Int. J. Advance Soft Compu. Appl, Vol. 10, No. 2, July 2018
ISSN 2074-8523

Topic Detection using Fuzzy C-Means with Nonnegative Double Singular Value Decomposition Initialization

Hamimah Alatas[1], Hendri Murfi[1], and Alhadi Bustamam[1]

[1]Department of Mathematics, Universitas Indonesia
e-mail: hamimah.alatas@sci.ui.ac.id, hendri@ui.ac.id, alhadi@sci.ui.ac.id

Abstract

Topic Detection or topic modeling is a process of finding topics in a collection of textual data. Detecting topic for a very large document collection hardly done manually. Therefore, we need an automatic method, one of which is a clustering-based method such as fuzzy c-means (FCM). The standard initialization method of FCM is a random initialization which usually produces different topics for each execution. In this paper, we examine a nonrandom initialization method called nonnegative double singular value decomposition (NNDSVD). Besides the advantage of non-randomness, our simulations show that the NNDSVD method gives better accuracies in term of topic recall than both random method and another existing singular value decomposition-based method for the problem of sensing trending topic on Twitter.

Keywords: *Topic detection, topic modeling, fuzzy c-means, initialization, singular value decomposition, Twitter.*

1 Introduction

Nowadays, information and communication technology are growing very rapidly. The internet is one of the proofs of these developments. People can get information via internet easily now. Moreover, with the increasing flow of information on the internet, many social networks have sprung up.

Social networks, i.e., social media, include textual data associated with the dissemination of information in a very large volume. One of the popular social media for information dissemination is Twitter. Twitter facilitates users to send or read text-based information known as tweets. Every day a variety of information on various topics spread by many people on Twitter. The information may talk about specific events [1] or in specific locations [2].

From Twitter, one could know what topics currently popular with reading the contents of tweets. However, due to time constraint, it is almost impossible for users to read the contents of all tweets or to know the topics of all tweets manually. Thus, an automatic topic detection method is required. The topic detection problem is also known as topic modeling. Topic modeling is a process used to analyze words in a collection of textual data to determine the topics in the collection, how the topics relate to each other, and how they change over time [3].

One of the topic modeling methods commonly used is a clustering-based method. Clustering is a data grouping technique which purposes are to group the data, so members of each cluster are more homogeneous and more like each other than with different cluster members. In topic detection problem, the center of groups or centroids is interpreted as topics [4].

The clustering-based method is a promising method for topic detection due to its simplicity and scalability for a large volume of data. The common clustering algorithm for topic detection is k-means [5, 6, 7]. K-means divides the data into k clusters, and each data is grouped to the nearest cluster. In other words, each data is a member of one cluster only. Therefore, k-means assumes that each data contains one topic only. This assumption is rather weak, because, in the real-world cases, data may have more than one topic.

We consider another clustering-based method for topic detection called soft clustering. Soft clustering is a clustering approach that groups data into multiple clusters based on a weighted parameter in the form of a fuzzy number. Therefore, soft clustering is a clustering approach that fulfills the assumption that a document may consist of several topics. The popular soft clustering method is Fuzzy C-Means (FCM) [8, 9, 10].

Generally, in the standard FCM method, the cluster center is initiated randomly. However, this random initialization usually produces different topics for each execution. Therefore, we consider non-random initialization methods. One of the non-random initialization methods of FCM for topic modeling is a singular value decomposition (SVD)-based method [11]. Given the word by tweet matrix, this method produces a word by latent semantic matrix. Next, the method uses the word by latent matrix to initialize the centroids by first converting its negative elements to zero.

In this paper, we examine another SVD-based initialization method called nonnegative double singular value decomposition (NNDSVD) [12]. In [12], the authors used NNDSVD to initialize the nonnegative matrix factorization algorithms. We adopt the author work to use the NNDSVD to initialize the FCM algorithm for topic detection. NNDSVD consists of two SVD processes. The first process decomposes the word by tweet matrix with SVD. Then, the second process describes the positive part of the word by the latent semantic matrix and the latent semantic by tweet matrix from the first process using SVD. Besides the advantage of non-randomness, our simulations show that the NNDSVD method gives better accuracies in term of topic recall than both random method and SVD-based method for the problem of sensing trending topic on Twitter.

The rest of the paper is organized as follows: Section 2 describes the related work. In Section 3 reviews methodology of FCM, SVD, and NNSVD. In Section 4, we describe the implementation of Fuzzy C-Means initialization method. Section 5 gives our case study and the simulation results. We give conclusions in Section 6.

2 Related Work

Topic detection is an important work especially to identify topics in discussion, lecture, or essay. Thus, many types of research have been dedicated to developing and improve topic detection. Allan, J. W. studied automatic topic detection method to help the topic-searching process in a large volume of data. The popular methods for topic detection are Nonnegative Matrix Factorization [13] and Latent Dirichlet Allocation [14, 3]. Non-negative matrix factorization (NMF) was proposed by Lee and Seung in 1999. NMF has become a widely used technique over the past decade in machine learning and data mining fields. The most significant properties of NMF are non-negative, intuitive and part based representative. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. For all the methods already discussed, there is an alternative method used for the topic detection, namely the clustering-based method [4]. There have been several papers that examine this topic detection using clustering.

Most clustering-based methods for topic detection are the k-mean algorithm [5, 6, 7]. Many papers have studied and discussed k-means as a topic detection method. In [5], k-means as the approach to the topic detection challenge which organized by the 2014 SNOW workshop is discussed. Other research has also studied how to improve the performance of k-means. For instance, paper [6] combined SVD and K-Means Clustering method for twitter topic detection and paper [7] discussed Batch Mini algorithm combination with k-means. However, the problem of k-means has always arisen, it assumes that every data has only one topic, so it does not produce the best result.

