**Analysis of fire accident factors on construction sites using web crawling and deep learning approach**

Jaehong Kim1, Sangpil Youm2, Yongwei Shan**\***3

1 Ph.D. Candidate, School of Civil & Environmental Engineering, Oklahoma State University, Stillwater, OK, [jaehong.kim@okstate.edu](mailto:jaehong.kim@okstate.edu)

2 Master Student, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, youms@iu.edu

**\***3 Corresponding authors, Associate Professor, School of Civil & Environmental Engineering, Oklahoma State University, Stillwater, OK, [yongwei.shan@okstate.edu](mailto:yongwei.shan@okstate.edu)

**Abstract**

The construction site is one of the industrial sites that can be exposed to fatal accidents. Fire accident prevention on construction sites is one of the most important components in site safety planning. However, because fire accidents have fewer frequencies than other types of accidents such as falls, plans related to fire safety on construction sites have been rarely studied. Without considering the characteristics of fire accidents, it is unreasonable to evaluate the risk-based only on the frequency of accidents. To fill the knowledge gap, this study was conducted to find factors related to accidents on the construction site using big data analysis. In this study, web-crawling was used to collect data related to accidents on construction sites in the past 20 years. Based on the collected data, the authors found the media exposure of keywords related to accidents and provided similarities between keywords through the deep learning approach. In particular, factors related to fire accidents were extracted and provided. The results were visualized by reducing the high-dimensional data into two-dimensional data that are easy to visualize through UMAP. In the case of fire accidents, the media exposure was higher than the actual frequency when compared to the fall accidents. This shows the possibility that fire accidents may have a greater impact on the surroundings than the actual frequency. This study contributes to the reduction of fire risks on construction sites by identifying the key factors related to fire accidents. It is also possible to intuitively check the factors related to each accident type and help to institutionalize ways to manage accident-related factors.

**KEYWORDS:** Construction sites, Safety, Fire accidents, Web crawling, Deep learning

**Introduction**

Recently, due to the development of various construction technologies, skyscrapers and large-scale construction projects are under construction around the globe. The development of construction technologies has shortened the construction duration, but the safety of the construction site is developing slowly. The Bureau of Labor Statistics (BLS) recently released the Census of Fatal Occupational Injuries (CFOI) in 2018, and the construction industry topped the list with 1,008 worker deaths. In addition, 5,250 fatal work injuries were recorded, a 2% increase from 2017. The fatal work injury rate was about 3.5 per 100,000 full-time equivalents (FTE) workers. As such, various types of accidents occur on construction sites, including fires. Fire accidents are greatly affected by the external environment, such as weather and surrounding buildings or hazard, making it difficult to control and prevent them. This is especially dangerous on construction sites because fire safety equipment such as sprinklers and fire alarms may not be completed depending on the progress of the construction (Hamid et al. 2003). Most construction accident-related research focuses on accidents with high frequency in the past. However, there are also limitations when analyzing only the frequent accidents. For example, fall accidents on construction sites are one of the fatal four accidents and but are less likely to lead to additional accidents to the surroundings. The frequency of fire accidents is much lower than that of fall accidents, but it can have a greater impact on the construction site. In other words, in order to improve the safety of the construction site, it is necessary to consider how the accident type affects the construction activity and the surrounding environment. The impact of each type of accident on the construction project and the surrounding environment involves various external conditions, and it is difficult to quantify these effects. For this reason, most construction accident studies focus on the frequency of accidents. To compensate for this limitation, this study investigated the media exposure level of related accidents under news article coverage. The news media often covers issues that have a great social impact rather than minor injuries on construction sites. Fatal accidents that can affect the surroundings are more likely to be exposed to the media. The lesson from the articles provided by the media can be learned. Moreover, the articles provided by the media are organized in a similar format, which is efficient for many researchers to use the data. To enhance construction safety, including fires on construction sites, this study collected articles on construction site accidents reported in the New York Times over the past 20 years. The web-crawling method was used for efficient and accurate data collection. Using collected articles, we analyzed the frequency of keywords related to construction site accidents. In addition, similarities and relationships between related keywords were analyzed through word embedding and network analysis. To intuitively visualize words that have a high relationship between fire accidents and fall accidents, the Uniform Manifold Approximation and Projection (UMAP) method was applied. In this study, the level of media exposure of construction site accidents was conducted to analyze accidents on construction sites and to present a new perspective to improve safety.

**Background**

***Construction site safety***

Although the construction industry always considers safety, the fatality rate on construction sites is always high. To reduce the fatality rate of construction sites various studies have been conducted to analyze accidents on construction sites, but there are limitations. The fatality rate of the construction industry was found to be the fourth highest after agriculture, mining, and transportation (Abdullah and Wern 2011). According to the Occupational Safety & Health Administration (OSHA), 20.5% of fatal workplace accidents occurred on construction sites in 2014 (Hosseinian and Torghabeh 2012; Zou and Zhang 2009). The OSHA provided that the main reasons for the construction worker's fatal accidents are by falls, struck by an object, electrocutions, and caught-in/between. Many studies related to accidents on construction sites are based on the frequency of accidents and are limited to analyzing the causes of accidents individually. One study suggested that the seven causes of accidents are lack of training, deficient enforcement of safety, no provided safety equipment, unsafe methods, unsafe site condition, no use of safety equipment, safety ignorance, and isolated sudden deviation (O'Toole 2002). Other studies have shown that construction accidents have occurred due to improper safety management due to a lack of safety awareness by managers (Zhou et al. 2008). In another study, the cause of the accident was identified by analyzing 40 deaths on the construction site from 2003 to 2008(Ling et al. 2009). The findings indicated that unskilled workers and lack of safety training are the main causes. Various causes related to accidents on construction sites such as lack of safety training and safety equipment were analyzed. It is also important to prevent fire accidents on construction sites, as they are likely to cause secondary accidents such as collapse, burial, and explosion. According to a related study, fires on construction sites are mainly caused by the transfer of sparks to the surrounding insulation during welding (Lee 2012). In order to evaluate the fire hazards on construction sites, there are also studies evaluating fire hazard tracking systems and related training (Hui et al. 2012). In addition, several studies have been conducted on effective evacuation in case of fire on the construction site (De-Ching et al. 2011; Ingason et al. 2010; Jeong et al. 2014). However, because these studies are based on case studies and specific projects, they are limited in application to general construction sites In order to prevent these accidents, it is an important approach to collect various data and analyze factors related to accidents on construction sites.(Aires et al. 2010).

