Text Mining CA Report

Li Jingchao(A0134565A), Zhang Fan(a0134457a), Ning Chao(A0134563H), Devendra Desale(A0134465E), Pan An(A0134556A) National University of Technology

This report is the project report for Textmining Course Accessment. The given questions described a situation where people are trying to find a pattern among large amount of cases of accidents in construction fields. In this paper we are going to explain our proposal and result for each problem.

Contents

Business Goal

Despite improvement in recent years, the construction industry remains the top contributor for workplace fatalities in Singapore. 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. These summaries can be analyzed 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. All of our tasks are finished with NLTK [1].

Findings

For each of the data analysis task some new information will be generated. The following discovery might be interesting:

- Three of the main reasons that causes fatal injuries to workers are:
  1. Struck by moving objects
  2. Caught in/between objects
  3. Falling from high places
- Large amount of fatal accidents involves falling, including falling objects and people falling.
- Almost all the accidents happen to employees and workers.

The most common activities were found as working on (something), operating on machines, installing the equipment. Current analysis can be extended to performed as per the type of accidents. Finding the common activities is very complex part, as the activities vary too much and even the similar activities are written in different ways. In many cases activity is not mentioned in generic way, this thing causes issue when we want to drill down to different activities.

Introduction

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. These summaries can be analyzed 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.

This paper proposed methods for text data mining in these tasks. These methods uses combinations of different methods in natural language processing:

1. SVM(for classification)
2. POS Tagging
3. Grammar and Regular Expression

Support vector machine [3] is used for basic accident case classification. Support Vector Machines are very specific class of algorithms, characterized by usage of kernels, absence of local minima, sparseness of the solution and capacity control obtained by acting on the margin, or on number of support vectors, etc.

POS tagging [2] is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. In this paper POS tagging is one of the key component for extracting information from large amount of messages.

1 Fatal Accidents

Practice the concepts and techniques we have learned in Text Mining elective. Perform text mining on Fatality and Catastrophe Investigation Summaries. Find out type of accidents (in terms of main causes) which are more common in fatal or catastrophic accidents. Find out kinds of objects which cause the accidents. Extract the more risky occupations in such accidents. Find out the common activities that the victims were engaged in prior to the accident.

Our method is to use SVM classification method to train and test a model, then use the model to do classification on type of accidents (in terms of main causes) on new data. Support Vector Machine can be applied not only to classification problems but also to the case of regression. Still it contains all the main features that characterize maximum margin algorithm: a non-linear function is leaned by linear learning machine mapping into high dimensional kernel induced feature space. The capacity of the system is controlled by parameters that do not depend on the dimensionality of feature space.

1.1 Data Preprocessing

The giving training data quality is not so good. Some of the labels are not correct. Before doing text mining work, we made some modifications in the two tables. For this task we mainly used MsiaAccidentCases.csv. Some of the types are not consistent with the titles and summary contents. We correct them according to the title and summary content. Besides there are other and others in the type list. We change them to other to make it uniform. We corrected almost 24 type of accidents of the records.

1.2 Classification

We use the MsiaAccidentCases.csv data to build the classification model. There are 3 columns of this data set. During the classification model building, we use only the cause and summary. Split the MsiaAccidentCases.csv data into 2 parts, 80% for training and 20% for % testing. Use TF-IDF Vectorizer from % the data features. Use SVM algorithm to train the classification model on the 80% of the MsiaAccidentCases.csv data. Then use the rest 20% of the data to do testing.

For SVM configuration we set $C = 5000$, $gamma = 0.0$, $kernel =' rbf'$. After the training we got an accuracy score of 0.542.

The following picture shows the confusion matrix.

Below is the classification report of the test data. We can see the precision, recall, fi-score and support value of the built model.

1.3 Prediction

Do classification about the accident cause of the osha.csv data. There are several columns of the osha.csv data. We use only the summary column. Transform the new data features to fit the already

	Caught in/between Objects	Collapse of object	Drowning	Electrocution	Exposure to extreme temperatures	Falls	Fires and Explosion	Other	Struck By Moving Objects	Suffocation	TEST DATA	nan
Caught in/between Objects	<5>	1	.	.	.	.	.	1	1	.	.	.
Collapse of object	.	<.>	.	.	.	.	.	1	.	.	.	.
Drowning	.	.	<.>	.	.	.	.	.	.	.	.	.
Electrocution	.	.	1	<.>	.	.	.	.	.	.	.	.
Exposure to extreme temperatures	.	.	.	.	<.>	.	.	.	.	.	.	.
Falls	1	.	2	1	1	<13>	.	.	.	.	.	.
Fires and Explosion	.	.	.	.	.	.	<1>	.	.	.	.	.
Other	.	.	.	.	.	.	.	<.>	.	.	.	.
Struck By Moving Objects	4	2	1	.	.	3	.	1	<6>	.	.	.
Suffocation	.	.	.	.	.	.	.	.	.	<1>	.	.
TEST DATA	.	.	.	.	.	.	.	.	.	.	<.>	.
nan	.	.	.	.	.	.	.	.	.	.	1	<.>

(row = reference; col = test)

