2018-03

# Mining predictors of success in air force flight training regimens via semantic analysis of instructor evaluations

Hwang, Jaesung

Monterey, California: Naval Postgraduate School


# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**MINING PREDICTORS OF SUCCESS IN AIR FORCE FLIGHT TRAINING REGIMENS VIA SEMANTIC ANALYSIS OF INSTRUCTOR EVALUATIONS**

by

Jaesung Hwang

March 2018

| | |
|---|---|
| Thesis Advisor: | Matthew Norton |
| Co-Advisor: | Su-Hwan Kim |
| Second Reader: | Robert Koyak |

THIS PAGE INTENTIONALLY LEFT BLANK

| REPORT DOCUMENTATION PAGE | | | Form Approved OMB No. 0704–0188 |
|---|---|---|---|
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503. | | | |
| **1. AGENCY USE ONLY** *(Leave blank)* | **2. REPORT DATE** March 2018 | **3. REPORT TYPE AND DATES COVERED** Master's thesis | |
| **4. TITLE AND SUBTITLE** MINING PREDICTORS OF SUCCESS IN AIR FORCE FLIGHT TRAINING REGIMENS VIA SEMANTIC ANALYSIS OF INSTRUCTOR EVALUATIONS | | **5. FUNDING NUMBERS** | |
| **6. AUTHOR(S)** Jaesung Hwang | | | |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)** Naval Postgraduate School Monterey, CA 93943-5000 | | **8. PERFORMING ORGANIZATION REPORT NUMBER** | |
| **9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)** N/A | | **10. SPONSORING / MONITORING AGENCY REPORT NUMBER** | |
| **11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____. | | | |
| **12a. DISTRIBUTION / AVAILABILITY STATEMENT** Approved for public release. Distribution is unlimited. | | **12b. DISTRIBUTION CODE** | |


**13. ABSTRACT (maximum 200 words)**

Most educational curricula have step-by-step learning objectives accompanied by some type of assessment that can be used to analyze student outcomes and trends. When these assessments are unstructured textual feedback, it is difficult to extract meaningful indicators that point to student success. In this thesis, we create a graphical representation of the text corpus of each individual student assessment in a flight-training program used by the Republic of Korea's Air Force. From it, we develop a coherent topic model, which allows us to characterize the training program. We then utilize the graphical representation of student assessments, together with the extracted topic model, to extract meaningful information from each assessment. This allows us to develop a statistical model to predict student outcomes. This information also allows us to quantitatively assess the importance of each topic, characteristics of instructor feedback and their connection to student success, as well as other factors. We apply our methodology to the criticism text written in the flight-training program student evaluations in order to construct a model that accurately predicts passing and failing based on extracted factors. We provide instructors and students recommendations for improving the success rate of the flight-training course.


| **14. SUBJECT TERMS** text mining, feedback analysis, semantic network, binary classification | | | **15. NUMBER OF PAGES** 105 |
|---|---|---|---|
| | | | **16. PRICE CODE** |
| **17. SECURITY CLASSIFICATION OF REPORT** Unclassified | **18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified | **19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified | **20. LIMITATION OF ABSTRACT** UU |

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release. Distribution is unlimited.**

# MINING PREDICTORS OF SUCCESS IN AIR FORCE FLIGHT TRAINING REGIMENS VIA SEMANTIC ANALYSIS OF INSTRUCTOR EVALUATIONS

Jaesung Hwang
Major, Republic of Korea Air Force
B.S., Republic of Korea Air Force Academy, 2007

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the

**NAVAL POSTGRADUATE SCHOOL**
**March 2018**

Approved by: Matthew Norton
Thesis Advisor

Su-Hwan Kim, Korea National Defense University
Co-Advisor

Robert Koyak
Second Reader

Patricia Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

Most educational curricula have step-by-step learning objectives accompanied by some type of assessment that can be used to analyze student outcomes and trends. When these assessments are unstructured textual feedback, it is difficult to extract meaningful indicators that point to student success. In this thesis, we create a graphical representation of the text corpus of each individual student assessment in a flight-training program used by the Republic of Korea's Air Force. From it, we develop a coherent topic model, which allows us to characterize the training program. We then utilize the graphical representation of student assessments, together with the extracted topic model, to extract meaningful information from each assessment. This allows us to develop a statistical model to predict student outcomes. This information also allows us to quantitatively assess the importance of each topic, characteristics of instructor feedback and their connection to student success, as well as other factors. We apply our methodology to the criticism text written in the flight-training program student evaluations in order to construct a model that accurately predicts passing and failing based on extracted factors. We provide instructors and students recommendations for improving the success rate of the flight-training course.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| AUC | area under curve |
| ISFN | individual student flight network |
| LASSO | least absolute shrinkage and selection operator |
| NLP | non linear programming |
| MST | minimum spanning tree |
| PN | program network |
| ROC | receiver operating characteristic |
| ROKAF | Republic of Korea Air Force |
| TF/IDF | term-frequency-inverse document frequency |
| TDM | term document matrix |
| WCSS | within-cluster sum of square |

THIS PAGE INTENTIONALLY LEFT BLANK

# EXECUTIVE SUMMARY

Military education or training is distinguished from general education in the sense that it is aimed at transferring skills for mission performance. These differences are also noticeable in the output of the evaluations. If the purpose is to convey knowledge, a student's achievement can be assessed on relatively objective criteria and the evaluation result can be easily structured to yield a standardized numeric value. These results provide useful information to the educator that can be used to measure student achievement or to analyze and improve the educational method. On the other hand, in the case of military training, the evaluation results often yield textual data and their use is limited compared to the standardized numeric assessments.

The purpose of our study is to create a graphical representation of the text corpus as well as each individual student assessment for students in the flight-training course used by the Republic of Korea's Air Force (ROKAF). We connect the graphical properties of the overall text corpus to a coherent topic model, which can help to summarize the entire training program. We then utilize the graphical representation of each student assessment, together with the extracted topic model, to extract features that provide a useful characterization of each student assessment. We then utilize these features within a supervised learning procedure to predict overall student outcomes, achieving high predictive accuracy on the applied data set. These features, however, are also easy to interpret in the context of the overall topic model and graphical representation, allowing us to quantitatively assess the importance of each topic, characteristics of instructor feedback and their connection to student success, as well as other factors that contribute positively (and negatively) to student success. This allows us to construct a statistical model that accurately predicts passing and failing based on extracted factors for participants in the ROKAF flight-training course. We are also able to provide instructors and students with recommendations for improving the success rate of the flight-training course.

Our first step is to preprocess the unstructured criticism text into structured form based on existing text mining theory. The preprocessing includes translation of the flight criticism text, which is a mixture of Korean and English text, completely into English,

integrating words with the same root, removing meaningless words, and correcting typographical errors and abbreviations.

The primary motivation of our study is to create a semantic network that can properly characterize abstract documents, specifically, the overall text corpus as well as each individual assessment. We create semantic networks that serve as characterizations of flight-training criticism text based on the relationship between words in the text. In particular, from the semantic network created from the entire criticism text, we develop a topic model of the entire flight-training program. We also develop a semantic network of individual student assessments to extract the factors that can predict whether the student will pass or fail the flight-training program.

The semantic network consists of key words from the entire criticism text. We group the words based on the topology of the semantic network and identify a set of words representing the topics of flight training. The identified topic model successfully represented the composition of the flight-training program. Table ES-1 displays the topic model that we developed and the words that compose the topic model.

Table ES-1. Topic in flight training and word composition of each topic

| Topic | Words |
|---|---|
| **LANDING** | LANDING, TIME, LATE SOMEWHAT, INSUFFICIENT, ALTITUDE, MAINTAIN, INADEQUATE, DIRECTION, NOT, DO, CAN, WHEN, SHOULD TRIM, APPROACH, AIMING, FINAL |
| **TAKE-OFF** | TAKEOFF, AFTER, BROLL, LEVEL, UP, DEGREE, LEFT, RIGHT, PITCH, LOW, NO, CHECK, PRACTICE, PITCH |
| **MANEUVER** | ROLL, OUT, RATE, LAZY, BANK, SPEED, HIGH, G, CONTROL, FLIGHT, PROCEDURE |
| **EMERGENCY** | STALL, POWER, RECOVERY, DESCENT, OPERATION, GROUND, GOOD, CHANGZHOU, POSTURE, HORIZONTAL, AIRSPACE, KEEP, DESCENT |

We also create a semantic network for each student's criticism text and extract the factors that affect passing and failing based on the network properties. These factors quantitatively describe characteristics such as an instructor's guiding style, topic emphasis, and the emotions associated with these topics. To evaluate the quality of the extracted

factors, we build a statistical model that predicts the passing and failing of the student. With the model, we are able to predict students' passing and failing with high accuracy using the extracted factors. From this analysis, we can provide the following recommendations for more effective flight training through the analysis of the features included in the prediction model.

First, instructor's guidance style should be different for each stage of training. In the initial phase of flight training, teaching each subject separately increases the training achievement of the student. From the middle stage of the flight training, each subject should be linked and should be discussed in a more comprehensive manner.

Second, the positivity and negativity of the evaluation from the instructor has influence over the student's final passing probability. However, positive evaluations do not always increase the probability of passing. In particular, the positive evaluation of landing at the end of the flight training tends to lower the student's probability of passing, possibly because this leads to student overconfidence regarding one of the most difficult, complicated, and important parts of the flight-training program.

Our methodology can be applied to other types of military trainings that yield unstructured assessment results and provides essential information about the composition and the trends of education.

THIS PAGE INTENTIONALLY LEFT BLANK

# ACKNOWLEDGMENTS

First of all, I would like to thank Professor Matthew Norton of the Naval Postgraduate School. He spent a lot of time in guiding my thesis while he was busy. Through his advice, I was able to solidify my research topic that had been somewhat abstract. I developed mathematical thinking and problem-solving skills through my research process, and I think this is due to his passion.

Also, I would like to express my sincere gratitude to Professor Robert Koyak of the Naval Postgraduate School; he gave key advice on the direction of the research as a second reader.

Lastly, I would like to thank Professor Su-Hwan Kim of Korea National Defense University for giving me the opportunity to start the research and to Professor Lyn Whitaker for teaching me practical and essential data analysis techniques.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

## A. BACKGROUND

Military education or training is distinguished from general education aimed at transferring knowledge in the sense that it is aimed at transferring skills for mission performance. These differences are also noticeable in the output of the evaluations. If the purpose is to convey knowledge, the student's achievement can be assessed on a relatively objective criterion and the evaluation result can be easily structured to yield a standardized numeric value. These results provide very useful information to the educator that can easily be used to measure student achievement or analyze and improve the educational method. On the other hand, in the case of military training, the evaluation results yield textual data and their use is limited compared to the standardized numeric assessments.

The purpose of this study is to propose a methodology that can extract useful insights from such unstructured textual feedback data. In particular, we would like to be able to target the predictors of student success and/or failure. Additionally, we would like to identify student and instructor trends that can provide insight into the current state of the program and offer suggestions for improving course components such as course structure, instruction methods, and student assessments. As the author of this paper is an officer of the Republic of Korea's Air Force (ROKAF) and has experience in flight training, we applied our methodology to flight criticism text written during the flight-training program of ROKAF for data collection and qualitative analysis.

The ROKAF operates three flight-training courses to train its pilots. Elementary flight training teaches pilots the basics of flying. Secondary flight training teaches difficult flying techniques using light-attack aircraft with turboprop engines. "High-level" flight training teaches basic combat capability. All courses assess students' flight skills, and students who do not meet the criteria (i.e., fail the program) are classified as ground officer.

The course with the most student failures (highest drop-out rate) is the secondary flight-training course, where practical in-flight training is carried out. The first, and arguably most critical, portion of the training program consists of a set of 10–11 training

flights followed by a test flight. Those who fail this first test flight drop out of the program, while those who pass continue onward with more training. This is a critical stage because a majority of dropouts are caused by this first test flight.

Leading up to this first test fight are 10–11 training flights. During each training flight, the instructor pilot will fly with the student pilot to instruct their flight technique and provide an assessment of the student's capabilities. The assessments for these training flights, however, are not the typical numeric scores you might expect. Instead, each student is only provided with written feedback from the instructor. Further complicating the analysis is the fact that each training flight is split between multiple instructors that not only provide feedback following their own personal methodology but also provide slightly different in-flight challenges for each student depending upon the natural flow of the flight itself.

This study is based on analysis of criticism/feedback data written in the training flights before the first test flight evaluation during the secondary flight course. When analyzing for all students, it is possible that there are variables that cannot be identified through criticism data such as age, past flight experience, and health status. For this reason, we limit analysis to the written critique data. Also, since the first training flight mainly consists of instructor demonstration, and no student flying tasks, criticism data is limited. Thus, we excluded the criticism of the first flight from the analysis. We analyze a total of 375 students' criticism data composed of 76514 sentences. From these 375, a total of 43 students were dropped (i.e., failed) after taking the first test flight assessment. Thus, for each student, there is criticism (text) data available for 11 different flights.

## B. OBJECTIVES AND APPROACH

After a course is administered and data is collected regarding the student's continuing performance (feedback data) and final outcome (pass/fail), educational data mining sets out to discover patterns and trends that provide insight into different aspects of the course, to improve the curriculum, instructor technique, and (in general) student outcomes. Fundamentally, we approach the educational data-mining task by using textual feedback data to create a topic model with a network topology, specifically a clustered

semantic network. At a high level, the clustered semantic network provides us with a model of the relationships between words in the flight training feedback vocabulary, with clusters revealing central topics as well as emotions connected to each topic. With this network created from the entire dataset, we are then able to represent each student assessment (given after each training flight) as a rich set of features relating to the semantic network. Specifically, each assessment constitutes a subgraph of the overall network (possibly disjoint) and belongs to a subset of topics within the overall semantic network, belonging to particular clusters. This allows us to characterize each assessment (or set of assessments) as related to a set of topics or emotions. We are also able to create a semantic network for each individual assessment, not only the overall data set, to characterize the properties of individual assessments. For example, does the assessment focus on one cluster? Or is the feedback unfocused, covering many topics at once?

The usefulness and validity of the clustered semantic network is highly dependent upon the design decisions (e.g., how to build the network, how many clusters to choose, how to do the clustering); as well as the criteria that drive the design decisions (what makes a network "good"?). In this study, design choices are driven by the goal of prediction, specifically prediction of student outcomes (success/failure). As mentioned before, a clustered semantic network allows us to represent every individual assessment as a set of features (e.g., frequency of occurrence of topics, emotions, properties of the induced subgraph, etc.). Therefore, we evaluate the quality of a semantic network by the predictive power of the features it produces. In other words, a topic model is "good" if it allows us to predict student outcomes from their individual assessments accurately. We find that many potential models have high predictive power. We further narrow our selection by using well-known cluster evaluation criteria paired with simple assessment of the visualization of the topic model.

Incorporating classification does much more than provide a data-driven evaluative criterion of the quality of a clustered semantic network. The classifier, through the use of feature selection, helps to reveal which characteristics have the most predictive power with regard to predicting student outcomes. Not only can these features represent particular topics or emotions, they can also represent the time at which this topic or emotion occurred,

meaning in which particular stage of the fight training (i.e., training flight 1 or 2 or 3 or 11). This provides us with a wealth of information about the predictors of student outcomes and how they relate to the specific structure of the course.

Critically, our method is also able to consider the time-series component of the feedback data beyond simple feature selection. Our general method need not be applied to the entire text corpus over all 11 training flights. It can be applied to any subset of text data. In our analysis, we created a semantic network for each individual student assessment as well as the entire text corpus. While utilizing the topic model from the overall text corpus, we also measure the characteristics of each individual student assessment by analyzing the properties of the semantic network generated from the individual assessment. Treating each semantic network as its own feature generator for a classifier, we were able to create classification models that focus only on individual flights and the assessments associated to assess how much predictive power each individual flight has with respect to student.

Overall, our methodology produces a topic model that provides insight into the overall flight training process, with these insights tied directly to student outcomes. Additionally, our model is able to reveal important characteristics of individual assessments. We find that the insights provided by our model can be used by instructors to reevaluate the structure of the training program, recruit well-suited students, and provide better-constructed feedback.

## C. STRUCTURE OF THESIS

In this study, we construct a tunable pipeline for creation of a clustered semantic network with tuning driven by a classification algorithm for predicting student outcomes from feedback data. In Chapter II, we briefly overview background material relevant to construction of our pipeline, including previous use of semantic networks, text processing, and clustering. In Chapter III, we detail our proposed methodology, along with the key questions we aim to answer. In Chapter IV, we discuss quantitative and qualitative results. In Chapter V we conclude with a brief discussion of conclusions and future work.

# II. THEORY AND LITERATURE REVIEW

In this chapter, we briefly review existing work that is related to our proposed analysis pipeline. We first comment upon the broad subject of text mining, which can be used as the most general categorization of this work, speaking also about the critical task of text preprocessing that is relevant to our pipeline. We then provide some context for our methodology, introducing some of the primary components of our data analysis pipeline. This includes the use of sematic networks, clustering techniques, and binary classification.

## A. TEXT MINING

The most general way to categorize this work is within the field of text mining. Recently, the techniques of text mining have been used in a variety of areas, and their definitions vary accordingly, but the simplest definition is "Data mining techniques applied to textual data or documents."

Hand et al. 2001 defined data mining as follows:

> Data mining is the analysis of (often large) observational data sets to find un-suspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (p. 1)

Under this definition, text mining is seen as a method of extracting useful patterns or knowledge from textual data or documents. These characteristics of text mining are well described by Zhang et al. 2015:

> Text mining, also known as knowledge discovery in textual database or text data mining, of which new interesting knowledge is created, is defined as the process of extracting previously unknown, understandable, potential and practical patterns or knowledge from the collection of massive and unstructured text data or corpus. (p. 681)

Text mining is a distinct subset of the broad subject of data mining, made distinct in that the primary data source is textual data that is generally un-structured data. Thus, preprocessing is a very important part in text mining. This holds true for our work as well. Specifically, we apply standard text preprocessing techniques to clean our data.

Maybury (1997) reviewed data preparation for text analysis, interpretation and transformation, information extraction, and information utilization. Vijayarani (2015) introduced the techniques that are widely used in text preprocessing, and Hemalatha (2012) presented the technique of preprocessing for efficient sentiment analysis. One of the most important tasks in text preprocessing is to transform un-structured text into structured data format via natural language processing. Indurkhya et al. (2010) discussed statistical language learning and natural language processing. Natural language processing includes word stemming and elimination of stop words (Vijayarani, 2015).

Stemming is one of the preprocessing technique for text data that reduces the complexity of a document by identifying and merging the words with the same root. Porter proposed an algorithm for morphological analysis of English text (Porter 1980) and developed a snowball stemmer. Bouchet-Valet (2014) implemented an "R" interface to Porter's snowball stemmer "SnowballC" package.

Pronouns and articles are not related to the contents of the document, so they are generally excluded from the text analysis. The process of excluding these words is called "Stop word removal." Luhn (1960) presented the concept and necessity of "Stop word" for the first time. Porter (2001) presented a snowball stop word list, which is a list of stop words for English. Tsz-Wai et al. (2005) suggested a method for automatic stop words list generation and Hassan et al. (2014) proved that dynamic generation of stop words is better than using a pre-compiled stop words list for Twitter sentiment analysis.

A typographical error caused by a human is one of the factors that makes the textual data un-structured. Kukich (1992) identified typographical errors by applying pattern recognition techniques, presenting an automatic correction algorithm. Joseph et al. (1984) proposed a typographical error correction algorithm for scientific and scholarly text.

The preprocessed text data is formatted in the form of a term document matrix. Miller (2013) presented a method of creating a term document matrix and utilized it in movie tagline analysis. Feinerer et al. (2017) implemented the "tm" package as a tool for the text analysis using R software.

Merkl (2002) discussed cluster analysis techniques for grouping documents into similar classes by searching for similarities between documents. Dumais (2002) examined potential semantic analysis and statistical approaches to extract relationships between terms in a set of documents.

## B. SEMANTIC WORD NETWORK

A major component of our analysis pipeline is the construction of a network-like representation of our text data. While the there are many approaches that exist to accomplish this task, we utilize the idea of a semantic network. Semantic networks are structured expressions of knowledge and relationships that are used in various fields. John (2015) defined and described a semantic network as follow:

> a graph structure for representing knowledge in patterns of interconnected nodes and arcs. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology, and linguistics. The Giant Global Graph of the Semantic Web is a large semantic network. (p. 1)

A semantic network is a network representation of objects with qualitative characteristics such as human thoughts or documents. Therefore, nodes constituting a semantic network represent a sub-concept constituting an abstract object, and arcs connecting nodes are the representation of relationship between the sub-concepts.

The most widely used technique for semantic network analysis is "WordNet." Miller et al. (1990) defined "WordNet" as:

> a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. (p. 1)

Tsang et al. (2004) proposed a method to calculate the semantic distance between words by comparing probability distributions over WordNet. Baccianella et al. (2010) modified the WordNet and proposed "SENTIWORDNET" for sentiment analysis and

opinion mining. Tingting et al. (2014) proposed text clustering methods based on semantic relationships among words using WordNet and lexical chains.

Network based analytics have also been used for stock market analysis. Bonnanno et al. (2003) analyzed the topology of a correlation based ***Minimum Spanning Tree(MST)*** using real market data. Onnela et al. (2003) studied the time dependent properties of the MST for portfolio optimization.

As will be discussed in detail in Chapter III, we utilize the MST algorithm in tandem with correlation and lift based distance measures to construct a semantic network. We then utilize clustering to extract the structure of a topic model.

## C. CLUSTERING AND EVALUATION

Creating a semantic network from text data provides insight into the relationships between words, potential emotional words, and groups of words. However, to extract a topic model from this network representation in a data driven manner, it can help to perform clustering. Then, not only do we have a map of word relationships, but also distinct groupings that can allow for a higher-level analysis of patterns and trends within the data. Two tasks are important in this stage: 1) the clustering method that is used and 2) the method for evaluating the "goodness of fit" of the cluster.

### 1. Clustering

Statistical clustering is a method for understanding the structure, or existence of, subsets of individuals within a population. It divides the entire population into relatively uniform groups and attempts to summarize the properties of each group.

The most widely used clustering methodology is *K*-means clustering. McQueen (1967) presented *K*-means clustering for the first time. Hartigan et al. (1979) presented a heuristic approach for *K*-means clustering. Huh et al. (2009) proposed a *K*-means clustering methodology that takes into account the weight of variables. Two problems in *K*-means clustering are that it is not applicable to categorical data and is vulnerable to outliers. As an alternative to this problem, Kaufman et al. (1987) proposed the *K*-medoids methodology.

If the object of clustering is a network, we can identify the network communities as clusters based on the degree of connectivity in the network. In this study, we utilize the following three algorithms as network clustering methods.

The fast-greedy algorithm (Clauset et al. 2004) is a method to quickly find a community by simply repeating the steps of nodes merging.

Edge betweenness algorithm (Newman and Girvan 2004) estimates the community by removing the edge with the highest betweenness. In the network with $m$ edges, the betweenness of edge $e$ is defined as $D_B(e)$ where $g_{i,j}$ the number of shortest paths from node $i$ to $j$ and $g_{iej}$ the number of shortest paths pass through the edge $e$ within shortest path from node $i, j$ . :

$$D_B(e) = \sum_i \sum_{j \neq i} \frac{g_{iej}}{g_{ij}}, e = 1, ..., m. \tag{1}$$

Walk trap community algorithm (Pons and Latapy 2004) identifies a community based on the probability of staying at specific node when starting from any node on the network and randomly walking to the neighboring node. This algorithm increases the size of individual communities by reducing the number of communities by reducing the number of communities sequentially by merging nodes.

