TEXT MINING

Continuous Assessment Report

Submitted To

Dr. Mun Kew

Ms. Fan Zhenzhen

Submitted By

Team

Abhinaya Hari	A0134585Y
Ashutosh Gaur	A0134613N
Karuna Kemmu	A0136595U
Kedar Khairnar	A0134584B
Vadivel D (TL)	A0134451N
Vidyut Singhania	A0136039H

Table of Contents

1. Executive Summary

1.1 Background

Real Estate development, or property development, is a multifaceted business, encompassing activities that range from the renovation and re-lease of existing buildings to the purchase of raw land and the sale of improved land. Despite improvement in recent years, the construction industry remains the top contributor for workplace fatalities. Each year, there are thousands of injuries and numbers of fatal accidents related to machine and equipment operation are used to happen on construction sites. A lot of these accidents involve the operator, but over half involve people on the ground - spotters, co-workers, laborers, shovel hands, passers-by and sidewalk superintendents who get too close. 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. Taking simple measures or adopting proper working procedures can prevent most accidents.

Various statutory bodies proposed to use leading indicators such as inspection findings, audit score and safety climate surveys to help construction-related organizations, e.g. large developers and contractors, forecast safety performance and improve safety risk controls proactively. 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.

1.2 Business Goals

The inspection report / feedback of any fatal or catastrophic accident happened at construction site can be analyzed using Text mining to identify safety indicators and appropriate measures to mitigate the identified risks and prevent the occurrence of similar accidents. Using Text mining technology, explore following: Types of accidents (main cause), kinds of objects cause the accidents, the more risky occupations in such accidents, the common activities that the victims were engaged in prior to the accident. Analyzing texts of accident reports can be instrumental in preventing the incidents at work place. By extracting useful information from the reports using text mining techniques and understanding what is in the report provides useful knowledge to take care of safety measures in construction sites.

Analyzing this result can improve an understanding of common accidents, objects associated and victim's profession etc. from different historical incidents thereby enabling the prevention of future accidents. This "Construction Accidents" Text Mining analysis relies on the amount of information available in the accident report summaries provided by Msia and OSHA datasets and text mining of such textual summaries will unveil the variables and their relationships that may not be evident through unstructured data.

1.3 Findings

Please find below summary after analyzing OSHA dataset using various Text Mining Tools and Techniques. Below figure contains Top Factors from each category.

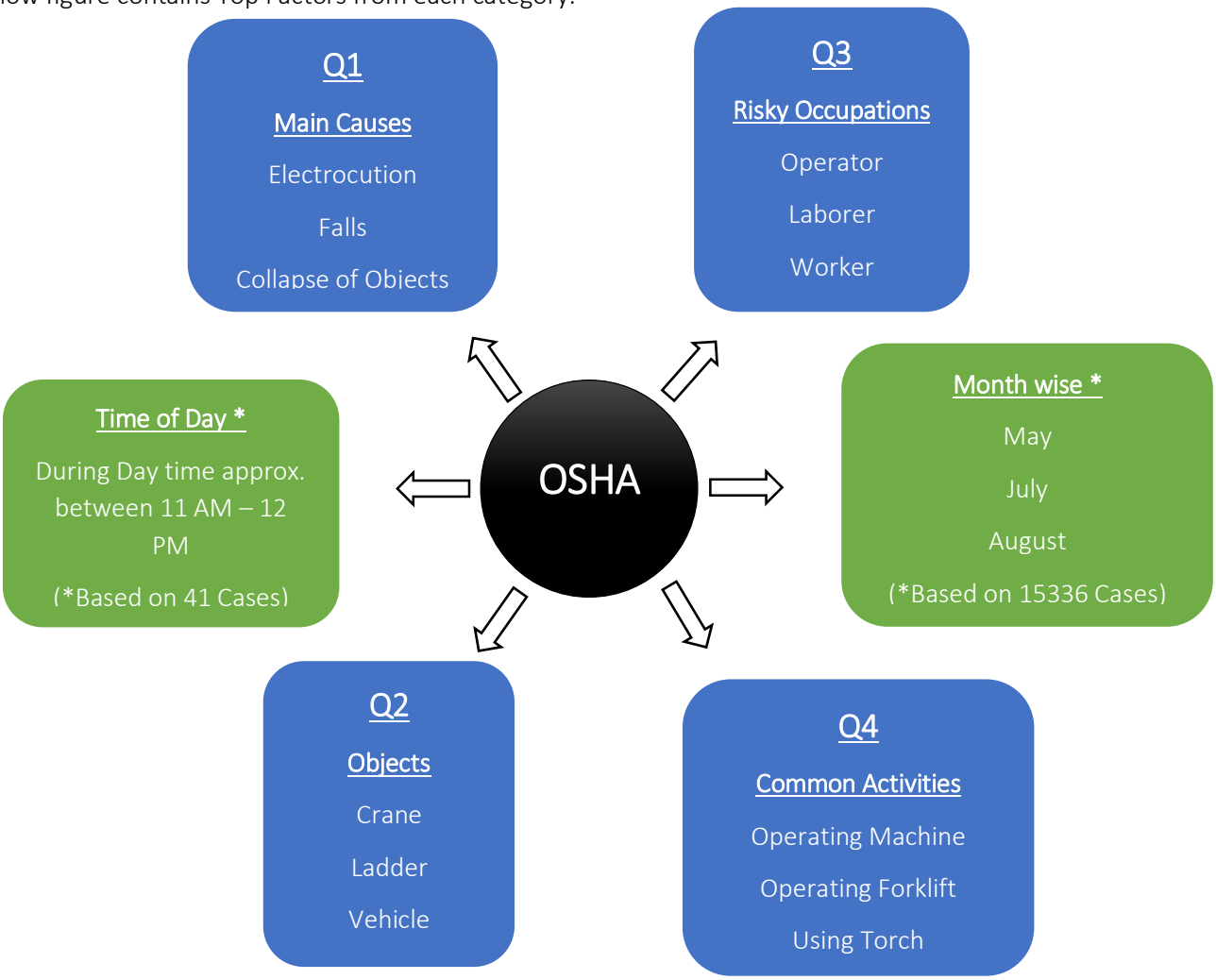

Figure 1: Text mining findings on OSHA dataset

2. Introduction

2.1 Purpose of Analysis

The main purpose of this analysis is to generate useful information from vast amount of Textual data available with us in the form of Msia and OSHA Datasets. By performing Text Mining on all these reports will show some association and causal relationship between different variables. Also, it will provide some information that is required in extracting interrelationship between large number of factors and sub factors in case of Construction related accidents.