***Web crawling application in construction research***

Web crawling is a technique for systematically browsing the web for the purpose of web indexing (Paul et al. 2017). It is often used for tracking web documents on the Internet to effectively collect the information the user needs. There are also studies to properly use online data for research purposes (Massimino 2016). This study provided guidelines on the skills and responsibilities required to collect online data (Because -> since) the data on the web is very huge, collecting web data manually can take a lot of time, and the accuracy can be reduced. However, web crawling technology automatically rotates the web server to repeatedly collect information that fits the purpose. These web crawling technologies are used in a variety of fields, especially in research involving decision models and prioritization (D’Haen et al. 2016; Guy et al. 2019). Recently, research on safety and security through web crawling has been conducted (Morgan et al. 2020). Web crawling technology has begun to be used not only for text but also for image analysis (Ali et al. 2018). In addition, research was conducted to utilize real-time data on the web rather than past data (Kim et al. 2019). To use web crawling in the research, researchers must set a clear target. Target is primarily a website, and researchers should make sure that web crawling technology is available on selected websites. This is very important because there are issues with data security. Next, the researcher determines the scope and frequency of data collection. Traditionally, the main purpose of using web crawling in the construction field consists of two main parts. The first is construction material management and optimization. To improve the efficiency of construction material management, related researches (used -> utilized) web crawling technology to collect relevant information and provide automated processes. (Hong et al. 2019; Yang et al. 2018). It also used web crawling to manage massive documents in construction projects. An example is a web crawling used to develop a system that collects text data with the latest information from the construction market and automatically assigns it to each applicable construction document (Moon et al. 2018). As above, the use of web crawling in the construction field was limited. Recently, this technology is used in various fields related to construction. An example is a study that collects a variety of geographic information on the web and provides a model to predict air emissions from heating (Lopez-Aparicio et al. 2018). However, few studies have analyzed the factors related to the safety of a construction site using web crawling. In this study, web crawling technology was used to find factors related to site safety, which may suggest a new approach to improving construction site safety. The authors used (python language ->a Python) and libraries to implement web crawler.

***Word embedding and network analysis***

Word embedding is a technique that provides a way to express similar words with the same meaning through data analysis (Yin and Shen 2018). This is a new way to represent words and documents and is one of the key breakthroughs in deep learning (Lai et al. 2016). Each word has a unique vector value through analysis using an embedding layer or hidden layer (Yoshioka and Dozono 2019). This method of analysis was mainly used as a new method for analyzing text or documents (Shao et al. 2017; Zhang et al. 2018). (This methodology overcomes the limitations of a one-hot encoding analysis, which was widely used for text analysis (Rodríguez et al. 2018). In the one-hot encoding method, since all words are composed independently, it is impossible to measure the similarity of each other. However, in the word embedding method, each word is expressed as a dependent relationship, and similarity can be measured (Rekabsaz et al. 2017). The word embedding method is currently used mainly in text analysis and natural language processing (NLP).) -> delete redundant with method part The main models using word embedding are Word2vec (Church 2017), GloVe (Hindocha et al. 2019), and fastText (Choi and Lee 2020), and Word2vec was used in this study. Word2vec Model was created and published in 2013 by a research team of Google. This model has been subsequently analyzed and updated by other researchers. Word2vec has very efficient performance and accuracy (Ombabi et al. 2017). In addition, it has a lot of google news pre-trained data, so it is suitable for transferring learning based on it (Khatua et al. 2019). In this study, network analysis was used to analyze not only the vector value of each word but also the relationship between linked words. Network analysis is one of the methodologies for finding the relationship between various types of raw data and basically consists of nodes and edges (Smith and Gorgoni 2018). Links between words can be expressed as graphs or structures, and relationships between words can be analyzed using them. Through the analysis of various centralities such as eigenvector centrality, degree centrality, and betweenness centrality, it is possible to check which nodes are important (Risselada et al. 2016). Also, the correlation between nodes can be analyzed by calculating the Jaccard coefficient that can measure the similarity of sample data (Bag et al. 2019). This network analysis is widely used in recent safety issues, such as pandemic research. There are studies to effectively prevent pandemic by analyzing the paths and conditions in which the pandemic spreads using network analysis (Sandhu et al. 2016).

**Methodology**

***Web crawling***

This study implemented a web crawling method for data collection. The web crawler is the main operator to conduct web crawling. The web crawler usually traverses web pages by using a recursive algorithm and then it goes over a certain range defined by researchers. The crawler stores data in a data structure that researchers can use efficiently for their studies (Mahto and Singh 2016). (In this study, the authors used Python 3.7. Since it has many libraries that can be easily applied to various areas, Python is considered as the most popular programming languages nowadays. Python also has an effective library for web crawling and has been widely used.)-> delete To begin with, the authors (need to )-> delete set the scope in which crawlers should travel. This research collected data from The New York Times, which is the top (3-> three) media company in terms of newspapers by circulation and thus it suffices representative of data needed in this study. In addition, some media companies restrict crawling or limit the amount of data for crawling while The New York Times has generous terms for crawling the data. (The New York Times provides a basic search environment that helps to narrow the range of search. This research utilized the search term “construction accident” and retrieved the most relevant 1,010 articles within 20 years. Among 1,010 articles, the authors only handle data formatted in text. Since different formats such as blogs and interactive documents consist of irregular structures, these formatted articles cannot be analyzed. )-> delete., redundant with the part of result

There are (two -> three) vital libraries for web crawling: Selenium, HTMLParser, and Beautiful Soup. Selenium is the set of tools that assist in the development of test automation for a web-based application (Mustika 2018). Selenium automatically traverses the web pages and stores data within the limit range. In this study, the authors stored the URL of each article from the web. Next, HTMLParser was used for parsing HTML files in each article. With the broken HTML, Beautiful Soup collected data the researchers desire to utilize. Beautiful Soup is a Python package that analyzes HTML/XML, as it extracts and edits information in web pages. It provides a friendly environment for building text analysis prototypes and mining the data (Zheng et al. 2015). In this research, the authors parsed the data from broken HTML and sorted out the part of the bodies, titles, and dates of the articles. The structure of the articles provided by The New York Times has a tag for body and title. It also provides a meta part that represents the metadata of the article and the meta part gives date information. Collected data were used for word embedding in the next part of this study.