## 2. Clustering Evaluation

The method used to evaluate the performance of $K$-means clustering is within-cluster sum of squares (WCSS). For $k$ clusters $(S_1, S_2 \ldots, S_k)$ from n-observations $(X_1, X_2 \ldots, X_n)$ with d-variables where $\mu_i$ is the mean of points in $S_i$, $K$-means clustering aims to minimize the WCSS defined as Equation 2.

$$WCSS = \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2}$$

Choosing the appropriate number of clusters is very important so that good data partitioning can be achieved. One of the criteria used to determine the optimal number of

clusters is Hartigan's rule (Hartigan et al. 1979). This rule compares the ratio of WCSS for clustering of $K$ and $K + 1$ clusters. Lander (2017) implemented Hartigan's rule in the "useful" package of "R." Gap statistic can replace the Hartigan rule (Tibshirani et al. 2001). This compares the dissimilarity in the cluster for the clustering result to the bootstrapped data.

Another criterion for evaluating clustering results is reproducibility. Reproducibility indicates the degree to which the clustering results match when multiple clustering is performed on the same data. One of the numerical indexes indicating the degree of reproducibility is "Rand Index" presented by Rand (1971). The "Rand index" is calculated as follows. When clustering is performed twice on the data consisting of $n$ objects, the clustering classification result for any two objects $a$ and $b$ is one of the following three cases.

1. $a$ and $b$ are classified into the same cluster in the results of the two-different clustering.
2. $a$ and $b$ are classified into the same cluster in one clustering result but are classified into different clusters in the other clustering result.
3. $a$ and $b$ are classified into the different clusters in the results of the two-different clustering.

The total number of pairs of objects in the data is:

$$\binom{n}{2} = \frac{n(n-1)}{2} \tag{3}$$

It is the first case and the third case that the clustering classification by the two-different clustering is identical. Let the number of object pairs in each case be $N1$, $N2$ and $N3$, respectively, then the number of pairs in which the cluster classification matches is expressed as "$N1 + N3$. Therefore, the "Rand index," which indicates the ratio of pairs in which the cluster classification match, is defined as Equation 4. If the clustering results are perfectly matched, the value of "Rand index" is equal to 1.

$$RandIndex = \frac{(N1 + N3)}{\binom{n}{2}} \tag{4}$$

Based on "Rand index," Gordon (1999) proposed a method for evaluating the reproducibility of clustering through data partitioning. Huh et al. (2004) proposed a method of applying the clustering rules calculated from different data to independent data and evaluating the reproducibility according to the similarity of two clustering results.

The result of deriving communities from the network are evaluated as modularity index *Q*. Newman et al. (2004) presented the Modularity Index as an indicator of the structural characteristics of community in the network. A network consisting of *n* nodes and *m* edges is defined as $N \times N$ adjacency matrix, and element of adjacency matrix $a_{i,j}$ has a binary value indicating whether nodes *i* and *j* are connected. $c_{i,}c_j$ indicate the community to which nodes $i, j$ belong. $\delta\left(c_{i,}c_j\right)$ has a binary value indicating whether node $i$ and $j$ belong to the same community. $k_1,...,k_n$ are the number of edges of individual nodes (degree), and the sum of $k_1,...,k_n$ is 2m in an undirected network. Then the expected value of $a_{i,j}$ under the perfect randomness is $\frac{k_i k_j}{2m}$ . Since the modularity represents the connection strength between nodes belong to the same community, the modularity of the network is defined as Equation 5.

$$Q = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{i,j} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \tag{5}$$

## D. BINARY CLASSIFICATION AND FEATURE SELECTION

When creating a clustered semantic network from text data, there are a multitude of options for constructing this process. For example, we must choose the number of words to eliminate during preprocessing, the between-word distance metric that is used to create the semantic graph structure, and the clustering method (as well as the number of clusters). Thus, one primary challenge in our analysis is model selection. In other words, in each stage of the analysis pipeline (i.e., preprocessing, network creation, clustering), there are

many methods to choose from as well as parameters to tune for each chosen method. Therefore, we must devise some evaluative criteria for selecting "the best" or a set of "best" pipeline parameters.

Fortunately, our data is also accompanied by binary labels for each student, i.e., pass or fail. Thus, we are able to utilize prediction to drive the model selection process which is now "supervised learning." This evaluative criterion is directly connected to our end goal as well, which is to devise a model which, at the very least, produces features that can predict student success or failure. Then, subsequent insights can be gained from the chosen topic model/semantic network itself, which has proven predictive power.

For performing binary classification, we utilize logistic regression with lasso regularization for feature selection with ***Area Under Curve*** **(*AUC*)**, which is the area under the ***Receiver Operating Characteristics*** **(*ROC*)** curve, as our evaluative method for classifier performance. The ROC curve represents the correlation between "True positive rate" and "False positive rate" for different decision thresholds. The AUC is the area under the ROC curve. Figure 1 shows the relationship between the ROC curve and the AUC value. We select AUC since it has been shown to work better than accuracy in evaluating the performance of classification models when the dataset is imbalanced. In general, only 10% to 20% of total students belong to the failure class so the ratio of passing and failing students in our dataset was very imbalanced. When constructing a model for high-dimensional data containing many variables, the methods to prevent over-fitting are essential and can be used for additional feature selection and related insights. A widely used method to prevent "over-fitting" for high dimensional data is regularization. Typical methods of regularization are ridge regression (Hoerl et al. 1970) and Least Absolute Shrinkage and Selection Operator (Tibshirani 1996).

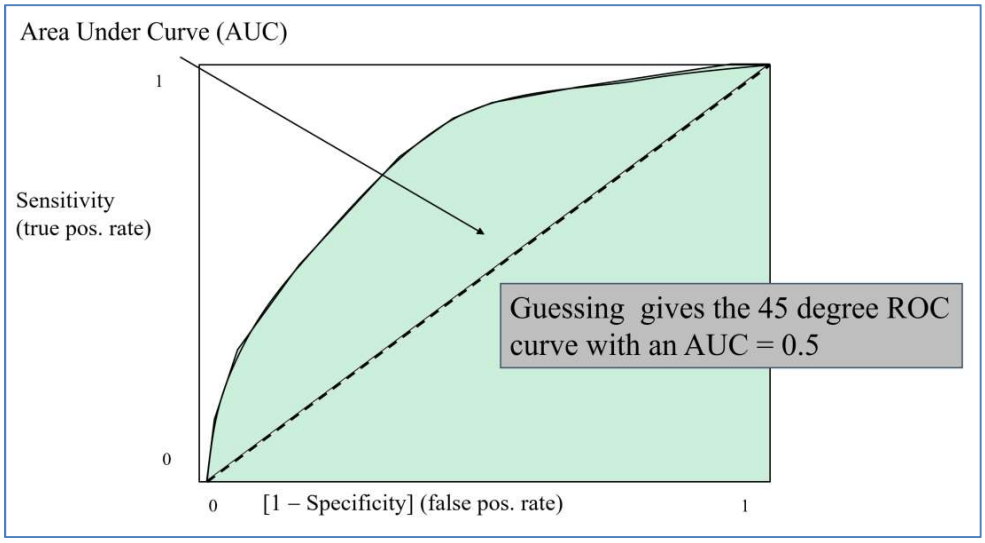


Figure 1. Area under curve. Source: Buttrey et al. (2017).

THIS PAGE INTENTIONALLY LEFT BLANK

# III. METHODOLOGY

## A. OVERVIEW

To understand our entire methodology, it helps to view our method as a pipeline that transforms individual student assessments (for each individual flight) into a vector of numerical features. These features are then used as input for a classification model that predicts the eventual success/failure of the student. The features associated with each individual assessment are generated from two primary sources: 1) The clustered sematic network (i.e., topic model) generated from the entire corpus of text data. We call this the ***Program Network* (*PN*)** since it is created from the data collected over the entire course of the training program for all students. 2) The semantic network generated from assessments for each individual student. We call this the ***Individual Student Flight Network* (*IFSN*)** since it is created from data only from a single assessment written for a single flight for one student. In other words, the MST methodology that will be described in the following paragraph is applied to only words within a single assessment.

Beyond the classification task, we would like the features generated for each assessment to have meaning. We accomplish this by generating a topic model via the PN. Furthermore, we characterize each individual assessment in the context of this topic model by representing each assessment as its own ISFN. The PN is always generated from the entire text corpus. After text preprocessing, which will be discussed in a later section, distances between words are measured according to one of two choices for distance metric. Using this, a semantic network is created and is further trimmed via the MST algorithm. Finally, to generate a topic model, clustering is performed using one of a few different clustering methodologies. The ISFN is always generated from either single or multiple assessments for only a single student. Similar to the PN, text is preprocessed, and distances between words are measured for only words within the individual's set of assessment. Then, a network is formed and trimmed as a MST. Clustering is only performed on the PN since it is the basis for the overall topic model.

To generate features from the PN for a single assessment, we look at how often each topic (cluster) from the PN occurs in the individual assessment. The frequency of occurrence of each PN topic within an assessment generates one feature for every topic (cluster). To generate features from the ISFN (which is created only from data from a single student and flight), we measure a set of numerical properties of the network that aim to capture its characteristics. We also measure properties of the ISFN that utilize the topic model derived from the PN. For example, if we use the topic model generated from PN, we can label each word in the ISFN as related to a particular topic. From this, we can look at how "close together" certain topic-words are within the ISFN. Properties like this help us analyze, in a quantitative way, the characteristics of each assessment in terms of the overall topic model (the PN). Specifically, as will be discussed in detail in later sections, we measure the density, modularity, closeness centrality, and the emotion associated with included topics in this way. These metrics allow us to answer questions for each assessment such as

1. How central is each topic to a particular student's Flight #2 assessment?
2. Of the topics mentioned in the assessment, which ones are associated with positive-emotion words and which are associated with negative-emotion words?
3. Does the instructor give focused feedback, with words clearly focused on one topic at a time? Or is the assessment unfocused, jumping from topic to topic in an unorganized manner?

Using these tools, we can build a vector of features for each individual student assessment for every flight (10 feature vectors for every student from training flights 2–11, where the ISFN for each student is created from only a single flight assessment). From this, we can formulate 10 separate classification tasks, with each focusing on an individual flight. This allows us to isolate the powerful, predictive characteristics from each individual flight. We can then combine these feature vectors to perform classification utilizing all flights at once. We do this in two ways. First, we can concatenate all 10 feature vectors together to create a single large feature vector for every student. Second, we can average all feature

vectors together since they have the same dimensionality and share meaning across dimensions. This allows us to isolate the most important features, overall, for the prediction of success/failure.

Overall, we need to set many parameters for each part of the pipeline. For preprocessing, we must decide what proportion of words to use. For creating the semantic network, we must decide how to measure distances between words. For clustering, we must decide which method to use and how many clusters should be created. We perform a two-stage model selection procedure to find the best parameter settings for the model. First, we perform the pipeline for every combination of settings and log the performance of the classification in terms of cross-validated AUC. Second, we isolate the top performing models and utilize a combination of cluster evaluation techniques and basic visual inspection to choose the best settings for generating the PN and ISFN. We found that many models achieve very high AUC and a secondary selection process is necessary to find models that are robust to slight changes in parameters.

## B. RESEARCH QUESTIONS

In section A, we presented an overview of the entire methodology. We now discuss the specific research questions that we aim to answer. Then, we present a more detailed account of the entire methodology.

### 1. How Accurately Can the Semantic Network Represent the Characteristics of Documents?

Our first conjecture is that the MST based semantic network is a valid methodology for representing text data and its underlying relationships. Under this conjecture, we will generate the PN and ISFN by calculating the distance between words based on both the co-occurrence frequency or lift value (to be discussed later) and applying the MST algorithm to the calculated distance matrix (for either the entire corpus (PN) or an individual assessment (ISFN)). The MST-based semantic network can be a useful tool to comprehensively describe the characteristics of text in terms of quantifying the abstract relationship between the words that compose the documents. However, we still need a clear

definition of "valid quantification" for an objective verification of this conjecture since a subjective interpretation of the resulting network is not an objective evidence.

By extracting predictors from the PN and ISFN to create feature representations of the assessments, we utilize our classification model to verify the validity of the network representation and overall topic model. If the MST-based semantic network adequately quantifies the criticism data, the network characteristics derived from the semantic network must also reflect the characteristics of the criticism data, specifically the student outcomes (pass/fail).

### 2. What Is the Topic Model of Flight Training?

Our second conjecture is that the clusters of words extracted from the overall PN provide a valid topic model for the flight training. We will develop and evaluate the topic model of the flight training by applying network community theory and the statistical clustering methods introduced in Chapter II on MST-based semantic network. The most important thing in evaluating clustering results is whether the set of selected words has a practical meaning as a topic. Although Chapter II covered several theories about the evaluation of clustering results, numerical evaluation results cannot be an absolute reference to the practical meaning of a cluster. For this reason, we use both quantitative and qualitative techniques to assess the appropriateness of the identified word clusters as a topic model.

### 3. What Are the Major Predictive Factors for Pass/Fail?

To determine if the clustered semantic network is "good," we measure its ability to predict student outcomes. Specifically, we utilize features that are extracted from our semantic network as input to a binary classification model. Also, we are able to utilize the binary classification model to assess which individual features have the most the predictive power. These features include topic information, specific words, network properties, as well as flight-specific information. For example, we can ask questions such as the following:

1. What is the most important topic for pass and fail among all flights?

2. What is the most critical word or topic-word relation for pass and fail?

3. Which training flight provides the best indication of pass and fail?

4. For each training flight, what are the most important topics for predicting pass/fail? Do important topics differ dramatically from flight to flight?

### 4. What Are the Recommendations?

We utilize the results of this study to advise instructors and students on flight training by answering the following questions:

1. What are the topics or maneuvers that can measure the student's potential and achievement overall?

2. What are the characteristics of feedback given to successful students? Unsuccessful student? Does this suggest that certain types of feedback are more helpful/less helpful to students? What teaching methods are recommended to improve student outcomes?

3. What should be the focus of each flight? Which test flight is the most important?

4. What should be emphasized when feedback is given for each topic?

## C. PIPELINE PART1: TEXT PREPROCESSING

The flight criticism text that is the subject of this study is not neatly organized for analysis. Therefore, the first step for text analysis is to formulate the unstructured text into a form that can be analyzed through preprocessing. Vijayarani (2015) asserted that preprocessing is essential for extracting useful knowledge through text mining, and described the typical preprocessing sequence as shown in Figure 2. Considering that the size of the flight critic text is not that large, the preprocessing was very important in this study.

In Figure 2, ***Term frequency-inverse document frequency*** **(*TF/IDF*)** is a method that weights the term frequency with respect to the total number of times the term appears

over all documents. We do not perform the TF/IDF processing because we estimate the weight of each term through the predictive power of passing and failing in the semantic network. In addition, considering the characteristics of the text data, we add two more steps: translation and typographical error / abbreviation handling.

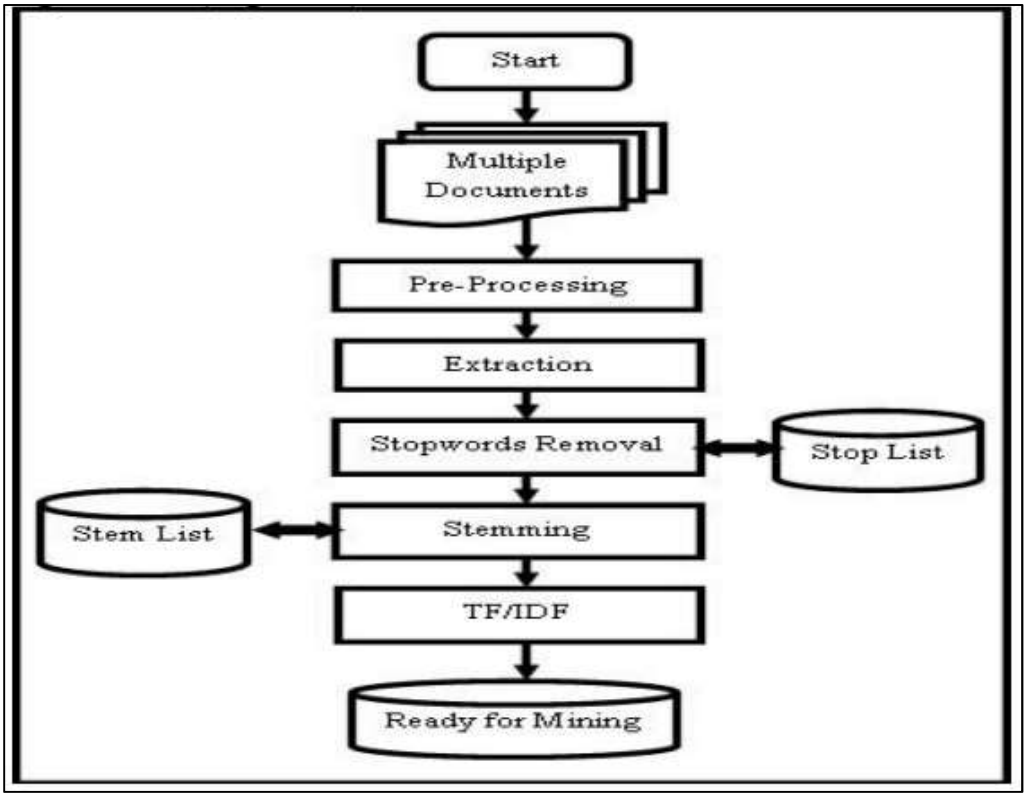


Figure 2. Text Mining Pre-Processing Techniques. Source: Vijayarani (2015).

### 1. Translation

Because the source of the criticism data is the flight-training course of the Republic of Korea Air Force, the main language used in the text is Korean. However, since most terms related to the actual flight task are in English, many of the main task-specific words that compose the document are written in English. The fact that different languages are present in the same document at the same time was a factor that increases the complexity of the document along with the variety of writing style. (For example, if the pronunciation of the English word is written in Korean, the computer recognizes the two words

differently.) We apply translation from Korean to English by using the Python API of Google Translator (SuHun 2017) to unify the format of the criticism text and to minimize the words complexity.

## 2. Natural Language Processing

Text analysis requires formalizing the text so that the computer can understand it. Miller (2013) suggested the bag-of-words approach and natural language processing as two main methods for this. As shown in Figure 3, analysts have to parse a corpus to create a common term, index, and matrix that the computer can easily analyze. The document consists of paragraphs, the paragraph consists of sentences, and the sentence consists of words. Natural language processing is a matter of collecting individual words and should be able to convey meaning. This section discuss the natural language processing method performed in this study, from word extraction to typographical error correction. Figure 4 describes the preprocessing by example.

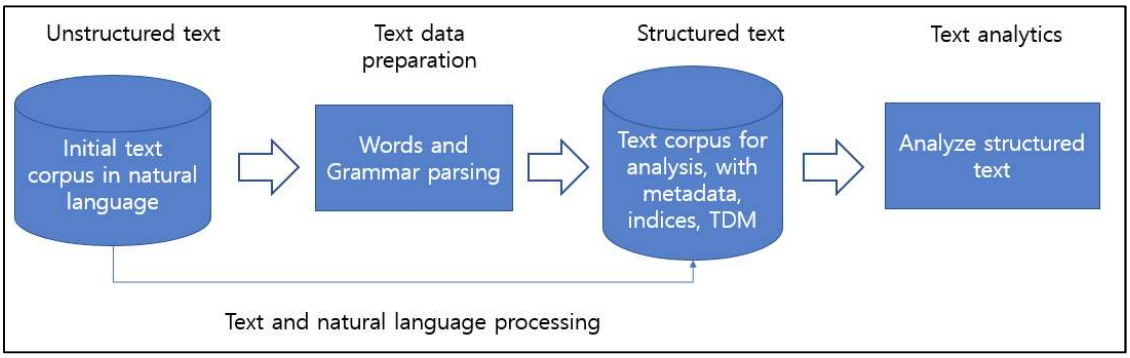


Figure 3. From text processing to text analysis. Source: Miller (2013).

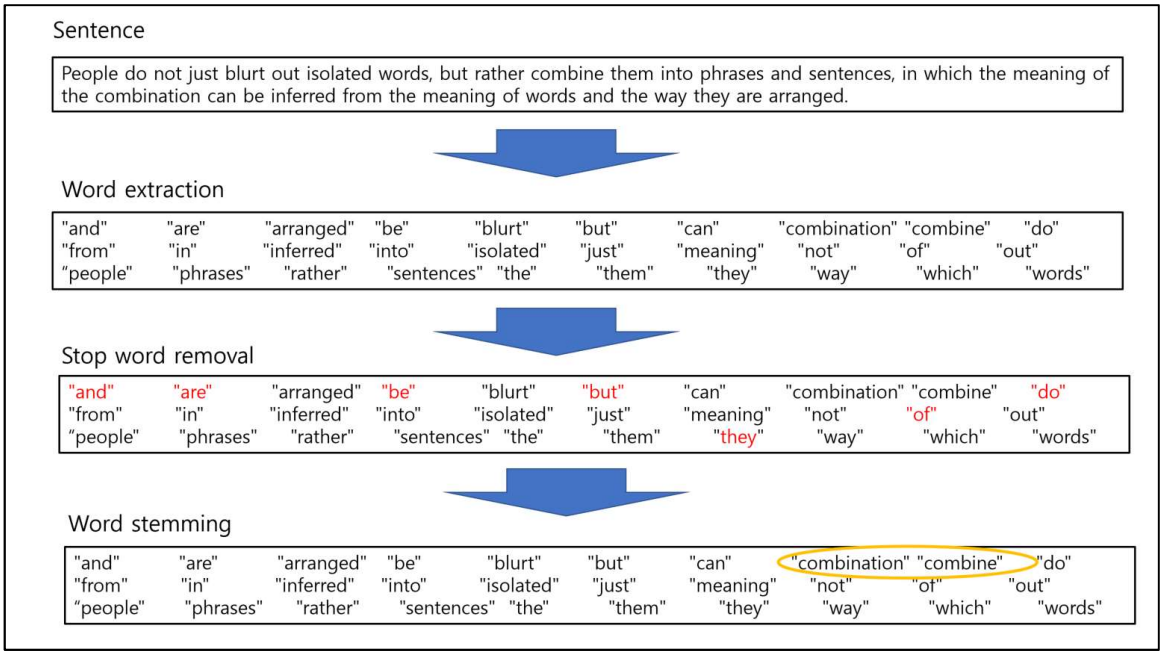


Figure 4. Text processing example

### *a. Word Extraction*

The document is divided into paragraphs and sentences, the sentence is divided into words, and the contents of the whole document are tokenized into unique individual words.

### *b. Stop Word Processing*

Some frequently occurring words are important for understanding the meaning of a document, However, prepositions, numbers, or articles such as “a,” “an,” and “the” occur frequently but are not related to the meaning of the document. Jeffrey (2008) mentioned that there might be a common stop word list, like the numbers and articles, but also stated that the stop word list could be different depending on the subject of the text. The latter case can also be found in the flight criticism text. In a general document, the word “G” has no significant meaning. However, it has an important meaning in the context of flight training because it is an abbreviation for “Gravity.”

In accordance with the general methodology, we remove numbers, punctuation marks, and special characters from the text. Although there are lists of stop words that are

widely used in English text analysis, we manually create our own list of stop words due to the uniqueness of our text data, being both multilingual and task specific.