2.2 Text Mining Goals

Text Mining is the process of analyzing text for the purpose of discovering and capturing semantic information with the ultimate goal of enabling knowledge discovery via either textual or visual access for use in a wide range of significant applications. Text Mining is the discovery of new, previously unknown information, by automatically extracting information from different written resources. In text mining the patterns are extracted from natural language text to generate new insights. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as relations between different factors, emotion, intensity and relevance. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents and reports. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups. One of the goal of Text Mining is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data.

To summarize, Text mining goals are:

1. Analyzing vast amount of textual data
2. Determining relationships among data
3. Group a large number of text according to their similarity
4. Analyze document to enable business to query both objective and subjective
5. Supports strategic and competitive Intelligence.

3. Text Mining Steps

We have been given below 2 Datasets for analysis using Text Mining Techniques.

Dataset	Msia	OSHA
Number of Records	• Training Records – 182 • Test Records – 53 • Total - 235	Total - 16323
Labelled	Yes • Total 3 Columns • Cause/Summary/Description	No
Structured	Yes	No

Table 1: Properties of Msia and OSHA Datasets

We have done Exploratory Analysis on Msia Dataset to find Number of causes and frequency of the same.

Cause	Number of Cases(Before Manual Analysis)
Caught in/between Objects	36
Collapse of object	5
Drowning	8
Electrocution	17
Exposure to Chemical Substances	2
Exposure to extreme temperatures	2
Falls	56
Fires and Explosion	4
Other	9
Others	1
Struck By Moving Objects	39
Suffocation	3
Grand Total	**182**

Table 2: Distribution in AS IS Msia Dataset

Then, we have manually analyzed **Msia (Train + Test = 235 Cases)** Dataset and generated additional columns to get idea of the kind of data present in dataset and if all cases are correctly categorized under respective causes.

Below are number of new columns we have generated for all 235 cases from Msia before going ahead with any Text mining task.

- Confirm that the case belong to your category, if not create a column for Correct Category
- What caused the accident (e.g. “Hit”, “Collapsed”)?
- What Work were they doing before the accident happened?
- Was any Machine involved? Which one?
- What is the Occupation of the victim?
- Age of the victim if mentioned?
- Which object was involved?
- Describe any important Keyword
- Was the accident Fatal (e.g. Killed or Died) Or Catastrophic?

After manual analysis of Msia Dataset, we have tried to correctly categorize some of cases which were put under incorrect cause in old dataset.

Please find below classification table with correctly classified cases.

Cause	Number of cases(Corrected)
Caught in/between Objects	14
Collapse of object	24
Drowning	5
Electrocution	17
Exposure to Chemical Substances	1
Exposure to extreme temperatures	2
Falls	52
Fires and Explosion	4
Other (Other + Others)	18
Struck By Moving Objects	43
Suffocation	2
Grand Total	**182**

Table 3: Distribution in correctly classified Msia Dataset

3.1 Answers to all Questions

Q. 1. Which type of accidents (in terms of main causes) are more common in fatal or catastrophic Accidents?

Steps taken to find out main causes in accidents cases.

The answer to this question forms a major portion of the entire project as we had to determine the probable cause for each incident of a much larger data set (Osha) based on the on a relatively miniscule one (MsiaAccidentAccidentCase). Hence, this question involves the automatic categorization (classification) of documents (each incident of the given data set) into pre-determined classes.

Since we had to initially train the classifiers on only 182 (manually) cleaned incidents from the training set (Msia) and then use these to determine the cause for each of the 16323 incidents of OSHA, we realized that any number of well-trained classifiers would certainly end up performing extremely badly at the said task. Furthermore, on closer analysis we noticed that Msia is a cleaned and well-structured data set when compared to OSHA, which contained numerous cases where the text is either extremely unclean, improperly formatted, not structured well or simply just a bag of certain keywords strung together. So, we realized that we should use classifiers which are known to be adept at working with such noisy and unclean data. Hence, we decided to proceed with a multi-step ensemble of iterative classifiers which would make leverage on the power of both, Linguistics and Statistical techniques, for the given task of classification.

1. Supervised Classifiers trained only on Msia

Initially, we followed the traditional method of training the classifiers only on the given train set (Msia) and then using them to determine the possible cause(s) for each cases of Osha. Since we had 11 pre-determined classes / causes, we proceeded to build one binary classifier for each such class and run them separately on the given data sets. Thus, each such classifier would only indicate whether the document in question belonged to the particular class it would be testing for. A point worth noting is that the given test set in Msia (comprising 53 rows) only contained 8 of the 11 existing classes.

a) Support Vector Machine (SVM) Classifiers

Since SVMs are known to perform well in noisy environments (generically as well with modification) [1], we decided to use this classification technique for the current problem. Also, SVM can perform well in distinguishing elements even for relatively small data sets. The SVM classifier leverages on its algorithm's ability to represent each provided input document as a vector whose dimension is, roughly, the number of distinct words present in it. Further, since various cases in OSHA contain extremely long sentences as well, the dimension of the hyperspace wherein the actual text classification is performed becomes extremely large – resulting in significantly more computational resources being expended. Hence, we tried to determine the number of optimal features to use for the SVM classifiers in a manner that we get good results while not expending too much time and computing power for the process. We discovered that with k=10 features we were getting reasonably acceptable results, for the balanced data sets, which satisfied both the aforementioned constraints [2].

Using SVM, we were able to determine 3812 incidents in OSHA as having only a single class. The remaining cases' causes will be determine by the subsequent Verbal and Noun Classifiers.

b) Naïve Bayes Classifiers

It is a statistical supervised learning technique based on the assumption that there exists an underlying conditional probability distribution behind the words in the document. Apart from being an exceptionally well known algorithm for the domain of text document classification, we decided to pursue this technique as well since it enables us to capture uncertainty regarding the document class based on the probability of words present for each class.
We constructed 11 Naïve Bayes classifiers on balanced data sets for the classification purposes, each of which will determine whether the document can belong to the particular class or not. However, this methodology performed quite poorly on the Msia test set of 53 rows, when compared with the corresponding SVM classifiers. We can attribute this to the extremely small training data set provided to the Naïve Bayes classifiers.

