OSHA Accident Report Analysis

Text Mining Assignment

Team Members:

Sheng Shutao A0150148M

List of contents

Executive Summary

Business goals

Our business goal is analyze OSHA accident summaries to identify occupations and workplace activities that face higher safety risks than others. Based on the result of analysis, construction project managers and safety professionals can then take appropriate measures to mitigate the identified risks and prevent the occurrence of similar accidents.

Four questions will be answer in this report:

1. Most common accident type in fatal or catastrophic accidents
   Struck by moving object, Falls and Caught in/between will be the most common accident types.
2. Most risky occupations
   Sheet metal worker, carpenter, driver are the most risky occupations.
3. Most risky human body parts
   Legs & arms are most likely been in risk.
4. Common activities that the victims were engaged in prior to the accident
   Operating/cleaning machine is most common, standing in dangerous area.

Introduction

Background

Construction industry remains the top contributor for workplace fatalities in Singapore. Similarly, poor construction safety performance can be observed in other countries. Construction accidents not only cause significant human suffering, they affect project progress and costs and the poor safety record damages the reputation of the industry and companies involved.
In construction industry, after a fatal or catastrophic accident happens, an inspection is conducted in response, generating a report including a Fatality and Catastrophe Investigation Summary. The summaries provide a complete description of the incident, generally including events leading to the incident and causal factors.

Business Goal

Analyze 'Fatality and Catastrophe Investigation Summary' to identify occupations and workplace activities that face higher safety risks than others. Based on the result of analysis, construction project managers and safety professionals can then take appropriate measures to mitigate the identified risks and prevent the occurrence of similar accidents.

Text Mining Goals

Four text mining goals will be answered in this report:

- Which type of accidents (in terms of main causes) are more common in fatal or catastrophic accidents?
- What are the more risky occupations in such accidents?
- Which parts of human body are more prone to be injured in such accidents
- What are the common activities that the victims were engaged in prior to the accident?

Data Understanding

We have 10 main categories and 1 other category:

Categories

Major categories are Caught in/between, Falls, Struck by, Electrocution. They are defined as ‘Focus Four Hazards’ by OSHA.

Caught in/between objects

Caught in-between hazards kill workers in a variety of ways. These include: cave-ins and other hazards of excavation work; body parts pulled into unguarded machinery; standing within the swing radius of cranes and other construction equipment; caught between equipment & fixed objects.

Falls

Fall hazards are present at most worksites and many workers are exposed to these hazards on a daily basis. A fall hazard is anything at your worksite that could cause you to lose your balance or lose bodily support and result in a fall. Any walking or working surface can be a potential fall hazard.

Struck by

Struck-by injuries are produced by forcible contact or impact between the injured person and an object or piece of equipment. Having said that, it is important to point out that in construction, struck-by hazards can resemble caught–in or –between hazards.

Electrocution

Electrocution results when a person is exposed to a lethal amount of electrical energy.

More categories

In our MsiaAccidentCases data, below hazard types are added:

- Collapse of object
- Drowning
- fires and explosions
- exposure to chemical substances
- exposure to extreme temperatures
- suffocation

Others

Any hazards which is not belows to above categories, will be categorized as ‘Others’.

Data files

- MsiaAccidentCases.xlsx (mention as msia data below)
  Small Categorized data set for training purpose.
  It got three columns, title, summary and category.

- Osha.xlsx (mention as osha data below)
  Large data set,
  It got columns, id, title, details, summary, pre analysis

Approaches

Question 1 Approach

Overall strategy

We have ten categories + others category.
At the beginning, all of the records in osha dataset will belongs to others category.
Then use models to categories the records to reduce the size of the others category.

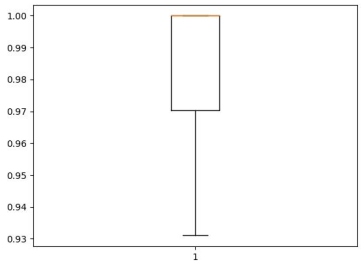
Step 1

Use osha pre-analysis data and domain knowledge mapping the existing categories to our categories.