### *c. Word-Stemming*

The characteristics or attributes of the text are related to the terms. A term is a collection of words that means something specific. The word collection related to the same concept or word stem is as follows. The words "marketer" and "marketing" occurred in the common word stem "market." Jeffrey (2008, p96) stated that "stemming is the process of removing suffixes and prefixes, leaving the root or stem of the word and often applied in the area of information retrieval, where the goal is to enhance system performance and to reduce the number of unique words." Porter (1980) proposed one of the most popular stemming algorithms. Figure 5 illustrates a general overview of Porter's stemming algorithm. Bouchet-Valet (2014) implemented an R interface for the Porter's word stemming algorithm named "SnowballC." We use "SnowballC" package for word stemming.

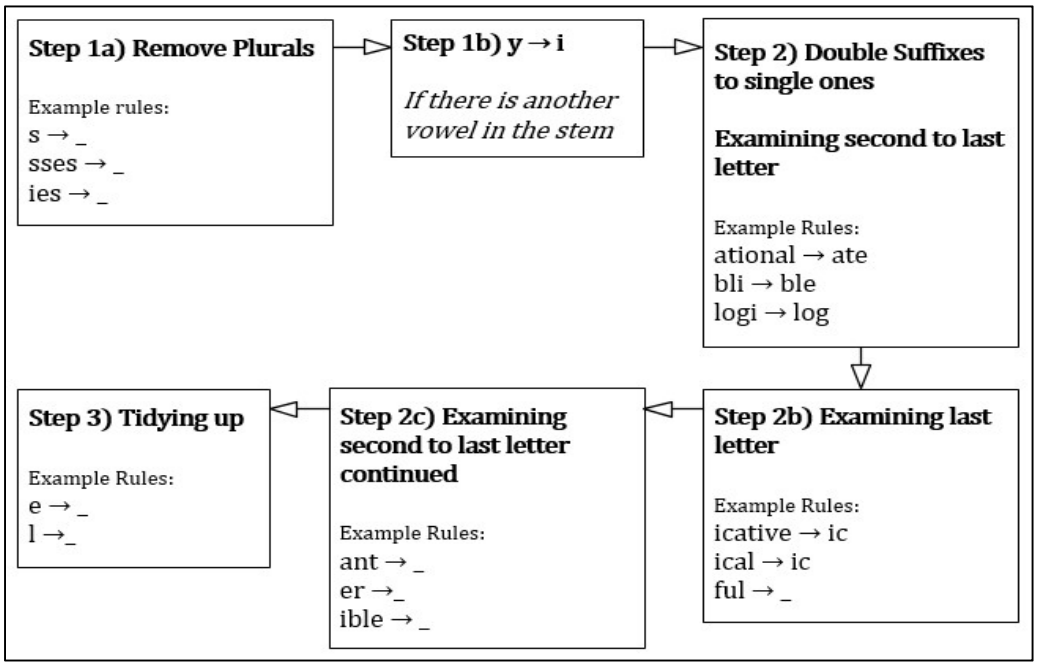


Figure 5. Porter's word stemming algorithm. Adapted from Porter (1980).

### *d. Typographical Errors and Abbreviation*

In analyzing criticism text of Korea's Air Force, typographical errors and abbreviation processing are as important as stop words processing and word-stemming. First, because the authors of the text do not use English as their native language, there are more mistakes in spelling and spacing than in normal English text. Second, flight instructors tend to use the abbreviations used in flight for writing criticism text and the usage pattern of abbreviation is not same for each instructor. For the abovementioned reasons, the case where the same words are treated as different due to typographical errors, spacing, and abbreviation is more than a general English text document. Since this inconsistency increases the complexity of the document and is directly related to the accuracy of the text analysis result, we identify a method to handle it.

Kukich (1992) presented a review of the automatic correction techniques for the three types of typographical errors, non-word error, spelling error, isolated-word error, and argued that the general application is limited because the presented techniques are only for certain types of errors. Since application of the general correction algorithm could potentially change the meaning of the text where it no longer reflects the context of flight criticism data, we implement a simple semi-automated algorithm for typographical errors and abbreviation processing without using the existing theories. The process described in the following steps and illustrated in Figure 6 gives a general overview of our typographical errors and abbreviation processing and the examples of typographical errors, abbreviations found through the algorithm.

- Step 1) From the documents, extract all unique words and count the occurrence of each word. Then, list the terms in descending order of appearance frequency.

- Step 2) Starting from the most frequent word, find the words that have similar character structure with selected frequent word. For example, to find typographical errors for term "PITCH," find the word start with character "P" and has the two of other characters ("I," "T," "C," "H").

- Step 3) Review the found word and modify to proper text.

- Step 4) Do steps 1–3 for the next most frequent word.

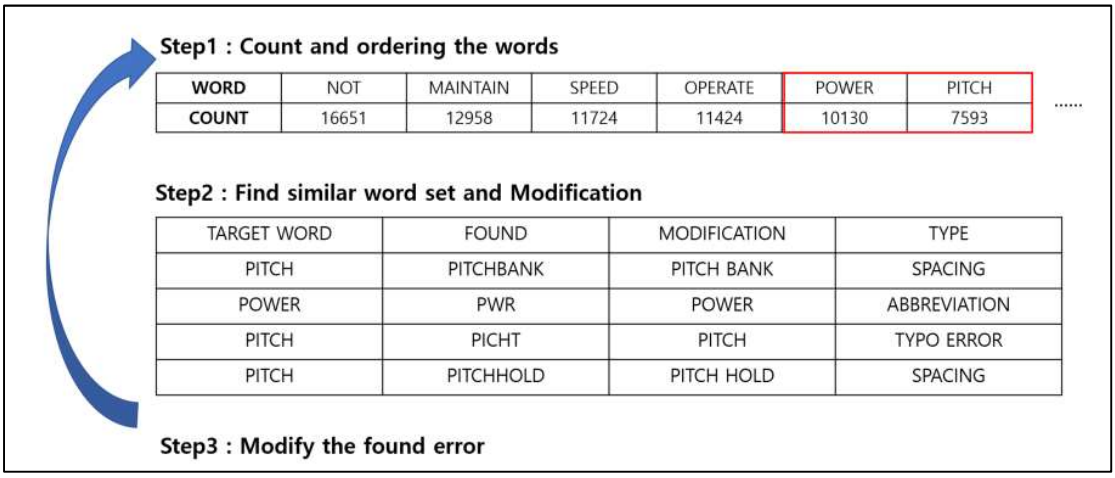


Figure 6. Typographical errors and abbreviation processing

## 3. Processing Results

We use the chart in Figure 7 to evaluate how well the text data are formatted after each preprocessing stage. In Figure 7, the horizontal axis represents the index of the word sorted by the frequency of occurrence in the entire set of assessments, and the vertical axis represents the ratio of the cumulative sum of the frequency of occurrence up to the word corresponding to the index in the entire set of assessments. The closer the curve is to the upper left corner, the less complex the words that compose the document. The comparison of the red line and the black line in the graph shows that the total number of unique words and complexity of the criticism data decreased as a result of the preprocessing.

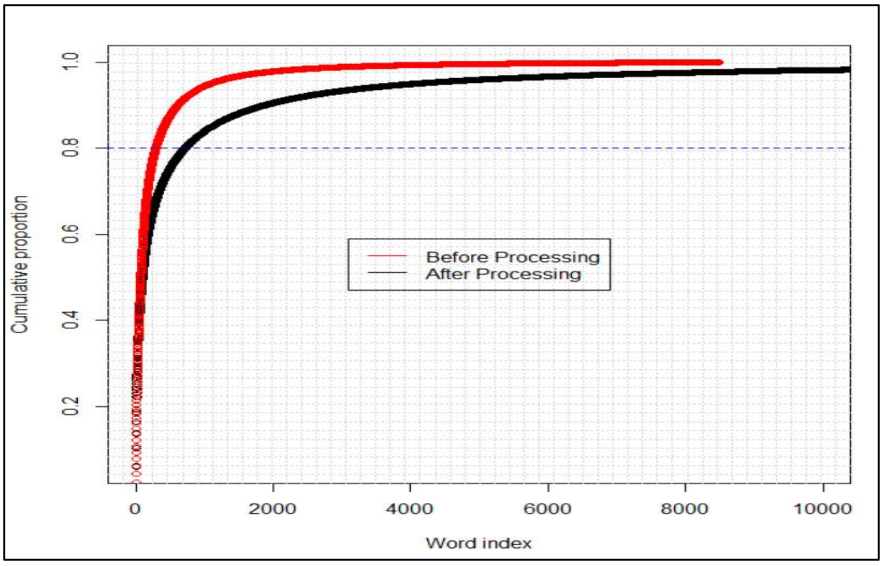


Figure 7. Word proportion comparison before and after processing

## 4. Term-Document Matrix

Miller (2013) presented a terms-by-documents matrix as a tool for text analysis. One of the main steps in the text analysis is to create a term-document matrix. In a document set, the column of the matrix corresponds to the word or word stem, and the row corresponds to the document. We define a document expressed in a row as a single sentence in the criticism data in order to analyze the co-occurrence frequency between words in detail. The numeric value of each cell in the term document matrix is the frequency count of term used in the document. Figure 8 shows the process of creating a term-document matrix.

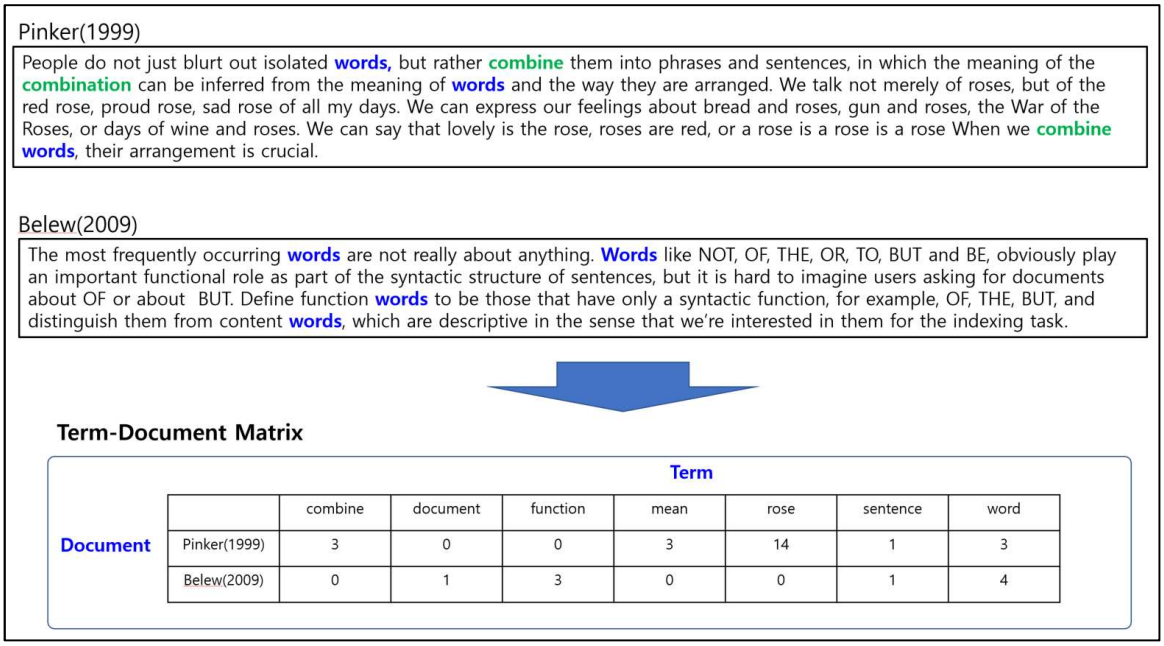


Figure 8. Creating term-document matrix.
Adapted from Miller (2013).

A typical text analysis will have more terms than the document, and the document-term matrix tends be a sparse matrix with most values of zero. As a result of preprocessing, 465333 words were extracted from 56453 sentences and 465333 words consisted of 7456 unique terms. Table 1 shows the distribution of all the terms composing the criticism text. The top 69 terms with the most frequent occurrences account for 50% of the total occurrences, and the top 323 terms account for 80% of the total.

Setting the appropriate scope of terms is very important for obtaining meaningful results in text analysis. By excluding terms that have a very low appearance frequency in the document from the term–document matrix, only the key words that convey the meaning of the document can be limited to the subject of analysis. However, if too many terms are excluded from the term document matrix, there is a risk that the semantics of the document cannot be extracted. Therefore, setting the appropriate scope of terms is very important for obtaining meaningful results in text analysis.

Although, the distribution of terms in the documents provides some reference in setting the scope of the terms for analysis, we could only estimate the approximate scope.

Because of this uncertainty, we perform the same analysis on term-document matrices covering 50% to 80% of the total occurrence terms and compared the results.

Table 1. Word proportion in total criticism text

| Proportion | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| # of Words | 5 | 13 | 25 | 43 | 69 | 105 | 175 | 323 | 691 | 7469 |

We classify the criticism text into three levels of documents and analyze them based on the term-document matrix corresponding to each level. Figure 9 describes three levels of documents and corresponding term-document matrix. The first level treats the entire flight criticism text as a single document and is used to develop the topic model of flight training by applying clustering techniques. At the second and third levels, documents are categorized by each student and flight number for each student. In this study, we use the term-document matrix for each of the documents classified at these levels to compare the criticisms of students who passed and those who failed and identified the factors that determine the pass-fail. In Miller's definition of term-document matrix, one row represents one document, whereas the term-document matrix in this study represent one document, and each row of document matrix represents the sentence constituting a document. This approach allows us to extract more detailed characteristics of the documents in terms of relationships between words.

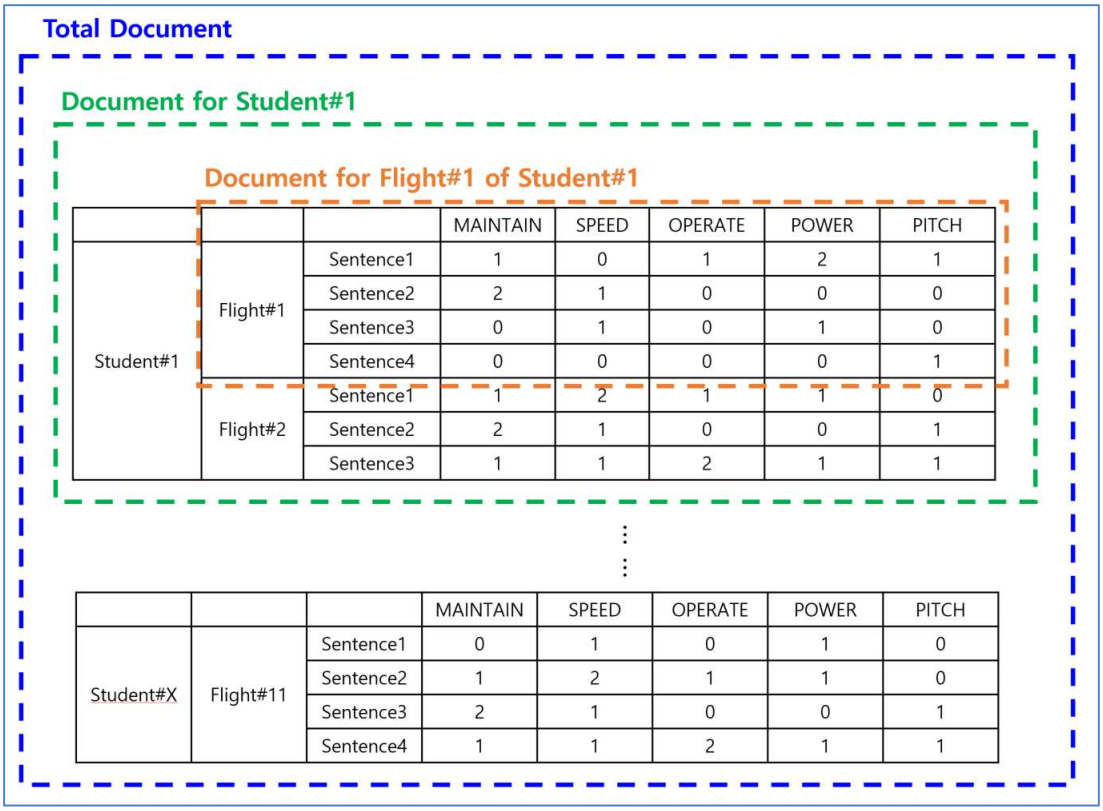


Figure 9. Document level of criticism text and corresponding TDM

## D. PIPELINE PART 2: SEMANTIC NETWORK

As mentioned in the "Research question" section (Chapter II, B.1), the basic premise of this study is that the MST generated network based on the document is valid as a semantic network that reflects the meaning and characteristics of the document. This section describes the method applied in this study to create a semantic network using the MST algorithm.

WordNet has been widely used and studied as a tool for the semantic analysis recently. Since the WordNet is intended to analyze the semantic network of large-scale texts, using "WordNet" for the small-scale data with limited topic such as flight criticism data is limited. On the other hand, Bonnanno et al. (2003) presented correlation-based MST of real data from daily stock returns of 1071 stocks for the 12-year period 1987–1998 as Figure 10. The node color is based on the Standard Industrial Classification system. From the figure, we can confirm that the network of stocks returns generated by applying the

correlation coefficient-based MST is clustered similar to industry classification. These results are similar in purpose to our study to generate a MST semantic network based on the document and to extract a topic model for the criticism data through the clustering.

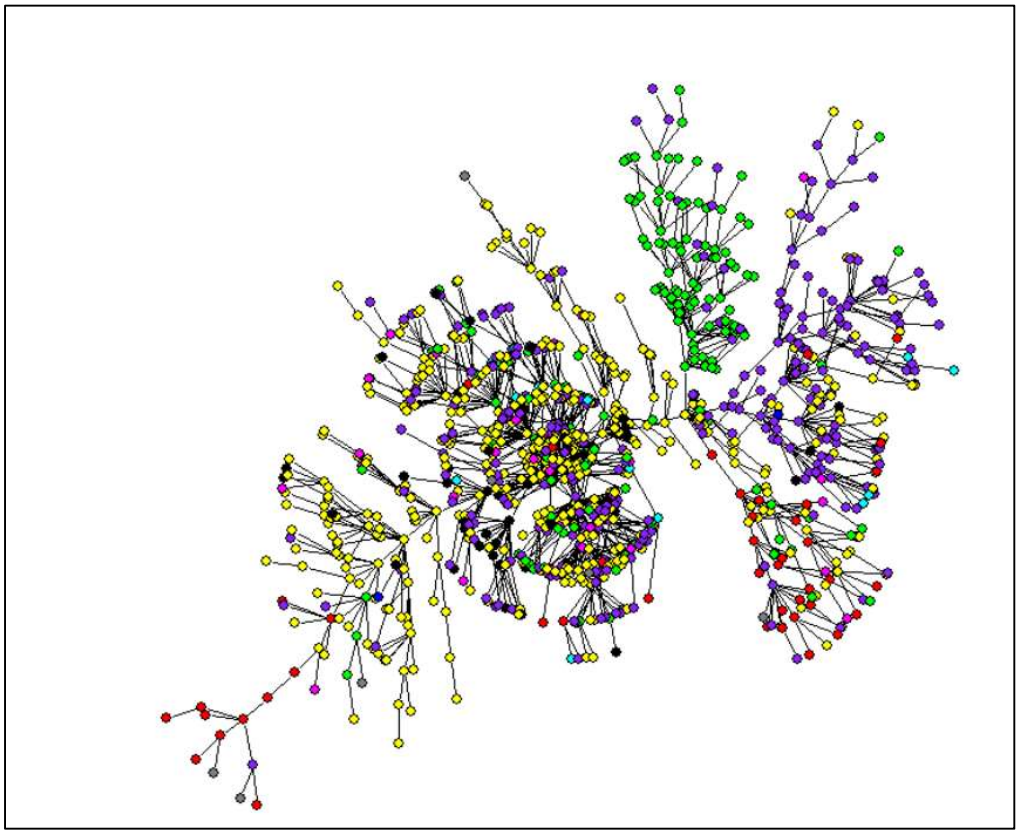

Figure 10. Correlation based MST of real data from daily stock return.
Source: Bonnanno et al. (2003)

### 1. Distance between the Terms

To use the MST algorithm, we need to begin with a network representation of our words. If we let each word represent a node, we then can connect each word with a weighted arc where the weight indicates the distance between each pair of words. To generate this network, we need a method for measuring the distance between words. We will discuss two options that we tested in this section.

### *a. Distance Based on the Correlation Coefficient*

Bonnanno et al. (2003) extracted a $N \times N$ correlation matrix from the trading data composed of $N$ assets traded simultaneously over a time period $T$. The correlation coefficient $\rho_{i,j}$ can then be associated with a distance between asset $i$ and $j$ as Equation 6.

$$d_{i,j} = \sqrt{2(1-\rho_{i,j})} \tag{6}$$

Similarly, we extract a $N \times N$ correlation matrix from the term-document matrix consisting of $N$ terms and $T$ sentences (documents). Each correlation coefficient $\rho_{i,j}$ can then be associated to a distance between word $i$ and $j$. In this study, we define the distance between words in a similar way so that the distance is specified by Equation 7.

$$d_{i,j} = (1-\rho_{i,j}) \tag{7}$$

### *b. Distance Based on Lift Value*

The distances between words calculated on the basis of correlation coefficients provides an intuitive and simple definition of distance but implies a potential error. It is highly likely that a word having a high frequency of occurrence has a high correlation coefficient with all other words irrespective of its importance. We also calculate and use the distance between words based on lift value. Miller (2013) introduced a "lift value" as an indicator for association rule analysis. Lift value quantifies the association strength of two events $A$ and $B$. In the Equation 8, denominator means the probability that events $A$ and $B$ will occur at the same time given that event $A$ and event $B$ are independent, and the nominator means the actual probability that event $A$ and event $B$ occur at the same time in the whole data set. That is, if there is no relationship between events $A$ and $B$, lift value will be 1, and the higher the connection strength, the higher the lift value will be.

$$Lift\_Value\_For\_A,B = \frac{P(AB)}{P(A)P(B)} \tag{8}$$

Similarly, by applying the definition of the lift value to the relationship between words in sentences, we can associate Lift value for word $A$ and word $B$ to a distance between word $A$ and $B$. This approach makes it possible to calculate the distance between words as Equation 9, without the influence of the occurrence frequency:

$$Dist\_between\_A,B = \frac{P(A)P(B)}{P(AB)} \tag{9}$$

In order to compensate for potential errors in the correlation-based distance, we use both correlation coefficient and the lift value to calculate the distance between words.

### 2. Build Minimum Spanning Tree

*Wikipedia* (Minimum spanning tree 2018) defined the MST in the following paragraph, and Figure 11 shows this definition well.

> A MST is a subset of edges of a connected edge-weighted (un)directed graph that connects all the vertices together, without any cycles and with the minimum possible total edge weight. The most distinctive feature of the MST is that it minimizes the sum of edge weights while connecting all the vertices in the graph.

In order to further solidify the meaning of the MST-based semantic network in text analysis, we define the document as a connected, undirected network with terms (words) as vertices and distance between terms as edge weights. By applying the MST to this graph, we can specify the implication of MST-based semantic network in text analysis as follows. When applied to the overall PN, it represents the common paths of thoughts over all documents through the core connection between the words that make up the document. When applied to the ISFN, it represents the flow of thoughts within each specific assessment. Additionally, it compactly represents the structural characteristics of the individual assessment. Recall that we create an MST-semantic network for the entire document in the PN and also a separate MST-semantic network for each individual student assessment in the ISFN.