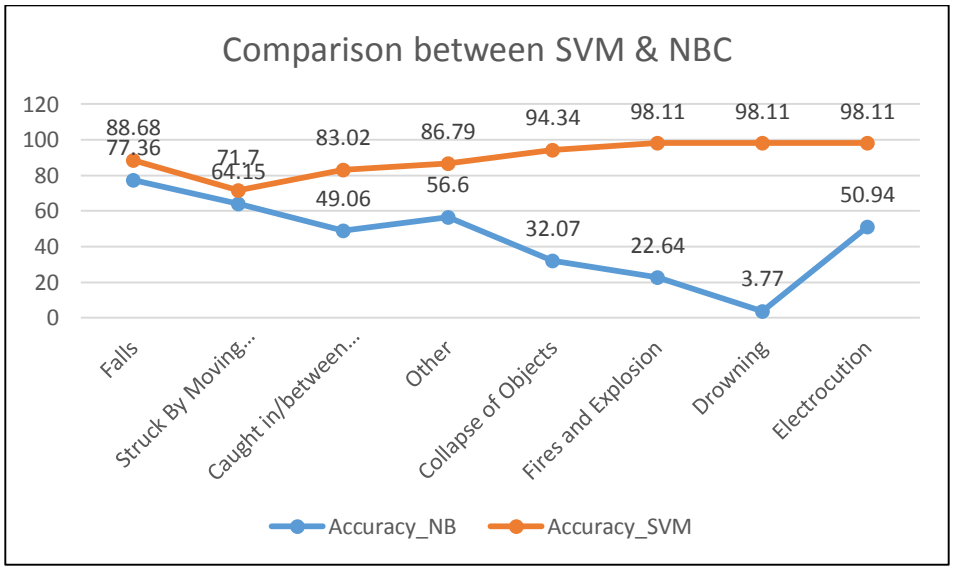

Figure 2: Comparison between SVM & NBC

Hence, we decided to proceed for the next part with only the SVM classifiers output.

Supervised Classifiers trained on Msia and OSHA (partial data set)

Based on the preceding part, we were able to determine the unique causes for only 3812 cases. Since, we must determine the unique causes for the remnant 12511 cases of OSHA, we decided to make use of certain linguistic features unique to each cause, such as particular verbs and nouns.

Verbal Classifiers

Since verbs are more probable to be unique for individual causes, we ran the verbal classifiers on the 12511 cases before the Noun classifier [this decision was validated later]. For this process, we trained the classifier on Msia as well as the 3812 cases of OSHA which were assigned individual cases by the SVM classifiers.

1) Extract all verbs present in sentences associated with each individual cause.

2) Determine frequency of occurrence of each such verb for the particular cause.
3) Find the top verbs, for each individual cause.
4) Further determine distinctly unique verbs associated with each cause.
5) Noticed that there were a lot of verbs which did not seem to make much sense when associated with the given cause (e.g. being, died, et al).
6) Finally create a list of such verbs for each cause such that if the word is present in a sentence, the sentence is likely to be of a particular class.

We ran the 11 verbal classifiers (each classifier for an individual cause represented the list of verbs which would generically be associated with the cause) on the 12511 cases which were either classified as having no class or multiple classes by all the SVM classifiers.

As a result of this, we were able to further classify 5250 more cases from OSHA as having a single unique cause. However, we are still left with 5627 cases which have multiple causes and 1634 cases with no cause assigned.

Noun Classifiers

We will now run the Noun Classifiers only on these remaining 7261 cases in order to determine their unique causes. We follow the same process as mentioned above for Verbal Classifiers, except that for Noun Classifiers, we search for and use nouns instead of verbs.

We first had to train the 11 Noun Classifiers on Msia and the (3812+5250) cases of OSHA classified by SVM and Verbs till now.

Since, we have already partially leveraged on the power of linguistics (verbal classifier) for classification, we end up finding only 890 more cases of Osha having single unique causes. For the remnant 6371 cases with multiple or no causes, we determined the unique cause by the following process:

1) Determine which class has maximum nouns (from the associated classifier list) present in the given sentence.
2) Assign the unique cause for the case to be the class with the maximum nouns present.

Files used:

Classifier.py

Q.2.What kinds of Objects cause the accidents?

- Given that there are 11 known causes of causing accidents at construction sites, it is very important to have an idea about what kind of objects are causing the accidents.
- Knowledge of such objects will make the workers take extra preventive measures to ensure those objects do not cause any serious harm.

Steps taken to find out Objects involved in such accidents cases.

1. Data is given as “Title case” and “Summary Case”
2. Title case is a headline briefly describing the nature of the freak accident
3. Summary case is the paragraph description which has the details about the accident

4. Title case texts (one liners) have been used to extract the objects (data = Title case)
5. NLTK Tokenizer was used to tokenize the data
6. All the text in title case after the word “By”, “With” and “Between” has been extracted.
7. POS tagging using NLTK was done for all the text obtained from the last step.
8. All the nouns among these extracted words form the list of objects that have caused the accidents.
9. Word cloud with the extracted text – gives an idea about the kind of words that were extracted to view objects

Files Used:

Osha_objectScript.py

Main_object.py

Wordcloud_object.py

Q. 3. What are the more risky occupations in such accidents?

Steps taken to extract Risky Occupations

1. Extracted the Summary from all the cases in ‘.txt’ files
2. Used Stanford core nlp for performing Natural language Processing:
   java -cp "*" -Xmx1024m edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators tokenize,ssplit,pos,lemma,ner,parse,dcoref
3. This performed all the following tasks and saved to xml file:
   - Tokenize
   - Sentence split
   - POS Tagger
   - Lemmatization
   - Named Entity Recognition
   - Parse Tree
   - Co-reference
4. From the xml files, extracted rules:
   a. For each word = ‘Employee’ which have POS as ‘NN’ or ‘NNP’,
      Extract the dependency – ‘dep’ which also has POS as ‘NN’ or ‘NNP’
   b. The dependency word is expected to be the occupation

Employee + (NN or NNP)