Step 2

Use msia data, use ‘title’ column, to build the customized multinomial naive bayes model with probability threshold.

- Check if the most informative terms make sense to determine if the model is health.
- Set the probability to a high level as 0.6 first, to get some high precision predictions. (precision is 98% during cross validation)



- Get a larger categorized data set.

Step 3

Combine the category result from step 1 & step 2. The result from step 2 will be use as premier if there is any different classification happen.

Step 4

Use the msia data + step 1 result + step 2 result, use title column, retrain the customized multinomial naive bayes model to get more terms.

- Check if the most informative terms make sense to determine if the model is health.

- Adjust the probability level, to get a reasonable ‘others’ category size.

Question 2 Approach

Step 1

Create a occupation dictionary, contain the occupation names and occupation related words.

Step 2

Check details column + summary column, if they contain the occupation name, then use the occupation.

Step 3

For the rest records, compare with the occupation related words, choose the most similar category.

Question 3 Approach

I created an human body dictionary, then use the osha data to create an world cloud of human body parts for osha data ‘Detail Case’ column.
Create a word cloud graph and analysis result with tableau.

Question 4 Approach

In osha data ‘Detail Case’ column, do POS tagging, then use the pattern **‘was’+Verb+Noun till the punctuation** to get what are they doing. The **Verb+Noun till the punctuation** will be extracted.
For example:
‘On June 21 2007 at approximately 10:00 a.m. Employee #1 the hooktender **was finishing the logging operations on a yarder skid road.** He had hooked a turn of logs to send to the landing. One of the logs hung up and swung out striking Employee #1 in the back. Employee #1 was hospitalized for treatment of his injuries. ’
The ‘**finishing the logging operations on a yarder skid road.**’ will be extracted.

Then create a word cloud graph.

Details

Step 1: Data Prepare

Msia data

- Data size is small
- Got empty rows
- Category data not balance
- Category name not standard, as ‘Other’ and ‘Others’ both exist
- The classification got error against the common sense, and some record should be well categorized, other than in ‘Others’ category.

Osha data

- Don’t have column names
- Some missing fields

Step 2: Data Cleaning

After data cleaning, got file msia_edit.csv and osha_edit.csv.

Msia data

- Remove empty rows
- Standard category names
    - Rename some ‘Other’ to ‘Others’
- Fix wrong categorized rows manually, reduce the ‘Others’ category record number

Title Case	Before	After
Died while in confined space	**Others**	**Suffocation**
Died been buried	**Others**	**Collapse of object**
Died being buried	**Others**	**Collapse of object**
Died struck by falling beam	**Falls**	**Struck by**
Died struck by falling object	**Falls**	**Struck by**
Died being crushed by pallet	**Collapse of object**	**Caught in/between Objects**
...	**...**	**...**

Osha data

- Name the columns by content: id, title, details, summary, pre analysis

Step 3: Apply Approaches

Apply the approaches described in section ‘Approaches’.
Usually steps will be data preprocessing, data mining, result visualization.

data preprocessing

1. Lowercase
2. Tokenize
3. Customize stopword removal
4. Stemming (not for all questions)
5. Lemmatization
6. Punctuation removal