***Word Embedding (Word2Vec)***

Word2Vec is one of the most powerful methods to implement word embedding. Since word2vec assigns a vector to each word, word2vec shows advantages over other word analysis techniques. Before the presence of word embedding algorithms such as word2vec, researchers usually used one-hot encoding to analyze words. One hot encoding converts each word to specific (associated with) values(numbers) that differ from every word. Even though this helped them to conduct NLP, the researchers could not answer the relation of each word because every word in one-hot encoding is independent. Word2vec resolve this problem, as it adds the embedding layer into the model. Word2vec follows a deep learning approach that probabilistically predicts (outputs -> words vectors) by using a hidden layer. It represented as follows,

, where is weighted matrix on hidden layer, d is the embedding dimension and V is vocabulary. Since Word2Vec conducts unsupervised learning that is trained on raw text data, it creates word embedding by figuring out the maximum likelihood of word prediction from their context (Ling et al. 2015). There are two models for implementing Word2Vec: Continuous Bag-of-Words model (CBOW) and skip-gram model. Word2Vec algorithm computes cosine similarity to find out similarity or dissimilarity between two vectors,

In this study, cosine similarity plays a significant role in analyzing words, which helps to compare the important words and construct relation among words. Continuous bag-of-words (CBOW) predicts word by considering context, which means target word is predicted from surrounded words of the target word and thus it has one output vector. The input vector is one-hot encoded. The weights between the input vector and the output vector are represented by V N matrix . The authors formulated k-th row of W as h.

There is different weight matrix ; between embedding layer to the output layer. With these (weights-> weight matrix), this model computes a score for each word,

h

Through these (formulae -> formula), the authors can calculate the posterior distribution of words in the vocabulary as follows (Rong 2014):

The skip-gram model works oppositely to CBOW. In the skip-gram model, this model predicts the context of words from a word. That is, skip-gram computes the possibility of the target words in several contexts from a word. The Skip-gram model also uses a weighted matrix, denoting the input vector of the word on the input layer, and it has the same definition of embedding layer outputs. Unlike the result in CBOW, the skip-gram model has multinomial distributions for the outputs. Even though two models both are widely used for conducting word2vec, these models can be optimized depends on the goal and direction of the research. The skip-gram model can be effective when the research tries to figure out fewer frequency words. On the other hand, CBOW is more suitable for studies that concentrate on high-frequency terms. (In this study),-> delete Since the authors focused on important keywords that can explain accidents on construction sites and frequently appeared words, the authors implemented Word2Vec with the CBOW model. There are several parameters for conducting word2vec. (The authors -> This study) restricted the minimum count of words to 200. In other words, the authors only considered words that appeared at least 200 times in overall articles. **(\*\*Even if such a word appears several times in an article, we count a word once.)-> this is wrong delete\*\*)** This model set the words vector into 400 dimensions which represent better accuracy of the model. Except for these two parameters, word2vec was operated with the default setting.

***Network analysis***

Networks are a versatile method to show and analyze simple and complex interactions among factors in articles and thus they are used for studies in diverse areas. The network representation is simple but rigid since many parts of a specific system are sorted out and concentrate on the interaction among its elements (Menczer et al. 2020). The network is represented with nodes and edges which connect nodes to each other. In this study, five keywords and similar words that have over 0.5 cosine similarity with the keywords are represented as nodes. The authors measured the Jaccard coefficient between each keyword. Jaccard index is applied to compare similarity (dissimilarity->delete), and the formula to represent it as follows.

Jaccard coefficient calculates the result of the division between the number of features that can be seen to all divided by the total number of features (Niwattanakul et al. 2013). The authors examined the interaction over five keywords and computed the Jaccard coefficient between two keywords over ten combinations of keywords set.

***Uniform Manifold Approximation and Projection (UMAP)***

With the result of the Word2Vec, the authors generated 400-dimensional word vectors. UMAP was used to visualize the vector (in-> into ) a low dimension space as UMAP is the state of the art technique for dimensional reduction. Dimension reduction creates low dimensional space without loss of structure in high dimensional space. UMAP has been widely used in various fields with larger sizes of data (McInnes et al. 2018). McInnes et al. (2018) describe UMAP as a theoretical view . Riemannian geometry and algebraic topology are theoretical ground that constructs UMAP. UMAP operates on weighted graphs, and it uses k-neighbors to cluster groups. UMAP has two phases with a graph learning algorithm. The weighted k-neighbor graph is created in the first phase, and a low dimensional layout of the graph is calculated in the second phase. UMAP is usually compared with the alternative dimensional reduction technique, t-SNE. Compared with t-SNE, UMAP performs faster and more efficiently as well as better preserves global structure. (In this study, the authors -> This study) conducted dimensional reduction with UMAP and visualize word vectors in a two-dimensional space. Every word expressed in a multi-dimensional vector was reduced to a two-dimensional vector, and then we annotated every dot in a two-dimensional graph. Since the authors focused on five keywords, the annotation of keywords had larger fonts and similar words with each keyword represented in the same color. In other words, the authors can easily examine keywords and similar words with keywords that are in the same cluster.

**Results**

***Preliminary analysis***

(Through web crawling, a total of 1,010 New York Times relevant articles were found. )- > delete?. Basically, “construction accident” was used as a search term on the New York Times web page, and the top 1,010 relevant articles were retrieved. Of all the data collected by web crawling, only document type articles were used for analysis, and 149 articles of the interactive document and blog type were excluded from the analysis. The article types excluded from the analysis are not valid for scraping because the structure is irregular. Therefore, a total of 861 articles were analyzed. After completing data cleaning, the authors classified articles according to composition. It was confirmed that the articles are generally composed of title, date, and body, and the text data included in the article was classified in consideration of this. The Beautiful Soup library was used to classify and scraped the parts needed for analysis. Table 1 shows the collected of source data.

