**NUS-ISS Masters of Technology - Knowledge Engineering**

**Choong Yue Lin (A0150392L), Ho Kok Loon (A0150144W), Lee Tin Onn (A0150375J), Lei Jun (A0150141B) and Vincent Leung (A0150347L)**

Text Mining Report :   
Analysis of Workplace Accidents

**30 October 2017**

Table of Contents

Analysis of Construction Fatality and Catastrophe Accidents 2

Data 2

Accident types 2

Occupation 4

Body part injured 5

Activity prior to accident 5

Conclusion 6

References 6

**Analysis of Construction Fatality and Catastrophe Accidents**

The objective of this project is to investigate the nature of construction accidents, such that to provide insights to forecast safety performance as well as improve safety risk controls for companies in the construction industry.

# Data

We look at two datasets of construction accidents cases.

In addition to the case summaries, the Malaysia Accident dataset also contains the labelled 11 causes of accidents. Comparably, the OSHA (Occupational Safety and Health Administration) dataset is labelled with the main injury sustained.

The Malaysia dataset is smaller, consisting of 235 case summaries, while the OSHA dataset consists of 16323 cases. The Malaysia dataset are all case summaries of fatalities, while the OSHA dataset is a combination of fatal and catastrophe incidents, which include severe injuries such as amputation and in-patient hospitalisation [1].

To build a model to classify the cause of accident, the Malaysia dataset was split into 77:23, with 182 cases for training and 53 cases for testing. One mislabeled classification was replaced with the superset group’s label by changing ‘Others’ to ‘Other’. Three misclassified summary case was identified and re-labelled with the correct classification.

# Accident types

We look into the main types of accidents that result in fatality, using the Malaysia dataset for classification training, different models are trained with tf-idf (term frequency-inverse document frequency), and the best performing model, with 73.58% prediction rate, SVM based on case title, is then used to classify the OSHA dataset.

|  | Accuracy on Title text | Accuracy on Summary text |
| --- | --- | --- |
| Naive Bayes | 69.81% | 37.73% |
| Decision Tree | 64.15% | 52.83% |
| SVM | 71.70% | 56.60% |

After performing classification on OSHA data using the trained SVM model, the cause was determined. The fatal accidents in the OSHA dataset was filtered by matching words that indicates a fatality, and we obtained a subset of 6816 cases.

From the subset of cases, we are able to determine the most common accident type that result in fatality or catastrophe is ‘struck by moving objects’ accidents, accounting for 37.73% of the cases, followed by ‘falls’, which accounts for 3326 of the 16323 cases, consisting of 20.38% of the cases.

Similarly, the most common cause of fatalities is ‘struck by moving objects’, though with greater proportion, accounts for 2774 cases, 40.70% of the fatal accidents. The second most common accident type that result in fatality is ‘falls’, consisting 1430 cases, 21% of the fatal accidents.

The results closely match those found from the Malaysia Accident cases, which two most common accident types are also ‘falls’ and ‘struck by moving objects’.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | OSHA | | OSHA Fatal | | Msia Accidents | |
| Cause | Count | Percentage | Count | Percentage | Count | Percentage |
| Caught in/between Objects | 1967 | 12.050481 | 830 | 12.177230 | 46 | 19.574468 |
| Collapse of object | 409 | 2.505667 | 206 | 3.022300 | 9 | 3.829787 |
| Drowning | 177 | 1.084359 | 154 | 2.259390 | 9 | 3.829787 |
| Electrocution | 1381 | 8.460455 | 409 | 6.000587 | 17 | 7.234043 |
| Exposure to Chemical Substances | 319 | 1.954298 | 106 | 1.555164 | 2 | 0.851064 |
| Exposure to extreme temperatures | 1499 | 9.183361 | 306 | 4.489437 | 3 | 1.276596 |
| Falls | 3326 | 20.376156 | 1430 | 20.980047 | 73 | 31.063830 |
| Fires and Explosion | 552 | 3.381731 | 348 | 5.105634 | 4 | 1.702128 |
| Other | 285 | 1.746003 | 146 | 2.142019 | 17 | 7.234043 |
| Struck By Moving Objects | 6159 | 37.732035 | 2774 | 40.698357 | 52 | 22.127660 |
| Suffocation | 249 | 1.525455 | 107 | 1.569836 | 3 | 1.276596 |

# Occupation

To find the main occupation of the injured, the text was filtered to remove non occupation related words. The remaining text is then used to match against a list of keywords that are associated with occupational profession.

| OSHA | | | Struck By Moving Objects | | |
| --- | --- | --- | --- | --- | --- |
| Occupation | Count | Percentage | Occupation | Count | Percentage |
| **construction worker** | 3636 | 22.275317 | **heavy vehicle driver** | 988 | 16.041565 |
| **heavy vehicle driver** | 1614 | 9.887888 | **construction worker** | 851 | 13.817178 |
| **chemical worker** | 1172 | 7.180053 | **machine operator** | 637 | 10.342588 |
| **machine operator** | 1141 | 6.990137 | **operator** | 463 | 7.517454 |
| **electrician** | 1074 | 6.579673 | **cleaner** | 266 | 4.318883 |
| **:** | : | : | **:** | : | : |

| Falls | | | Caught in/between Objects | | |
| --- | --- | --- | --- | --- | --- |
| Occupation | Count | Percentage | Occupation | Count | Percentage |
| **construction worker** | 1720 | 51.713770 | **construction worker** | 326 | 16.573462 |
| **heavy vehicle driver** | 148 | 4.449790 | **machine operator** | 302 | 15.353330 |
| **installer** | 140 | 4.209260 | **heavy vehicle driver** | 247 | 12.557194 |
| **cleaner** | 116 | 3.487673 | **operator** | 214 | 10.879512 |
| **carpenter** | 99 | 2.976548 | **cleaner** | 97 | 4.931368 |
| **:** | : | : | **:** | : | : |