Data mining technology used

1. Customize naive bayes classification
2. Pos tagging
3. Regular expression
4. Customize text similar score

Result visualization

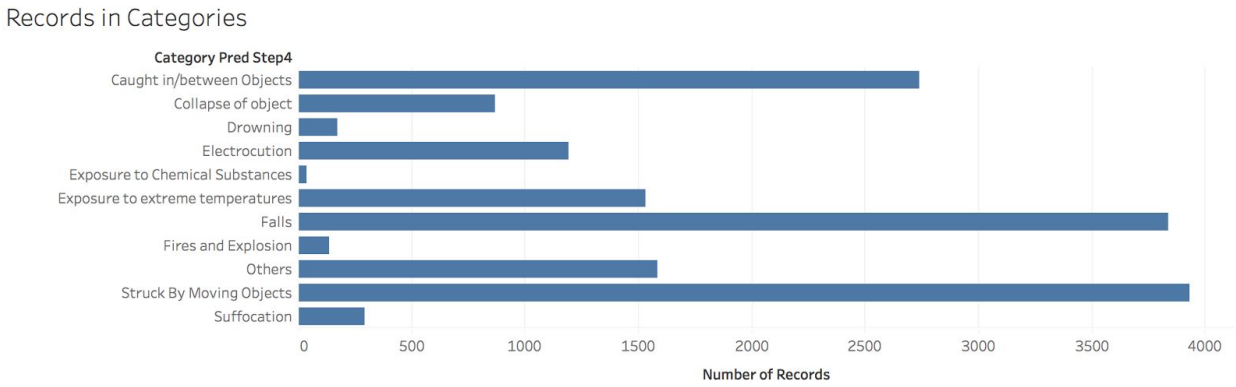
1. Word cloud
2. Tableau

Conclusions

Answers to data mining questions

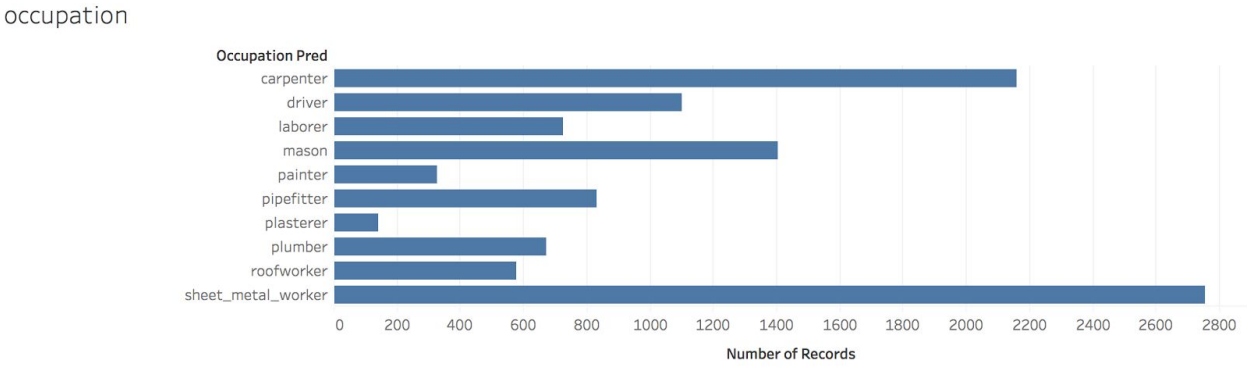
1.Most common accident type in fatal or catastrophic accidents

During all categories, Stuck by moving objects, Falls, Caught in/between are most common accident type.
According to naive bayes' most related terms, Exposure to extreme temperatures & Fires and Explosion is quite similar, which is reasonable. They are very common accident also.
Then Electrocution is also very common.



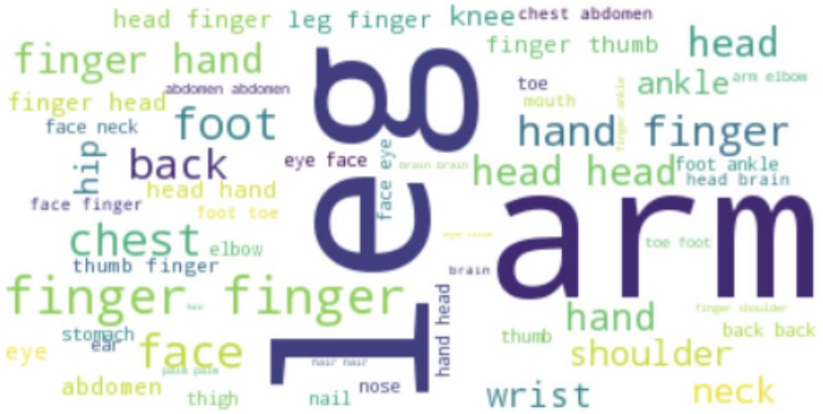
2.Most risky occupations

According to my algorithm, sheet metal worker is the most dangerous position as there are a lot danger during the work. It make sense.
The carpenter also got risky, I check the data, because they are always working with machine.
Driver also risky, after check the data, I find they are alway in the sense when loading and unloading goods which might be dangerous.