**Table 1. Source data information**

|  |  |
| --- | --- |
| **Category** | **Information** |
| Issued duration | 2000.01.01 – 2019.12.31 |
| News source | The New York Times |
| Total number of articles | 1010 |
| The number of relevant articles | 861 |
| Total number of words | 453,283 |

***(Basic statistical analysis) -> Exploratory data analysis (preliminary + basic statistical analysis)***

Based on the selected data analyzed through preliminary analysis, the Natural Language Toolkit (NLTK) library was used to conduct the Natural Language Process. By using the NLTK library, the authors removed 'English stopwords' and changed all words to lower case. The definition of 'English stopwords' is unnecessary words that are filtered before and after the processing of natural language data. This is the data processing required for further analysis such as word2vec, and the minimum preprocessing is performed within the range that does not transform the original data as much as possible. As one of the basic statistics, frequency analysis was conducted for each keyword. The frequency of words appearing in each article was measured, and even if repeated twice or more in one article, it was measured once. Among the words satisfying this condition, five keywords related to this study were selected. The five keywords were fire, fell, collapsed, building, and people. The selection criteria considered the types of accidents on construction sites and keywords representing fire. In many studies, three major elements of building fire are defined as a fire, building, and people. In addition, fell and collapsed are the most frequent types of accidents on construction sites. All keywords should have sufficient frequency to be used in this study.

According to the analyzed results, buildings and people were the most frequent. The fire, fell, and collapsed keywords related to the type of accident on the construction site showed relatively similar frequency. Since fell and collapsed are similar words, it may be reasonable to compare fire with the frequency of combining them. When comparing the combined numbers of fell and collapsed, the fire has a frequency of 27% and fell and collapsed have a frequency of 73%. This result differs significantly from the Bureau of Labor Statistics (BLS) analysis of construction site accident frequency. According to the BLS report, among the accidents on the construction site, fell and collapsed accidents account for about 40% of all accidents, and fire accidents are 2%. When converted to 100%, fire accidents have a frequency of 5% in fell-related accidents. There is a large gap in 27% of the results of this study and 5% of the BLS report. The possible reasoning about this difference can be explained by the characteristics of the media that the authors described in the introduction section. Fire-related accidents may be exposed to the media more than fell-related accidents, which may explain that fire accidents on construction sites have a greater impact than fell-related accidents.



**Figure 1. Frequency of keywords**

***Word embedding with Word2vec***

In this study, cosine similarity was used to provide a similarity between keywords. (Cosine similarity is a concept mainly used to measure the similarity between words, and Word2vec also uses it. In order to express the distance between vectors in multidimensional space as a cosine value, this concept was used.) - > delete redundant with method part A total of 10 combinations were made by classifying each of the 5 keywords selected above, and the similarity of each was calculated. In order to confirm the reliability of this study, cosine similarity among the most similar words among the analyzed words was first checked. The words chosen by the authors are fell and collapsed. The dictionary meaning between these two words is very similar. Therefore, if the results of this study are reliable, the similarity between these words should be high. The similarity between the two words was 0.951, which was calculated much higher than the similarity between the other keywords. Through this, the authors can confirm the reliability of this study. In the results related to the fire keyword, the fire keyword was more similar to the building keyword than the people keyword. The similarity between fire and building was 0.525, and the similarity between fire and people was 0.331. It can be interpreted that fire showed a much higher degree of similarity to words related to the building than words related to people. For the fell and collapsed keywords, the similarity between each keyword and the people and building keywords was very similar. In the case of the fell, the similarity with the building keyword was 0.443, and the similarity with the people keyword was 0.486. This result contrasts with the fire keyword and fell has a higher similarity to the people keyword. Also, the similarity between people and buildings was calculated as a negative value, indicating that the correlation between the two keywords is not great. The following table shows the similarity between keywords.

**Table 2. Similarity between keywords**

|  |  |
| --- | --- |
| **Relationship between keywords** | **Similarity** |
| Fire and Building | 0.525 |
| Fire and People | 0.331 |
| Fire and Collapsed | 0.691 |
| Fire and Fell | 0.728 |
| People and Fell | 0.486 |
| People and Collapsed | 0.510 |
| People and Building | - 0.144 |
| Building and Collapsed | 0.554 |
| Building and Fell | 0.443 |
| Fell and Collapsed | 0.951 |

The table below shows five selected keywords and 20 words with high similarity. It was written based on the cosine similarity values, listed in order of similarity. In the results, the fell and collapsed keywords were calculated to be most similar to each other. This shows the reliability of this study, as well as the calculation of cosine similarity between keywords. In terms of similarity word results related to the type of accident on the construction site, such as fire, fell, and collapsed keywords, the words Monday and Friday are included in the list. This shows results consistent with research on the construction industry's 'The distribution of injuries'. In a related study, the day of the week where injuries were most likely to occur in the field was found to be Monday, which is consistent with the results of this study (Wigglesworth 2006). In addition, unlike the results of other keywords, the fire keyword result showed a high degree of similarity with the word “night”. Generally, the possibility of spread increases when a fire occurs at night. (Because ->This is Because ) most commercial buildings are less likely to stay after working hours, and even residential buildings have difficulty recognizing fire during sleep. In most fire accidents, the probability of the spread of a fire tends to increase as the initial detection of the fire is delayed. In particular, a construction site may have a small number of employees staying overnight for the monitoring or may be empty. The characteristics of these construction sites can lead to the rapid spread of fires in fire accidents. Also, during the construction phase of the building, it is more difficult to recognize fires because safety equipment such as fire and smoke alarms have not been completed (Hamid et al. 2003). This was highlighted once again through the results of this study. There are many words that have a meaning of ‘administration’ or ‘inspection’ in words with high similarity to the building keyword. The top five words with the most similarity to the building keyword are department, inspectors, issued, commissioner, and city, and these words tend to have a common meaning. These results show that articles related to accidents on construction sites mainly deal with building inspection and management issues. Also, words with high similarity to the people keyword have many words related to behavior.

**Table 3. Top 20 list of similar words of keywords**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fire** | **Fell** | **Collapsed** | **Building** | **People** |
| police | collapsed | fell | department | died |
| investigators | side | debris | inspectors | injured |
| authorities | floor | floor | issued | killed |
| hospital | debris | ground | commissioner | men |
| injuries | ground | west | city | accident |
| chief | street | steel | cranes | authorities |
| night | Friday | street | contractor | debris |
| floor | avenue | death | violations | train |
| man | death | side | tower | injuries |
| dead | morning | site | crane | dead |
| Monday | worker | Friday | investigation | workers |
| Friday | injuries | march | office | ground |
| driver | Monday | wall | equipment | Monday |
| investigation | injured | injuries | site | crash |
| truck | dead | authorities | district | march |
| officials | wall | Monday | charges | residents |
| death | authorities | avenue | investigators | members |
| worker | steel | tower | company | Tuesday |
| fell | crash | Tuesday | officials | Thursday |
| working | died | injured | mayor | cars |

***Network analysis***

Network analysis is an analysis method that describes the relationship of data with nodes and edges. By using this network analysis, it is efficient to interpret the relationship between nodes of data. In this study, keywords are nodes, and words with a cosine similarity of 0.5 or higher are connected by edges. Nodes depict five keywords and similar words, and the number of nodes is 136. Edges represent connection among words, there are 353 edges in the network. Table 4 shows the number of nodes each keyword has, and the collapsed keyword has the most nodes. There was no significant difference in the number of nodes each keyword had.

