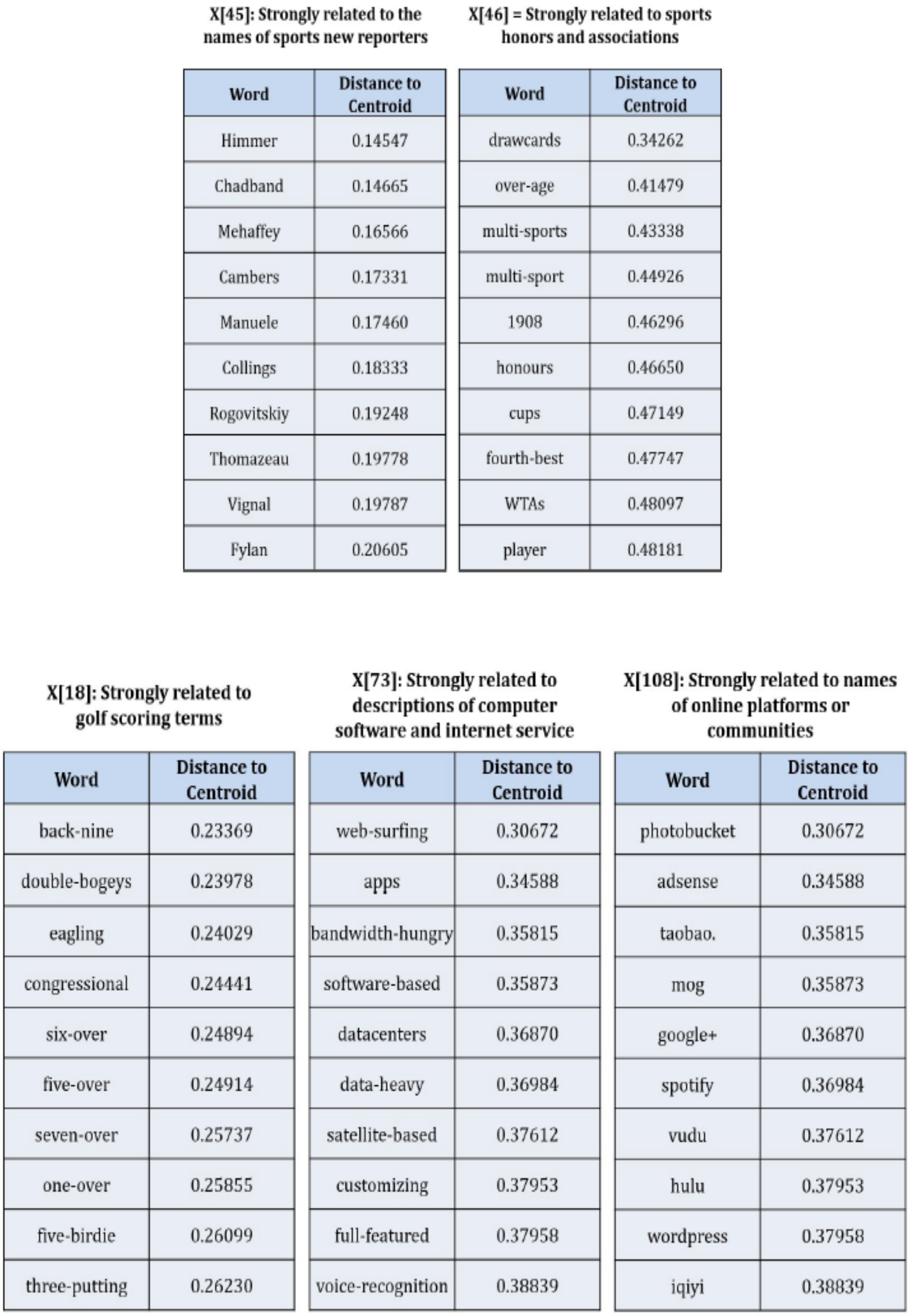
. Table 5.2: Training set and test set for decision tree

| Class | Total Number of Documents | Training Set | Test Set |
|---|---|---|---|
| Sports | 25,500 | 20,500 | 5,000 |
| Technology | 25,500 | 20,500 | 5,000 |