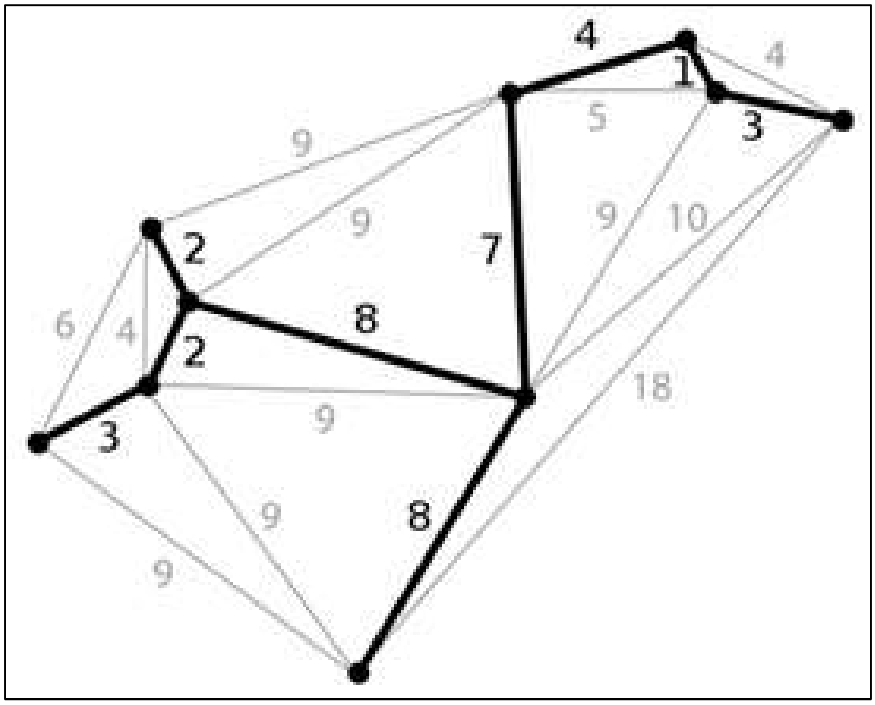


Figure 11. Minimum spanning tree as a subset of connected graph. Source: Minimum spanning tree (2018).

As mentioned in Chapter II, typical algorithms for finding MST are the Prim's algorithm (Prim 1957) and Kruskal's algorithm (Kruskal 1956). Both algorithms estimate MST that connects all nodes with the shortest distance weight without a cycle. However, the Kruskal's algorithm identifies the MST by adding the edge corresponding to the shortest path, and the Prim's algorithm identifies the MST by adding the closest node. We use the "igraph" package (Csardi et al. 2006) of R to generate the MST-based semantic network from the distance matrix. By default, the "igraph" package uses Prim's algorithm to build the MST.

Figure 12 and Figure 13 show the MST-based semantic networks of the entire criticism text based on 69 words, which accounts for 50% of the words occurring in the entire criticism text across all assessments. The former is generated from the correlation coefficient-based distance matrix, and the latter is generated from the Lift value-based distance matrix. The two figures show that there is a clear difference in structure between the two graphs generated by applying different distance definitions to the same text data.

The difference between the two MSTs is clear for the words located at the center of the network. As discussed in the previous section, when MST is generated based on the correlation coefficient, frequently occurring words tend to be located at the center of the tree, whereas when the MST is generated based on the lift value, this tendency differs. It would be reasonable to apply the correlation-based MST considering the topic model estimation through the clustering technique which will be discussed later. However, before comparing the prediction accuracy by predicting the acceptance of students based on the characteristics extracted from the network, we cannot determine which definition of the distance is more appropriate. Therefore, we perform the analysis using both definition and compare the results.

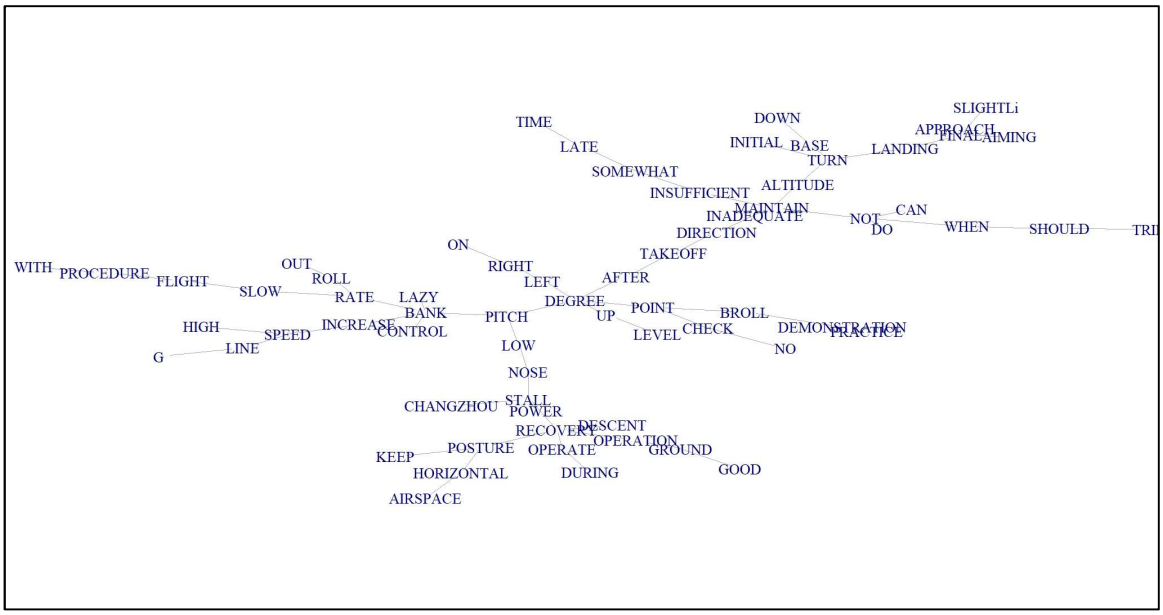


Figure 12. Correlation based MST of total flight criticism text

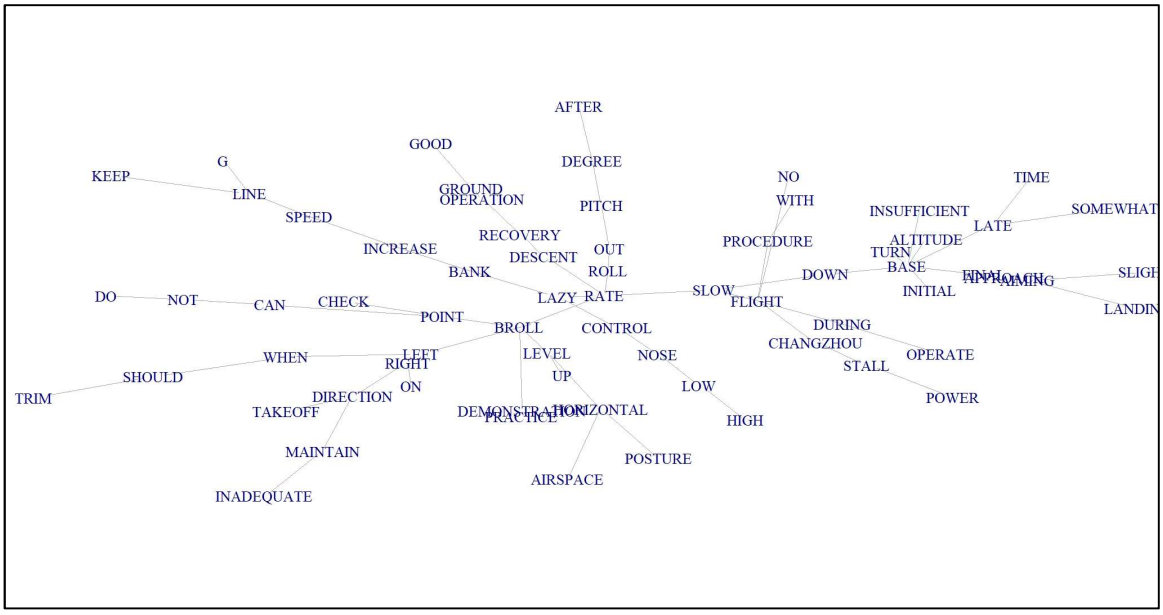


Figure 13. Lift based MST of total flight criticism text

## E. PIPELINE PART 3: CLUSTERING

If the MST generated from the document represents the flow of thought embedded in the document and the structural characteristics of the document, the cluster extracted from MST can be regarded as the topics of the document. We develop the topic model that constitutes the flight training through the clustering of the MST generated from the whole document (i.e., the PN).

To extract a topic model, we attempt to use *K*-means clustering as well as network community algorithms as clustering methods. We also vary the possible number of clusters (topics) of flight training from 3 to 6.

As mentioned in the overview at the beginning of Chapter III, we perform a two-stage model selection procedure to find the best parameter settings for the model. First, we perform the pipeline on every combination of settings and log the performance of the classification in terms of cross-validated AUC. Second, we isolate the top performing models and utilize a combination of cluster evaluation techniques, including reproducibility, within cluster sum of square (WCSS), and modularity, as well as basic visual inspection to choose the best settings for generating the PN and ISFN. We find that

models generated from the most of settings have high AUC values. In particular, as the number of clusters increases, the AUC tends to be higher because the variables included in the model increase as well. Therefore, a secondary selection process is necessary to find models that are robust to slight changes in parameters.

### 1. *K*-means Clustering

In general, the *K*-means clustering algorithm uses Euclidean distances but occasionally Manhattan or Minkowski distances are used. Because the distance function is used for cluster estimation, all variables used for clustering must be numeric. However, it is impossible to directly apply the *K*-means algorithm to the MST since the MST extracted from the document specifies only the relative connection between the nodes but does not specify an absolute numeric value (i.e., a Euclidean distance between all nodes). In order to apply the K-means algorithm to MST, we use the network layout algorithm. Network layouts are simply algorithms that return Euclidean coordinates for each node in a network. The most typical algorithm among various algorithms is developed by Kamada (Kamada et al.1988). The latter optimizes the coordinates of each node pair so that the distance on the graph reflects the path distance. We utilize the Kamada algorithm to MST to estimate the coordinates of each word in the two-dimensional Euclidean space and calculate the distances from the estimated coordinates to perform clustering.

### 2. Network Community

Chapter II introduced the methodologies of dividing the network into communities, such as the edge betweenness community, the random walk community, the fast-greedy community and modularity index as an evaluation indicator for the derived community. We also utilized these methods for comparison to *K*-means.

## F. PIPELINE PART 4: CLUSTERING PERFORMANCE EVALUATION

As already mentioned, we utilize a number of quantitative characteristics to evaluate the quality of clusters generated to assist in model selection. We will detail some of these quantitative indicators in this section.

### 1. Quantitative Evaluation

**Hartigan's rule-based evaluation:** A widely used criterion for determining the optimal number of clusters is Hartigan's Rule. This basically compares the ratio of the within-cluster sum of square (WCSS) to the clustering of K and K+1 clusters. In this study, we used Hartigan's rule to estimate the ratio of words to be analyzed and the appropriate number of clusters

**Modularity based evaluation:** The result of deriving the community from the network is evaluated as "Modularity index." A larger positive value for the "Modularity index" indicates better cluster performance. Figure 15 shows the variation of the "Modularity index" according to the clustering methodology and the number of clusters. In this study, we use the "Modularity index" as an index to compare the clustering methodologies and to estimate the appropriate number of cluster.

Figure 14 shows the result of applying Hartigan's rule to each network cluster generated from 50% to 80% of the whole criticism text. The clustering performance was higher when using a smaller number of words and the number of clusters was estimated to be between 4 and 6.

Figure 15 shows the performance of each clustering methodology according to the number of clusters based on "Modularity Index." As with Hartigan's rule, we can estimate that the number of appropriate clusters is 4 to 6.

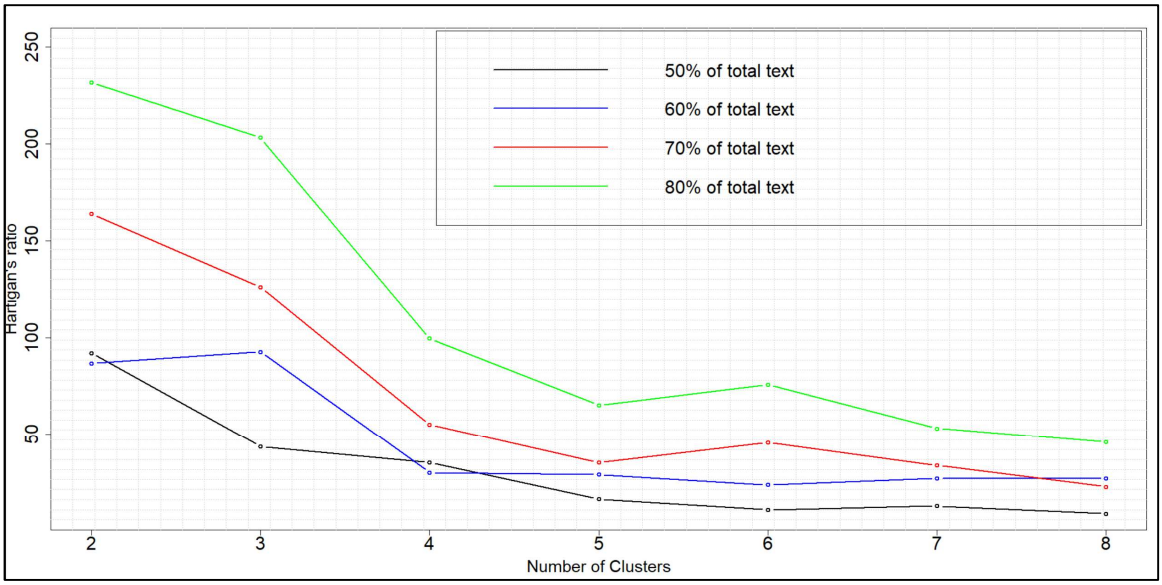


Figure 14. Hartigan's ratio per number of clusters and word proportion

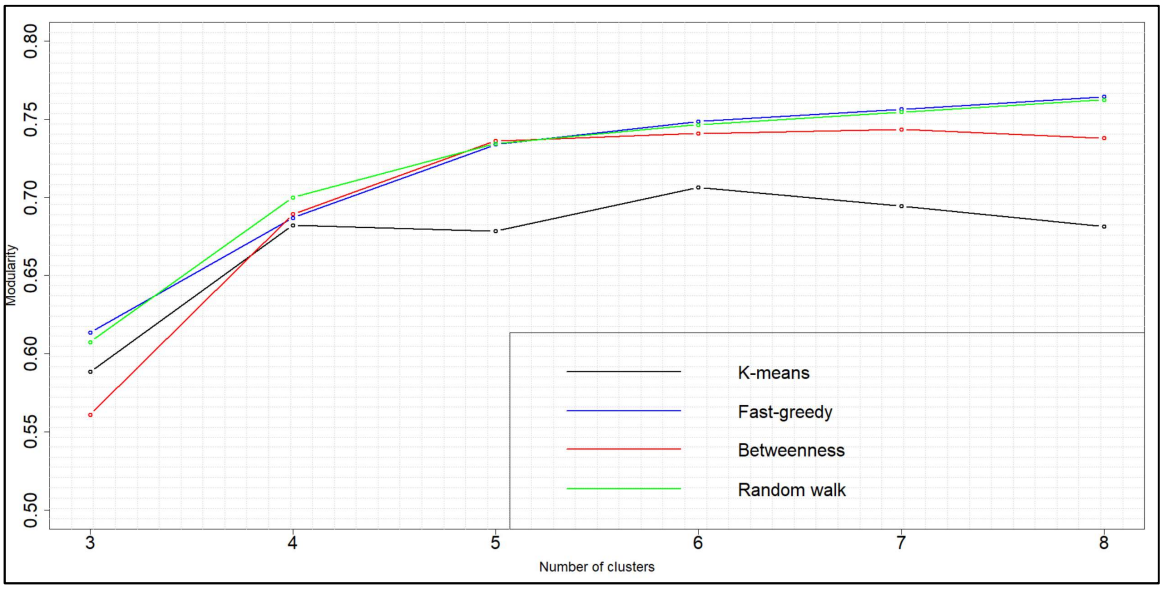


Figure 15. Modularity index per number of clusters and clustering method

**Reproducibility based evaluation:** In Chapter II, we reviewed the concept of reproducibility in clustering and the "Rand index," an indicator of reproducibility. Because the clustering algorithms depend on "random starting point," the clustering results are different each time. However, if the number of clusters is adequate, clustering results will

be similar regardless of the starting point. Accordingly, we used "Rand index" to identify the proper scope of terms and the number of clusters. We used the methodology suggested by Huh et al. (2004) for the reproducibility evaluation with the following steps.

1. Divide the data to be clustered into three and named data sets 1,2, and 3
2. Perform clustering on data set 1 to create clustering rule 1.
3. Perform clustering on data set 2 to create clustering rule 2.
4. Classify each object of data set 3 according to rule 1 and rule 2.
5. Calculate the "Rand index" by comparing the cluster classification results of data set 3 from clustering rule 1 and rule 2.

A high "Rand index" means high reproducibility of clusters under the given conditions. Figure 16 shows the reproducibility of each network cluster generated from 50% to 80% of the whole criticism text. Similar to other clustering evaluation methods, the reproducibility is high when the number of clusters is 3 or 4.

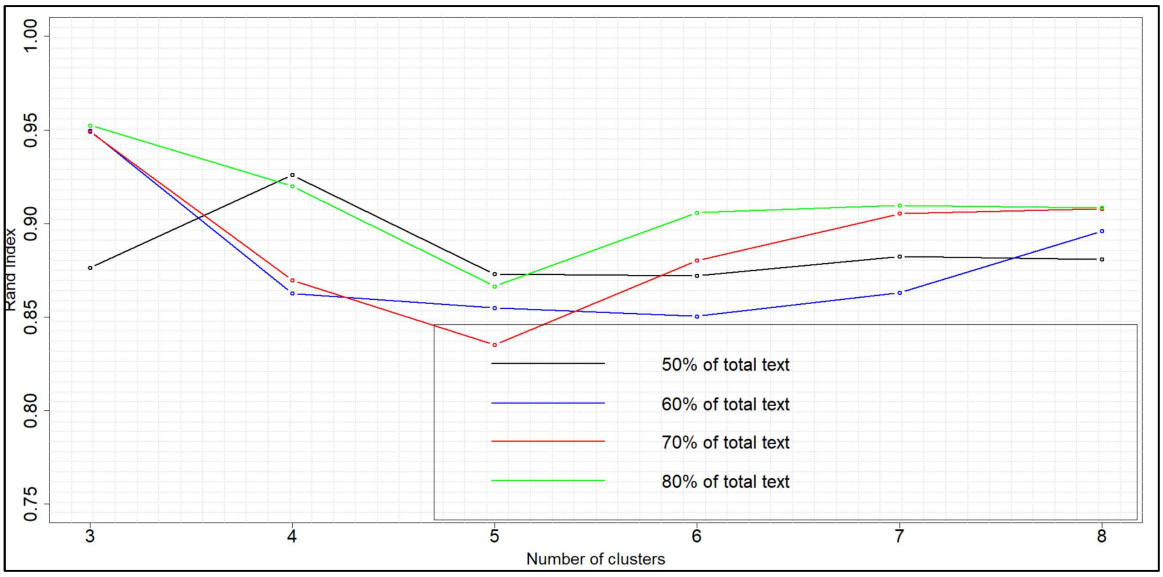


Figure 16. Rand index per number of clusters and word proportion

## 2. Qualitative Evaluation

Although the methodologies applied in the previous section provide a reference standard for the selection of appropriate clusters and topic model, a qualitative analysis is inevitable to assess whether the topic model had practical meaning.

Figure 17 shows the clustering results for a network defined by correlation distance and a network defined by a Lift value-based distance. As mentioned in the previous section, we could visually confirm that correlation distance-based networks are more suitable for clustering.

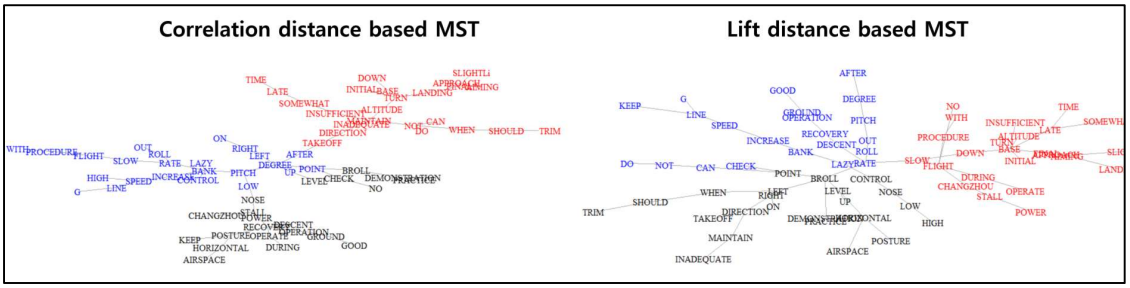


Figure 17. Comparison of MSTs based on correlation and lift distance

Figure 18 shows the result of estimating three topic models by applying K-means clustering and network community methodologies to a correlation distance-based network that accounts for 50% of the total criticism data. Even if the same number of clusters is applied to the same network, the topic models are slightly different according to the clustering methodology.

