Construction Industry   
Text Mining Analysis

Institute of Systems Science,

National University of Singapore

Team PT 03

Anusuya Manickavasagam (A0163300Y)

Chinnasubbareddygari Mohan Reddy (A0163433L)

Gello Mark Vito (A0163448Y)

Isaac Varun Kumar (A0163230W)

Kesavan Sridharan (A0163207M)

Nandha Kumar (A0163346E)

# Executive Summary

This study attempts to dissect the given construction industry data. It aims to discover insights and provide quantitative explanations of what can be deduced in the construction industry. Using text mining techniques and established Python libraries, the proponents try to mine and interpret the data through supervised and unsupervised models.

The data contains of two sets, the first is categorized and labeled, while the other contains a lot more data but is uncategorized. Using the first set and some data from the second set, the training data was built to come up with a model that is used to categorize the second set. By building the models, a series of datamining techniques were applied on the data such as stop words removal, tokenization, POS tagging, and etc. Another goal of this study is to produce the type of jobs and body parts that are prone to be injured. The activities performed before the accident is also another target goals. The top category for these are extracted through synsets (identifies the data and return sets of synonyms) which the proponents further analyzed by classifying the result hypernyms.

There were various models that were made but the most efficient model that performed well is the classification model. Using this model, the proponents were able to categorize the datasets and identify that “Falling” is the category that played dominantly against the accidents reported. This should give an tighter due diligence to construction industries and give extra attention when handling heights, as well as ensuring protective devices or equipment.

By mining the data, the proponents also discovered that drivers, laborers, and foremen are more prone to accidents compared to other occupation. These results imply that these type of jobs require more safety measures. It was also found out that hand, foot, and the back are the body parts which are most likely to be injured. By putting standards on protective equipment that covers these body parts, the accidents can surely be lessen. Activities performed before encountering the accident were also ranked, “operating” tops the list followed by “installing” and “cleaning”. These type of jobs, body part, and activities can to combined to form series of conclusion such as laborers and foreman while doing cleaning or operating are prone to falls.

Lastly, the approach done when doing text mining on the data were only limited to some techniques such as classification, regression, word to vector analysis, clustering, and etc. The proponents were not able to improve greatly the performance of the models due to time constraint. The word2vec model looks promising but would require more time in tuning the model. Other advanced text mining techniques were not included in this study as well which are worth to try. Some of these are Keras, TensorFlow, LTSDM Recurrent Neural Networks, and etc.

# Introduction

## Purpose of Analysis

This study aims to identify the most common pain points in the construction industry. The construction industry is a booming field which serves as a foundation of a growing economy. However, accidents are prone and usually leads to deaths as well. This translates to loss of revenue and confidence to the construction firms. To help mitigating these risks, the proponents of this study used different data mining methodologies to identify the factors contributing to these risks so that safety measures can be recommended to be put in place.

## Business and Text Mining Goals

The data used in this study are from two files: MsiaAccidentCases.xslx and osha.xslx. The former contains a more structured dataset with entries categorized from eleven causes of accidents:

* Caught in/between objects
* Collapse of object
* Drowning
* Electrocution
* Exposure to chemical substances
* Exposure to extreme temperatures
* Falls
* Fires and explosions
* Struck by moving objects
* Suffocation
* Others

The latter contains more entries which are not classified into groups. Our aim is to do text mining on the MsiaAccident.xlsx dataset and derive a model which can correctly categorize if given a certain accident entry. After arriving at a suitable model, we apply the same model on the osha dataset and classify the accidents into the eleven categories above.

## Tools used

### **Python 3.6 (Spyder 3.2.3)**

* Jupyter Notebook

## Model Building and Testing

### Preprocessing

Upon analyzing the given files, it was discovered that the files required clean up and data rationalization. It was observed that Malaysia Accident data file contained some misclassifications as well. The following pre-processing and massaging of the input data had to be performed

* **Manual Re-classification** – Visual inspection of the document showed misclassifications. These misclassifications were manually fixed.
* **Excluding industry specific words** - The domain being accidents, the document contains large references to words such as “accident”, “died”, “victim”, “employee”, “worker”, etc. We quickly realized that these words are polluting the model. And hence were added to the stop words list.
* **Input selection** – Both data sheets contains columns that we felt were irrelevant to the given problem statement. Such columns were ignored for processing and for model building. We focused on **“summary”** columns only.
* **Feature Section and Further clean up:** Tokenization and filtering of the “summary” columns gave a huge terms list even after stop word removal. The following methodology was followed to safely remove unwanted words but at the same retain the concept described in each document.
  + **Part of speech tagging** was done and it was discovered that words that were verbs, adjectives, adverbs gave better model and associations.
  + The tokens of the document after filtering were lemmatized.

### Preprocessing Algorithm

### The data from the MsiaAccident.xlsx dataset was extracted and stored in a dataframe.

### After extraction, the following pre-processing techniques were applied on the data.

### Dropping of rows with all columns = ‘na’

*dataFrame=dataFrame.dropna(how='all')*

* + - 1. Stop Words Removal

Words which do not hold significant information were discarded to optimize the model accuracy.