Dependence --> ‘dep’

```
    <Timex tid="t1" type="TIME">2011-06-16T06:30</Timex>
    <sentiment>Neutral</sentiment>
  </token>
  <token id="9">
    <word>Employee</word>
    <lemma>Employee</lemma>
    <CharacterOffsetBegin>46</CharacterOffsetBegin>
    <CharacterOffsetEnd>54</CharacterOffsetEnd>
    <POS>NNP</POS>
    <NER>O</NER>
    <Speaker>PER0</Speaker>
    <sentiment>Neutral</sentiment>
  </token>
  <token id="10">
    <word>#</word>
    <lemma>#</lemma>
    <CharacterOffsetBegin>55</CharacterOffsetBegin>
    <CharacterOffsetEnd>56</CharacterOffsetEnd>
    <POS>#</POS>
    <NER>MONEY</NER>
    <NormalizedNER>£1.0</NormalizedNER>
    <Speaker>PER0</Speaker>
    <sentiment>Neutral</sentiment>
  </token>
```

```
<token id="13">
  <word>driver/helper</word>
  <lemma>driver/helper</lemma>
  <CharacterOffsetBegin>61</CharacterOffsetBegin>
  <CharacterOffsetEnd>74</CharacterOffsetEnd>
  <POS>NN</POS>
  <NER>O</NER>
  <Speaker>PER0</Speaker>
  <sentiment>Neutral</sentiment>
</token>
```

```
<dep type="dep">
  <governor idx="9">Employee</governor>
  <dependent idx="11">1</dependent>
</dep>
<dep type="det">
  <governor idx="13">driver/helper</governor>
  <dependent idx="12">a</dependent>
</dep>
<dep type="dep">
  <governor idx="9">Employee</governor>
  <dependent idx="13">driver/helper</dependent>
</dep>
<dep type="case">
  <governor idx="17">Fabricators</governor>
  <dependent idx="14">with</dependent>
```

5. From the occupation list, removed the non-relevant bigrams containing ‘male’ or ‘female’
6. Created a word cloud for frequency for the remaining words

Files used:

main_occupation.py

occupation.py

wordcloud_occupation.py

Q. 4. What are the common activities that the victims were engaged in prior to the accident?

Steps taken to extract common Activities victims were doing prior to the accident.

1. Extracted the Summary from all the cases in '.txt' files
2. Used Stanford core nlp for performing Natural language Processing:java -cp "*" -Xmx1024m edu.stanford.nlp.pipeline.StanfordCoreNLP –annotators tokenize,ssplit,pos,lemma,ner,parse,dcoref
3. This performed all the following tasks and saved to xml file:
    - Tokenize
    - Sentence split
    - POS Tagger
    - Lemmatization
    - Named Entity Recognition
    - Parse Tree
    - Co-reference

4. From the xml files, extracted rules:

If POS = 'VBG'(*Verb, gerund or present participle*), then extract the dependency – dobj (*direct object*)

From the list of extracted VBG + dobj combination, removed the non-relevant bigrams containing 'him' or 'employee'

VBG + Dependence --> 'dobj'

```
<token id="19">
  <word>operating</word>
  <lemma>operate</lemma>
  <CharacterOffsetBegin>107</CharacterOffsetBegin>
  <CharacterOffsetEnd>116</CharacterOffsetEnd>
  <POS>VBG</POS>
  <NER>O</NER>
  <Speaker>PER0</Speaker>
</token>
<token id="20">
  <word>a</word>
  <lemma>a</lemma>
  <CharacterOffsetBegin>117</CharacterOffsetBegin>
  <CharacterOffsetEnd>118</CharacterOffsetEnd>
  <POS>DT</POS>
  <NER>O</NER>
  <Speaker>PER0</Speaker>
</token>
```

```
<dep type="det">
  <governor idx="21">forklift</governor>
  <dependent idx="20">a</dependent>
</dep>
<dep type="dobj">
  <governor idx="19">operating</governor>
  <dependent idx="21">forklift</dependent>
</dep>
<dep type="compound">
  <governor idx="25">Truck</governor>
  <dependent idx="23">Linde</dependent>
</dep>
<dep type="compound">
  <governor idx="25">Truck</governor>
  <dependent idx="24">Lift</dependent>
</dep>
```

5. Created a word cloud for frequency for the remaining words

Files used:

main_activity.py

activities.py

wordcloud_activity.py

4. Tools and Libraries used

Software used:

- Excel
- Python
- Excel Text Mining Add in
- Tableau

Important Python Libraries used:

- Lxml,
- etree
- pandas
- nltk
- pytagcloud
- sklearn
- collections
- csv
- numpy
- re

5. Conclusion

- We have done Text Mining on OSHA dataset using Excel Text Mining Add in and extracted below important insights from all maximum possible cases from **OSHA** dataset. Below are some of the fields generated for Text analysis.

Text	Topic Form	Topic Categor	Ra	Type	Theme	Frequenc	Mentions
On April 11 2013	avalanche	concept	1	Top>Event>NaturalDisaster	Top>NaturalSciences	2	avalanche, avalanche

Figure 3: Text analysis using Excel Text Mining Add in

- All results in graphical format are mentioned below along with tools used and dataset used for the same.
- Then, we went ahead with **answering all 4 questions** using different Text Mining tools and techniques. You can find all infographics created from extracted information from uncleaned **OSHA** Dataset.

All infographics created below from extracted valuable information answers main 4 important questions as well as some of the hidden insights from very large OSHA dataset.

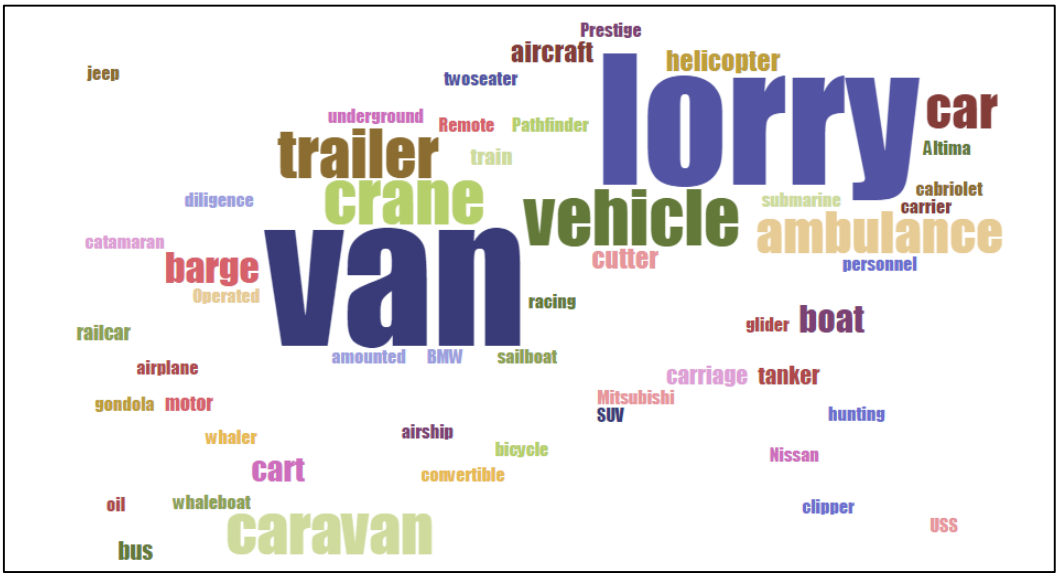

Figure 4: Type of Vehicle involved in number of cases
Tool used: Excel Text Mining Add in
Dataset: OSHA