**Rule 1:** IF (X[45] ≤ 1.5) AND (X[46] ≤ 3.5) AND (X[18] ≤ 1.5) THEN (Y = Technology)
**Rule 2:** IF (X[45] ≤ 1.5) AND (X[46] ≤ 3.5) AND (X[18] > 1.5) THEN (Y = Sports)
**Rule 3:** IF (X[45] ≤ 1.5) AND (X[46] > 3.5) AND (X[73] ≤ 5.5) THEN (Y = Sports)
**Rule 4:** IF (X[45] ≤ 1.5) AND (X[46] ≤ 3.5) AND (X[73] > 5.5) THEN (Y = Technology)
**Rule 5:** IF (X[45] > 1.5) AND (X[73] ≤ 6.5) AND (X[108] ≤ 7.5) THEN (Y = Sports)
**Rule 6:** IF (X[45] > 1.5) AND (X[73] ≤ 6.5) AND (X[108] > 7.5) THEN (Y = Technology)
**Rule 7:** IF (X[45] > 1.5) AND (X[73] > 6.5) AND (X[46] ≤ 5.5) THEN (Y = Technology)
**Rule 8:** IF (X[45] > 1.5) AND (X[73] > 6.5) AND (X[46] > 5.5) THEN (Y = Sports)

Figure 5.9: Constructed decision tree



X[45]: Strongly related to the names of sports new reporters

| Word | Distance to Centroid |
|---|---|
| Himmer | 0.14547 |
| Chadband | 0.14665 |
| Mehaffey | 0.16566 |
| Cambers | 0.17331 |
| Manuele | 0.17460 |
| Collings | 0.18333 |
| Rogovitskiy | 0.19248 |
| Thomazeau | 0.19778 |
| Vignal | 0.19787 |
| Fylan | 0.20605 |

X[46] = Strongly related to sports honors and associations

| Word | Distance to Centroid |
|---|---|
| drawcards | 0.34262 |
| over-age | 0.41479 |
| multi-sports | 0.43338 |
| multi-sport | 0.44926 |
| 1908 | 0.46296 |
| honours | 0.46650 |
| cups | 0.47149 |
| fourth-best | 0.47747 |
| WTAs | 0.48097 |
| player | 0.48181 |

X[18]: Strongly related to golf scoring terms

| Word | Distance to Centroid |
|---|---|
| back-nine | 0.23369 |
| double-bogeys | 0.23978 |
| eagling | 0.24029 |
| congressional | 0.24441 |
| six-over | 0.24894 |
| five-over | 0.24914 |
| seven-over | 0.25737 |
| one-over | 0.25855 |
| five-birdie | 0.26099 |
| three-putting | 0.26230 |

X[73]: Strongly related to descriptions of computer software and internet service

| Word | Distance to Centroid |
|---|---|
| web-surfing | 0.30672 |
| apps | 0.34588 |
| bandwidth-hungry | 0.35815 |
| software-based | 0.35873 |
| datacenters | 0.36870 |
| data-heavy | 0.36984 |
| satellite-based | 0.37612 |
| customizing | 0.37953 |
| full-featured | 0.37958 |
| voice-recognition | 0.38839 |

