MiniProject: Mining Accident Reports

## Project Objective

Employers are required to report any serious work-related injuries and death to the authority. This information helps employers, workers and the authority to evaluate the safety of a workplace, understand industry hazards, and implement worker protections to reduce and eliminate hazards.

In this mini-project, assume you are engaged by a client to perform text mining on the accident reports to help find answers to the following questions:

1. What are the major types of accidents reflected in the reports?
   * No labels, supervised or non-supervised?
   * Clustering or Topic modelling?
   * All data or partial data?
2. Which type of accidents are more common?
   * Frequency of doc wrt topic
3. Can we find out the more risky occupations in such accidents?
   * Information Extraction, how to identify “occupations” words?
4. Which part of the body is injured most? (Optional)
   * Information Extraction, how to identify “body” words?

The dataset is in file “osha.txt“.

## Data understanding and cleaning

Load the data file into R. – read.delim(), header=FALSE

e.g. textdata <- read.delim("osha.txt", header=FALSE, sep="\t", quote = "", stringsAsFactors = FALSE)

Explore your data.

* How many records do you have? How many variables?
* Examine the first few records in the datasets.
* What information does the dataset contain?
* Which fields are useful for your study?
* How long are the reports generally?
* How’s the data quality?
* What are the contents of the reports roughly? [ Create a word cloud for the dataset ]
  + Vectorsource, corpus, DTM
  + Term frequency summary
  + Wordcloud

**Data Exploration**

* How many records do you have? How many variables?
* What information does the dataset contain?

There are 10,000 records, with 3 variables – Report ID, Title of report, Report description (containing details of accidents – location, occupation and type of injury)

* Which fields are useful for your study?

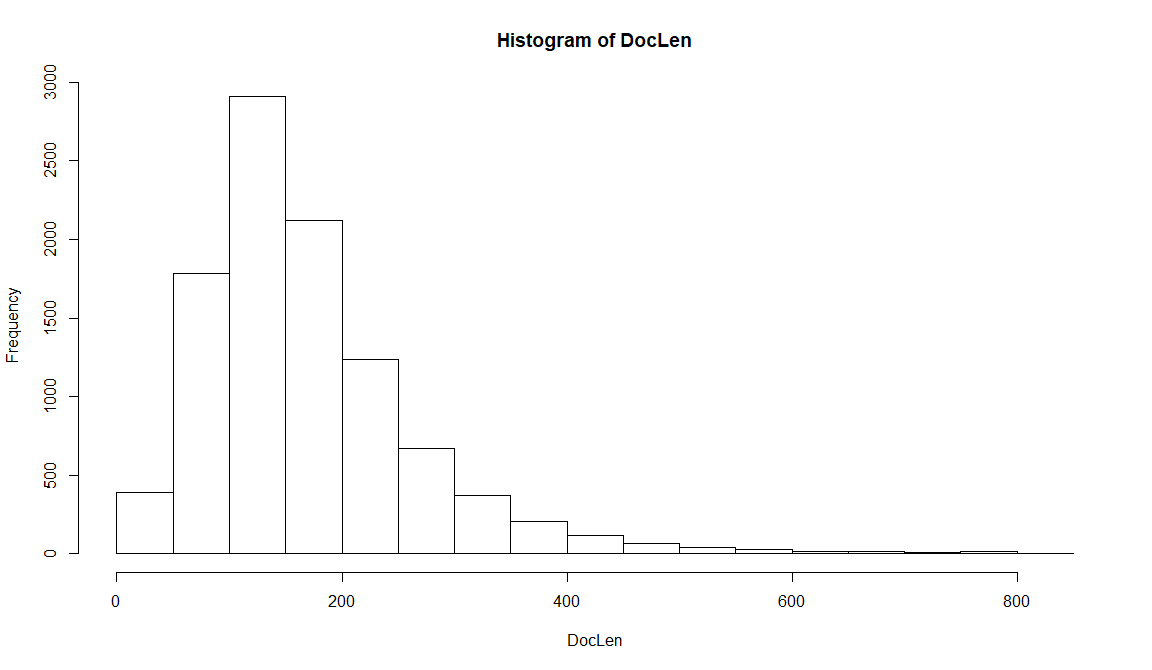
Report description is the most useful as it contains most information about the injury.