**Table 4. Network analysis information by keywords**

|  |  |
| --- | --- |
| **Keywords** | **Connected Nodes** |
| Collapsed | 91 |
| Fell | 88 |
| Fire | 80 |
| Building | 50 |
| People | 50 |
| Total nodes: 136 / Total edges: 353 | |

In addition, the Jaccard coefficient between each keyword was calculated and provided. The Jaccard coefficient values between keywords are shown in Table 5. The Jaccard coefficient is a statistical value used to measure the similarity and diversity of sample data. Through this, the network of each keyword can be expressed as one unified network. This network graph is visualized in Figure 2. This entire network has different sizes for nodes and annotations based on degree. This means that the more nodes are connected (the higher the degree), the larger the size of the node. Among the keywords in this study, collapsed has the largest degree. In addition to the five keywords, the word in which the node size is noticeably larger is'death'. This shows that the (death word -> word “death”) has a high degree besides the five keywords in the network in Figure 2.

**Table 5. Jaccard Coefficient between keywords**

|  |  |
| --- | --- |
| **Relationship between keywords** | **Jaccard Coefficient** |
| Fire and Building | 0.262 |
| Fire and People | 0.262 |
| Fire and Collapsed | 0.513 |
| Fire and Fell | 0.541 |
| People and Fell | 0.484 |
| People and Collapsed | 0.424 |
| People and Building | 0.020 |
| Building and Collapsed | 0.205 |
| Building and Fell | 0.160 |
| Fell and Collapsed | 0.772 |

A close up of a logo

Description automatically generated

**Figure 2. The network of keyword's similar words**

***Visualizing with UMAP***

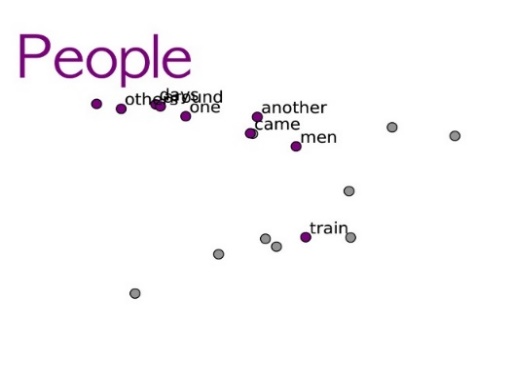
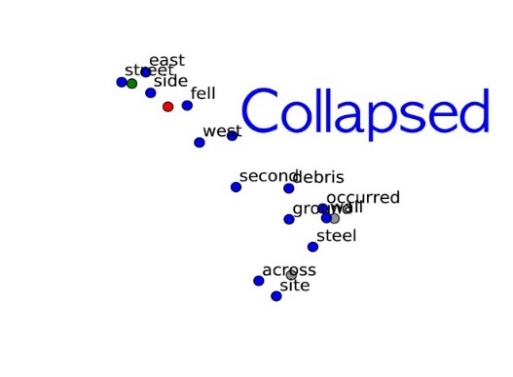
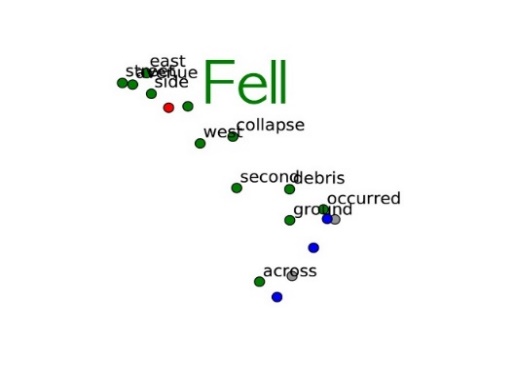
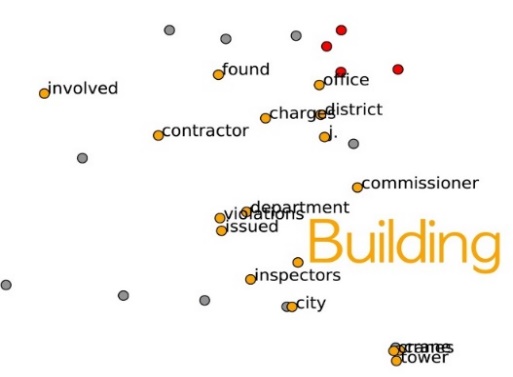
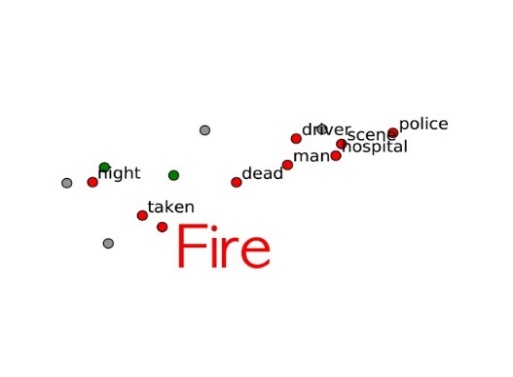
In this study, the results analyzed through Word2vec were visualized in UMAP. Figure 3 shows the overall UMAP graph for this study. In order to increase the discernment of the graph, each keyword is expressed in a different color, and the range of each keyword is indicated by gradation. The gray dots on the UMAP are words with low similarity to keywords. As shown in the figure below, fell and collapsed were almost overlapped on the UMAP. This means that the similarity between the two keywords in UMAP is very high. In the case of the fire keyword, the range appears wider than other keywords, and there is an intersection with the building keyword. For the fell and collapsed keywords, the people keyword appeared to be closer than the building keyword.

A picture containing text

Description automatically generated

**Figure 3. Word2vec result overview with UMAP**

The figure 4 shows the UMAP visualization limited to each keyword. Through this, it is possible to know how similar words with each keyword are expressed at the UMAP and the distance between each word. By using this, this study can provide information about the relation among words in two-dimension plane according to the similarity of words.