* nltk.corpus.stopwords.words('english') – nltk library of stop words
* Custom stop words are added by observing the data:
  + Time related words such as year, morning, afternoon, months, etc.
  + Numerical related abbreviation such as 4th, street number, floor number, etc.
    - 1. Word Tokenize
* nltk.work\_tokenize – nltk library to tokenize words (bag of words) for analysis
* Word Lemmatize
* nltk.WordNetLemmatizer – nltk library to extract the significant portion of the word
  + Further identify the word as POS (part-of-speech) and further filtering accordingly
    1. Finally, we came up with the words which were the basis of the models we attempted to ace:

### Model Building

#### Word2vec + Logistic Regression

*“* (Raja)

* *A specific DL approach to learn vector representations of words*
* *A w2v model should be learnt first by passing in a large text collection”*
* For the Word2vec, the data we got from Preprocessing are used to feed the model.
* Using gensim library, we built the model using 1 iteration.
  + *trained\_model = gensim.models.Word2Vec(train\_data, min\_count=1)*
* The dataset is divided into 2 parts. After this, we get the vectors of each entry using the trained\_model.
  + Positive – rows which we want the model be familiarized with for categorization
  + Negative – rows in which we try helping the model to distinguish data into our desired category
* The derived samples are usually array representations of words converted into vectors hence the model name Word2vec.
* These vectors are the model output by executing the command:

*vector = trained\_model[word]*

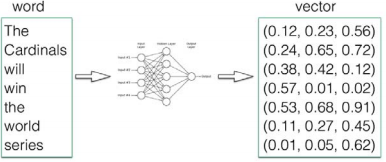
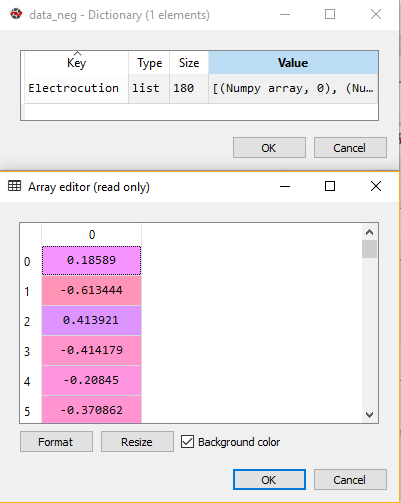


Figure Word2vec Illustration 1

Figure Word2vec Illustration 2

* After achieving the vectors through our deep learning trained\_model, the dataset is split into train and test in a 70/30 proportion. The train data are then fed to fit a basic LogisticRegression model which is then used to validate the test data.
* This model however only yields around **50+% accuracy**. There are a lots of improvements that can be done for this and it is worth a try to explore more the method. Due to time constraint, the team was not able to explore further for this model.

#### SVM using GridSearch

#### Clustering

* The Summary case from the Malaysia dataset are extracted.
* **preprocessDocument()** function is applied over the Summary case.
* To get the best value of threshold, max\_df and min\_df are changed in the Tfidvectorizer.
* Normalizer is applied to reduce the dimensions
* The variance explained in the SVD step is obtained as 91%
* The KMeans method is applied for clustering process
* The clusters size range from size 1 to size 5
* On analyzing various sizes of clusters we arrive at the best cluster size at 5.
* Cluster size 5 classifies as follows:
  + *#Cluster 0: electrical related incidents*
  + *#Cluster 1: falling incidents*
  + *#Cluster 2: Machine injuries*
  + *#Cluster 3: fire and explosion*
  + *#Cluster 4: Vehicle strucks*

### Wordnet and Pos tagging

For answering the questions risky occupations in accidents, human body parts prone to accidents and the common activities the victims were engages prior to accidents wordnet and POS tagging were used.

**Human Body parts prone to accidents:**

* **ExtractBodyPart()** function- synsets was used for extracting bodypart. Specifically, synset was applied over the “external\_body\_part.n.01” and the lowest\_hypernyms was extracted. So the bodyparts like hand,legs,head was extracted when applied over the Summary case in Malaysia and osha dataset when a body part was found in a sentence/document.

**Risky occupations in accidents:**

* **getNouns** function- The tokens which are formed after tokenizing as a part of document preprocessing are checked for noun.
* **Worker() function-** After that the noun are checked if they belong to the wn.synsets('worker'). If there are more than one occupation in a document then they are appended. The entire function is applied on the Summary case of the Malaysia and osha dataset.

**Activities victims were engaged prior to accidents:**

* **getNouns** function – The tokens which are formed after tokenizing as a part of document preprocessing are checked for ‘VBG’ part of speech.
* The function is then applied to each of the document and the result set stored in a list.
* The final result is stored in an excel and the result is analyzed.
* Many terms were obtained like ‘being’,’using’ which does not fulfil our requirement. Keeping only the activities like ‘cleaning,installing’ and so on the remaining words were removed.

### Miscellaneous Questions

Before you start

* What are the objectives for this text mining project?

The objective of this project is to do text mining on the Malaysia dataset and build a classification model which can classify the accident categories correctly.