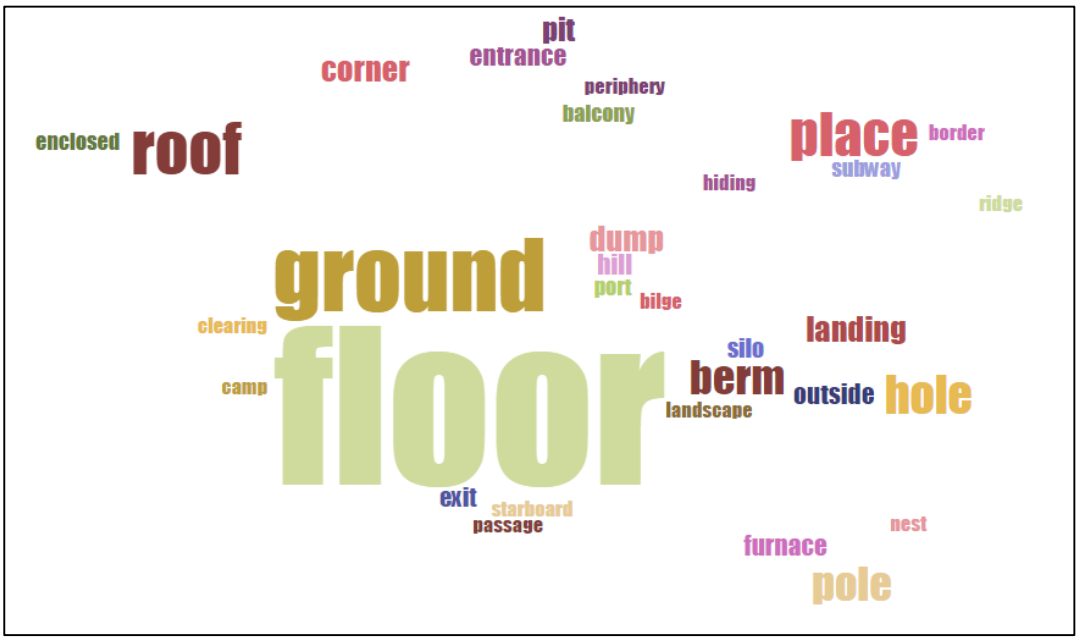

Figure 5: Locations where incidents occurred
Tool used: Excel Text Mining Add in
Dataset: OSHA

Figure 6: Body part injured
Tool used: Excel Text Mining Add in
Dataset: OSHA

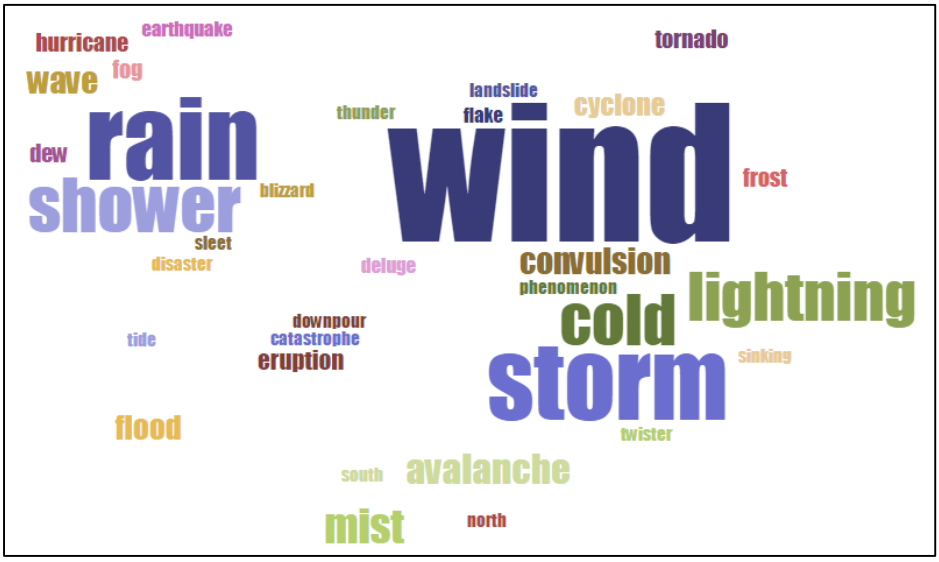

Figure 7: Weather conditions
Tool used: Excel Text Mining Add in
Dataset: OSHA (Out of 16323 cases, 631 cases have mentioned Natural disaster as Topic)