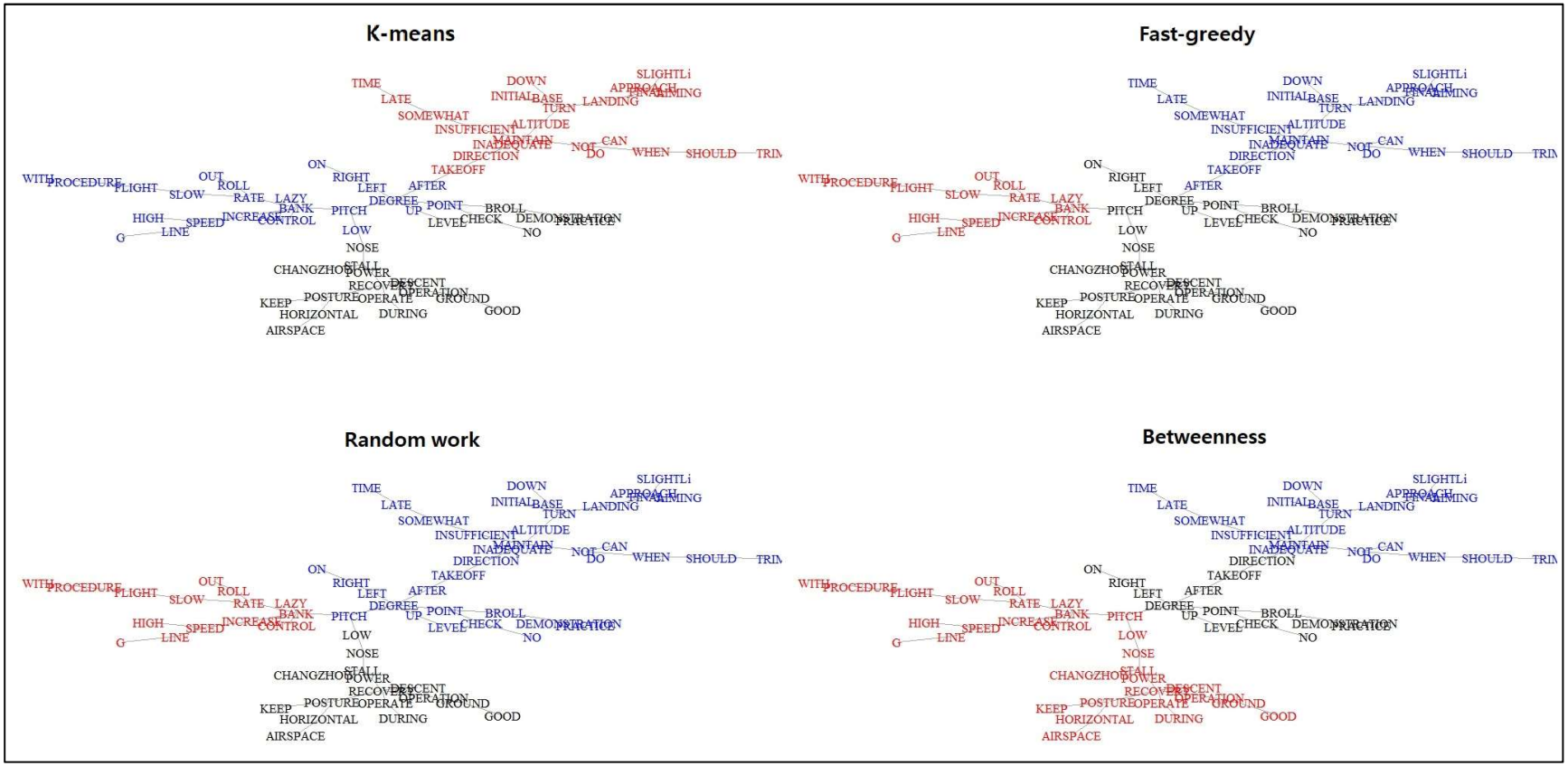


Figure 18. Comparison of outputs from different clustering methods

## G. PIPELINE PART 5: EXTRACTING FEATURE FROM SEMANTIC NETWORKS

As shown in Figure 9, we classified the documents into 3 levels, overall criticism text, student-specific criticism text and student-flight specific criticism text. The term document can be generated from each level of document, and the MST-based semantic network can be generated from the associated term-document matrix. One of the advantages of representing a document as a semantic network is that we can infer the characteristics of the document through quantitative indicators that represent structural characteristics of the network. We utilize the quantitative indicators extracted from the semantic network to identify the major factors that determine the acceptance and rejection. This section explains our methodology for extracting quantitative indicators and the qualitative interpretation of each indicator.

### 1. Network Density

One quantitative feature that we measure for each individual student is the density of the ISFN. The network density is an index indicating the global characteristics of the network. The density is defined as a ratio of the number of edges to the number of possible edges in a network with "n" nodes. When the number of nodes in the network is n, the number of possible edges in the network is $n(n-1)$ . Let $g_i$ be the number of edges that node $i$ has in an undirected network, the density can be calculated as Equation 10.

$$density = \frac{1}{2}\frac{\sum_{i=1}^{n} g_i}{n(n-1)} \tag{10}$$

The fact that the density value of the semantic network is high implies that the links between the words constituting the document are evenly distributed. On the other hand, the fact that the density value of the semantic network is low implies that the links between words in a document are concentrated in a specific word pair. When focusing on only a single flight, we form the ISFN for every student from only a single assessment for that flight. For the two methods that utilize every flight, we still utilize the characteristics measured for each flight's individual ISFN, combining by averaging or by concatenating.

### 2. Modularity

Modularity is another measure of the structure of a network that we measure for each ISFN. Using the definition introduced in Chapter II, we say that two words (nodes) belong to the same community if they belong to the same topic (cluster) under the topic model implied by the clustered PN.

As specified by Equation 5, we can calculate the modularity index from network adjacency matrix and the cluster (community) of the nodes constituting the network. The high modularity index value means that when the network is divided into a given cluster information, the connection strength between nodes in the same cluster is strong, and the connection between nodes in the different clusters is weak. The "Modularity Index" extracted from the semantic network indicates how well the topic that make up the document are separated within the document. In other words, the high Modularity index of the criticism data can be interpreted as the fact that the training for the flight of the student was clearly divided into each subject. On the other hand, the low Modularity index of criticism data can be interpreted as the fact that the training was conducted in a comprehensive manner.

### 3. Topic Structure

The clusters identified in the PN in this study represent the topics that constitute the document, and each topic is represented by a set of words corresponding to the topic. By analyzing the topology of each topic, or the distribution of topic-based words, within the ISFN semantic network, we can extract indicators that represent the structural characteristics of the individual assessments in the context of the derived topic model.

#### *a. Proportion and Frequency of Topics*

The simplest feature associated with each individual student-flight assessment is the proportion of words within the assessment that belong to each topic from our topic model. For a topic model with *K*-clusters (*K*-topics), we generate k features for every assessment indicating the frequency with which the chosen topic is mentioned in the assessment.

### *b. Closeness Centrality of Topics*

Closeness centrality of each node represents how close each node is to the center of the network. In other words, high closeness centrality means that the node is close to the center of the network. Let $d(i,j)$ be the distance between node $i$ and node $j$ in a network composed of $N$ number of nodes, the closeness centrality of node $i$ can be defined as Equation 11.

$$C_i = \frac{n-1}{\sum_{i \neq j} d(i,j)}, i = 1, \cdots, n. \tag{11}$$

We define the centrality of each topic within the ISFN semantic network as the average of the closeness centralities for words within the assessment that belong to the topic. This average closeness centrality of each topic can be interpreted as degree of importance. Note that the words within the topic are defined using the clustered PN, while the closeness centrality for those words is calculated using the ISFN. Thus, this generated one flight-assessment feature for every topic found via clustered PN.

### *c. Emotional Degree of Topics*

For emotional analysis on text data, we must define a list of emotional words while considering the characteristics of the target text. The instructor's evaluation of the flight is expressed through positive or negative words such as "GOOD" or "BAD." In addition, words such as "FAST" and "HIGH" are very important in the flight criticism text. We use the words that are highly likely to express the evaluation of the instructor, among the frequently occurred words in the criticism text, and used it as a word list for emotional analysis. The list of identified emotional words are in Table 2.

Table 2. Emotional words in flight criticism text

| Category | Emotional words for the category |
|---|---|
| **POSITIVE** | "GOOD" |
| **NEGATIVE** | "NOT," "INSUFFICIENT," "NO," "MISS," "INADEQUATE," "POOR" |
| **OTHERS** | "LOW," "SLOW," "HIGH," "FAST," "LATE" |

Considering that the connections between words in the MST-based ISFN semantic network represent the flow of thoughts embedded in the criticism text, the distances between the emotional word and the words representing the topic can be regarded as a measure of the corresponding emotion associated with the topic within the individual assessment. Under this concept, for every assessment (ISFN), we associate a numerical score for every (topic, emotion-word) pair to indicate the strength of each emotion connected with the topic within the ISFN. Specifically, this study used the mean value of the distance between each emotional word and the words in the ISFN that also belonged to the cluster/topic in the PN network as a quantitative measure of the emotion on the topic. Figure 19 shows the result of identifying four topics in the MST-based PN semantic network generated from the whole criticism data. For the convenience of explanation, let's say that the identified topics are "TOPIC RED," "TOPIC GREEN," "TOPIC BLUE," "TOPIC BLACK." The intensity of the positive evaluation of each topic can be inferred from the distance between "GOOD" and other words. The distance is defined as "**the number of edges required to connect the "GOOD" with target word**." Note that this definition is different from the distance defined for MST based semantic network generation. (e.g., the distance between "GOOD" and "GROUND" is 1, and the distance between "GOOD" and "GROUND" is 2.). By applying our methodology to the PN network, we can estimate that the evaluation of "TOPIC RED" which is closest to "GOOD" (or which has the strongest connection strength with) is the most positive. Note that we only use the PN network for illustrative purposes and we measure distances on the ISFN.

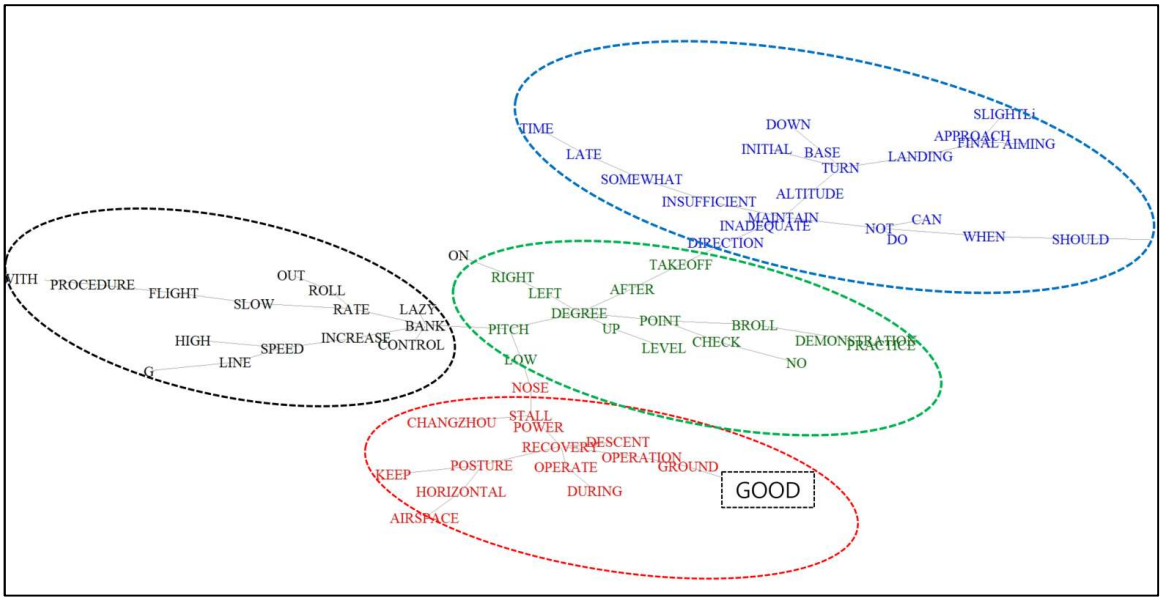


Figure 19. Positive emotional distance for each topic of flight training

## H. PIPELINE PART 6: CLASSIFICATION AND EXPERIMENTAL SETTING

Figure 20 describes the process for building models that predict students' passing and failing based on the criticism text. The model building process is divided into three steps. First, we build the PN and the associated topic model of flight training from the overall criticism text. Second, we generate the ISFN and extract predictors based on each flight-student specific term-document matrix. Finally, we build the model for predicting passing and failing with the extracted predictors by applying two supervised learning techniques.

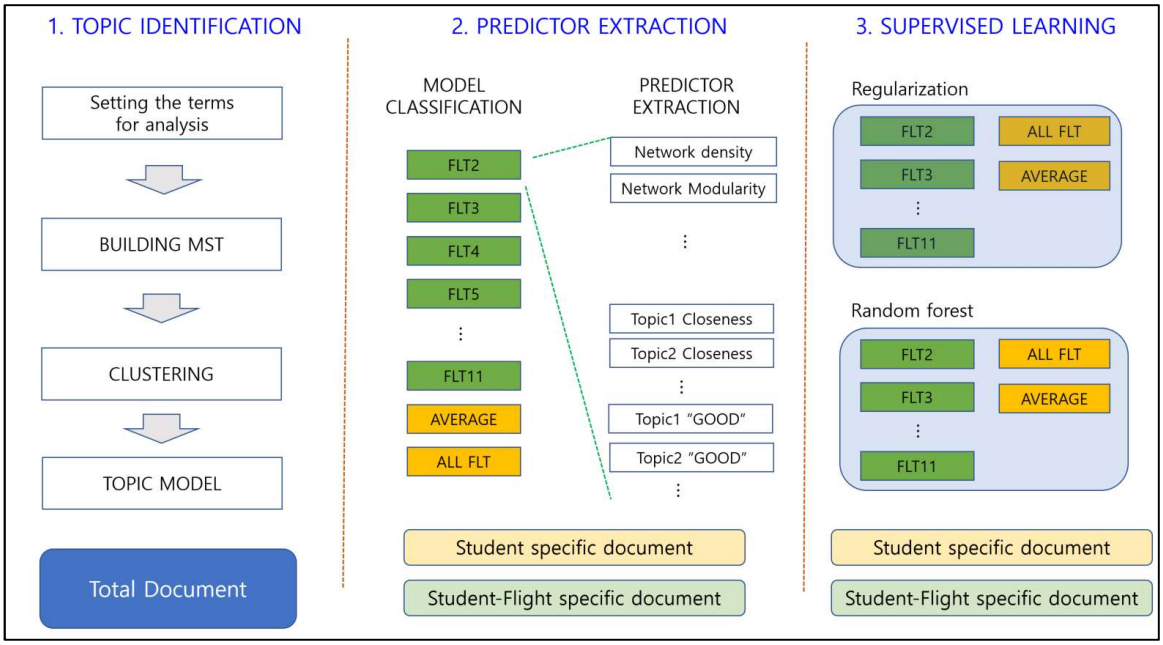


Figure 20. Modeling building process for flight criticism text analysis

## 1. Topic Model Identification

We identify a topic model for the entire document via the PN. At each step of topic model building, parameters and methodologies for semantic network generation and clustering should be decided first. Figure 21 shows the parameters to be determined for each step and the possible options for each parameter and methodology.

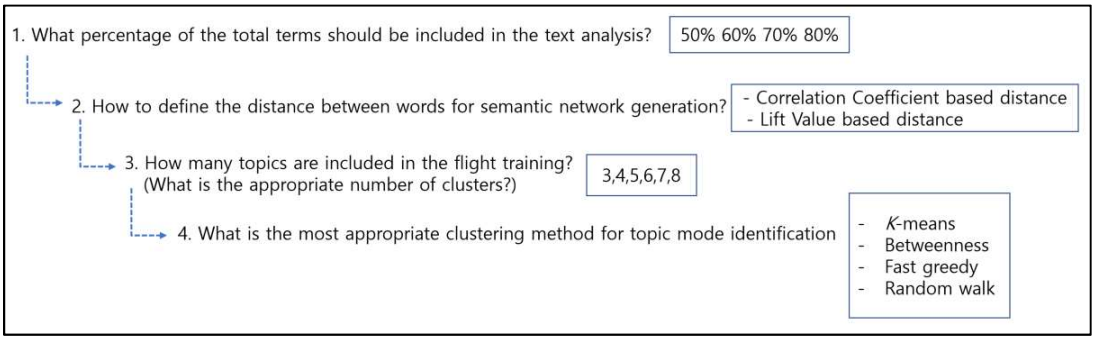


Figure 21. Parameters and methodologies to identify the topic model

By applying all of the options presented, a total of 144 cluster sets were yielded as a possible topic model. In order to find the most reasonable topic model, we need to choose parameters that best reflect the characteristics of the flight criticism text. However, it is almost impossible to find optimal parameters for unstructured text analysis. Although we can establish the methodologies and parameters for the topic model estimation using rules of thumb for clustering or prior knowledge related to the flight training, randomness and subjectivity can eventually lead to bias in the fitted model. To avoid such bias and to perform objective analysis, we fit independent prediction models for each possible topic model based on different methodologies and different parameters. We then look at the top performers in terms of predictive performance to select our eventual topic model.

## 2. Model Classification and Extracting Predictors

### *a. Model Classification*

The criticism text is written separately for each flight performed by each student. We apply the methods presented in the previous chapter to create an ISFN for each flight of each student and extracted structural and emotional characteristics and fitted an independent classification model for each flight except the first flight. Each model has the same types of predictors. Through the analysis and the comparison of the prediction model based on different flights, we identify the weight of each flight and the main passing and failing factors for each flight. In addition, we construct the "Whole" model that includes the characteristics extracted from all flights as predictors and the "Average" model with the mean value of characteristic from all flights. These two different models were used to identify whether the time-sequential manner of flight training has a significant difference as a factor in determining passing and failing.

### *b. Extracting Predictors*

We utilize the structural and emotional characteristics extracted from the semantic network as a predictor of passing and failure. The method of extracting characteristics from the clustered semantic network and the interpretation of the extracted characteristics were described in detail in the previous chapter. Predictors are categorized into two categories: global characteristics, and topic-specific characteristics. "Network Density" and

"Modularity" correspond to global characteristics. Proportion of topics in the document, centrality and emotional distance correspond to "topic-specific" characteristics. In this study, the number of emotional words for flight criticism text is set to 12, so the total number of predictor extracted for each topic is 14, including "Proportion" and "Closeness centrality."

All types of models have the same predictors. When the topic model is developed based on three clusters, each model for flight#2 through flight#11 has 44 predictors. However, when the characteristics extracted from each flight are included in one model, the total of 440 predictors are included.

For the convenience of interpretation of results, we named predictor according to the type of model and the characteristics of variables. Figure 22 shows the three types of predictor naming conventions of this study. "Example1" means "Network Density" which is a global characteristic (Network Property) extracted from the second flight. "Example2" represents the proportion of the "TOPIC1" in the second flight. "Example3" represents the distance from emotional word "GOOD" to "TOPIC1" in the second flight.

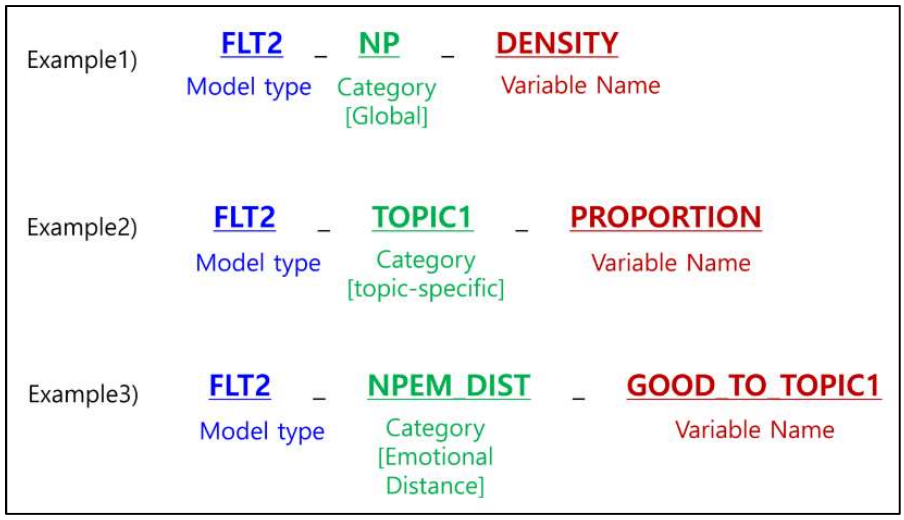


Figure 22. Naming rules for predictors in flight criticism text analysis

## 3. Supervised Learning

This chapter explains the methodology of training the model using the features extracted from the semantic network and evaluating the performance of the trained model. In particular, the model constructed in this study is based on high dimensional data with each assessment having many extracted features. In order to prevent "Over-fitting" from high-dimensionality and to identify the factors that are critical to passing and failing, we build a model with regularization for feature selection.

### *a. Train / Test Set Sampling*

In order to train and evaluate the model, we should divide the data into training set and test set. As mentioned in Chapter I, the data used in this study consist of 332 students who passed the flight test and 43 students who failed the flight test. Ertekin (2013) proposed an oversampling technique for the minority class and its application in supervised learning. Ertekin (2013) evaluated the advantages of oversampling in terms of prediction performance. In the flight criticism data, the minority class is students who failed the flight test. This imbalance of response variables not only leads to a reduction in prediction performance but also to a biased model. The purpose of our model is not only to predict passing and failing, but to extract factors that affect passing and failing through the selected variables in the fitted model. Because the biased model cannot sufficiently reflect the characteristics of the failing students, we perform an oversampling of the failing students to extract characteristics of both passing and failing students. This oversampling additionally allows us to guarantee that the cross-validation process yields valid data splits, with sufficient examples from each class within each split. We extend the data of failed students through random sampling with replacement to compensate for the imbalance of response variables. Figure 23 illustrates the process of extending data and sampling training set and test set.

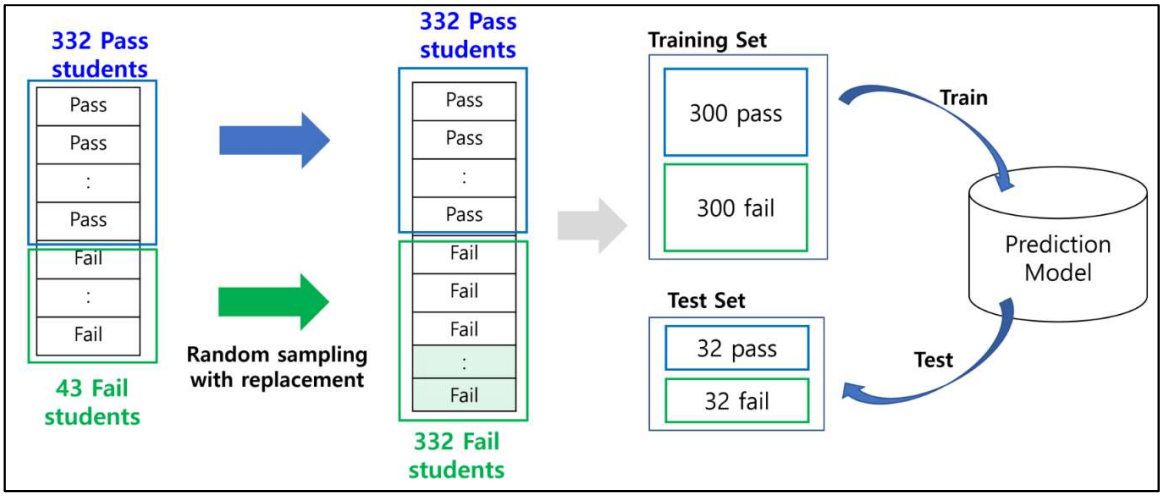


Figure 23. Training and test set sampling process for flight criticism text analysis

### *b. Standardization of Predictors*

The reason for constructing the passing and failure model in this study is to analyze the weight of selected variables in the model to identify the factors that are crucial to passing and failure. We develop the weight of each variables from the coefficient value assigned to each variable in the finally fitted model. When the scale of each predictor is different, the scale of the coefficient value would also different. To compare the impact of all features equivalently, we standardize predictors to compare the coefficient value assigned to the predictors on the same scale. Through the standardization, predictors are transformed so that the mean is 0 and the standard deviation is 1. Let the mean of predictor $x_1, x_2, \cdots, x_n$ with n number of observation is $\bar{x}$ and the standard deviation is $s_x$, the standardized predictor $z$ is expressed as Equation 12.

$$z_i = \frac{x_i - \bar{x}}{s_x} \tag{12}$$

### *c. Prediction and Regularization*

We utilize logistic regression with Lasso regularization for variable selection. For data consisting of N observations with predictive variables $x_t$ and binary response variable $y_t$, a generalized linear model is fitted to minimize the likelihood by choosing the

coefficient values $\beta$ . The formula for maximizing likelihood in generalized linear model is specified by Equation 13.

$$\max_{\beta} L(\beta) = \prod_{i=1}^{N} p(x_i, \beta)^{y_i} \left(1 - p(x_i, \beta)\right)^{1-y_i}$$
$$where: p(x_i, \beta) = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \tag{13}$$

In order to avoid overfitting and to select variables, Lasso fits the model by applying an L1 penalty to the likelihood. Model fitting formula for Lasso is specified by Equation 14.

$$\min_{\beta} -\ln L(\beta) + \lambda \|\beta\|_1 \tag{14}$$

In the Equation 14, $\|\beta\|_{l1}$ limits the size of coefficients. Some coefficients can become zero and eliminated with $\|\beta\|_1$ penalty. Therefore $\lambda$ value determines the amount of reduction. We implement Logistic regression with Lasso using "glmnet" R package (Friedman J et al. 2010).