* What’s the data?

The Masia.xls file was used for building model and Osha.xls was used to validate the model.

How’s the data quality? Any errors in the data? Are the labels correct? Is the information in the data sufficient?

Text Mining

* Which text mining techniques are required here? Which questions are more challenging?

The text mining techniques that are used here are document classification, clustering, pos tagging, word2vec model

* Do I need to build classification models? What are the categories? One category per document, or a number of most likely categories? How to achieve that? Is the amount of labeled data enough? Any alternative approach?

Yes, we have built a classification model to classify into categories

* Are there any terms that should be excluded? Any synonyms?

We have excluded the stopwords and customized stopwords which we felt were not needed in the document.

* What kind of information or entities need to be extracted? Any ready-made package to do that? What clues from the text can be used for extraction?

Occupation and body parts were extracted using wordnet. The most frequent activities was extracted using pos tagging.

* What’s next after extraction?

After extraction we did a frequency counting and found the most risky occupations, frequent activities before accident and risky body parts.

# Validation on the OSHA dataset

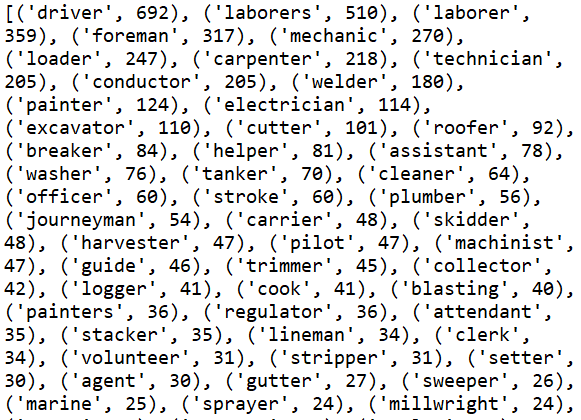
## Conclusion

### Which type of accidents (in terms of main causes) are more common in fatal or catastrophic accidents?

As per above model we can conclude that more common accident is happening due to “Falling”.

### What are the riskier occupations in such accidents?

The most risky occupations are driver,laborers and foreman.



### Which parts of human body are more prone to be injured in such accidents?

The human body part prone to injury are hand,foot and back.

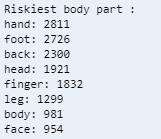
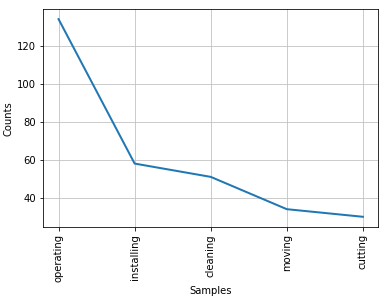


Figure 3 Body Parts Prone to Accident

### What are the common activities that the victims were engaged in prior to the accident?

After applying the part of speech tagging, specifically for ‘VBG’ we were able to extract what the victim was doing just prior to the accident. The most common activities are operating, installing, cleaning, moving and cutting.



# What was not found out?

We were not able to achieve the accuracy on classifying the accidents as the data contains too much of noise and we were not able to achieve a proper trained model.We were trying to identify the main sentences in each of the document in order to improve the overall accuracy of the model, but the task proved to be quite challenging.

## What actions can be done or taken out?

Perform aspect mining to identify the aspects in the document and apply the model to increase the accuracy and more relevant features.

## How the results can be used (describe an implementation plan)

From the above results we can determine the most common accident is "Falls". So, we can take precautionary measure to the people who are working on the riskiest occupation in these categories.  
We can create safety equipment’s and procedures on the riskiest body parts on these categories to prevent/reduce the casualty. We can also narrow down and focus on more safety measures to the specific activities people were doing.

## Further recommended research

We need to improve upon the accuracy level to classify correctly into the relevant categories of accidents. Further, we can try finding out if there was a supervisor/ manager in each accident case and try to analyze if the accidents happened due to lack of supervisors/managers on the spot/area.

# References

Raja. (n.d.). *IVLE.* Retrieved from https://ivle.nus.edu.sg: https://ivle.nus.edu.sg/workbin/file\_download.aspx?workbinid=efa6d0eb-6466-4415-8c60-31d21a906e73&dwFolderId=726c2bf6-66d5-4dc8-b6bb-eb9883690917&dwFileId=8cfb13c3-c294-4a5e-8108-1d40025016ae

# Appendix

[Figure 1 Word2vec Illustration 1 4](file:///C:\Users\GelloMark\Desktop\Google%20Drive\GitHub\KE5205-TextMining-CA\TeamPT03%20-%20Construction%20Industry%20Text%20Mining%20Analysis%20v8.docx#_Toc497098966)

[Figure 2 Word2vec Illustration 2 4](file:///C:\Users\GelloMark\Desktop\Google%20Drive\GitHub\KE5205-TextMining-CA\TeamPT03%20-%20Construction%20Industry%20Text%20Mining%20Analysis%20v8.docx#_Toc497098967)

[Figure 3 Body Parts Prone to Accident 6](#_Toc497098968)