Figure 1: *Confusion Matrix*

	precision	recall	f1-score	support
Caught in/between Objects	0.50	0.62	0.56	8
Collapse of object	0.00	0.00	0.00	1
Drowning	0.00	0.00	0.00	0
Electrocution	0.00	0.00	0.00	1
Exposure to extreme temperatures	0.00	0.00	0.00	0
Falls	0.81	0.72	0.76	18
Fires and Explosion	1.00	1.00	1.00	1
Other	0.00	0.00	0.00	0
Struck By Moving Objects	0.86	0.35	0.50	17
Suffocation	1.00	1.00	1.00	1
TEST DATA	0.00	0.00	0.00	0
nan	0.00	0.00	0.00	1
avg / total	0.73	0.54	0.60	48

Figure 2: *Classification Report*

trained TF-IDF. Then use the trained model on the new data. A part of the classification results are shown as follows.

Index	Type	Size	Value
0	string_	1	Struck By Moving Objects
1	string_	1	Falls
2	string_	1	Struck By Moving Objects
3	string_	1	Caught in/between Objects
4	string_	1	Falls
5	string_	1	Falls
6	string_	1	Falls
7	string_	1	Struck By Moving Objects
8	string_	1	Caught in/between Objects
9	string_	1	Falls
10	string_	1	Falls
11	string_	1	Falls
12	string_	1	Struck By Moving Objects
13	string_	1	Struck By Moving Objects
14	string_	1	Falls
15	string_	1	Caught in/between Objects
16	string_	1	Falls
17	string_	1	Struck By Moving Objects
18	string_	1	Falls
19	string_	1	Struck By Moving Objects

Figure 3: *Classification Result*

2 Objects of Accident

In this task, we used some different tools in nltk to help us find out the statistics of different objects causing an accident.

The approach we use to tackle this problem is to train a regular expression chunker which extracts the objects that cause the accidents.

2.1 Data Preprocessing

After examine the text data in Title Case column, we found that, there is certain syntactic structure of the text, i.e. the target objects which cause the accident appear after a proposition or to. E.g. Employee Injures Self With Knife, the object causes the accident is knife, and it appears after a proposition with. Therefore, we can use this pattern to extract the target objects.

Also there are a lot of error messages ' InspectionOpen DateSICEstablishment Name', we eliminated these messages with some simple script.

We found that the first character of each word in the sentence is capitalized and it causes problems when we try to POS tag the text. So normalizing each word to small letter is required.

2.2 POS Tagging

After data exploration and preprocessing, we POS tag each record of Title Case to get the POS tags so that we can use regular expression to extract the target objects later.

After data exploration and preprocessing, we POS tag each record of Title Case to get the POS tags so that we can use regular expression to extract the target objects later. Below is an example of POS tagged record, as can be seen, the target object forklift (NN) is after by (IN).

2.3 Regular Expression

After POS tagging, we build a regular expression-based chunk parser, for each record of Title Case, we extract the target object.

We calculate the frequency of each object and list the top objects (as shown below) with the highest frequency.

From the result, we can see that the most frequent object is fall which actually is not an object, so we sort the top 11 from the result, and got the top 10 objects and their frequency respectively.

```
('fall', 1825)
('ladder', 295)
('roof', 220)
('truck', 165)
('explosion', 160)
('machine', 155)
('forklift', 140)
('scaffold', 84)
('tree', 82)
('vehicle', 80)
('flash fire', 76)
```

Figure 4: *Top Objects*

The final top 10 objects and there frequencies are

ladder, roof, truck, explosion, machine, forklift, scaffold, tree, vehicle, flash fire.

In order to provide a more intuitive way to view our result we made a wordnet, shown as Figure. 5.

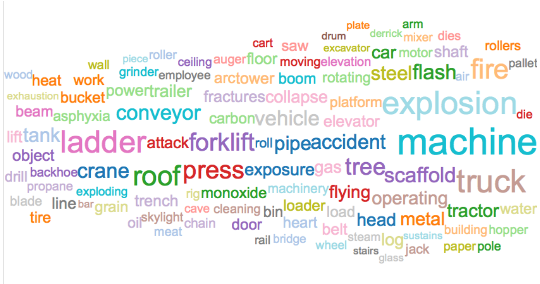

Figure 5: *Common Objects That Causes Accidents*

Due to the complexity of the syntactic structure, one simple regular expression cant identify all target objects in records, as we can see from the result, the top one object found by the chunk parser is fall which appears 1825 times, however, fall may or may not be considered as an object which causes the accident, so manual examination of the result is quite necessary. To further improve the accuracy, we can analyze the syntactic structure of each record more carefully and come up with more sophisticated regular expression or try more complex parser, e.g. dependency parser, etc.

3 Occupations

Finding occupations that are more risky contains work of finding highly frequently appearing subject of all accident descriptions.

3.1 Data Preprocessing

We made some modifications in the two tables. For this task we mainly used osha.csv. Some of the types are not consistent with the titles and summary contents. We correct them according to the title and summary content. Besides there are other and others in the type list. We change them to other to make it uniform.