Moreover, paper [8], [9], and [10] discussed and explained clustering method using Fuzzy C-Means (FCM). From these research, it is concluded that FCM can solve the one-topic problem of k-means.

There is the development of topic detection using FCM. Mursidah and Murfi have developed FCM algorithm for topic detection by using the Singular Value Decomposition (SVD)-based initialization problem [11]. SVD is a matrix factorization method that factorizes a $M \times N$ matrix A into a $M \times M$ orthogonal matrix U, a $M \times N$ pseudodiagonal matrix Σ, and a $N \times N$ orthogonal matrix V^T [15]. The orthogonal matrix U from SVD will be used to initialize the centroids of FCM. In advance, FCM will converting the negative elements from matrix U to zero. Besides a non-random advantage, their simulation also shows that the accuracy of FCM with the SVD-based initialization is better than that of FCM with the random initialization. However, this strategy loses some information by converting negative elements to zeros. This research will use Non-Negative Double Singular Value Decomposition (NNDSVD) initialization [12].

3 Sensing Trending Topics on Twitter

In general, the processes of sensing trending topics on Twitter as follows: firstly, users provide a set of words or location information to filter tweets containing at least one of them. Moreover, users determine a time slot for detecting trending topics, e.g., every 10 minutes. At the end of each time slot, the system gives topics in the time slot. Given tweets, the processes of topic detection start by extracting features from the tweets. Next, we apply fuzzy c-means (FCM) to detect topics from the tweets (Section 3.1). The standard method to initialize the centroids of FCM is random initialization. An alternative non-random initialization is singular value decomposition (SVD)-based initialization (Section 3.2).

3.1 Fuzzy C-Means

Fuzzy c-means (FCM) is a clustering method that allows each data point to belong to multiple clusters with varying degrees of membership. The advantage of using fuzzy c-means is clustering a dataset into c clusters where each data may update more than one centroids.

Suppose A = $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n\}$ is a dataset and Q = $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_c\}$ is a center of clusters or centroids. Each data has a variable indicator $M = [m_{ik}], m_{ik} \in [0,1]$ which denotes the degree of membership of data $\boldsymbol{a}_i$ to cluster $\boldsymbol{q}_k$. The cumulative weighted distance between data and centroids are expressed as following objective function:

$$J(A, Q, M, c, w) = \sum_{i=1}^{c} \sum_{k=1}^{n} m_{ik}^{w} \, \|\boldsymbol{a}_k - \boldsymbol{q}_i\|^2 \tag{1}$$

where c is number of clusters ($2 \leq c \leq n$), n is number of data, w is degrees of fuzziness ($w > 1$), $M = [m_{ik}]$ is a membership matrix where the entry m_{ik} is the membership level between i-th data and k-th centroid [8]. The membership degrees' values have the following constraints:

$$0 \leq m_{ik} \leq 1 \tag{2}$$

$$\sum\nolimits_{i=1}^{c} m_{ik} = 1 \tag{3}$$

$$0 \leq \sum\nolimits_{k=1}^{n} m_{ik} \leq n \tag{4}$$

According to the Lagrange multiplier theory, the Lagrange function of J can be formed as follows:

$$L = \sum_{i=1}^{c} \sum_{k=1}^{c} (m_{ik})^{w} \, \|a_k - q_i\|^2 + \lambda \left(\sum_{k=1}^{c} (m_{ik}) - 1 \right) \tag{5}$$

where λ represent Lagrange multiplier [8]. The optimal conditions of the objective function will be achieved if parameters m_{ik} and $\boldsymbol{q}_i$ is optimum. The optimum values of m_{ik} and $\boldsymbol{q}_i$ are obtained iteratively by finding the differentiation of the Lagrange function J for each parameter and setting the differentiation equals to zero. In each iteration the value of m_{ik} and $\boldsymbol{q}_i$ are as follows:

$$\begin{gathered} m_{ik} = \frac{\left(\frac{1}{\|\boldsymbol{a}_k - \boldsymbol{q}_i\|^2} \right)^{\frac{1}{w-1}}}{\sum_{i=1}^{c} \left(\frac{1}{\|\boldsymbol{a}_k - \boldsymbol{q}_i\|^2} \right)^{\frac{1}{w-1}}} \\ \boldsymbol{q}_i = \frac{\sum_{k=1}^{n} m_{ik}^{w} \, \boldsymbol{a}_k}{\sum_{k=1}^{n} m_{ik}^{w}} \end{gathered} \tag{6}$$

Fig. 1 describes the algorithm of fuzzy C-means method where we will compute m_{ik} and $\boldsymbol{q}_i$. First we input A, c, w, T for max iteration and ε for error then we initialize $\boldsymbol{q}_i$. After that we compute m_{ik} and J_t to get update value $\boldsymbol{q}_i$ and back to compute over and over until $t > T$ or $|J_t - J_{t-1}| < \varepsilon$.