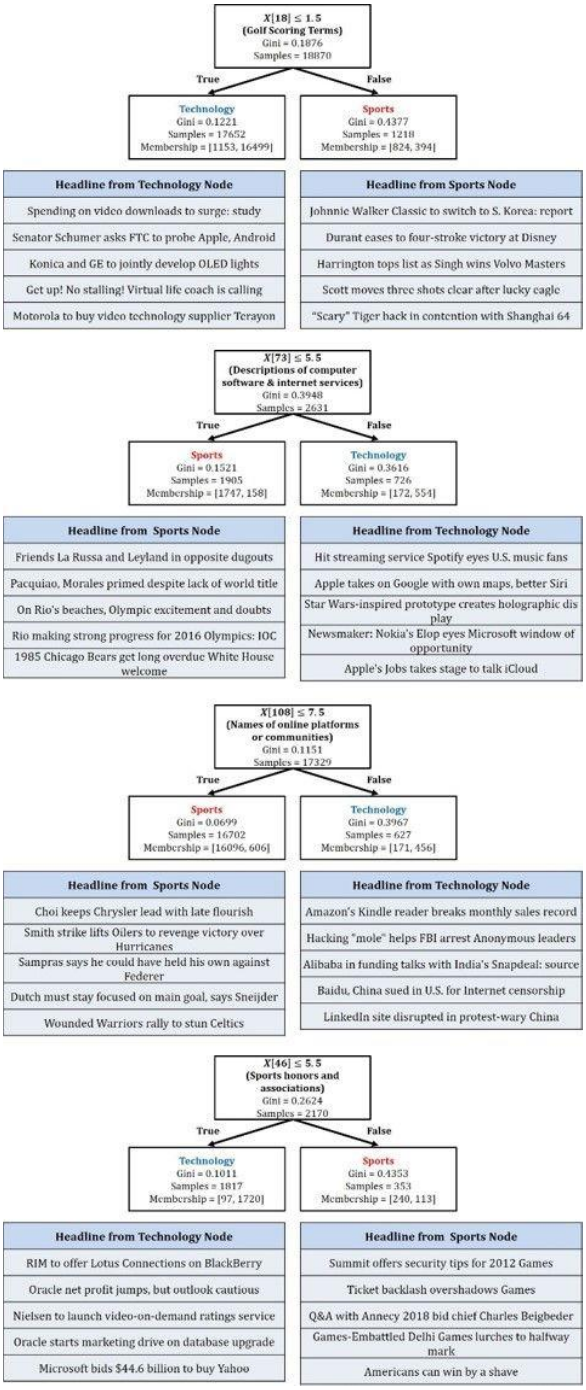
X[108]: Strongly related to names of online platforms or communities

| Word | Distance to Centroid |
|---|---|
| photobucket | 0.30672 |
| adsense | 0.34588 |
| taobao. | 0.35815 |
| mog | 0.35873 |
| google+ | 0.36870 |
| spotify | 0.36984 |
| vudu | 0.37612 |
| hulu | 0.37953 |
| wordpress | 0.37958 |
| iqiyi | 0.38839 |

Figure 5.10: Concept clusters of each nodes

Unlike a decision tree generated from doc2vec vectors, this decision tree generated from the bag-of-concepts document vectors provides an intuitive explanation behind the tree. As each node of the tree represents a specific concept, we can understand the operating logic and the intrinsic characteristics of the classifier and the dataset.

Figure 5.10 lists some concepts that the decision tree uses to classify Sports class from Technology class. First splitting node (root) occurs at 45th concept cluster of the document vectors. Top 10 words in this cluster that are closest to the centroid seem to indicate that this concept cluster contains the names of people. Through exploring Reuter website, we have discovered that these words are indeed the names of the reporters, who mainly write sports articles. Consequently, it becomes evident that this classifier considers the names of the reporters as an important criterion for differentiating two classes. Next, we will look at the splitting nodes prior to the leaf nodes. Looking at the left most splitting node, we find that if the value in the 18th feature of a document vector is less than 1.5, corresponding document belongs to Technology class, while if it is bigger than 1.5, it belongs to Sports class. This decision rule becomes intuitively clear if we look into the corresponding concept cluster. From Figure 5.10, we can identify that the terms strongly related to golf scores are clustered into this concept. This node, consequently, classifies the documents according to the occurrences of the golf scoring terms. Looking at the actual headlines of the documents that are being classified at this node (Figure 5.11), we indeed see that this node successfully manages to classify golf-related articles from other articles.



| Headline from Technology Node |
|---|
| Spending on video downloads to surge: study |
| Senator Schumer asks FTC to probe Apple, Android |
| Konica and GE to jointly develop OLED lights |
| Get up! No stalling! Virtual life coach is calling |
| Motorola to buy video technology supplier Terayon |

| Headline from Sports Node |
|---|
| Johnnie Walker Classic to switch to S. Korea: report |
| Durant eases to four-stroke victory at Disney |
| Harrington tops list as Singh wins Volvo Masters |
| Scott moves three shots clear after lucky eagle |
| "Scary" Tiger back in contention with Shanghai 64 |

| Headline from Sports Node |
|---|
| Friends La Russa and Leyland in opposite dugouts |
| Pacquiao, Morales primed despite lack of world title |
| On Rio's beaches, Olympic excitement and doubts |
| Rio making strong progress for 2016 Olympics: IOC |
| 1985 Chicago Bears get long overdue White House welcome |