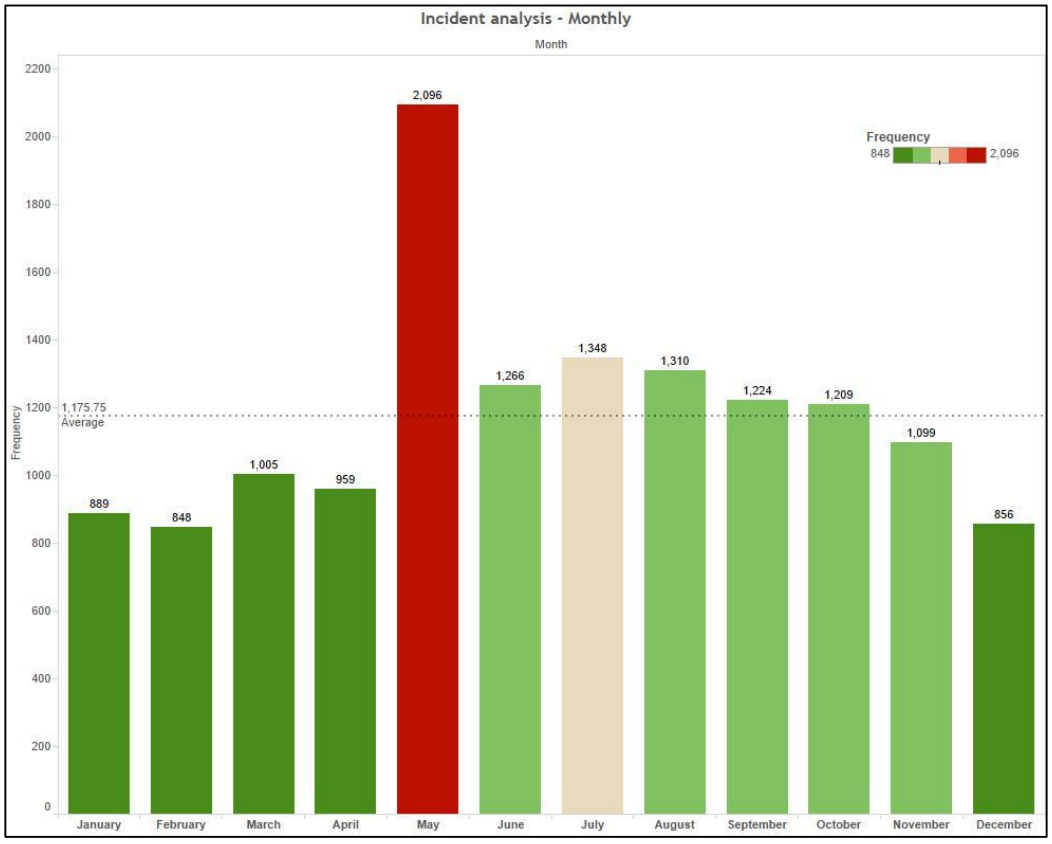

Figure 8: Monthly Incident Analysis
Tool used: Tableau / Excel Text Mining Add in
Dataset: OSHA

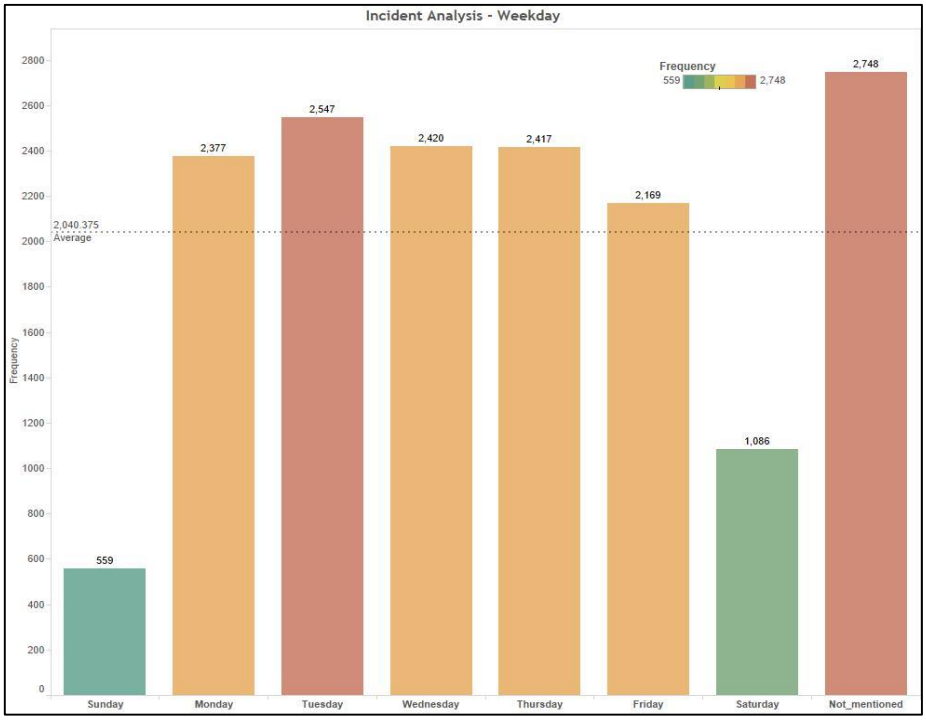

Figure 9: Weekday Analysis
Tool used: Tableau / Excel Text Mining Add in
Dataset: OSHA

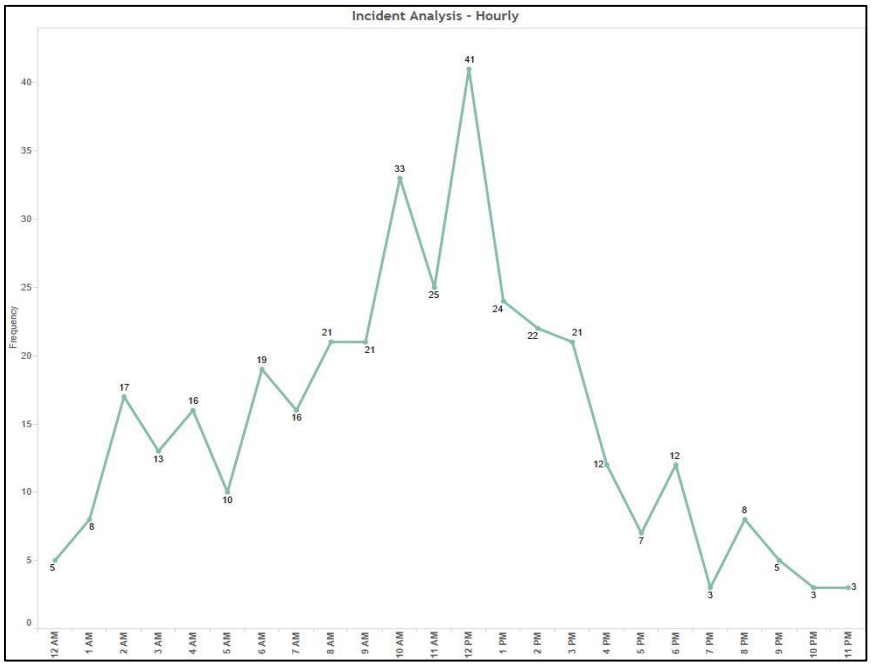

Figure 10: Time of Day Analysis
Tool used: Tableau / Excel Text Mining Add in
Dataset: OSHA (Out of 16323, 300 cases were having Time mentioned in summary)