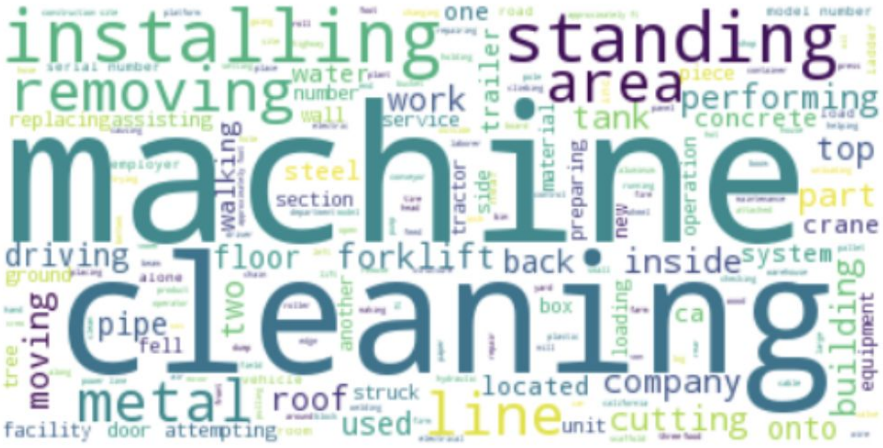
3.Most risky human body parts

From the word cloud below, the leg & arm is the most risky human body parts.
And the head, finger or foot is also important.



4.Common activities that the victims were engaged in prior to the accident

From the word cloud below, a lot of the activities related to machine, it is obvious.
A lot of victims are doing cleaning, it is interesting, after check the osha data, I find a lot of victims are cleaning the machine when the hazard happen. Cleaning process might be a black spot of the safety rules.
And some people just standing there when the hazard happen.



Recommend actions based on result

Below actions are recommended based on results above:

- Sheet metal worker should be trained well for safety.
- Improve the machine safety level for carpenter's tools.
- Improve the driver's safety requirement. Train them to keep safe when loading and unloading goods.
- In question 3 we saw a lot of hazard is injured leg and arm. We should put more efforts to invert new material working suites to protect legs & arms.
- More safety requirement should be applied to cleaning machine, may be after work, when workers are relaxing.
- Block the dangerous area should be an safety requirement, and remind workers stand in safety area when resting.
- The worker's safety equipment as working helmet, working boots, working suites need to be strictly. As there are so many 'Struck by' happen, even more than the Falls.
- The safety requirement for worker working from height need to keep very restrict. As so many Falls happened.

Further recommended research

More domain knowledge will help to make better result.
Our categories is different from the OSHA's, some category is too similar, and some categories are quite similar, which will affect the document category. Use a good category structure will help.
Some knowledge about the construction occupation can be enhance. As currently we don't have the training data for occupation recognize, I use my customized dictionary to do the prediction. The quality of my dictionary is very important, if the quality can be improve, the result will improve.
During the assignment, I tried some approaches such as dependency analysis, it is not working well, even worse than regular expression, more investigation required for this writing pattern recognize.

List of references

- https://www.osha.gov/dte/outreach/construction/focus_four/
- https://www.osha.gov/dte/outreach/construction/focus_four/falls/falls_ig.pdf
- https://www.osha.gov/dte/outreach/construction/focus_four/caught/caught_iorb_ig.pdf
- https://www.osha.gov/dte/outreach/construction/focus_four/struckby/struckby_ig.pdf
- https://www.osha.gov/dte/outreach/construction/focus_four/electrocution/electr_ig.pdf
- https://en.wikipedia.org/wiki/List_of_construction_trades