| Headline from Technology Node |
|---|
| Hit streaming service Spotify eyes U.S. music fans |
| Apple takes on Google with own maps, better Siri |
| Star Wars-inspired prototype creates holographic dis play |
| Newsmaker: Nokia's Elop eyes Microsoft window of opportunity |
| Apple's Jobs takes stage to talk iCloud |

| Headline from Sports Node |
|---|
| Choi keeps Chrysler lead with late flourish |
| Smith strike lifts Oilers to revenge victory over Hurricanes |
| Sampras says he could have held his own against Federer |
| Dutch must stay focused on main goal, says Sneijder |
| Wounded Warriors rally to stun Celtics |

| Headline from Technology Node |
|---|
| Amazon's Kindle reader breaks monthly sales record |
| Hacking "mole" helps FBI arrest Anonymous leaders |
| Alibaba in funding talks with India's Snapdeal: source |
| Baidu, China sued in U.S. for Internet censorship |
| LinkedIn site disrupted in protest-wary China |

| Headline from Technology Node |
|---|
| RIM to offer Lotus Connections on BlackBerry |
| Oracle net profit jumps, but outlook cautious |
| Nielsen to launch video-on-demand ratings service |
| Oracle starts marketing drive on database upgrade |
| Microsoft bids $44.6 billion to buy Yahoo |

| Headline from Sports Node |
|---|
| Summit offers security tips for 2012 Games |
| Ticket backlash overshadows Games |
| Q&A with Annecy 2018 bid chief Charles Beigbeder |
| Games-Embattled Delhi Games lurches to halfway mark |
| Americans can win by a shave |

Figure 5.11: Headlines of documents in each leaf node

Similar results follow for other remaining nodes. Through these nodes, we can understand that golf and sports associations are two major concepts that this classifier uses to differentiate the documents in Sport class from those in Technology class. Similarly, we realize that computer software related terms and the names of online platforms are two major concepts that this classifier utilizes to identify the documents in Technology class from those in Sports class. As shown by Figure 5.11, the headlines of the articles that are distinguished at these nodes further substantiate the importance of these concepts in the classifier as they appear to be strongly related to their corresponding concept clusters. Although similar classification task can be carried out by doc2vec, it, unlike the proposed method, fails to provide any intuitive explanation behind the operating logic of the classifier or the unique characteristics of the given dataset.

# Chapter 6. Conclusion

This paper proposes the bag-of-concepts method for representing document vectors, through which the advantages of the bag-of-words method and doc2vec are integrated to overcome their weaknesses. Utilizing semantic similarity of the word vectors, the proposed method uses concepts as basis for representing document vectors, thereby more effectively capturing true defining characteristics of the documents. Furthermore, the proposed method maintains the low dimensionality of doc2vec, while providing intuitive representation interpretability at the same time. With intuitive representation interpretability, we can acquire more explicit and profound understanding of the document vectors and their differences.

If the proposed method is applied in specific text mining task such as document classification task, we can furthermore comprehend the operating logic and unique characteristics behind the built models. Consequently, even those who aren’t experts in text mining and data mining can easily understand and accept the constructed model and its constituting vectors. Due to these representation interpretability and model explainability, the proposed method can be applied in solving various real business problems, in which document representation itself is not the only issue. Information retrieval system is also another field, in which the bag-of-concepts can be applied. Previously, query expansion based on semantic similarities have been used to calculate more accurate distance between a given query and documents. Through representing the documents with the bag-of-concepts method, simple information retrieval techniques can be used to retrieve matching documents without relying on query expansion.

In this paper, the labels of the concept clusters have been manually determined. In future works, however, we will explore ways to label the concept clusters semi-automatically or automatically, providing more objective labels for the concept clusters. Furthermore, we will also compare the impacts of various clustering algorithms in the quality of the generated concept clusters. With further exploration, we hope that the bag-of-concepts will establish itself as a fundamental building block for solving various text mining problems arising from real business environment.

# Bibliography

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval volume 463. ACM press New York.

Bansal, M., Gimpel, K., & Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. In ACL (2) (pp. 809-815).

Bing, L., Jiang, S., Lam, W., Zhang, Y., & Jameel, S. (2015). Adaptive concept resolution for document representation and its applications in text mining. Knowledge-Based Systems, 74, 1-13.