Algorithm 1. Fuzzy C-Means

```
Input  : A, c, w, T, $\varepsilon$
Output : $\boldsymbol{q}_i$
1. set t = 0
2. Inisialize $\boldsymbol{q}_i$
3. Compute $m_{ik}$
4. Compute $J_t$
```

```
5. t = t + 1
6. Update $\boldsymbol{q}_i$
7. Update $m_{ik}$
8. Compute $J_t$
9. If $t > T$ or $|J_t - J_{t-1}| < \varepsilon$ then stop
   else go to step 5
```

Fig 1: Algorithm of FCM

3.2 Singular Value Decomposition-Based Initialization

In general, FCM algorithm uses a random initialization where $\boldsymbol{q}_i$ is initialized randomly. However, this method will have a different result in each run. This chapter will explain on how to do a non-random initialization that is SVD-based initialization.

SVD is a matrix factorization method that factorizes a $M\ x\ N$ matrix A into a $M\ x\ M$ orthogonal matrix U, a $M\ x\ N$ pseudo diagonal matrix Σ, and a $N\ x\ N$ orthogonal matrix V^T [15]. In general, SVD steps can be explained as below:

1. Calculate the matrix AA^T.
2. Form matrix U whose columns are orthonormal eigenvectors $[u_1, u_2, \dots, u_M]$ corresponding to the eigenvectors of AA^T.
3. Form matrix Σ with its main diagonal values is the singular values of AA^T. Those singular values are sorted from the largest to the smallest.
4. Form matrix V from matrix A and matrix U.

An illustration of matrix factorization using SVD can see in Fig 2

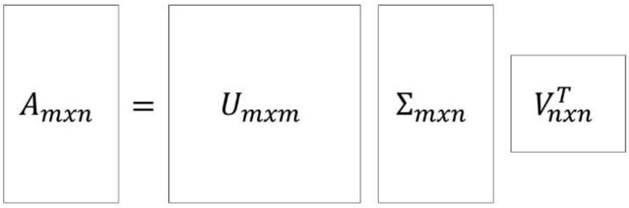


Fig 2: An illustration of SVD

The singular value of the matrix Σ is sequenced from the largest to the smallest, so by taking the first p row and p column of the matrix Σ it can produce the best possible of p-rank approximation for matrix A. The mxn matrix Σ is replaced by a pxp matrix Σ that has the most significant singular values. The retrieval of p row and p column in Σ matrix not only elithe the minates the zero values, but also eliminates some relatively small singular values [15]. This process called Truncated SVD. An illustration of Truncated SVD can be seen in Fig 3.

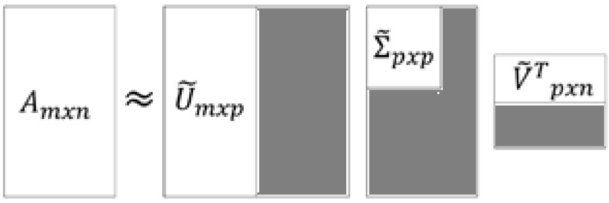


Fig 3: An illustration of Truncated SVD

Using truncated SVD, the size of an orthogonal matrix U is mxp, where p can be set equal to the number of clusters. This orthogonal matrix U is then being used as the initialization of cluster centers in the FCM algorithm. However, because SVD method could generate a matrix with negative elements, then these negative elements are modified to zeros [11].

4 Fuzzy C-Means with NNDSVD Initialization

This chapter will explain the proposed initialization method, i.e., nonnegative double singular value decomposition (NNDSVD). Firstly, it will explain the connection of chapter 3.1 and this chapter as the introduction.

NNDSVD is a two-process SVD method. The first process of this method is deciphering the matrix A with the SVD. Then, the second process is to describe the positive part of the matrix U and V in the first process with SVD [12]. In general, NNDSVD steps can be explained as below:

1. Calculate the main triplet of matrix A, namely $A = [\sigma_k, \boldsymbol{u}_k, \boldsymbol{v}_k]$
2. Form the matrix $\{C^{(j)}\}_{j=1}^{k}$ obtained from the pair of vectors, that is,

$$C^{(j)} = \boldsymbol{u}_j \boldsymbol{v}_j{}^T \tag{7}$$

3.
4. Extract positive parts from each triplet C_+ from C and $[\boldsymbol{u}_+, \boldsymbol{v}_+]$ from $[\boldsymbol{u}, \boldsymbol{v}]$
5. The expansion of the singular value of $C +$ and C is:

$$\begin{aligned} C_+ &= \mu_+ \hat{x}_+ \hat{y}_+^T + \mu_- \hat{x}_- \hat{y}_-^T, \\ C_- &= \epsilon_+ \hat{x}_+ \hat{y}_-^T + \epsilon_- \hat{x}_- \hat{y}_+^T, \end{aligned} \tag{8}$$

to initialize the non-negative matrix W and H.

The result of the NNDSVD algorithm is the non-negative matrix W and H. The W matrix of the NNDSVD results will be used for the initialization of the FCM algorithm. Fig. 4 describes the detailed algorithm of NNDSVD.

NNDSVD can also be used for initialization process [12]. In [12], the authors used NNDSVD to initialize the nonnegative matrix factorization algorithms. In this research, we used NNDSVD to initializing the cluster centers in the FCM algorithm. The result of NNDSVD algorithm is a matrix W with the size of mxk and matrix H with size of kxn. If k is set equal to the number of clusters, then the matrix W can be used to inisialize the cluster centers. Different from SVD, the elements of matrix W of NNDSVD has non-negative values so there is no need to convert those to zeros.