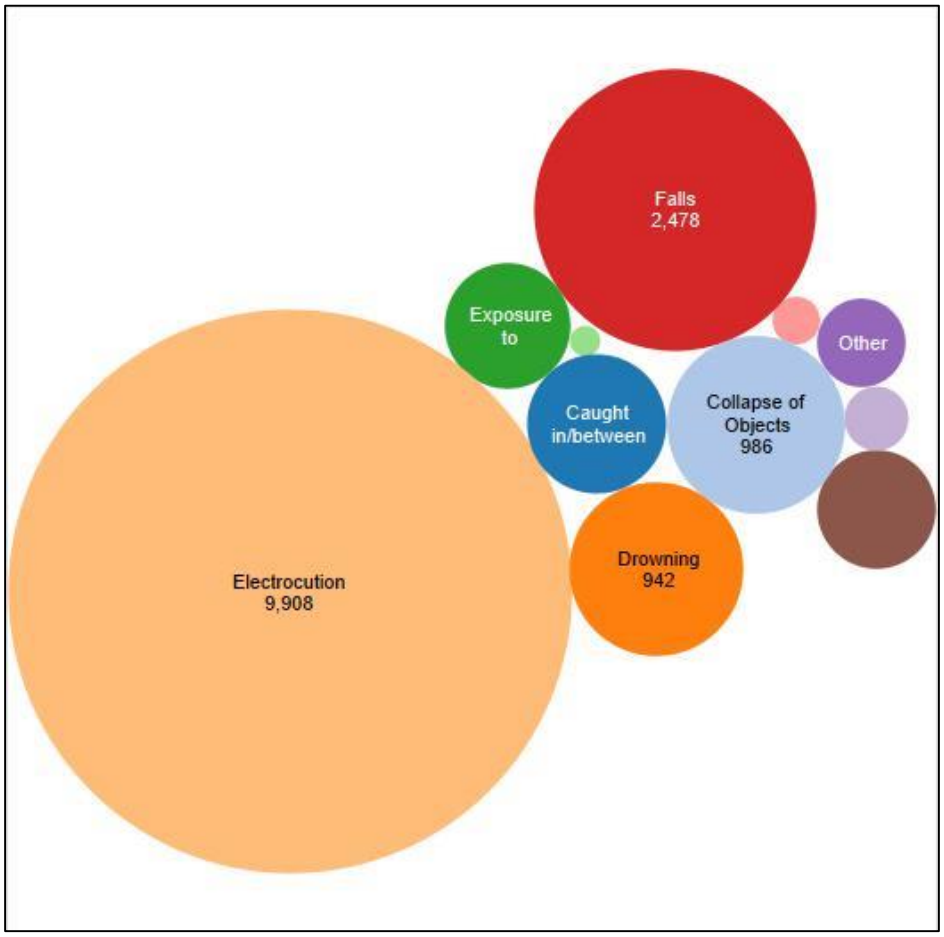

Figure 11: Q.1 Top Causes
Tool used: Python / Tableau
Dataset: OSHA

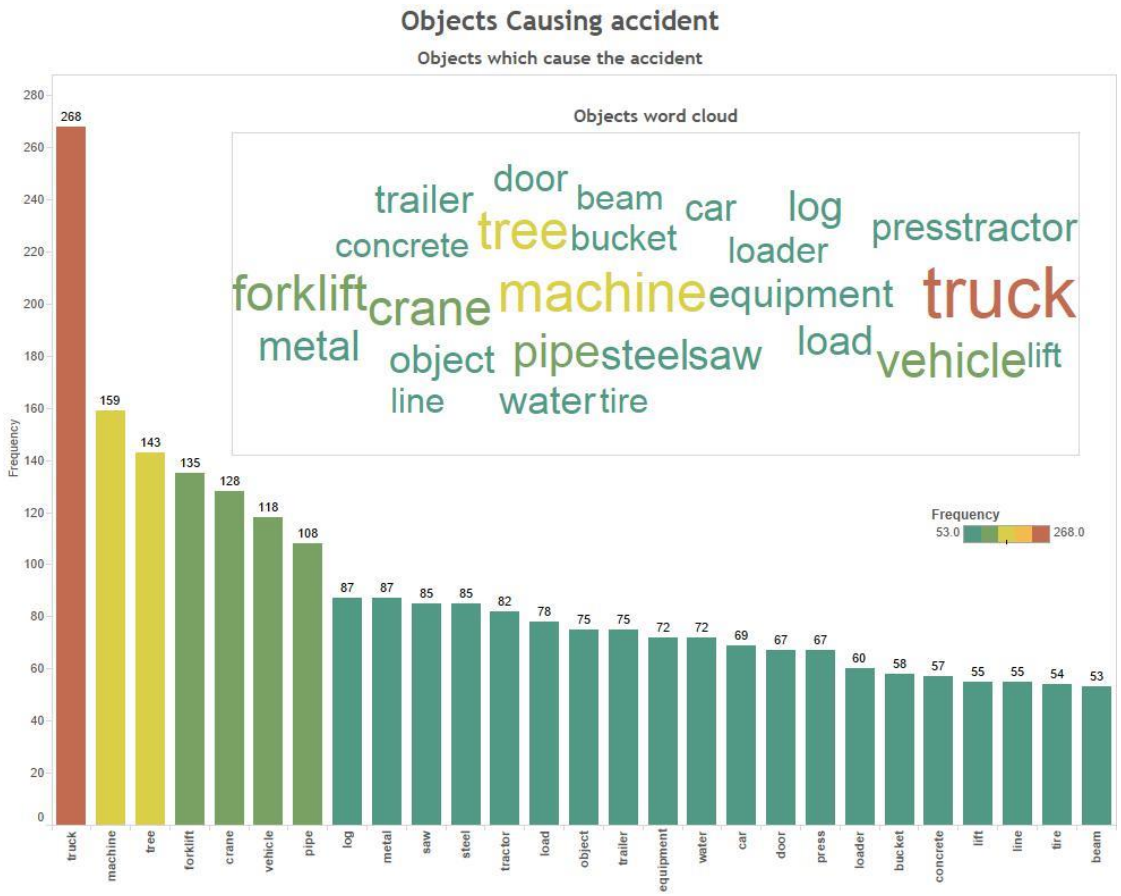

Figure 12: Q.2 Objects caused the accidents

Tool used: Python / Tableau

Dataset: OSHA

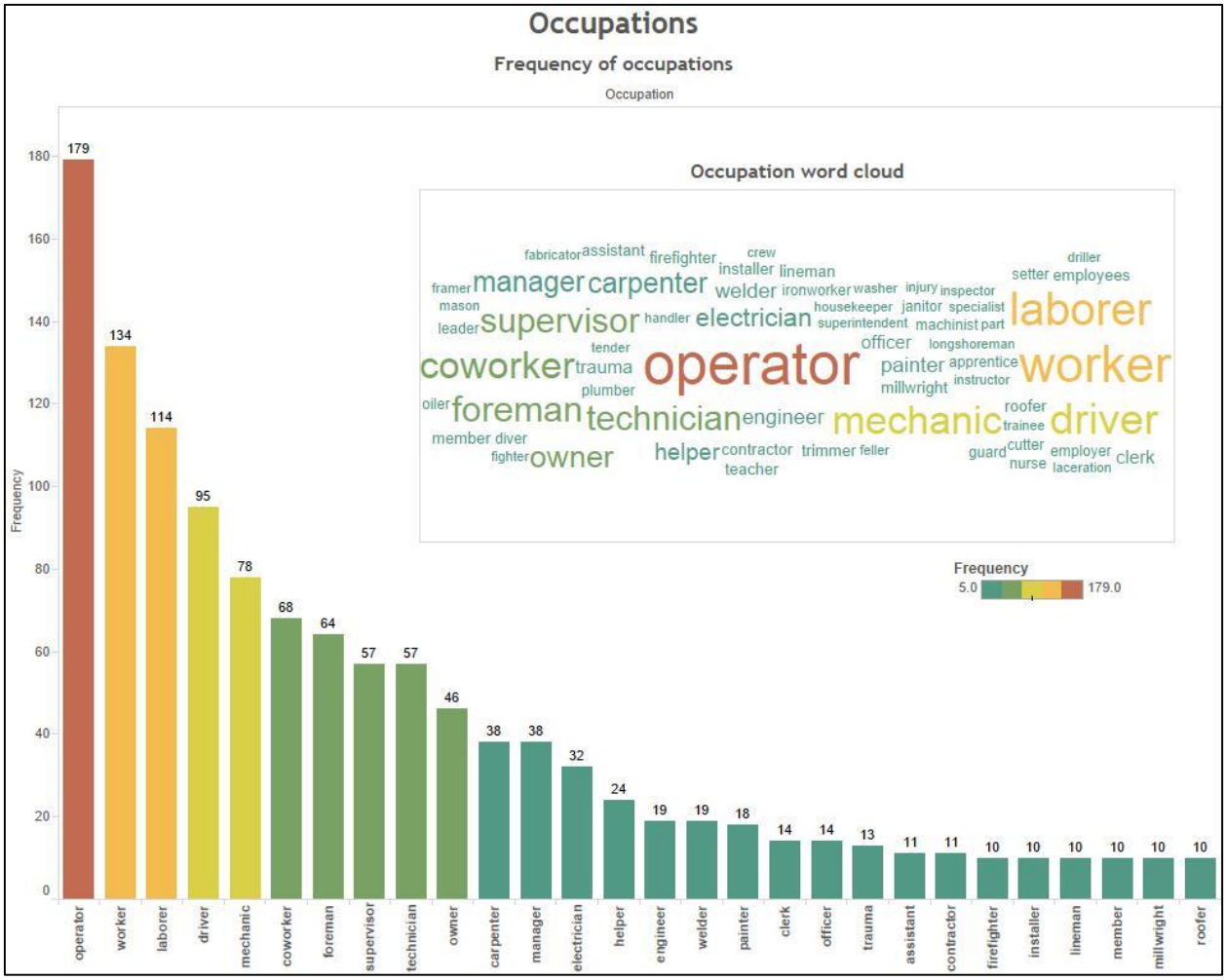

Figure 13: Q.3 Risky Occupations

Tool used: Python / Tableau

Dataset: OSHA

Figure 14: Q.4 Activities before Accidents
Tool used: Python
Dataset: OSHA

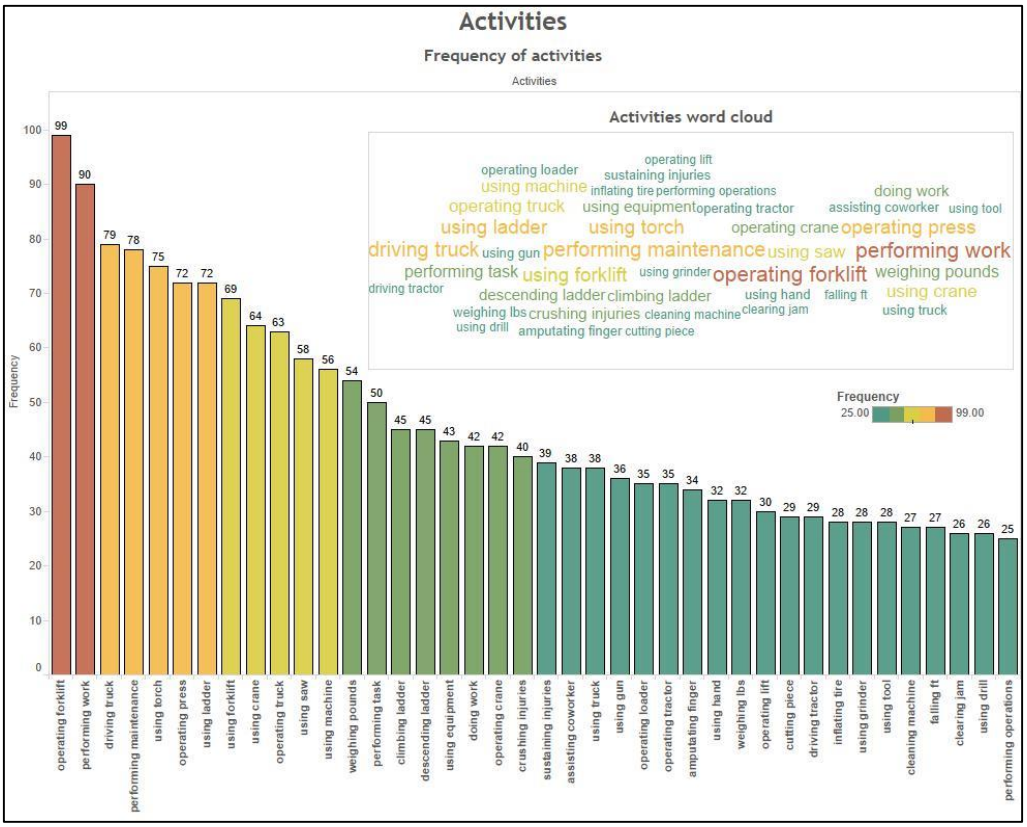

Figure 15: Q.4 Activities before Accidents
Tool used: Python / Tableau
Dataset: OSHA

References

1. Biggio B., Nelson B., Laskov P., Support Vector Machines under Adversarial Label Noise, 2011. http://www.jmlr.org/proceedings/papers/v20/biggio11/biggio11.pdf
2. Guduru N., Text Mining with Support Vector Machines and Non-Negative Matrix Factorization Algorithms, 2006. http://datamining.cs.uri.edu/theses/Thesis-Neelima-Guduru.pdf
3. Manning C., Raghavan P., Schütze H., Naïve Bayes Classification, Introduction to Information Retrieval.http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html
4. NLP: http://nlp.stanford.edu/software/dependencies_manual.pdf
5. POS Tag: http://cs.nyu.edu/grishman/jet/guide/PennPOS.html
6. NLTK: http://www.nltk.org/book/

List of Tables

List of Figures