**Figure 4. Subgraphs for keywords**

***Factors related to fire accidents***

Articles related to fire accidents were classified in order to in depth analyze the factors related to fire accidents on construction sites. Through the classification of the season in which the article was written, the most articles related to fire accidents on construction sites were written in summer. According to the BLS report, which investigates the distribution of injuries throughout the year in workplaces, more injuries occur during the summer season than at other times of the yea(Pierce 2013)r (Pierce 2013).(Pierce 2013) In addition, the two sources had a common feature that injuries were much less frequently reported near the end of the calendar year. Table 6 shows the distribution of articles and injuries related to fire accidents on construction sites by season.

**Table 6. Distribution of articles related to fire accidents and injuries, by season**

|  |  |  |
| --- | --- | --- |
| **Division (season)** | **Collected articles** | **The Bureau of Labor Statistics** |
| Spring (Mar-May) | 29.6% | 24.8% |
| Summer (Jun-Aug) | 31.2% | 27.4% |
| Fall (Sep-Nov) | 16.0% | 24.2% |
| Winter (Dec-Feb) | 23.2% | 23.6% |

In this study, the topics and major factors of articles dealing with fire accidents on construction sites were classified. In the collected articles, the biggest factor in the fire accident was an explosion caused by a chemical gas leak. Depending on the progress of the construction project, there is a possibility that gas pipes will be exposed to the site, and safe control of gas pipes is an important factor in improving fire safety. For this, a modified building code and on-site manual will be required. In addition, building-related factors such as building code violations, lack of regular inspection, and inadequate fire safety systems account for about 40%. (Because ->This is because) explosions and fires caused by chemical gases can also be partially controlled with the building and fire code, many fire causes on construction sites are deeply related to the building keyword. Factors related to the behavior of people on construction sites, such as carelessness and welding activities, accounted for a relatively small proportion. It can be confirmed that these results are consistent with the results of the analysis through Word2vec in this study. Table 7 shows the main factors of articles dealing with fire accidents on construction sites. Among the factors in the table, those expressed in gray color are related to the building keyword.

**Table 7. Distribution of major factors in fire accidents**

|  |  |
| --- | --- |
| **Factor** | **Percentage** |
| Explosion related chemical gas | 20.0% |
| Violation of building and fire code | 16.8% |
| Lack of building and site inspection | 11.2% |
| Inappropriate fire safety system | 10.4% |
| Carelessness | 8.8% |
| High wind | 5.6% |
| Absence of an evacuation plan | 5.6% |
| Activities related to demolition | 5.6% |
| Welding | 4.8% |
| Activities related to renovation | 3.2% |
| Etc. | 8.0% |

**Discussion**

This study explored the relationship between factors related to accidents on construction sites through web crawling and deep learning approaches. First, keyword analysis was conducted with 861 articles collected by web crawling in this study. Since the authors set the search condition when using web crawling as a “construction accident,” most articles focus on accidents on construction sites. For the next step of data analysis, five keywords related to accidents and safety on the construction site were selected, which are fire, fell, collapsed, building, and people. Since the frequency of these keywords must be sufficient for further analysis of the collected data, the authors checked the frequency.(Add + Even if such a word appears several times in an article, we count a word once.) The smallest frequency was fire, measured 192 times. The keyword with the highest frequency was building, which was measured 514 times. Of the 5 keywords, the minimum frequency was 192, which was sufficient for further analysis. The most interesting thing about frequency analysis is that the frequency of fire is relatively high. As mentioned earlier, the frequency of fire is 192 times, which is not much different from the frequency of fell and collapsed. The frequency of fire was about 10% smaller than fell and about 20% smaller than collapsed.  In the BLS report, the frequency of fire accidents is investigated within 2% of the total construction accidents, so it differs from the results in this study. The difference between the frequency of actual accidents and the frequency of media exposure can be used as evidence to confirm the great impact of fire accidents on construction sites. Of all the accidents on the construction site, fire accidents have a low frequency, and thus fire accidents are excluded from the list of major accidents on the construction site. OSHA announces "Fatal Four" on the construction site every year. The "Fatal Four" section of this report includes accident types related to falls, struck by object, electrocutions, and caught-in/between. As OSHA publishes reports and statistics related to 'Fatal Four' every year, many stakeholders on the construction site can check this and reflect it on the construction site. However, (because -> as) the report is based on the frequency of accidents and deaths, there are limitations. Fire accidents are excluded from this “Fatal Four,” but the results of this study confirmed that the fire accident had a similar impact on the media as the main types of accidents on the construction site. Types of accidents in which the media frequency is higher than the actual accident frequency, such as fire accident, means that the impact of one accident frequency is greater than that of other accident types. This result is in line with the fire accident characteristics. Fire accidents are more likely to lead to additional accidents than other types of accidents on the construction site. Due to the spread of fire, it can affect the surrounding buildings and roads. Currently, most research and reports related to the safety of construction sites focus on the frequency of accidents. If researchers consider the characteristics of the accident types and their surrounding influence along with the frequency of the accident, this could be a new approach to improving construction site safety.

In this study, the number of connected nodes of each keyword was found through network analysis. The fire had 80 connected nodes, which is about 12% less than the collapsed with the most nodes. The fire keyword has the same coefficient between buildings and people and fell and collapsed keywords have higher coefficients with people than building keyword. This study also conducted an analysis using Word2vec, one of the word-embedding models. According to the results of this study, the fire keyword showed higher similarity to the building keyword than the people keyword. This can be explained by the fact that fires on construction sites have a higher relationship with building factors. This is in line with that in this study, the violation of the building code and the lack of regular inspection and training appeared as the main factors of fire on the construction site. Modified building and fire codes and periodic inspections considering the fire risk characteristics of the construction site can be the most effective way to reduce fire accidents. In the case of fell keyword, the similarity with the people was higher than the building. This shows that in the case of a fall accident, the impact of people's activity or behavior may be greater than that of building factors. These results can be used as relevant data when establishing safety regulations on construction sites. In the case of explosions, which account for the largest proportion of fire accident-related articles collected in this study, additional modifications to the building and fire code will be required. There are many activities where chemical gas is used on construction sites, and gas pipes may be exposed to the site. Cracks and damage of exposed pipes is the biggest reason for explosion accidents on construction sites. Many articles also point to no major regulatory changes associated with this fact. For the safety of the construction site, it is essential to modify and strengthen the building and fire codes related to the management and inspection of chemical gases according to the construction stage. In addition, in the case of fell-related accidents, it may be effective to establish rules that affect people's activity or behavior.