### *d. Cross Validation*

To evaluate the predictive performance of the model we use the AUC. We select AUC as opposed to classification accuracy due to the imbalanced nature of our data set. AUC has been shown to be a better measure of performance in these situations.

We select the $\lambda$ parameter for regularization through 5-fold cross validation with evaluation based upon AUC. Figure 24 shows the cross-validated AUC for different $\lambda$ values. In the figure, the number of variables included in the model decreases as $\lambda$ increases. There are two main criteria for selecting the optimal lambda value. The simpler method is to choose a $\lambda$ that maximize the AUC. The "1se" rule is a method of selecting $\lambda$ within one standard error range of maximum AUC value. In Figure 24, number of selected predictors is 44 and when the "1se" rule is applied, the number of selected predictors is 38.

If there is no significant difference in the performance of the model, a simpler model should take the priority. Therefore, we apply the "1se" rule to select $\lambda$ and fit model.

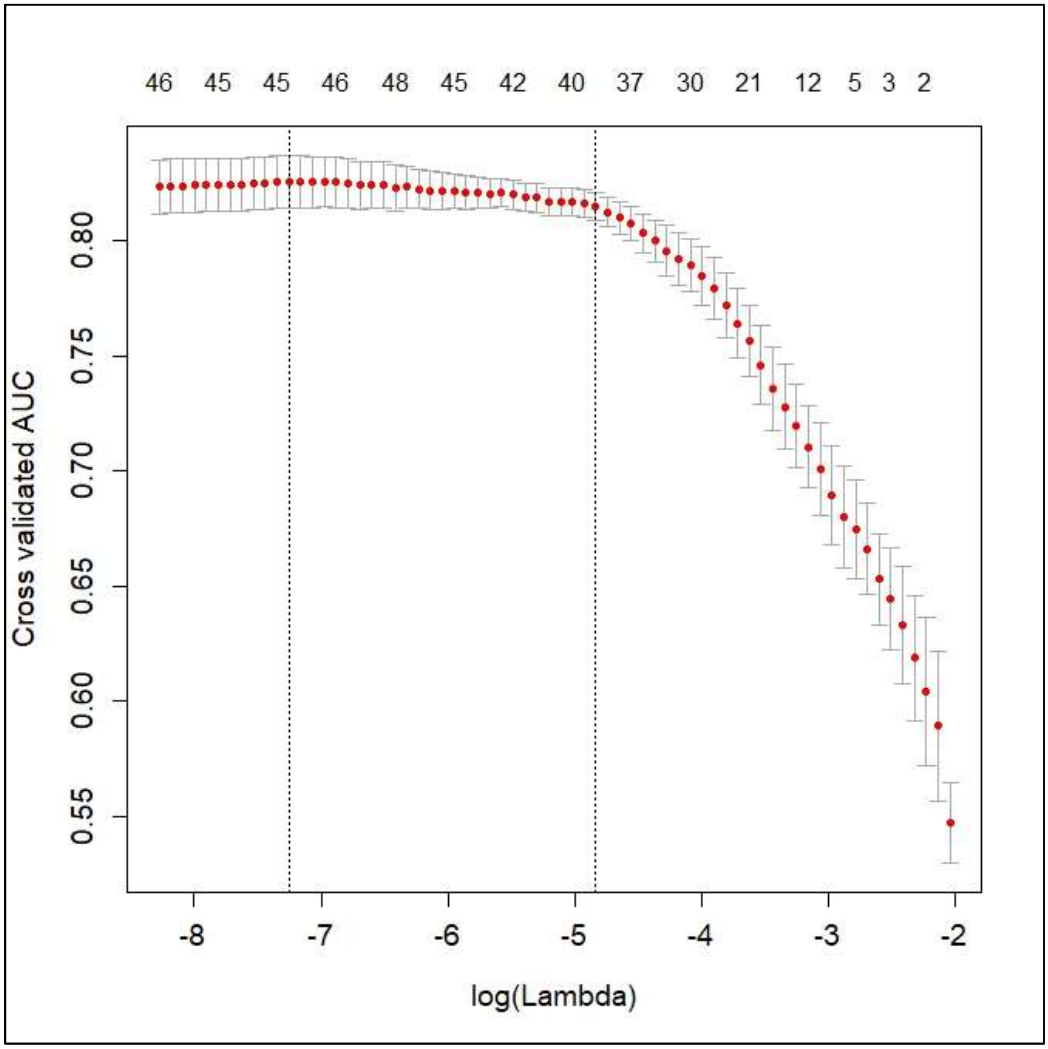


Figure 24. Cross validated AUC of passing failing prediction per $\lambda$

THIS PAGE INTENTIONALLY LEFT BLANK

# IV. RESULT ANALYSIS

Our primary premise is that the predictive power (with respect to student outcomes) of the features extracted from the semantic network model can be a measure of how well the model represents the characteristics and topics of the flight-training program. Under this premise, this chapter quantitatively compares the definitions of word-to-word distance, the number of cluster, the clustering methodology, and the importance of each flight based on the performance of the model evaluated with AUC. Also, through the qualitative analysis, we select the appropriate topic model and identify the major factors for predicting passing and failing and extract the characteristics of individual flights.

## A. QUANTITATIVE ANALYSIS

A total of 144 topic models can be created through the experimental design of this study through the different combinations of methods and parameters. We run the prediction pipeline for all 144 settings, logging the predictive performance with respect to AUC. In this section, we will discuss the variation of predictive performance across the different factors that comprise the overall pipeline.

### 1. Number of Terms to Include for Preprocessing

In order to estimate the number of terms to be included in the analysis, we compare the predictive performance of the model using terms covering 50% to 80% of the total terms in the document. Figure 25 shows the distribution of cross validated AUC values for all 144 models with respect to these parameter settings. Because the average value of AUC for the models that use 50% of terms (only 69 terms) is more than 0.8, we can argue that performing analysis on only 69 most frequently occurred words might be sufficient. The performance of the model is not greatly improved even if it is based on a wider range of words. Considering that number of words to analyze exponentially increases as the percentage of terms increases, it is reasonable to conclude that only 50% of terms are needed.

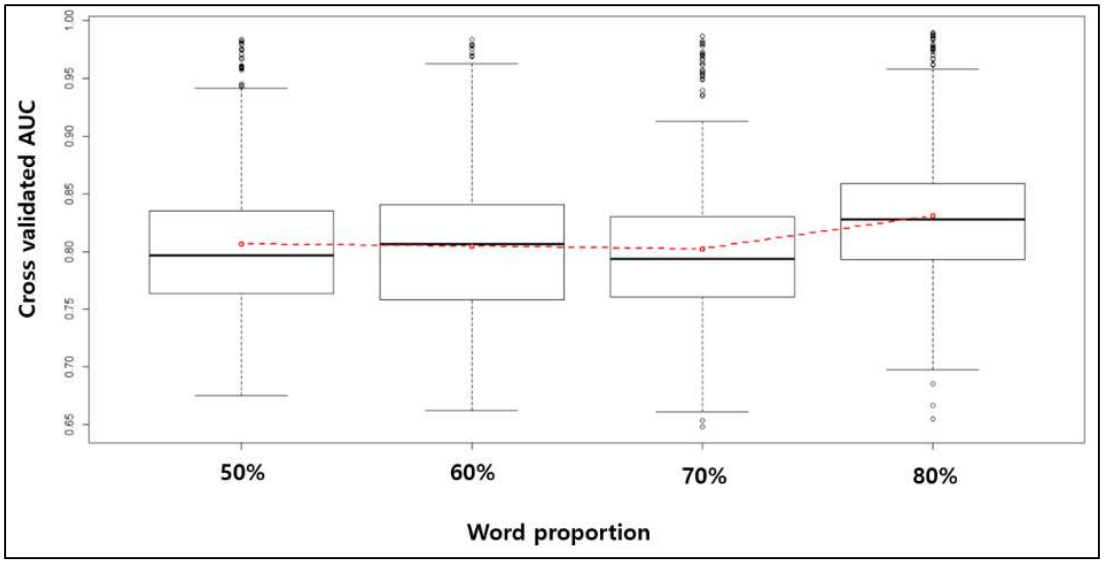


Figure 25. Five folds cross validated AUC per scope of terms

## 2. Word Distance Definition

To determine which distance measure is appropriate to measure the distance between words, we compare the predictive performance of all 144 models, isolating the network models based on the correlation coefficient distance and the network models based on the lift value. Figure 26 shows the distribution of cross validated AUC for all 144 models differentiated only with respect to the distance measure. Overall, the AUC value of correlation distance-based models is slightly larger than that of lift value-based models. In Chapter III, we identified that correlation-based distance is potentially better because the structure of the generated semantic network is more suitable for clustering. The performance comparison results are consistent with our initial inference. Based on these results, we apply correlation-based distance to generate the PN for topic model estimation and IFSN for predictor extraction.

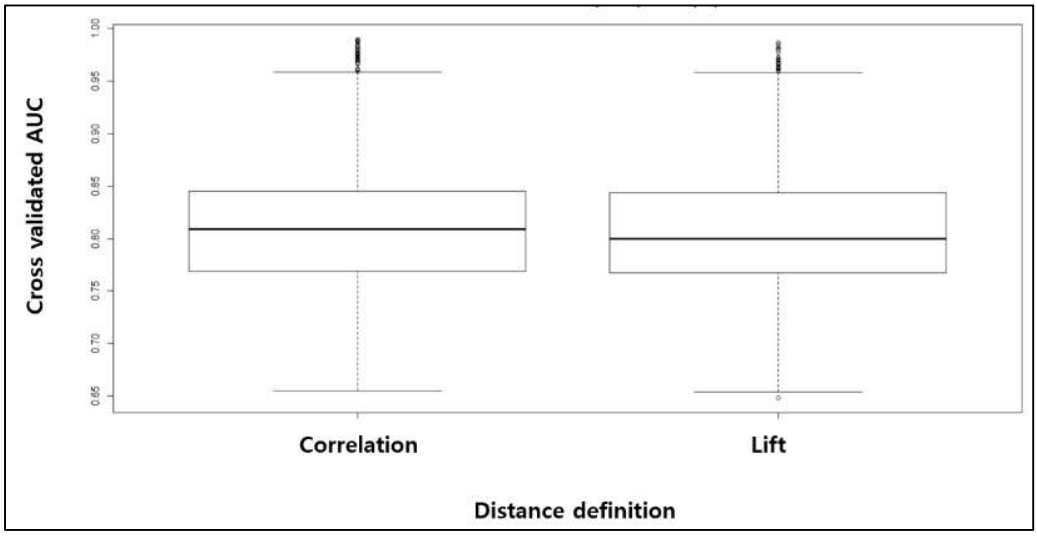


Figure 26. Five folds cross validated AUC per distance definition

### 3. Number of Clusters

In order to estimate the appropriate number of topics (clusters), we compare the predictive performance of models with different numbers of clusters. Figure 27 shows the distribution of cross validated AUC values of all 144 models from the different number of clusters. The performance of the model does not increase significantly (in terms of the best performing models with AUC around .97) even if the number of clusters is increased. In our experimental setting, whenever the number of clusters increased by one, the number of predictors included in the model increased by 14. We can argue that the explanatory power of the additional predictors is not strong because the performance of the best model does not increase significantly as the number of clusters increases. This leads us to conclude that between 3 and 5 clusters are appropriate for maximal AUC performance. These results are consistent with the result of clustering performance evaluation using Hartigan's rule, Modularity index, and reproducibility.

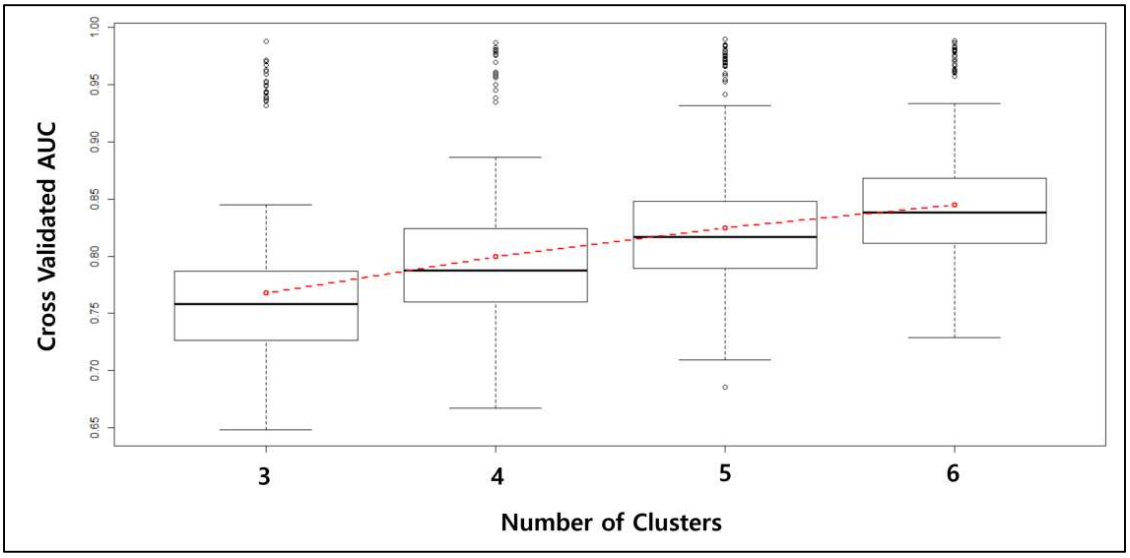


Figure 27. Five folds cross validated AUC per number of clusters

### 4. Clustering Method

In order to identify the most appropriate clustering method, we compare the predictive performance of the model using different clustering methods. Figure 28 shows the distribution of cross validated AUC values of all 144 models differentiated with respect to the different clustering methods. The performance of the models is very similar. However, looking at the variance of the evaluated performance, we see that the "K-means" methodology yields more stable results across changes to other parameters.

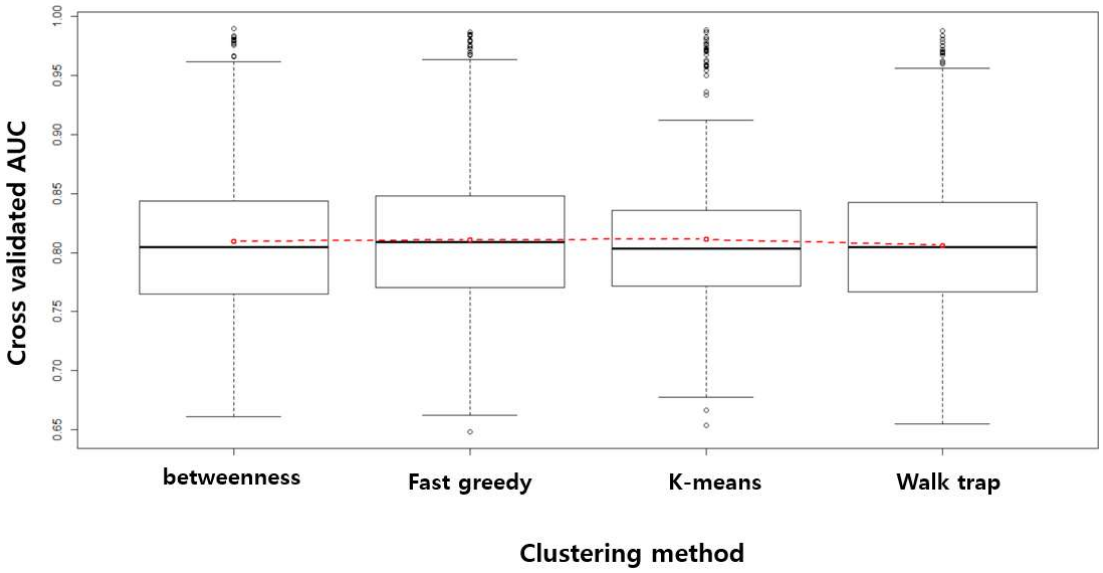


Figure 28. Five folds cross validated AUC per clustering method

## 5. Weight Analysis of Each Flight

Of the 144 models produced, we also are able to differentiate between them with respect to which flight (2-11) was used to generate the ISFN and accompanying features. Our pipeline produced classification models that utilize data only from individual flights, or data from all flights together (the "ALL" and "AVERAGE" models). We can then look at the predictive performance of these models to estimate the predictive power of each flight individually. Figure 29 shows the distribution of cross validated AUC values of all 144 models differentiated with respect to the flight that was used to generate the ISFN's and the corresponding features for each assessment.

"FLT2" refers to the second training flight, "FLT3" refers to the third training flight, etc., and the box plot for each flight is the distribution of the AUC values when the model is fit based only on the criticism data for that flight. The "Average" model is the average of all characteristics extracted from each individual flight, and the "All" model has all of the characteristics extracted from each flight concatenated into one large set of predictors. The difference in the performance of these models means that the importance of each flight is different.

It is intuitive to think that the influence of a flight carried out later in the program has greater influence than an earlier training flight. However, we see surprisingly that the flight with the most predictive power is the second flight (Flight #2). The analysis for this notable feature will be done in the next section through a detailed analysis of the structure of the model after selecting the topic model.

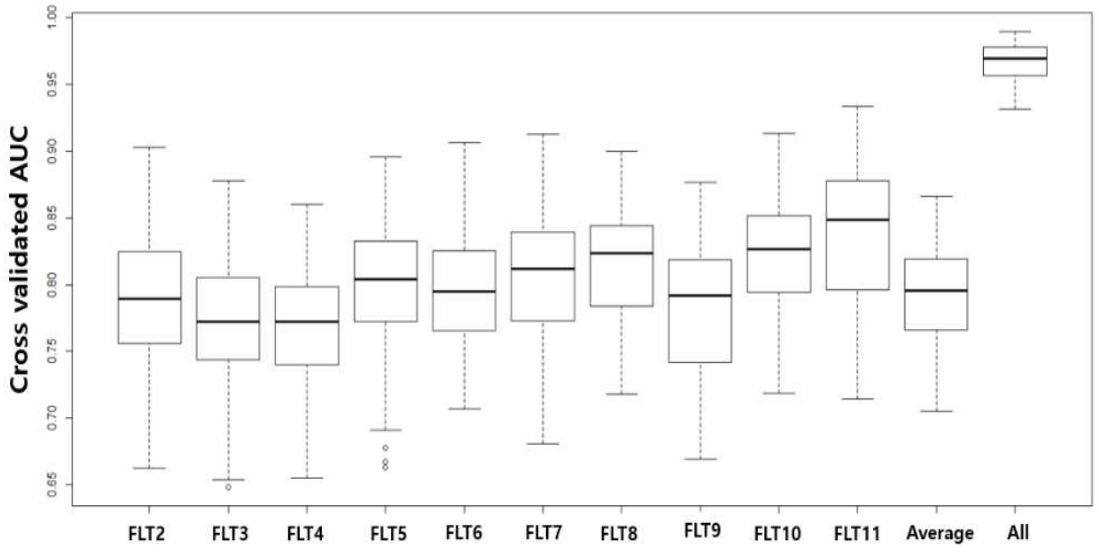


Figure 29. Five folds cross validated error per model type

## B. QUALITATIVE ANALYSIS

### 1. Topic Model

So far, we have utilized AUC values to indicate the ideal parameters and method choices for creating our model generation pipeline. For example, we saw that AUC, Hartigan's rule, modularity index and reproducibility all suggest we use 3–5 clusters. However, this evaluation criterion only allows us to narrow down the potential set of models. To select the best topic model among these top quantitative performers, we turn to qualitative evaluation (i.e., trial-error visual inspection). We see that we can locate among these topic models a robust topic model that clearly encapsulates the core components of flight training.

### *a. Topic Model Selection Based on Reproducibility*

One of the results identified through the qualitative analysis is that there is a significant difference in the results of clustering according to the methodology applied even if clustering is performed on the same data. However, when the number of clusters is set to 4 with number of terms corresponding to 50% of entire document, the results from different clustering methodologies were almost same. Figure 30 shows the results of creating a semantic network and performing clustering for the number of terms equal to 50% of the entire document with 4 clusters. The similarity of these results is strong evidence for the validity and robustness of the topic model. Additionally, through visual inspection and analysis, we find that we can also give very practical meaning to each cluster, which we discuss in the next section.

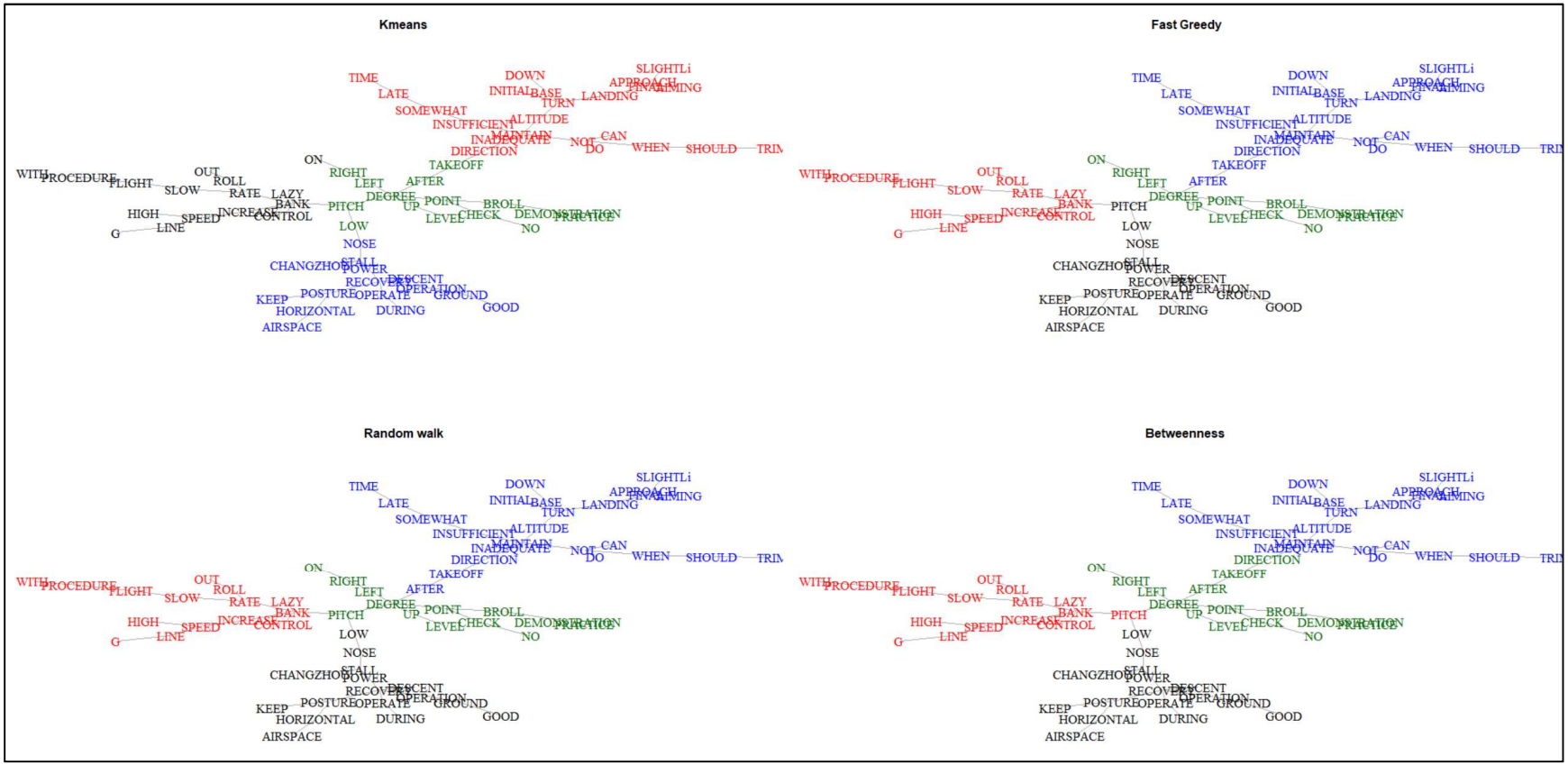


Figure 30. Clustered semantic network from different clustering method for number of cluster 4 on 50% proportion of terms

### *b. Composition of Topic Model*

Table 3 lists the topics, and the lists of words that make up each topic. The composition of these topics is very closely associated with the composition of actual flight training. Figure 31 shows the relationship between flight training and the topic model. The topic "Emergency action" deals with recovery in the event of an aircraft losing power or entering the stall at ground, airspace or takeoff and landing stages.