Algorithm 2. Nonnegative Double Singular Value Decomposition

```
Input: a nonnegative mxn matrix $A$, an integers $k < min\ (m, n)$
Output: a nonnegative $mxk$ matrix $W$ and a nonnegative $kxn$ matrix $H$
1. count $k$ triplet main $A$: $[U, S, V]$
2. initialization $w_{p1} = \sqrt{s_{11}} x u_{p1}\ and\ h_{1q} = \sqrt{s_{11}} x {v^T}_{1q}$ for $p = 1, \dots, m$
   and $q = 1, \dots, n$
3. for $j = 2{:}\, k$
            $x = u_{pj}, y = v_{qj}\ for$ p = 1, … , m and q = 1, …, n
   find $x_+, y_+, x_-, y_-$
   find $||x_+||, ||\ y_+||, ||x_-||, ||y_-||$
   $\mu_+ = ||x_+||||\ y_+||\ and\ \mu_- = ||x_-||, ||y_-||$
   If $\mu_+ > \mu_-$
     u = $\frac{x_+}{||x_+||}$, v = $\frac{y_+}{||y_+||}$, $\Sigma = \mu_+$
   else
     u = $\frac{x_-}{||x_-||}$, v = $\frac{y_-}{||y_-||}$, $\Sigma = \mu_-$
   end
   $w_{pj} = \sqrt{s_{jj} x \Sigma} x u\ and\ h_{jq} = \sqrt{s_{jj} x \Sigma} x v^t$ for $p = 1, \dots, m\ and\ q =$
   $1, \dots, n$
4. end
```

Fig 4: Algorithm of NNDSVD

5 Experiments and Results

In this section, we describe first the general stages of fuzzy c-means implementation with NNDSVD initialization as shown in Fig 5. Then, we present the simulations and discuss the results.

5.1 Data Preparation

In this research, the evaluation of this method focuses on a scenario of sensing real-world topics as discussed in Section 3. We use three large datasets collected from Twitter, for which the sets of ground truth topics have been produced by examining news stories appearing in the mainstream media. The datasets are the English FA CUP Final which is the English domestic football season, the Super Tuesday (ST) primaries which is a part of the presidential nomination race of the US Republican Party, and the US Elections that took place in November 2012 [1]. The datasets are textual data with JSON format that needs to be adjusted before retrieving the content. The description of the datasets can be seen in Fig 6, Fig. 7, and Fig. 8. From Fig. 6, we see that the FA CUP dataset consists of 13 document collections from 13-time slots. Fig. 7 describes the US Election dataset that has 26 document collections from 26-time slots. Fig. 8 shows the information on the US Super Tuesday dataset that consists of eight document collections from eight-time slots.

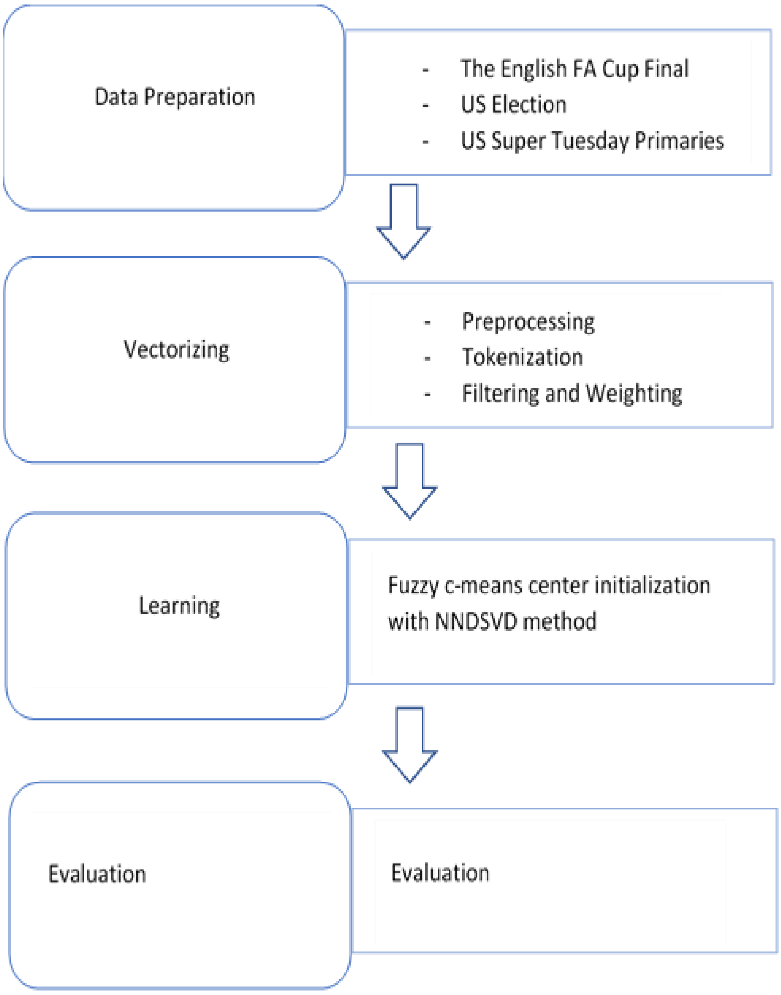

Data Preparation
- The English FA Cup Final
- US Election
- US Super Tuesday Primaries
Vectorizing
- Preprocessing
- Tokenization
- Filtering and Weighting
Learning
Fuzzy c-means center initialization with NNDSVD method
Evaluation
Evaluation

Fig 5: General stage of fuzzy c-means implementation with NNDSVD initialization method