3.2 POS Tagging and Results

For this task the POS tagging is almost the same with the one in Chapter. 2. Here we reduce the process of explaining: when extracting subject of an accident, the subject(noun) will be considered as the occupation.

After we run the test we got the following occupations as the most frequently involved in accidents:

$$Employee, Worker, Operator, Carpenter,$$

$$Driver, Mechanic, Installer,$$

$$Foreman, Zookeeper$$

There might be inaccuracy with the result, further research will cover how to increase the accuracy.

4 Common Activities

For finding the common activities we are using the summary column in osha dataset. Which provides detailed information about the incident which caused the accident, including what happened prior to the accident.

4.1 Data Preprocessing

As we know the activities are mentioned in the first sentence of every summary case, we are going to limit the number of lines to the first, also it will help us to reduce the complexity involved in computation and POS tagging. Furthermore, we can avoid dealing with secondary or detailed information is given later sentences. We also removed the empty rows as they wont be contributing towards the results.

4.2 POS Tagging and Chunking

To get the part of sentence which involves the user activity we are going to perform chunking, Grammar for chunking the activity, and represented in past continuous tense, and ends with noun phrase.

n our chunk there are many english stopwords which are causing the activities to be non consistent, we will be removing all english stopwords.

For indicating the activity we will only look into the sentences which have words with *.ing. This gives us the chunks which contain actions along with the things which they were envolved with

4.3 Finding Activities

In the end we will be only extracting the actions the persons were involved in by using POS tagging and grammar as

$$grammar = r"""< VBG|JJ > +"""$$

This gives us all the actions the they performed.

The following word nets are to show the word frequency:

nde ▲	Type	Size	Value
0	tuple	2	('working laborer', 46)
1	tuple	2	('performing maintenance', 44)
2	tuple	2	('working firm', 37)
3	tuple	2	('working coworker', 30)
4	tuple	2	('working company', 30)
5	tuple	2	('working roof', 27)
6	tuple	2	('operating forklift', 25)
7	tuple	2	('assisting coworker', 22)
8	tuple	2	('working inside', 22)
9	tuple	2	('working construction site', 22)
10	tuple	2	('working machine operator', 21)
11	tuple	2	('using forklift', 19)
12	tuple	2	('working part', 18)
13	tuple	2	('working employer', 18)
14	tuple	2	('working facility', 17)
15	tuple	2	('clearing jam', 15)

nde ▲	Type	Size	Value
0	tuple	2	("'working", 2315)
1	tuple	2	("'operating", 727)
2	tuple	2	("'using", 424)
3	tuple	2	("'installing", 287)
4	tuple	2	("'performing", 251)
5	tuple	2	("'cleaning", 223)
6	tuple	2	("'removing", 212)
7	tuple	2	("'standing", 137)
8	tuple	2	("'driving", 130)
9	tuple	2	("'assisting", 110)
10	tuple	2	("'replacing", 93)
11	tuple	2	("'walking", 93)
12	tuple	2	("'cutting", 90)
13	tuple	2	("'moving", 87)
14	tuple	2	("'repairing", 85)
15	tuple	2	("'unloading", 72)

Figure 6: *(Left)Common Action Involved Together with Relavent People. (Right)Common Actions before Accidents.*

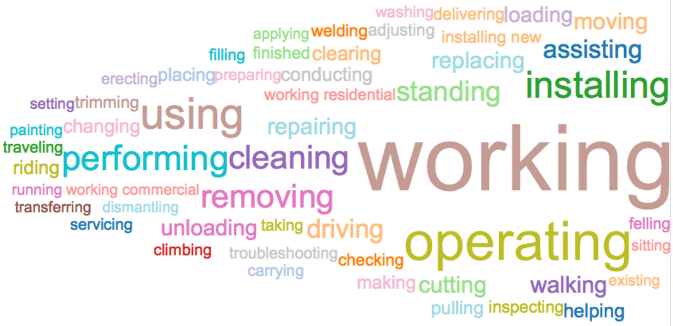

Figure 7: *Common Activities before Accident*

Figure 8: *Common Activities*

5 Conclusion

Use SVM algorithm to build classification model to find out type of accidents (in terms of main causes) which are more common in fatal or catastrophic accidents of the osha data.

The common activities in which person was involved could be found. Also we can get the things involved and other culprit for the accidents by the chunking.

However, even though we found common actions, but they are not summarized very well. the number of actions in which we can clearly pinpoint what person was doing, but we fail to get comparable numbers when we are trying to pinpoint the.

While capturing only if we asked the operator to put a generic action, which could be a list of 30-50 activities which are predefined and then we can also further ask the subcategory. This will resolve the issue of inconsistent text values.

We can import the results provided here into visualization or other analytical tool to further analyse the issues. Current results can also be used as things during which worker should be careful with and we can also recommend the precautions to be taken.

Try with more than one sentence from summary case to see the results. Also try using the n-gram approach while performing the chunking.

References

[1] NLTK Project. Natural language processing toolkit, 2015. [Online; accessed 24-Oct-2015].

[2] Wikipedia. Part of speech tagging, 2015. [Online; accessed 19-Oct-2015].

[3] Wikipedia. Support vector machine, 2015. [Online; accessed 2-Nov-2015].