Table 3. Topics in flight training and word composition of each topic

| Topic | Words |
|---|---|
| **LANDING** | LANDING, TIME, LATE SOMEWHAT, INSUFFICIENT, ALTITUDE, MAINTAIN, INADEQUATE, DIRECTION, NOT, DO, CAN, WHEN, SHOULD TRIM, APPROACH, AIMING, FINAL |
| **TAKE-OFF** | TAKEOFF, AFTER, BROLL, LEVEL, UP, DEGREE, LEFT, RIGHT, PITCH, LOW, NO, CHECK, PRACTICE, PITCH |
| **MANEUVER** | ROLL, OUT, RATE, LAZY, BANK, SPEED, HIGH, G, CONTROL, FLIGHT, PROCEDURE |
| **EMERGENCY** | STALL, POWER, RECOVERY, DESCENT, OPERATION, GROUND, GOOD, CHANGZHOU, POSTURE, HORIZONTAL, AIRSPACE, KEEP, DESCENT |

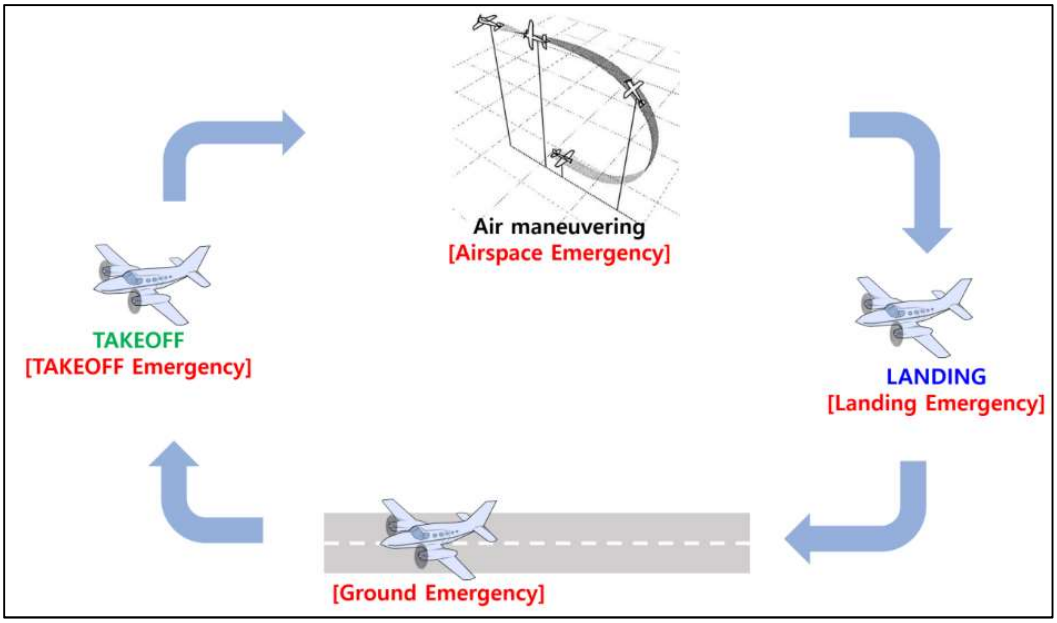


Figure 31. Relationship between flight training and topic model

### *c. Derived Characteristics of Each Topic*

From the topology of each topic within the semantic network, we can derive key characteristics associated with each topic within the context of flight training. Figure 32 shows the topology of each topic within the network.

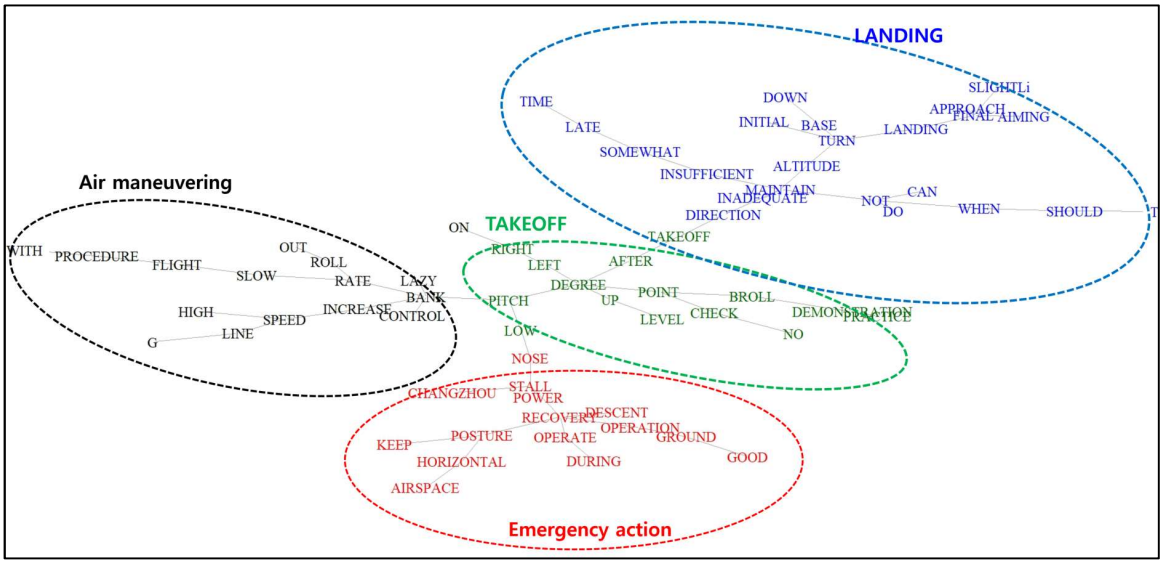


Figure 32. Topology of topics within flight criticism semantic network

**LANDING:** Most criticism in flight training focuses on landing. Most negative words such as “INSUFFICIENT,” “INADEQUATE,” “NOT” and “LATE” belong to the landing topic cluster. Highlights during the landing phase include altitude maintenance, time management and direction maintenance. Also, the use of “TRIM” is emphasized because the use of “TRIM” minimizes manipulation to maintain and direction.

**TAKE-OFF:** “TAKE-OFF” is located at the center of the semantic network, which means that there are many sub-topics that are shared with other topics. Direction maintenance, for example, is emphasized during both takeoff and landing phases, and aircraft attitude control, such as “PITCH” and “BANK” is emphasized in both the air maneuver and take-off phases. During the take-off phase, the inspection of the aircraft condition is emphasized. After the takeoff, it is necessary to change the operation mode of the automatic flight control system. Since there was historically a case of an accident

related to this, it can be assumed that this is why the check of the condition of the aircraft is emphasized.

**AIR MANEUVERING:** Emphasis of air maneuver is on aircraft attitude control, speed management and procedural compliance. In particular, speed management and adaptation to gravity are emphasized to prevent falling into a "BLACK OUT" state due to gravity force applied to the pilot during air maneuvers.

**EMERGENCY RESPONSE:** Training of emergency measures applies to all stages of flight. The word "CHANGZOU" means traffic pattern for landing in Korea. The topic includes the recovery procedure for stall, and emergency glide landing when the aircraft loses power. This topic includes ground engine start-up and restart procedures. The evaluations for these sub-topics are mostly positive. This is because the students have a relatively smaller number of things to consider than other topics.

### 2. Analysis of Major Passing and Failing Factors

We presented the quantitative characteristics extracted from the semantic network and the practical meaning of each characteristic in Chapter III. These characteristics are variables that predict pass/fail outcomes within the fitted model and the effect of each factor on pass/fail is quantified through the coefficient assigned to the corresponding variable. Since we standardized all predictors, the magnitude of a coefficient indicates the explanatory power of the predictor, and the sign indicates whether the probability of success has an increasing or decreasing relationship with the predictor. Formally, a model coefficient is an estimate of the logarithm of the odds ratio for success when the predictor is fixed at a particular value, and when the predictor is increased by one standard unit.

#### *a. Centrality of the Topics*

The high degree of centrality of a particular topic means that the words that make up the topic are centrally located within the ISFN, which means that the topic is focused on in flight training.

In the same context, if the coefficients assigned to the centrality of the topic within the fitted model have a positive value, this means that the student's probability of passing

increases when the instructor focused on the corresponding topic. On the other hand, if the coefficients assigned to the centrality of the topic have negative value this means that the probability of passing decreases when the instructor focused on the corresponding topic. Figure 33 shows the coefficient values assigned to the centrality of each topic as flight training progresses.

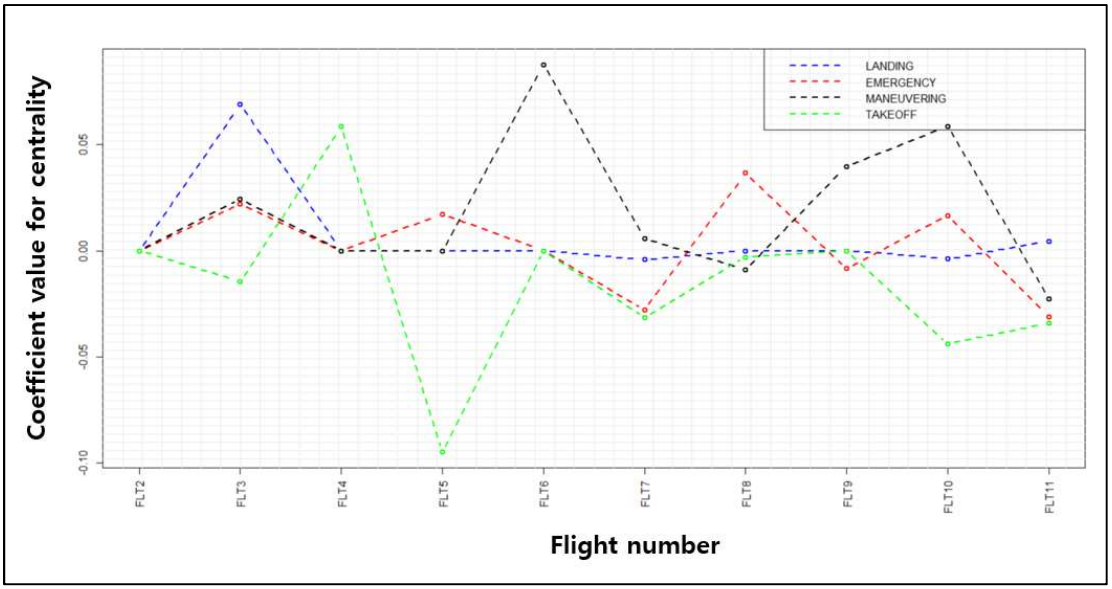


Figure 33. Coefficient for topic centrality

Figure 33 shows that the topics of focus for each flight are different. In the early stages of flight training, the emphasis on education in takeoff and landing has a positive effect, while in mid and late flight training, focusing on maneuvering increases the probability of passing. Also, focusing on the take-off portion in the later stages of flight training results in lowering the probability of passing. These results can be interpreted as follows. Since take-off and landing are essential items in flight training, focusing on take-off and landing has a positive impact at the beginning of flight training. Focus on take-off at the end of training, however, could indicate that the student has not mastered the basics. However, one of the most important items in flight training is aircraft control capability, which is improved through maneuver training. In addition, since the improvement of aircraft control capability leads to the improvement of take-off and landing, focusing on

maneuvering training has the most positive effect after mid-term. From a similar point of view, it is desirable to focus on training for maneuver rather than take-off from mid-term, even if the student's take-off skill is insufficient.

The most noticeable feature in Figure 34 is that the higher the "TAKEOFF" centrality, the lower the passing probability. Because all variables are standardized in fitting the model, they have a distribution with a mean of 0 and a standard deviation of 1. Let the coefficient value assigned to the "TAKEOFF" centrality be $\hat{\beta}$, the value of the intercept be $\hat{\beta}_0$ and let the values of all the other features be constant. The estimated log-odds for success comparing a fixed value of "TAKEOFF" to incrementing it by one standard unit is given by $\hat{\beta}$ and the probability of success is given by Equation 15. Figure 34 shows relationship between the "TAKEOFF" centrality value and the passing probability. For Flight #5, when the other features are the same, the passing probability with the lowest "TAKEOFF" centrality is very high as 87%, on the other hand it is very low as 13% with the highest "TAKEOFF" centrality.

$$\begin{aligned} &\hat{\eta} = \widehat{\beta_0} + \widehat{\beta} \times TakeOff\_Centrality, \\ &\widehat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} \end{aligned} \tag{15}$$

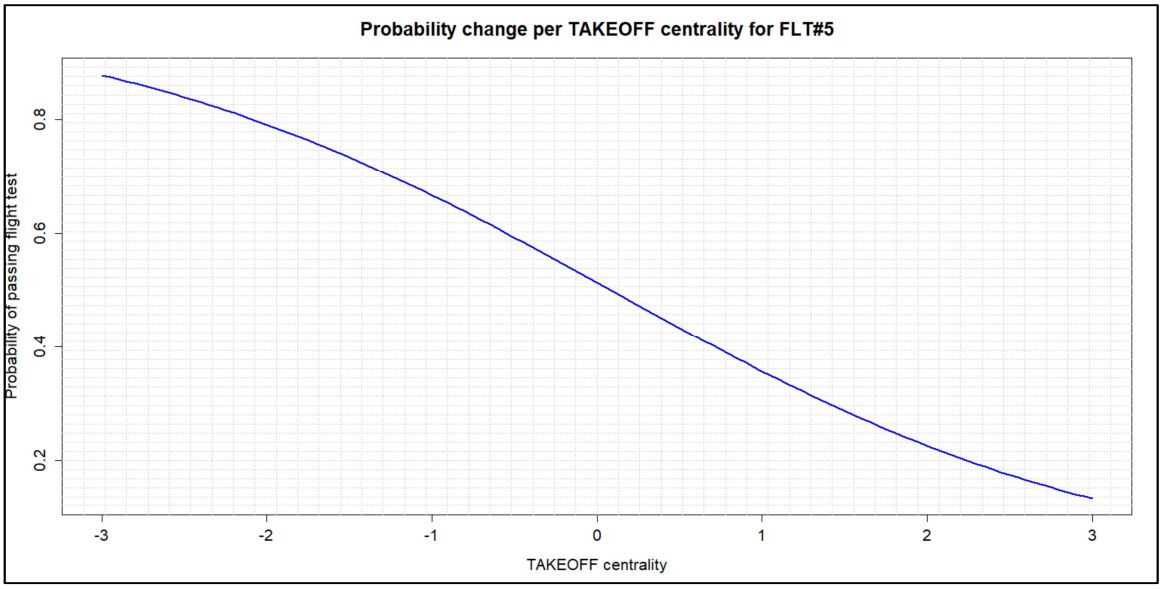


Figure 34. Passing probability per "TAKEOFF" centrality

### *b. Effect of Modularity and Network Density*

As mentioned in Chapter III, high modularity of the ISFN means that the flight training assessment reflected organized and focused instruction for the flight session. Conversely, high density reflects a more comprehensive manner of instruction, where many topics are closely intertwined within the assessment. Figure 35 shows the change of the coefficient values assigned to the density and modularity as flight training progresses.

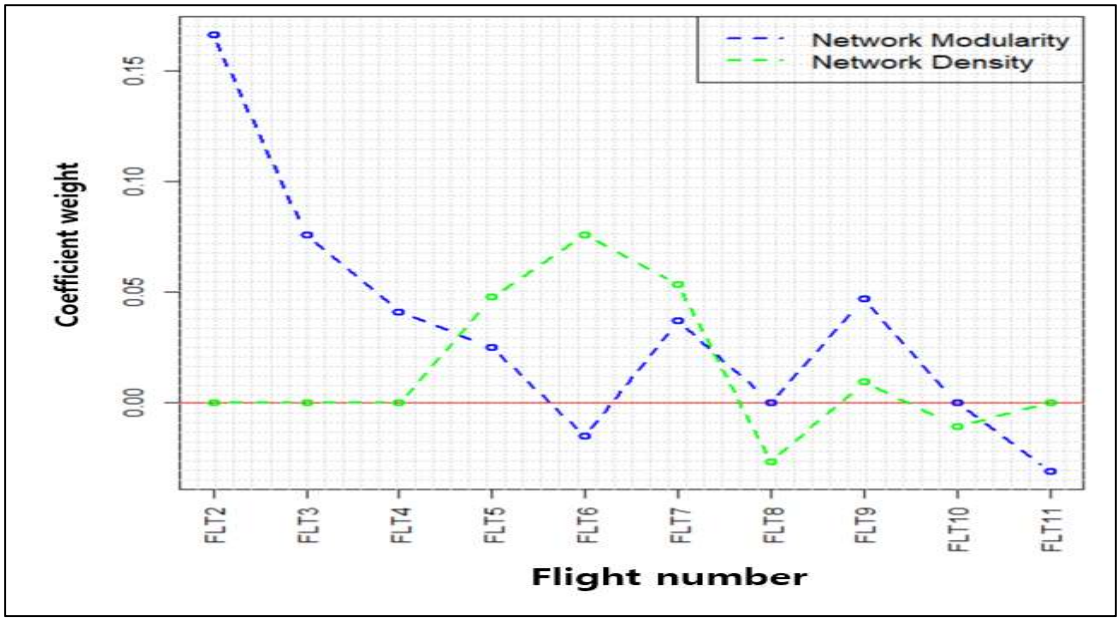


Figure 35. Coefficient value for density and modularity

As the definition of density and modularity are opposite, the effect on passing also works in reverse. In the early stage of flight training, the higher the modularity the higher the probability of passing. On the other hand, in the middle of training the higher the density, the higher the probability of passing. This makes sense, because at the initial phase of flight training, students are unfamiliar with the aerial situation and characteristics of the aircraft, so comprehensive awareness is almost impossible and focused instruction is needed as opposed to comprehensive instruction that tries to address too many topics at once. Therefore, separately instructing each topic is shown to improve the outcomes of the training in the early stage of the training. However, since the test flight is almost always a comprehensive course from take-off to landing, teaching students from a comprehensive perspective enhances the outcomes of the training if given after the students have adapted to the aerial situation to some extent.

Figure 35 shows that the high modularity value has the most positive impact on passing probability in Flight #2. When values of the other features are constant, the passing probability estimated according to modularity can be calculated by applying Equation 15. Figure 36 shows the impact of modularized instruction on passing probability more

quantitatively. If other characteristics are the same and the modularity is the lowest, the passing probability is 24%, but if the modularity is high, the passing probability is 75%.

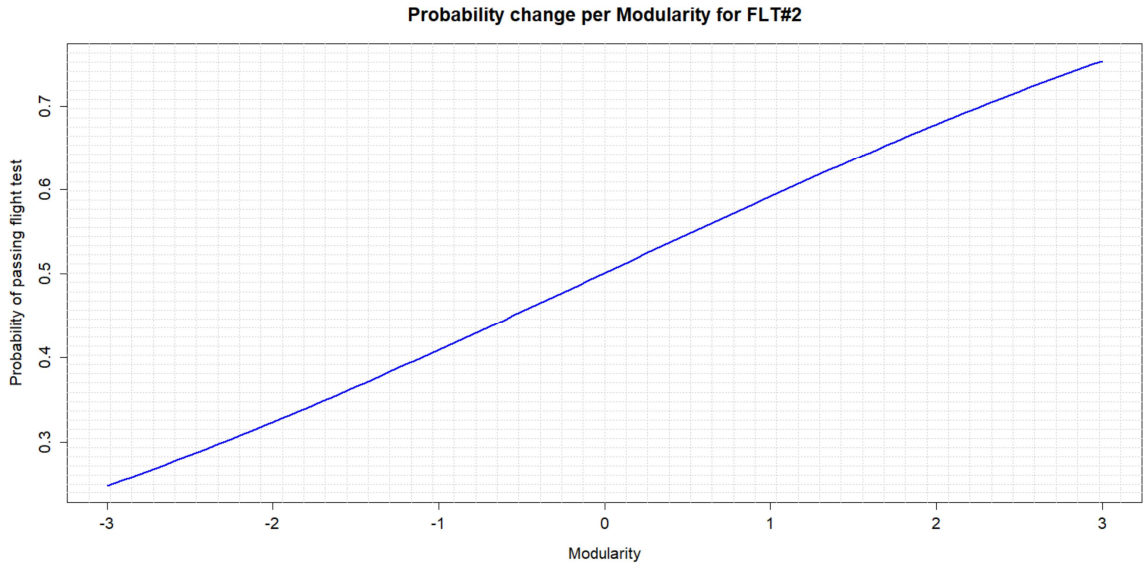


Figure 36. Passing probability per Network modularity

### *c. Effect of Positive/Negative Distance of the Topics*

The emotional distance for each topic is calculated as the mean of the shortest distances between the word representing positive or negative emotions and the words constituting the topic within the ISFN. The interpretation of the coefficient is a bit different from the preceding factors since the values of the variables are distance. For example, a positive coefficient value for the distance between the word "GOOD" and the topic "LANDING" means that a positive comment for "LANDING" has a negative impact on the passing probability. Figure 36 shows the change of the coefficient values assigned to the emotional distance for the positive word "GOOD" as flight training progresses. Figure 37 shows the change of the coefficient values assigned to the emotional distance for the negative word "INSUFFICIENT" as flight training progresses.

In Figure 37, positive evaluation at the early stage of flight training negatively affects the final passing probability, but positive evaluation at the latter stage of flight training is likely to have positive effect on the passing probability. What is noteworthy is

that the positive assessment of air maneuver increases the probability of passing throughout the flight training. Positive evaluation of air maneuver means that the student has good spatial perception ability and aircraft control ability. Because these abilities are also important factors in all subjects of the flight training, a positive assessment of air maneuver improves the passing probability at all stages of flight training. On the other hand, positive assessments of landing have a negative impact on the probability of passing. This may be because this gives the student too much confidence in landing, and over-confidence in the landing phase, where there are many variables such as wind direction and wind speed, can lead to dangerous situations.