# Body part injured

| OSHA | | | Struck By Moving Objects | | |
| --- | --- | --- | --- | --- | --- |
| Bodyparts | Count | Percentage | Bodyparts | Count | Percentage |
| **hand** | 3019 | 11.788824 | **hand** | 1388 | 13.769841 |
| **foot** | 2682 | 10.472881 | **finger** | 1201 | 11.914683 |
| **head** | 2250 | 8.785974 | **head** | 925 | 9.176587 |
| **finger** | 1936 | 7.559842 | **foot** | 796 | 7.896825 |
| **back** | 1596 | 6.232184 | **hand** | 1388 | 13.769841 |
| **:** | : | : | **:** | : | : |

Of 118 types of body parts and 25609 injuries, the most common body part injured is hand, comprising of 11.79% of injured body parts. Followed by foot, and head, accounting for 10.47% and 8.79% respectively.

| Falls | | | Caught in/between Objects | | |
| --- | --- | --- | --- | --- | --- |
| Bodyparts | Count | Percentage | Bodyparts | Count | Percentage |
| **foot** | 1044 | 18.943930 | **hand** | 569 | 16.705813 |
| **head** | 692 | 12.556705 | **finger** | 448 | 13.153259 |
| **leg** | 399 | 7.240065 | **foot** | 288 | 8.455666 |
| **back** | 338 | 6.133188 | **arm** | 250 | 7.339988 |
| **hand** | 232 | 4.209762 | **head** | 240 | 7.046389 |
| **:** | : | : | **:** | : | : |

The most common body parts that are injured are appendages such as hand and foot.

# Activity prior to accident

The activity prior to the accident is identified by identifying the verb in each summary, and checked against an activity list. Tools used in performing this extraction are as follows :

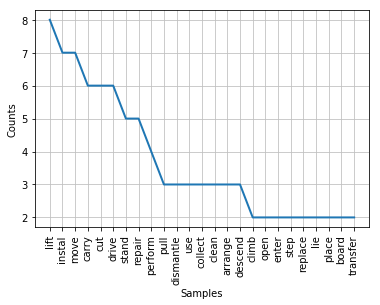
1. NLTK Regular Expression Parser to extract the phrases containing the verbs
2. NLTK lemmatizer to find the base word of the verbs
3. Simple word filters to further fine tune the accuracy of the activities extracted.

The most common activity prior the accident is the operating machinery. Followed by moving objects, standing in precarious positions and cleaning. 28260 distinct activities were extracted for OSHA dataset, whereas, only 138 activities could be extracted from the Malaysia dataset.

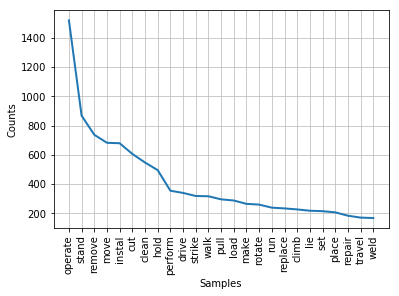
Challenges in extracting activities from the dataset:

1. Malaysia dataset only have <Summary> column which is a shorter description of the accident whereas Osha has a more detailed <Description> column which provides a more detailed description. Thus the activities extracted from Malaysia dataset is limited.
2. In some sentences, for example, the actually activity is "handling the work of maintaining", thus the actual activity is maintenance work but because the regular expression parser is designed to extract the verb activity resulted in "handling" being captured instead of "maintaining" and results in a less than accurate description of the activity extracted from the dataset. Further fine tuning were required to address this issue by only extracting the last verb which is “maintain”.

| Top 5 activities prior to accident | |
| --- | --- |
| Malaysia accidents | OSHA |
| Lifting | operating |
| installing | standing |
| moving | removing |
| carrying | moving |
| cutting | cutting |



*Frequency distribution and word cloud for activity in Malaysia dataset*



*Frequency distribution and word cloud for activity in Malaysia dataset*

| OSHA | | |
| --- | --- | --- |
| Struck By Moving Objects | Falls | Caught in/between Objects |
| operating | installing | operating |
| moving | standing | moving |
| cutting | removing | standing |
| standing | climbing | cleaning |
| removing | walking | rotating |

# Conclusion

The most common accident in workplace accidents is ‘struck by moving objects’. And the most risky occupation that results in fatal or catastrophe accidents is construction worker. And the most common part of the human body that is injured is hands. And the most common activity engaged in prior to such accidents is operating machinery.

# References

1. UNITED STATES DEPARTMENT OF LABOR. 2017. *Severe Injury Reports | Occupational Safety and Health Administration*. [ONLINE] Available at: <https://www.osha.gov/severeinjury/index.html>. [Accessed 22 October 2017].