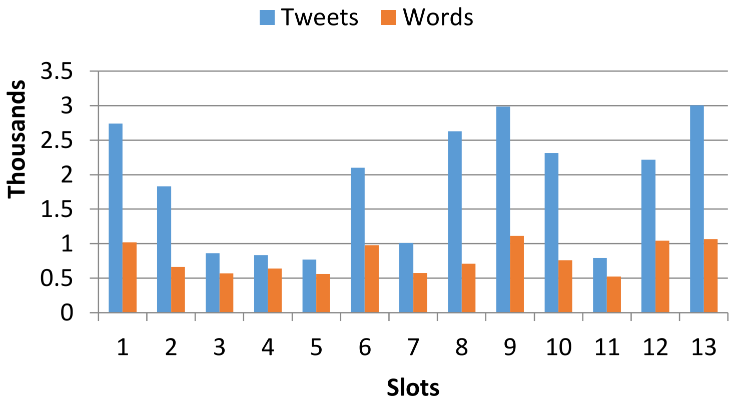


Fig 6: The FA Cup Dataset

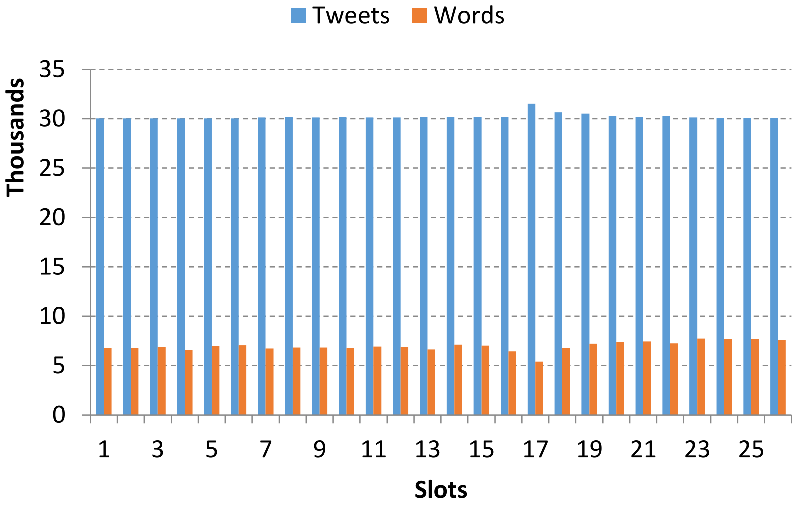


Fig 7: The US Election Dataset

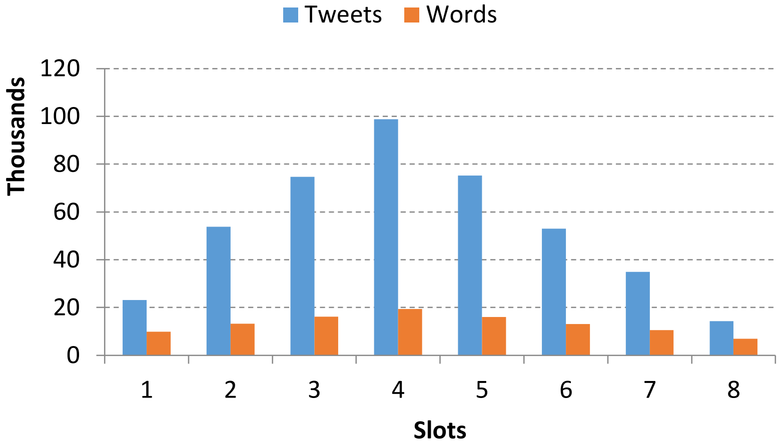


Fig 8: The Super Tuesday Dataset

5.2 Data Preparation

Before the data can we used for simulation, the data must pass several steps of preparation. The preparation includes converting to lower case, converting "www. *" or "HTTP?: // *" to "URL", converting @username to AT_USER, removing additional white spaces, replacing #word to word, and converting unstructured data into structured data.

5.3 Tokenization

The next step is tokenization. Tokenization is the process to get the tokens (words) of tweets by using spaces as separators between words. The collection of the words then is stored in a list called a word dictionary.

5.4 Filtering

The third stage is word filtering. The word dictionary is filtered to select potential words. If a word in the word dictionary is similar to a word contained in the stopwords, it will be deleted from the word dictionary.

5.5 Weighting

The fourth stage is word weighting. The words are scored by weighting scheme. The weighting of words is very influential in determining the similarity between documents with keywords [16]. If the weight of each word can be properly pinpointed, it is expected that the result of the similarity of the text will produce a good document ranking. Suppose the document vector $\boldsymbol{d} = (d_1, d_2, \dots, d_n)$ where d_i is denoted for the *i*-word weight of the document $\boldsymbol{d}$. function $f_i(\boldsymbol{d})$ is a binary weighted function with

$$f_i(\boldsymbol{d}) = \begin{cases} 1, & if\ word\ i\ is\ on\ document\ \boldsymbol{d} \\ 0, & if\ word\ i\ is\ not\ on\ document\ \boldsymbol{d} \end{cases}$$

Re-weighting needs to be done if there are less important words that often appear to cover important information. The term frequency-inverse document frequency (*TFIDF*) is one of the common weighting schemes which can be defined as follows

$$f_i(\boldsymbol{d}) = \frac{tf_i}{df_i}, \forall i$$

where $f_i(\boldsymbol{d})$ is the weight of the word *i* in document ***d***, *tfi* is the number of occurrences of the word *i* in document ***d***, and *dfi* is the number of documents containing the word *i* [16].

When the documents are tweets, then the tweet vectors are set as columns of matrix called the word-tweet matrix. The topic modeling work on this word-tweet matrix to extract topics on the corresponding datasets.