The list of words with high similarity to the three keywords related to accidents (fire, fell, collapsed) commonly includes Monday and Friday. This result is consistent with statistical data related to accidents on construction sites. According to a related study, workers' injuries on construction sites were the highest on Monday (Wigglesworth 2006). Unlike other types of accidents on construction sites, the word 'night' has a high degree of similarity in the fire keyword. Construction sites tend to have few occupants after work hours compared to other building types. This issue can reduce the possibility of early detection of a fire on a construction site. Additional fire and smoke detection system or monitoring system is required to immediately detect night-time fires on construction sites. In addition, regulations must be supplemented so that such fire safety equipment can be installed effectively and compulsorily on construction sites. Also, words with high similarity between the fire and building keywords have words related to ‘inspection’ in common. The results of this study remind us once again that periodic and detailed inspection is essential to improve the safety of construction sites. By providing UMAP visualization graphs, the results of the research can be expressed in graphs, not texts, which can enhance the understanding of related researchers. In the whole UMAP graph with all keywords marked, the five keywords the authors selected are concentrated in similar locations. This shows that the five keywords are organically related and can affect each other. In detail, the fire keyword was expressed close to the building keyword in UMAP, and some ranges overlapped.

**Conclusions and Recommendations**

Although many studies have been conducted on the safety of construction sites, related studies have limitations in analyzing based on the statistical data. Because of this, the main focus of these studies is limited to accident types such as falls, which have a high frequency on construction sites. In the case of fire accidents on construction sites, it has a low frequency, but the impact of one fire accident may be greater than that of other accident types. When analyzing accidents on construction sites, it is necessary to consider not only the frequency of accidents but also various factors such as the surrounding impact and probability of spread. In order to accurately analyze the influence of factors related to accidents on construction sites, it is necessary to consider (not only using existing statistical data but also introducing and analyzing new methodologies. -> not just using existing statistical data but introducing and analyzing new methodologies.) In this study, the frequency of media exposure as well as the frequency of accidents on construction sites was investigated to find the impact and factors of each type of accident. New approaches such as web crawling, network analysis, and word embedding using deep learning were introduced and used in this study. Through this study, it was possible to investigate the media exposure level of keywords related to accidents on construction sites and to identify factors affecting accidents on construction sites. The authors found that fire accidents accounted for only 2% of all accidents at construction sites, but the media exposure accounted for about 25%. This shows that the impact of a fire accident on a construction site can be greater than that of other types of accidents. It also means that fire safety parts should not be overlooked when providing codes and manuals to improve the safety of construction sites. In addition, this study suggested that the main points of articles dealing with fire accidents on construction sites are (deeply-> highly) related to the building keyword. Building and fire code violations, lack of regular inspections, and an incomplete fire safety system were major factors in fire accidents on construction sites. This is not a result of the careless and accidental behavior of workers on construction sites, but factors that can prevent accidents with appropriate regulations and on-site safety systems. By exploring the relationships between factors for fire accidents on construction sites where various factors are compounded, it is possible to make meaningful data in developing safety-related regulations that consider the characteristics of construction sites.

However, it is necessary to analyze by adding various keywords as well as the main keywords analyzed in this study. In addition, analysis of other media that can investigate the frequency of media exposure is required in addition to the news article. Future studies should (account for -> explain/introduce) (other data -> different data) involved in the fire safety range since the main factors might be different depending on the collected data. In this study, an analysis was conducted to find the relationship and similarity between representative keywords, but analysis of sub-words may be required based on sufficient data collection. Further studies may consider various types of accidents and factors of the construction site. Therefore, future studies (can ->will/might be able to) analyze the factors related to the safety of construction sites by subdividing them, and through this, fire safety on construction sites (can -> will/might) be considered on a larger scale. (Also -> Furthermore), new methodologies in other fields, such as web crawling and deep learning, must be actively introduced to suggest ways to improve the safety of construction sites.

**References**

Abdullah, D., and Wern, G. C. M. "An analysis of accidents statistics in Malaysian construction sector." *Proc., International Conference on E-business, Management and Economics*, IACSIT Press Honk Kong, 1-4.

Aires, M. D. M., Gámez, M. C. R., and Gibb, A. (2010). "Prevention through design: The effect of European Directives on construction workplace accidents." *Safety science*, 48(2), 248-258.

Ali, R., Ali, A., Khatak, A. M., and Aslam, M. S. "Large Scale Image Dataset Construction Using Distributed Crawling with Hadoop YARN." *Proc., 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, IEEE, 394-399.

Bag, S., Kumar, S. K., and Tiwari, M. K. (2019). "An efficient recommendation generation using relevant Jaccard similarity." *Information Sciences*, 483, 53-64.

Choi, J., and Lee, S.-W. (2020). "Improving FastText with inverse document frequency of subwords." *Pattern Recognition Letters*.

Church, K. W. (2017). "Word2Vec." *Natural Language Engineering*, 23(1), 155-162.

D’Haen, J., Van den Poel, D., Thorleuchter, D., and Benoit, D. F. (2016). "Integrating expert knowledge and multilingual web crawling data in a lead qualification system." *Decision Support Systems*, 82, 69-78.

De-Ching, H., Shen-Wen, C., Chien-Hung, L., Po-Ta, H., Yi-Ting, S., and Huei-Ru, S. (2011). "A study for the evacuation of hospital on fire during construction." *Procedia Engineering*, 11, 139-146.

Guy, I., Schwartz, I., and Radinsky, K. (2019). "Search system for providing web crawling query prioritization based on classification operation performance." Google Patents.

Hamid, A. R. A., Yusof, W. Z. W., and Singh, B. S. B. J. (2003). *Hazards at construction sites*, Universiti Teknologi Malaysia.

Hindocha, E., Yazhiny, V., Arunkumar, A., and Boobalan, P. (2019). "Short-text Semantic Similarity using GloVe word embedding."

Hong, S.-H., Lee, S.-K., and Yu, J.-H. (2019). "Automated management of green building material information using web crawling and ontology." *Automation in Construction*, 102, 230-244.

Hosseinian, S. S., and Torghabeh, Z. J. (2012). "Major theories of construction accident causation models: A literature review." *International Journal of Advances in Engineering & Technology*, 4(2), 53.