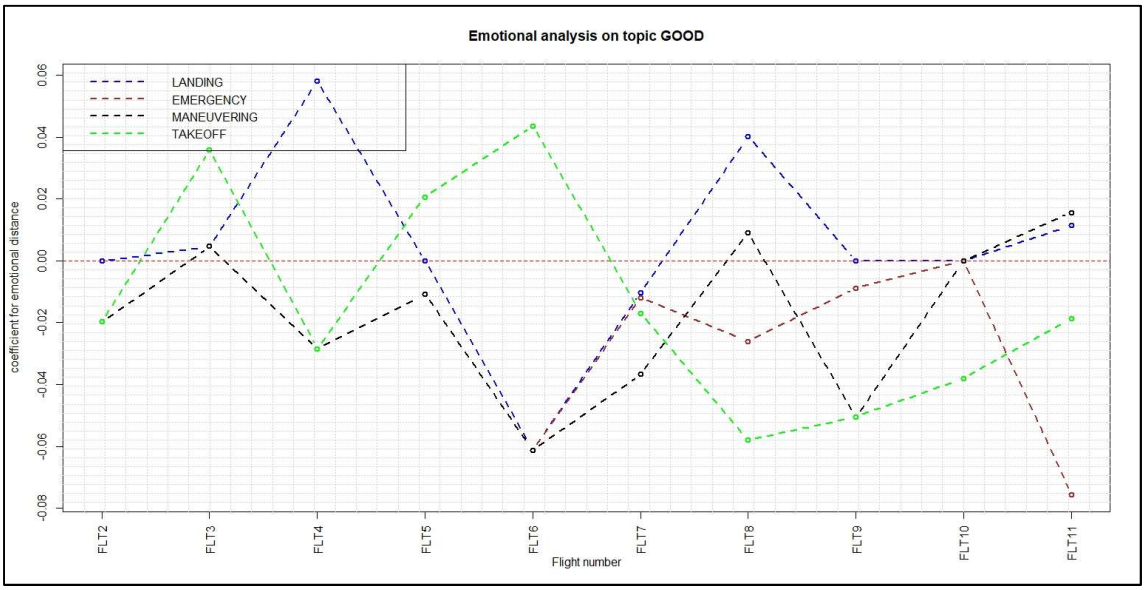


Figure 37. Coefficient value for the positive emotional distance

A notable feature in Figure 37 is that positive feedback on takeoff (distance between take-off topic and positive emotional word “GOOD” is close) in the final phase of flight training (Flight #8) increases the passing probability while positive feedback on landing in the final phase of flight training has a negative impact on passing probability. Figure 38 shows this feature more quantitatively. In Flight #8, the probability of passing increases from 34% to 66% when the distance between positive word “GOOD” and the landing topic

gets farther. On the other hand, if the distance between the takeoff topic and the positive word “GOOD” gets farther, the passing probability decreases from 72% to 28%.

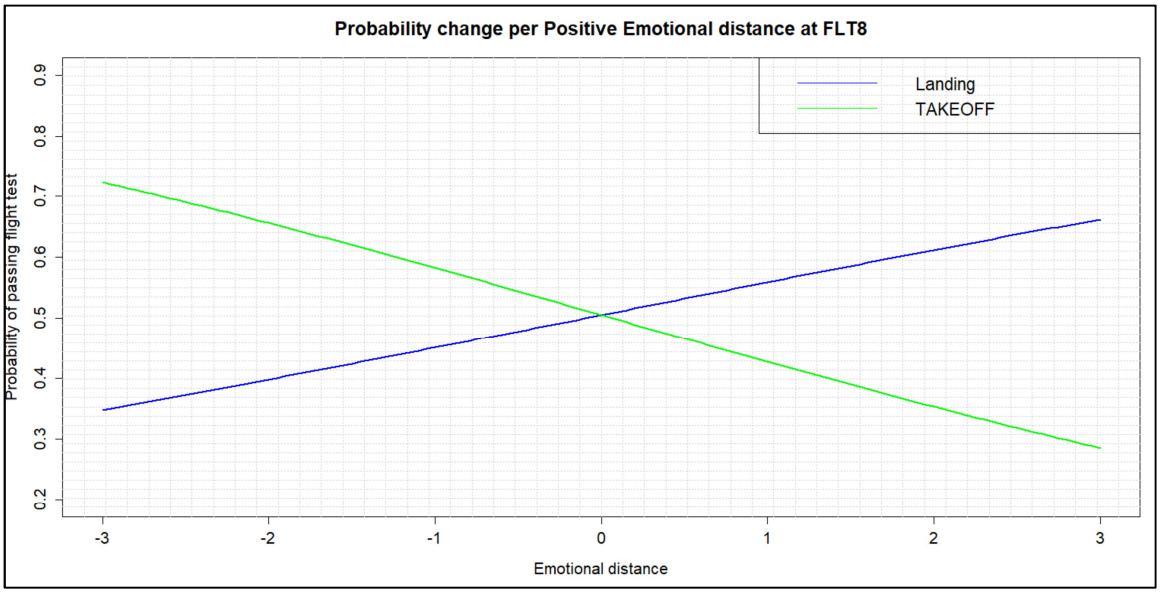


Figure 38. Impact of positive feedback on landing and take-off for Flight #8

Figure 39 shows the effect of the distance between the subject and the negative word “INSUFFICIENT” for each training flight. Overall, negative feedback at the initial phase of flight training has a rather positive impact on passing probability, but a negative feedback at the final phase of flight training lowers the passing probability. This tendency is prominent in the feedback of the take-off subject. Figure 40 shows these results more quantitatively. In Flight #2, the probability of passing decreases from 72% to 27% when the distance between negative word “INSUFFICIENT” and the takeoff subject increase. On the other hand, for Flight #9 the probability of passing slightly increases from 46% to 54% when the distance between the negative word “INSUFFICIENT” and the takeoff subject increase. These results suggest that at which point, the instructor should give a negative or positive feedback in order to maximize the effect of training. It is good to give details of what to fix in early phase of flight training, but since the student is accustomed to flight in the latter part of the training, providing unconditional negative feedback has a negative impact on the passing probability.

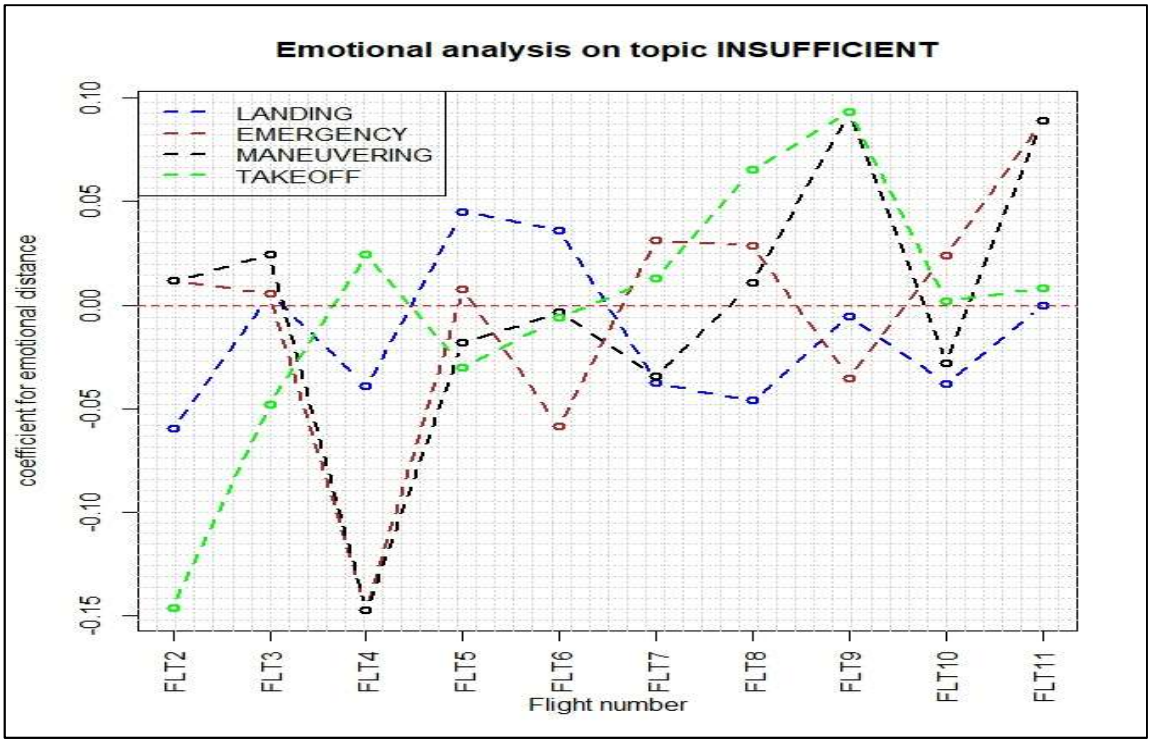


Figure 39. Coefficient value for the negative emotional distance

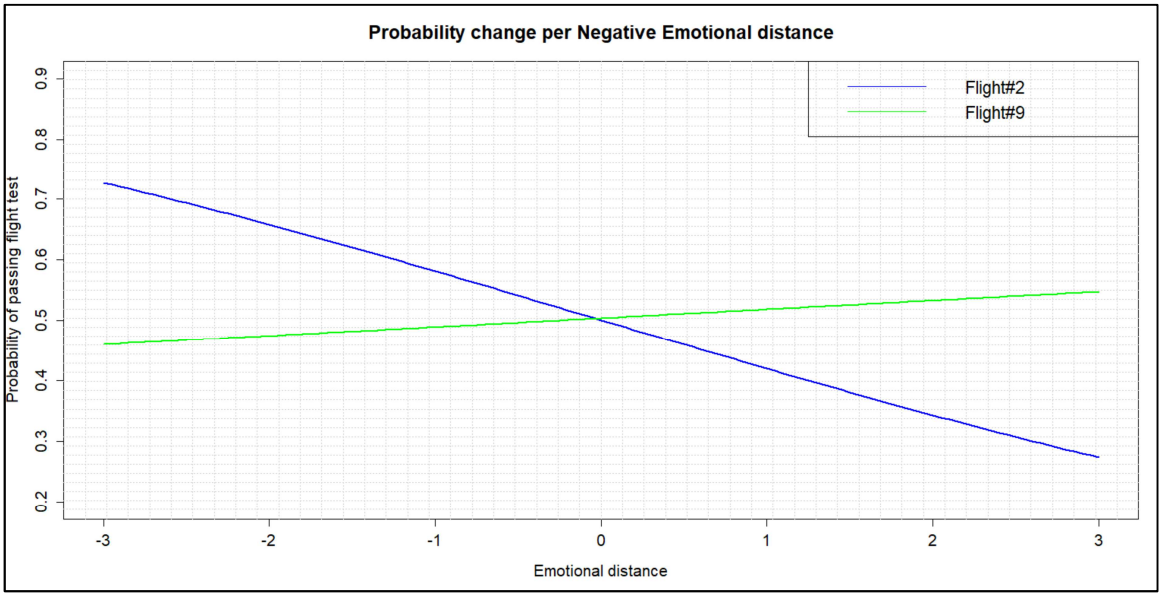


Figure 40. Impact of negative feedback on take-off for flight#2 and flight#9.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. CONCLUSIONS

## A. SIGNIFICANCE OF THE STUDY

In this study, we applied text mining techniques and the MST - based semantic network technique to the flight criticism data and were able to extract the quantified success factors. Based on the identified factors, the acceptable passing/failing prediction model predicted the students' passing and failing with very high accuracy. In addition, the topic model estimated from the MST-based semantic network was very similar to the actual flight training, and the factors and the recommendations derived through qualitative analysis were reasonable enough. These facts demonstrate that the MST-based semantic network method presented in this study is valid for educational text analysis. In the course of the study, we obtained the following significant knowledge.

In the data preprocessing process, we applied Google's automatic translation algorithm (SuHun 2017) to convert documents composed of two languages into one language. As a result, the complexity of the translated document was reduced, and the performance of the final model was also improved. This shows the applicability of the automatic translation algorithm to the preprocessing process of text mining.

We applied the simplest method to extract the semantic network from the document. The distance between words was calculated based on the frequency of co-occurrence of words in the sentences, and the semantic network was constructed by applying the MST algorithm. Despite the simple approach, the contents of the documents and the structural characteristics of the documents were very well represented through the semantic network. From this point of view, the MST-based semantic network methodology is expected to be used as a text mining technique that summarizes the contents of a document compactly and extracts its characteristics.

We estimated the subject model that composed documents by applying four different clustering techniques to MST - based semantic network. We had to choose the scope of terms to include and the number of clusters for the topic model selection. We utilized some statistical clustering performance indicators such as, reproducibility,

modularity and WCSS, However, in order to select a topic model with a practical meaning, we had to perform a qualitative analysis by looking at the group of words and finding the actual meaning. Finally, the selected topic model was consistent with actual flight training and could be given meanings “takeoff,” “landing,” “aerial maneuver,” “emergency action.” Also, from the finally selected topic model, we found the fact that clustering result from different clustering methods yields similar result if the clustering result has a practical meaning. Through this, it was identified that the consistency between different clustering methodologies, that is, the reproducibility of clustering methodology, can be used as an evaluation criterion of clustering performance. The composition and characteristics of the flight training derived from the clustered semantic network are:

- Flight training consists of “takeoff,” “landing,” “aerial maneuver,” and “emergency action.”
- For the take-off phase, it is important to control the aircraft nose and check the condition of the aircraft.
- For the landing phase, time management and maintaining attitude during the runway approach is important
- In airborne maneuver, procedure execution and bank control are important factors.

We extracted quantitative factors based on network characteristics from the clustered semantic network and gave meaning to each factor related to flight education. Based on the extracted factors, we identified which factors are also associated with successful education and compare the impact of each factor and its proportion in the model. Through this, the following knowledge could be derived.

- Each flight has a different weight for the overall training results. Particularly, second flight is important because it is critical to evaluate students’ spatial perception ability and spatial perception ability is a very important factor for training results.

- Instructor guidance style should be different for each stage of training. In the initial phase of flight training, students should understand the characteristics of each subject, so teaching each subject separately will increase the student's probability of passing. In the middle stage of flight training, each subject should be linked and education should be performed from a comprehensive viewpoint. The most emphasized part from the middle stage is the part of the aircraft's posture control, which is a key capability in all subjects.

- The positive and negative evaluation of the instructor influences the student's final acceptance probability. However, positive evaluations do not always increase the probability of acceptance. Especially the positive evaluation of landing during the later training flight tends to lower the student's probability of acceptance; This may be because positive criticism gives the student excessive confidence in landing.

In order to derive meaningful knowledge from the fitted model, additional qualitative analysis is required. Even though we could extract quantified factors for successful flight training by applying the methodology presented in this study, prior understanding of the flight training was essential to give meaning to the quantified factors and to interpret the results. This suggests that a subject matter expert with a solid understanding of the flight curriculum is required to derive meaning knowledge with methodology presented in this study.

## B. FUTURE WORK

Similar data for flight training are available in the form of voice records saved for every flight. Voice records of all flights are recorded for the purpose of an air accident investigation. In future work, it may be possible to extract textual data from these voice recordings with the use of well-known neural network Non Linear Programming (NLP) methods. This would provide more words and a larger vocabulary with which to build our semantic network, potentially revealing additional relationship between topics, as well as additional performance indicators.

As mentioned in Chapter I, most military training aimed at acquiring skills yield the textual evaluation results. In order to solidify the validity of the methodology presented in this study, it is necessary to verify whether the meaning knowledge can be derived even when applying the same methodology to evaluation results of other training courses. Furthermore, we expect that the application to other text data which dealing with specific subjects such as meeting minutes, work reports, as well as evaluation text will also yield meaningful results.

In this study we used “LASSO” as a binary classification technique for ease of predictor selection and impact comparison. However, there are many useful and validated binary classification techniques such as Random forest and Support vector machine. Therefore, it would be useful to study other classification techniques summarizing educational text by comparing the results of binary classification techniques other than LASSO.

In selecting the topic model of flight training, we found that meaningful clusters could be identified by comparing the clustering results from different clustering methods. This suggests that the similarity between clustering results from different clustering methods can be a criterion for clustering evaluation. In order to verify this, it is necessary to compare the accuracy of selecting a cluster based on similarity and the accuracy of selecting a cluster by applying the existing methodology to multivariate data including categorical variables.

# LIST OF REFERENCES


Ahuja RK, Magnanti TL, Orlin JB (1993) *Network Flows: Theory, Algorithms, and Applications* (Prentice Hall, Upper Saddle River, NJ).

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *7th Conference on International Language Resources and Evaluation* (Malta). http://www.lrec-conf.org/proceedings/lrec2010/pdf/2769_Paper.pdf.

Bonnanno G, Caldarelli G, Lillo F, Mantegna RN (2003) Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E* 68(4), http://link.aps.org/doi/10.1103/PhysRevE.68.046130.

Bouchet-Valat M (2014) SnowballC: Snowball stemmers based on the C lib stemmer UTF-8 library. Accessed Aug 8, 2014, https://CRAN.R-project.org/package=SnowballC.

Buttrey S, Whitaker L (2017) Advanced Data Analysis. Class notes, Operations Research, Naval Postgraduate School, August 1, Monterey, CA.

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks, *Physical Review E* 70(6), https://doi.org/10.1103/PhysRevE.70.066111.

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Accessed June 29, 2015, https://igraph.org .

Newman M, Girvan M (2004) Finding and evaluating community structure in networks, *Physical Review E* 69(2), https://doi.org/10.1103/PhysRevE.69.026113.

Dumais ST (2002) Latent semantic analysis. Cronin, ed. *Annual Review of Information Science and Technology* (Information Today, Medford, NJ), 189–230.

Ertekin S (2013) Adaptive oversampling for imbalanced data classification, *Information Science and Systems* (Springer, NY), 261–269

Ester M, Kriegel, HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceeding of 2nd International Conference on Knowledge Discovery and Data Mining* (Portland, OR), 226–231.

Feinerer I, Hornik K (2017) Text Mining Package. R package version 0.7-3. Accessed Dec 6, 2017, https://cran.r-project.org/web/packages/tm/index.html.

Friedman J, Hastie T, Tibshirani R (2010) regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), http://www.jstatsoft.org/v33/i01/.

Gordon AD (1999) Clustering validation and description, *Classification* (Chapman and Hall, Boca Raton FL), 183–211.

Hand DJ, Mannila H, Smyth P (2001) *Principles of Data Mining: Adaptive Computation and Machine Learning* (MIT Press, Cambridge, MA).

Hartigan JA, Wong MA (1979) A K-means clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1), http://www.jstor.org/stable/2346830.

Hassan S, Fernandez M, He Y, and Alani H (2014) On stop words, filtering and data sparsity for sentiment analysis of Twitter. *9th International Conference on Language Resources and Evaluation* (Reykjavik, Iceland). http://oro.open.ac.uk/40666/1/292_Paper.pdf.

Hemalatha, Varma GPS, Govardhan A (2012) preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science* 1(2), www.ijettcs.org/Volume1Issue2/IJETTCS-2012-08-14-047.pdf.

Hoerl E, Kennard W (1970) Ridge regression: Biased estimation for nonorthogonal problems. *American Society for Quality* 12(1):55-67.

Huh MH, Lee YG (2004) Reproducibility assessment of k-means clustering and application, *Journal of Korean Statistical Society* 17(1):135-144.

Huh MH, Lim YB (2009) Weighting variables in kmeans clustering. *Journal of Applied Statistics* 36(1), https://doi.org/10.1080/02664760802382533.

Indurkhya N, Damerau F (2010) *Handbook of Natural Language Processing,* 2nd ed. (Chapman and Hall/CRC, Boca Raton, FL).

Lander JP (2017) Useful: A collection of handy, useful function. Accessed June 7, 2017, https://CRAN.R-project.org/package=useful.

Jeffrey L (2008) Text data mining: Theory and methods, *Statistics Surveys 2*, https://doi.org/10.1214/07-SS016.

John S (2015) Semantic networks, Mar. 1, 2015, http://www.jfsowa.com/pubs/semnet.htm

Joseph JP, Antonio Z (1984) Automatic spelling correction in scientific and scholarly text. *Communication of the ACM* 27(4), https://doi.org/10.1145/358027.358048

Kamada T, Kawai S (1988) An algorithm for drawing general undirected graphs, *Information processing letters* 31(1):7-15.

Kaufman L, Rousseeuw PJ (1987) Clustering by means of Medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods* (North Holland), 405–416.

Kukich K (1992) Techniques for automatically correcting words in text, *Association for Computing Machinery Computing Surveys* 24(4), https://doi.org/10.1145/146370.146380.

Kruskal JB (1956) On the shortest spanning subtree of a graph and the traveling salesman problem, *Proceedings of the American Mathematical Society* 7:48-50.

Luhn HP (1960) Key word-in-context index for technical literature (kwic index). *Journal of the association for information science and technology* 11(4), https://doi.org/10.1002/asi.5090110403.

McQueen J (1967) Some methods for classification and analysis of multivariate observation. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (University of California), 281–297.

Maybury MT (1997) *Intelligent Multimedia Information Retrieval* (AAI Press/MIT Press, Menlo Park, CA/Cambridge, MA).

Merkl D (2002) Text mining with self-organizing maps. Klosgen W, Ytkow JM, eds. *Handbook of Data Mining and Knowledge Discovery* (Oxford University Press, Oxford, United Kingdom), 903–910.

Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3(4), https://doi.org/10.1093/ijl/3.4.235.

Miller TW (2013) *Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R* (FT Press Analytic, Upper Saddle River, NJ).

Minimum spanning tree (2018) *Wikipedia*. Accessed Feb 1, 2018, https://en.Wikipedia.org/wiki/Minimum_spanning_tree.

Newman M, Girvan M (2004) Finding and evaluating community structure in networks, *Physical Review E* 69(2), 10.1103/PhysRevE.69.026113.

Onnela JP, Chakraborti A, Kaski K (2003) Dynamics of market correlations: taxonomy and portfolio analysis. *Physical Review E* 68(5), https://journals.aps.org/pre/abstract/10.1103/PhysRevE.68.056110.

Pons P, Latapy M (2006) Computing Communities in Large Networks Using Random Walks. *Journal of American Statistical Association* 10(2):191-218.

Porter MF (1980) An algorithm for suffix stripping, *Program* 14(3), https:// doi.org/10.1108/eb046814.

Porter MF (2001) Snowball: A language for stemming algorithms. Accessed October, 2001, http://snowball.tartarus.org/texts/introduction.html.

Prim R (1957) Shortest connection networks and some generalization, *Bell System Technical Journal* 36(6): 1389–1401.

Rand WM (1971) Objective criteria for the evaluation of clustering methods, *Journal of American Statistical Association* 66:846-850.

SuHun Han (2017) Googletrans 2.2.0 documentation. Accessed Oct 2, 2017, http://py-googletrans.readthedocs.io/en/latest/#.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal. Statis. Soc B*., 58(1):267-288.

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Royal. Statis. Soc B*., 63(2):411-423.

Tingting W, Yonghe L, Huiyou C, Qiang Z, Xianyu B (2014) A semantic approach for text clustering using Wordnet and lexical chains. *Expert Systems with Applications* 42(4), https://doi.org/10.1016/j.eswa.2014.10.023.

Tsang V. and Stevenson S (2004) Calculating semantic distance between word sense probability distributions. *Proceedings of the 8th Conference on Computational Natural Language Learning* (Boston, MA), 81–88.

Tsz-Wai R, He B, Ounis I (2005) automatically building a stop word list for an information retrieval system. *Journal of Digital Information Management* 3(1), http://www.dirf.org/jdim/abstractv3i1.htm#01.

Vijayarani (2015) Preprocessing techniques for text mining—an overview. *International Journal of Computer Science & Communication Networks* 5(1), http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf.

Zhang Y, Chen M, Liu L (2015) A review on text mining. *2015 6th IEEE International Conference on Software Engineering and Service Science* (Beijing). https://doi.org/10.1109/ICSESS.2015.7339149.

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California