Cao, L., & Wang, F. (2015). Robust latent semantic exploration for image retrieval in social media. Neurocomputing, 169, 180-184.

Cui, Z., Shi, X., & Chen, Y. (2016). Sentiment analysis via integrating distributed representations of variable-length word sequence. Neurocomputing, 187, 126-132.

Dai, A. M., Olah, C., Le, Q. V., & Corrado, G. S. (2014). Document embedding with paragraph vectors. In NIPS Deep Learning Workshop.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391.

Harris, Z. S. (1954). Distributional structure. Word, 10, 146-162.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.

Huang, A. (2008). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (pp. 49-56).

Jiang, S., Bing, L., Sun, B., Zhang, Y., & Lam, W. (2011). Ontology enhancement and concept granularity learning: keeping yourself current and adaptive. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1244-1252). ACM.

Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. In Journal of Machine Learning Research (pp. 37-53).

Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006). Detecting spam blogs: A machine learning approach. In Proceedings of the National Conference on Artificial Intelligence (p. 1351). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 volume 21.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. Journal of Machine Learning Research, 9, 85.

Manning, C. D., & Schtze, H. (1999). Foundations of statistical natural language processing volume 999. MIT Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In Proceedings of workshop at international conference on learning representations (pp. 1-12).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Ren, M., Kiros, R., & Zemel, R. S. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems (pp. 2935-2943).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24, 513-523.

Sayeedunnissa, S. F., Hussain, A. R., & Hameed, M. A. (2013). Supervised opinion mining of social network data using a bag-of-words approach on the cloud. In Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012) (pp. 299-309). Springer.

Sedding, J., & Kazakov, D. (2004). Wordnet-based text document clustering. In Proceedings of the 3rd workshop on robust methods in analysis of natural language data (pp. 104-113). Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, 141-188.

Wu, L., Hoi, S. C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. Image Processing, IEEE Transactions on, 19, 1908-1920.

Xing, C., Wang, D., Zhang, X., & Liu, C. (2014). Document classification with distributions of word vectors. In Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA) (pp. 1-5). IEEE.

Xue, B., Fu, C., & Shaobin, Z. (2014). A study on sentiment computing and classification of sina weibo with word2vec. In Big Data (BigData Congress), 2014 IEEE International Congress on (pp. 358{363). IEEE.

Zhong, S. (2005). Efficient online spherical k-means clustering. In Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (pp. 3180{3185). IEEE volume 5.

# 초 록

텍스트마이닝을 위해 주로 두 가지 문서표현 방법론이 사용된다. Bag-of-Words (BoW)의 경우, 문서 내 단어의 빈도수로 문서를 표현하기 때문에 직관적이고 해석 가능한 문서 벡터를 생성해준다. 하지만, 단어의 수가 증가할수록 차원의 저주가 발생하여, 문서 간의 유사도를 제대로 보존해주지 못한다는 단점이 있다. 또한, BoW는 모든 단어를 독립적으로 고려하기 때문에, 의미가 비슷한 단어 간의 관계를 반영하지 못한다. 최근 제안된 doc2vec은 신경망 알고리즘을 활용하여, 문서 간의 유사도를 효과적으로 보존하는 문서 벡터를 생성해준다. 하지만 생성된 문서 벡터들이 직관적이지 않고 해석할 수 없다는 단점이 있다.

본 연구에서는 Bag-of-Concepts (BoC)라는 새로운 문서표현 방법론을 제시한다. BoC는 word2vec으로 생성된 단어 벡터를 군집화하여, 비슷한 의미의 단어들을 하나의 개념으로 표현한다. 그리고 각 문서 벡터를 개념들의 빈도수로 표현한다. BoC에 Concept frequency – inverse document frequency와 같은 적절한 가중치 법을 적용한다면, 기존의 문서표현 방법들보다 훨씬 효과적으로 문서 간의 유사도를 표현할 수 있다는 것을 확인했다. 그리고 생성된 문서 벡터와 이를 활용해서 학습된 텍스트마이닝 모델 또한 직관적으로 해석 가능하다는 장점이 있다.

**주요어**: Bag-of-Concepts; 해석 가능한 문서표현법; word2vec 군집화
**학 번**: 2014-22642