5.6 Learning

In this stage, we determine the optimal number of topics. We select the optimal number of topics from the number set of $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ for three datasets. First, we set the parameter *w* or fuzzy degree to 1.1 The learning process starts by centroid initializations of FCM algorithm using the random numbers, SVD or NNDSVD. After centroid initializations, the tweets in the form of the word-tweet matrix are grouped using the FCM algorithm.

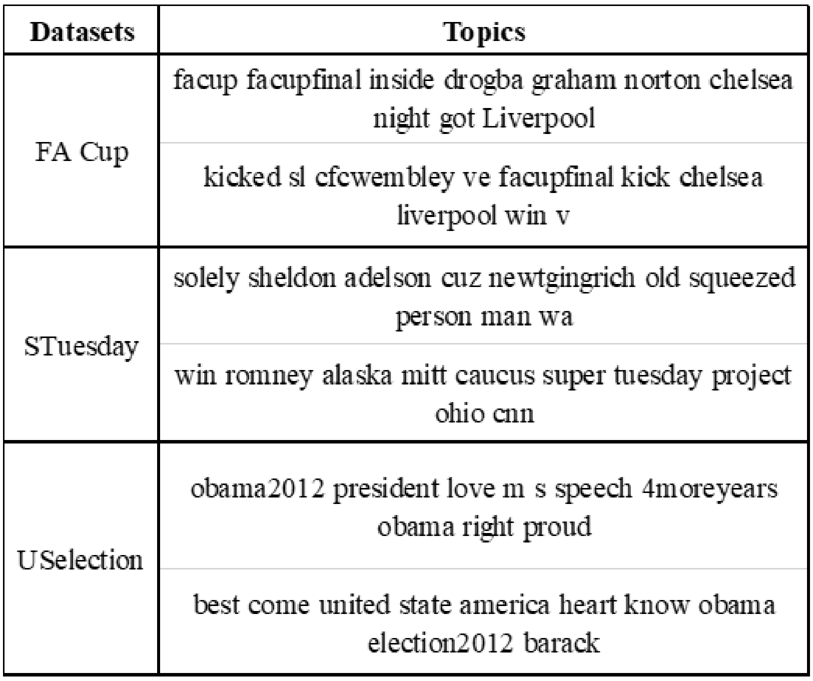
5.7 Evaluation

The last stage of the simulation process is to calculate the accuracies of the extracted topics. The accuracies of the extracted topics are measured by considering the ground truth topics which are created by professionals in the field. This study uses a measurement unit called topic recall, that is, the percentage of ground truth topics successfully detected by the FCM algorithm. The values of the topic recall range from 0 to 1. The higher the topic recall value, the more ground truth topics are detected by the learning method. Otherwise, if the topic recall is small, then only a few topics match the ground truth topics.

5.8 Results and Discussion

Topics are taken from each cluster by sorting the weights or values of the centroids from the largest to the smallest. Then, we look for words that correspond to those values. The most often word can be seen from the values of the centroid, the greater the value contained in the centroid, the more often a word appears. We sort the values on the centroid and take top ten highest scores of words to represent the topics. Some examples of topics that extracted by FCM algorithm are shown in Table 1.

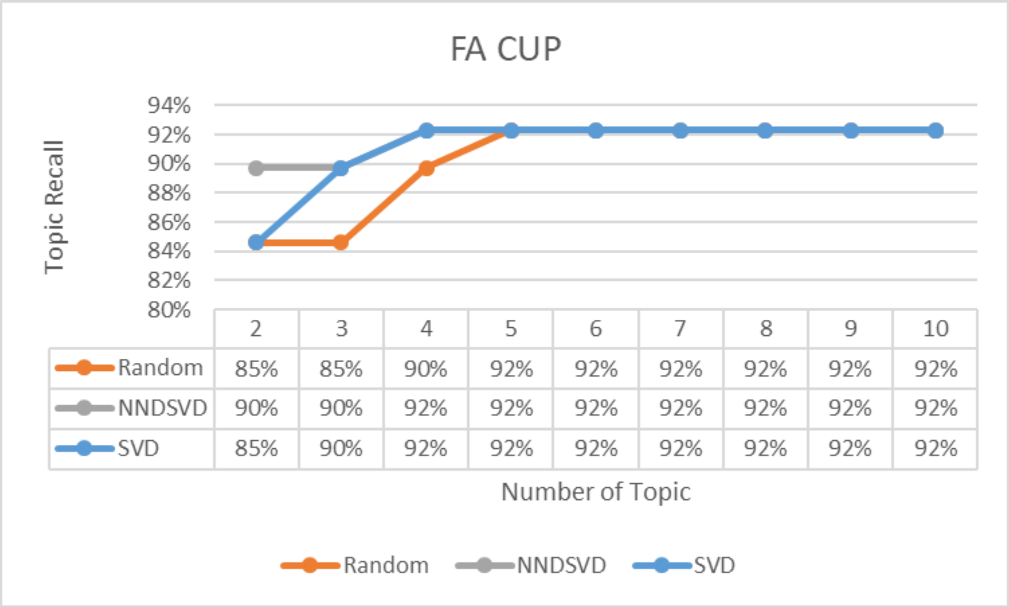
Table 1: Examples of topics extracted by FCM algorithm



Datasets	Topics
FA Cup	facup facupfinal inside drogba graham norton chelsea night got Liverpool
	kicked sl cfcwembley ve facupfinal kick chelsea liverpool win v
STuesday	solely sheldon adelson cuz newtgingrich old squeezed person man wa
	win romney alaska mitt caucus super tuesday project ohio cnn
USelection	obama2012 president love m s speech 4moreyears obama right proud
	best come united state america heart know obama election2012 barack

