Latent Dirichlet Allocation (LDA) Model and kNN Algorithm to Classify Research Project Selection

M A Safi‘ie1,2, E Utami1, H A Fatta1
1 Department of Informatics Engineering, Postgraduate Programme, AMIKOM University, Yogyakarta
2 Department of Informatics Engineering, Faculty of Mathematics and Natural Sciences, University of Sebelas Maret, Surakarta

E-mail: safiie99@mipa.uns.ac.id

Abstract. Universitas Sebelas Maret has a teaching staff more than 1500 people, and one of its tasks is to carry out research. In the other side, the funding support for research and service is limited, so there is need to be evaluated to determine the Research proposal submission and devotion on society (P2M). At the selection stage, research proposal documents are collected as unstructured data and the data stored is very large. To extract information contained in the documents therein required text mining technology. This technology applied to gain knowledge to the documents by automating the information extraction. In this articles we use Latent Dirichlet Allocation (LDA) to the documents as a model in feature extraction process, to get terms that represent its documents. Hereafter we use k-Nearest Neighbour (kNN) algorithm to classify the documents based on its terms.

1. Introduction
Proposal documents are collected at the selection are still in the category of unstructured data and very large, in order to extract the information contained its documents is required text mining technology. This technology can provide solutions such as processing, organizing/grouping and analyzing unstructured text in very large quantities. Text mining technology can be applied to these problems because the documents has characteristic unstructured text, the volume of data is stored very large. This technology can be done to gain knowledge of text by automating the extraction of information. The techniques commonly used in text mining technology are summarizing text, information extraction, text categorization, text groupings, text tracking, and etc. The proposed of this research is to get the performance of LDA model and kNN algorithm to classify proposal documents.

1.1 Text mining
Text Mining also known as Intelligent Text Analysis, Text Data Mining or Knowledge Discovery in Text (KDT), its generally refers to interesting extracting processes and non-trivial information and knowledge of unstructured text. In the Data Mining area the tools is designed to handle structured data, but in the Text Mining area, its can work with unstructured or semi-structured data sets. Text Mining is a new invention by computer from unknown information, automatically information extracted from different written sources[1]. The key is linking together to form new facts or new hypotheses to be explored further with conventional way experiments.
Text Mining is similar to Data Mining, which is about Knowledge Discovery, so in the process of its knowledge discovery has stages as in Data Mining as described in Figure 1. The stages are document collection, pre-processing, data transformation, pattern matching and evaluation. Text Mining also uses the same technology as Natural Language Processing (NLP) such as information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization and question answering.

**Figure 1. Knowledge discovery process[2]**

1.2 *Porter Stemmer Algorithm*

Stemming is a process that provides the mapping of variations in word morphology into its basic word (stem)[3]. This process is also called merging, based on the assumption that terms with the same stem usually have the same meaning. Stemming Porter algorithm is a combination of words proposed by Martin Porter. The algorithm is based on the idea that English words are mostly made up of a combination of compact imbalances.

**Figure 2. Stemmer Porter algorithm design[3]**
1.3 Latent Dirichlet Allocation (LDA) Model

Latent Dirichlet Allocation is a probabilistic, generative model for searching for latent semantic topics in large data text collections[4]. Each topic found is characterized by a certain distribution of the words of the document. Each document is characterized as a random mix of topics that indicate proportion at a time in the document for each topic. The random mix of these topics is a short description, not only expressing the semantic content of a document but also a principle approach to describing the document quantitatively. The key in the LDA model is a premise that words contain strong semantic information from document text.

Figure 3. Graphical representation of LDA[4]

1.4 K-Nearest Neighbour (kNN) Algorithm

The kNN algorithm is done by searching for the k group of objects in the closest training data (similar) to the object in the new data/data testing. The kNN algorithm is a method for classifying objects based on learning data closest to the object. To define the distance between two points ie the point in the training data (x) and the point in the test data (y) then use the Euclidean formula, as shown in the equation:

\[ D(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \]  

The kNN algorithm is done by searching for the k group of objects in the closest training data (similar) to the object in the new data / data testing. The kNN algorithm is a method for classifying objects based on learning data closest to the object. To define the distance between two points ie the point in the training data (x) and the point in the test data (y) then use the Euclidean formula, as shown in the equation: With D is the distance between the points in the training data x and the data testing point y to be classified, where \( x = x_1, x_2, ..., x_i \) and \( y = y_1, y_2, ..., y_i \).

2. Experimental

The steps of this research, are as follows:

a. Data collection: manually collecting P2M document, the data collected from the system and has format *.pdf.

b. Convert pdf to text: the collected document data then parsed and converted to raw text. In addition on converting documents, its also stores the attribute of the document into the database.

c. Text Cleanup: is the process of removing punctuation, symbols or character-character contained in the text and its not required to the next process.

d. Tokenizing: clean text data then changed to lowercase letters which subsequently breaks the text into its constituent words.

e. Stopwords Removal: the stage to remove words that do not represent text, stopwords are usually additional words. The removal of these words to obtain terms representing the document in other other words is called filtering.

f. Stemming: is the stage to change the words to its roots word, by removing affix (affix) either in the form of prefix, suffix, confix and particle.
g. Information Extraction: terms that represent the text document then using the LDA model is searched for probability, the document has the highest probability of representing a particular theme / topic.

The documents will be experimented with several test scenarios described in table 1 below:

| No | Training Data | Testing Data | Description |
|----|---------------|--------------|-------------|
| 1  | 100%          | 100%         | The overall data will be data learning and testing |
| 2  | 75%           | 25%          | 75% data will be used as data learning and 25% data will be used as data testing |
| 3  | 50%           | 50%          | 50% data will be used as data learning and the remaining 50% used as data testing |
| 4  | 25%           | 75%          | 25% data will be used as data learning and the remaining 50% is used as data testing. |

3. Result and Discussion
The results of research testing scenario are shown in Table 2, total documents were used are 20 samples with 5 categories of topics. Based on the table 2, the highest values of classification were obtained at 20 sample documents but the value decreases with the lower number of data sample.

| No | Data set    | Result (%) |
|----|-------------|------------|
| 1  | 20 Dokumen  | 85,00      |
| 2  | 15 Dokumen  | 73,33      |
| 3  | 10 Dokumen  | 70,00      |
| 4  | 5 Dokumen   | 60,00      |

4. Conclusion
This study have following conclusions:
1. The high value of accuracy is obtained by the increasing use of documents as training data and the accuracy will decrease if less data training.
2. Higher accuracy is achieved when the number of documents topics in the LDA model are closer to the actual number of documents topics.

References
[1] Gupta V, Science L C, and Lehal G S 2009 A Survey of Text Mining Techniques and Applications 1(1) 60–76
[2] Agarwal C, Charu 2015 Data Mining: The Text Book (New York : Springer International Publishing)
[3] Tala Fadillah Z 2003 A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia
[4] Hu D J 2009 Latent Dirichlet Allocation for Text, Images, and Music (San Diego: University of California)