Hui, L., Yongqing, W., Shimei, S., and Baotie, S. (2012). "Study on safety assessment of fire hazard for the construction site." *Procedia Engineering*, 43, 369-373.

Ingason, H., Lönnermark, A., Frantzich, H., and Kumm, M. (2010). "Fire incidents during construction work of tunnels."

Jeong, M.-J., Lee, M.-G., and Ham, E.-G. (2014). "Assessment of Fire Evacuation Safety for Building Construction." *Journal of the Korean Society of Safety*, 29(6), 119-124.

Khatua, A., Khatua, A., and Cambria, E. (2019). "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks." *Information Processing & Management*, 56(1), 247-257.

Kim, Y.-A., Kim, G.-H., Kim, H.-J., and Kim, C.-G. (2019). "Design and Implemention of Real-time web Crawling distributed monitoring system." *Journal of Convergence for Information Technology*, 9(1), 45-53.

Lai, S., Liu, K., He, S., and Zhao, J. (2016). "How to generate a good word embedding." *IEEE Intelligent Systems*, 31(6), 5-14.

Lee, S.-R. (2012). "An Experimental Study on the Fire Risk at Welding· Cutting Process." *Fire Science and Engineering*, 26(3), 60-66.

Ling, F. Y. Y., Liu, M., and Woo, Y. C. (2009). "Construction fatalities in Singapore." *International Journal of Project Management*, 27(7), 717-726.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. "Two/too simple adaptations of word2vec for syntax problems." *Proc., Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299-1304.

Lopez-Aparicio, S., Grythe, H., Vogt, M., Pierce, M., and Vallejo, I. (2018). "Webcrawling and machine learning as a new approach for the spatial distribution of atmospheric emissions." *PloS one*, 13(7).

Mahto, D. K., and Singh, L. "A dive into Web Scraper world." *Proc., 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 689-693.

Massimino, B. (2016). "Accessing online data: Web‐crawling and information‐scraping techniques to automate the assembly of research data." *Journal of Business Logistics*, 37(1), 34-42.

McInnes, L., Healy, J., and Melville, J. (2018). "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426*.

Menczer, F., Fortunato, S., and Davis, C. A. (2020). *A First Course in Network Science*, Cambridge University Press.

Moon, S., Shin, Y., Hwang, B.-G., and Chi, S. (2018). "Document management system using text mining for information acquisition of international construction." *KSCE Journal of Civil Engineering*, 22(12), 4791-4798.

Morgan, J., Tietje, R., Wang, D., and Pattabhi, T. (2020). "Web Threat Investigation Using Advanced Web Crawling." Google Patents.

Mustika, N. R. (2018). "Automated Black Box Testing using Selenium Python." *International Journal of Computer Science and Software Engineering*, 7(9), 201-204.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. "Using of Jaccard coefficient for keywords similarity." *Proc., Proceedings of the international multiconference of engineers and computer scientists*, 380-384.

O'Toole, T. (2002). "Construction site safety roles Journal of Construction Engineering and Management 128."

Ombabi, A. H., Lazzez, O., Ouarda, W., and Alimi, A. M. "Deep learning framework based on Word2Vec and CNNfor users interests classification." *Proc., 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT)*, IEEE, 1-7.

Paul, S., Mitra, A., and Dey, S. (2017). "Issues and challenges in web crawling for information extraction." *Bio-Inspired Computing for Information Retrieval Applications*, IGI Global, 93-121.

Pierce, B. (2013). "The Seasonal Timing of Work-Related Injuries October 2013." *Joint Statistical Meetings*, 2371-2381.

Rekabsaz, N., Lupu, M., and Hanbury, A. "Exploration of a threshold for similarity based on uncertainty in word embedding." *Proc., European Conference on Information Retrieval*, Springer, 396-409.

Risselada, H., Verhoef, P. C., and Bijmolt, T. H. (2016). "Indicators of opinion leadership in customer networks: self-reports and degree centrality." *Marketing Letters*, 27(3), 449-460.

Rodríguez, P., Bautista, M. A., Gonzalez, J., and Escalera, S. (2018). "Beyond one-hot encoding: Lower dimensional target embedding." *Image and Vision Computing*, 75, 21-31.

Rong, X. (2014). "word2vec parameter learning explained." *arXiv preprint arXiv:1411.2738*.

Sandhu, R., Gill, H. K., and Sood, S. K. (2016). "Smart monitoring and controlling of Pandemic Influenza A (H1N1) using Social Network Analysis and cloud computing." *Journal of Computational Science*, 12, 11-22.

Shao, L., Zhang, H., Jia, M., and Wang, J. (2017). "Efficient and effective single-document summarizations and a word-embedding measurement of quality." *arXiv preprint arXiv:1710.00284*.

Smith, M., and Gorgoni, S. (2018). "An introduction to network analysis." *Networks of International Trade and Investment: Understanding globalisation through the lens of network analysis*, 1.

Wigglesworth, E. (2006). "Occupational injuries by hour of day and day of week: a 20‐year study." *Australian and New Zealand journal of public health*, 30(6), 505-508.

Yang, S., Wi, S., and Kim, S. (2018). "Development Methodology of Web Crawling Based on Physical Properties DB of Building Materials for the Efficiency of Building Energy Simulation." *한국생활환경학회지*, 25(4), 467-475.

Yin, Z., and Shen, Y. "On the dimensionality of word embedding." *Proc., Advances in Neural Information Processing Systems*, 887-898.

Yoshioka, K., and Dozono, H. "The Classification of the Documents Based on Word Embedding and 2-layer Spherical Self Organizing Maps." *Proc., Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 357-361.

Zhang, Z., Huang, Y., Zhu, P., and Zhao, H. "Effective character-augmented word embedding for machine reading comprehension." *Proc., CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, 27-39.

Zheng, C., He, G., and Peng, Z. (2015). "A Study of Web Information Extraction Technology Based on Beautiful Soup." *JCP*, 10(6), 381-387.

Zhou, Q., Fang, D., and Wang, X. (2008). "A method to identify strategies for the improvement of human safety behavior by considering safety climate and personal experience." *Safety Science*, 46(10), 1406-1419.

Zou, P. X., and Zhang, G. (2009). "Comparative study on the perception of construction safety risks in China and Australia." *Journal of construction engineering and management*, 135(7), 620-627.