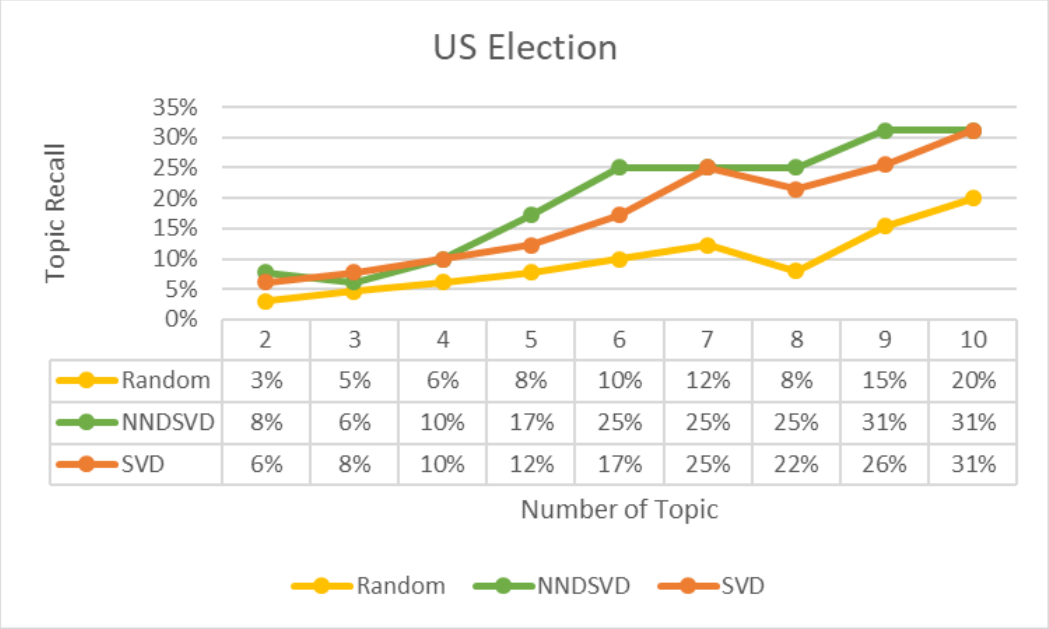
Once the topics are extracted by the learning process, their accuracies are measured based on their topic recalls. This topic recalls range from 0 to 1. The higher the topic recall is, the better the extracted topics.

In this simulation, we conduct a comparative study of initialization of FCM. We compare the random-based initialization, SVD-based initialization, and the NNDSVD-based initialization. The FCM method with random initialization is executed five times, and the final topic recall is an average of five topic recalls. We consider that the FCM with the random initialization gives different topics every algorithm executed. On the other hand, the FCM with the SVD-based and NNDSVD-based initialization gives the similar topics every algorithm executed. Next, we compare topic recalls for all datasets as described in Fig 9, Fig. 10, and Fig. 11.



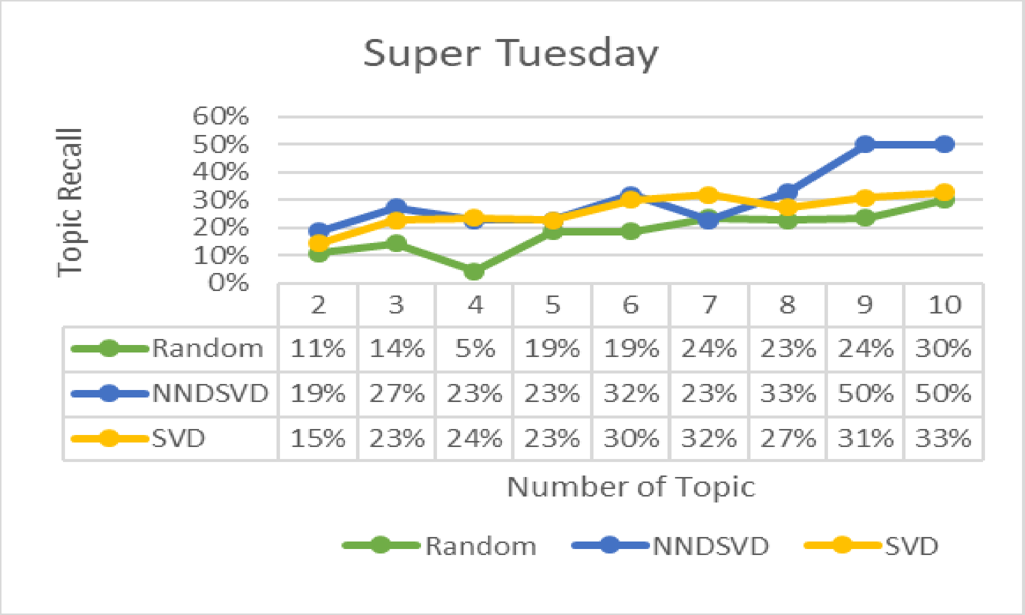
	2	3	4	5	6	7	8	9	10
Random	85%	85%	90%	92%	92%	92%	92%	92%	92%
NNDSVD	90%	90%	92%	92%	92%	92%	92%	92%	92%
SVD	85%	90%	92%	92%	92%	92%	92%	92%	92%

Fig 9: The comparison of topic recall between the random-based initialization, SVD-based initialization, and the NNDSVD-based initialization for the FA Cup dataset



	2	3	4	5	6	7	8	9	10
Random	3%	5%	6%	8%	10%	12%	8%	15%	20%
NNDSVD	8%	6%	10%	17%	25%	25%	25%	31%	31%
SVD	6%	8%	10%	12%	17%	25%	22%	26%	31%

Fig 10: The comparison of topic recall between the random-based initialization, SVD-based initialization, and the NNDSVD-based initialization for the US Election dataset



	2	3	4	5	6	7	8	9	10
Random	11%	14%	5%	19%	19%	24%	23%	24%	30%
NNDSVD	19%	27%	23%	23%	32%	23%	33%	50%	50%
SVD	15%	23%	24%	23%	30%	32%	27%	31%	33%

Fig 11: The comparison of topic recall between the random-based initialization, SVD-based initialization, and the NNDSVD-based initialization for the Super Tuesday dataset

From Fig. 9, we see that NNDSVD-based initialization, SVD-based initialization, and random-based initialization work well in the FA Cup dataset. All three methods reach topic recall of 0.923.

Fig. 10 shows that for the US Election dataset, the random-based initialization produces topic recall of 0.2 which means 20% of the ground truth topics detected when it extracts ten topics. The SVD-based initialization reaches topic recall of 0.312 which means 31.2% of ground truth topics found when it extracts ten topics. Otherwise, the NNDSVD-based initialization reaches topic recall of 0.312 which means 31.2% of ground truth topics when it extracts nine and ten topics. We can see that NNDSVD-based initialization reaches better accuracy than the random-based initialization and similar with the SVD-based initialization.

From Fig. 11 which shows the simulation for the US Super Tuesday dataset, we see that the random-based initialization produces topic recall of 0.3 which means 30% of the ground truth topics detected when it extracts ten topics. The SVD-based initialization reaches topic recall of 0.327 which means 32.7% of ground truth topics detected when it extracts ten topics. Otherwise, the NNDSVD-based initialization reaches topic recall of 0.5 which means 50% of ground truth topics detected when it extracts 9 and ten topics. From these results, we conclude that that NNDSVD-based initialization gives better accuracy than the others for the Super Tuesday dataset.

6 Conclusion

The standard method of centroid initialization in the FCM algorithm is random. When we use for detecting topics, then this random initialization usually produces different topics for each execution. In this paper, we examine a nonrandom initialization method based on nonnegative double singular value decomposition (NNDSVD) besides another existing nonrandom initialization method based on singular value decomposition (SVD). We evaluate the accuracy of the methods for the problem of sensing trending topic on Twitter. Besides the advantage of non-randomness, our simulations show that the NNDSVD-based initialization method gives better accuracies in term of topic recall than the random-based initialization method for all three datasets. Moreover, the NNDSVD-based initialization reaches better accuracies that the SVD-based initialization method for one out of three datasets.

References

[1] Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., and Goker, A. (2013). Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia,* vol. 15, no. 6, pp. 1268-1282.

[2] Sitorus, A. P., Murfi, H., Nurrohmah, S., Akbar, A. (2017). *Internasional Journal of Electrical and Computer Engineering*, vol. 7, No. 1, pp. 330-336.

[3] Blei, D. M. (2012). Probabilistic topic models. *Communication of the ACM*, vol. 55, No. 4, pp. 77–84.

[4] Allan, J. W. (2002). *Topic Detection and Tracking: Event-Based Information Organization*. Springer Science and Business Media, LLC.

[5] Petkos, G., Papadopoulos, S., Kompatsiaris, Y. (2014). Two-level Message Clustering for Topic Detection in Twitter. *Proceeding of The SNOW Data Challenge*. Seoul: Korea.

[6] Nur'aini, K., Najahaty, I., Hidayati, L., Murfi, H., & Nurrohmah, S. (2015). Combination of Singular Value Decomposition and K-Means Clustering Methods for Topic Detection on Twitter. *Proceeding of Internasional Conference on Advanced Computer Science and Information System*. Depok: Indonesia.

[7] Fitriyani, S. R., Murfi, H. (2016). The K-Means with Mini Batch Algorithm for Topic Detection on Online News. *Proceeding of the 4th International Conference on Information and Communication Technology*, Bandung: Indonesia.

[8] Bezdek, J. C., Robert, E., William, F. (1984). FCM: The Fuzzy C-Means Clustering Algorithm. *Computers & Geosciences*, vol. 10, no. 2-3, pp.191- 203.

[9] Yu, J., Cheng, Q., Huang, H. (2004). Analysis of the Weighting Exponent in the FCM. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34

[10] Nock, R., Nielsen, F. (2006). On Weighting Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no.8.

[11] Mursidah, I., Murfi, H. (2017). Analysis of Initialization Method on Fuzzy C-Means Algorithm Based on Singular Value Decomposition for Topic Detection. *Proceeding of Internasional Conference on Informatics and Computational Science*. Semarang: Indonesia.

[12] Boutsidis, C., Gallopoulos, E. (2008). Svd based initialization: a head starts for nonnegative matrix factorization. *Pattern Recognition*, vol. 41, pp. 1350-1362.

[13] Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, vol. 401, pp. 788–791

[14] Blei, D. M., Ng, A. Y., Jordan, M. I., Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993-1022.

[15] Jacob, B. (1990). *Linear Algebra*. New York: W. H. Freeman and Company.

[16] Manning, C. D., Schuetze, H., Raghavan, P. (2008). *Introduction to Information Retrieval*, Cambridge University Press.