* How long are the reports generally?

Reports ranged from 3 words to 848 words, with median of 149 words.

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.0 107.0 149.0 170.1 210.0 848.0



* How’s the data quality?

The data quality is generally good, as words are spelt in full and do not seem to have any spelling mistakes. This will allow us to extract the information readily, and the count (from the DTM report) is reflective of the actual number of occurrences.

* What are the contents of the reports roughly? [ Create a word cloud for the dataset]
  + Vectorsource, corpus, DTM
  + Term frequency summary
  + Wordcloud

1. What are the major types of accidents reflected in the reports?
   * No labels, supervised or non-supervised?
   * Clustering or Topic modelling?
   * All data or partial data?
2. Which type of accidents are more common?

To find out the major types of accidents, there are several approaches. One way is to determine the accident based on the count of the type of accident. Another approach is to base it on the topics – different words that are associated with a particular accident (topic modelling). As there are no labels available, therefore non-supervised method of topic modelling was used. Topic modelling (as compared to clustering) was chosen so that we are able to sense the association of words in order to identify the type of injury.

**Common accidents based on topic modelling:**

Based on the type of vocabularies that are commonly found in each topic, burns seems to be the most common, followed by vehicular accidents and falls. However, this method is rather subjective due to the interpretation of association of words in order to suggest a conclusion for each topic (type of accident).

Topic Number 6 1 4 5 2 7 8 10 9 3

Freq (descending) 1513 1452 1157 1153 1146 903 739 707 672 558

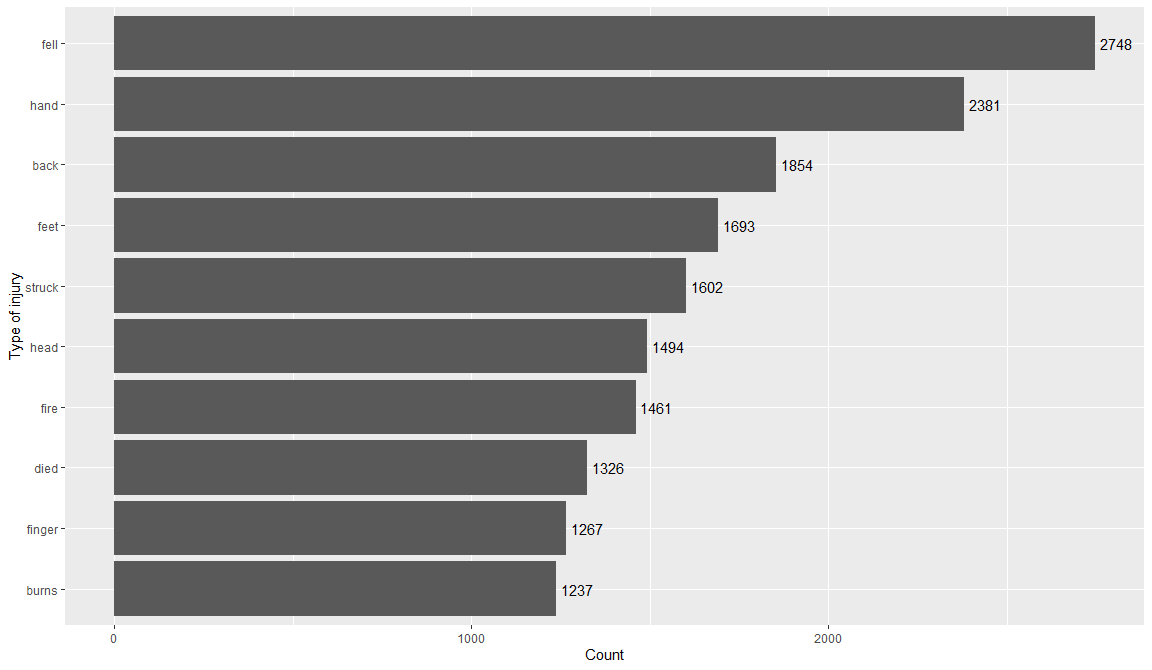


**Common accidents based on DTM (Global frequency):**

Falls are the most commonly cited accident, followed by struck and fire.

Hand, back and feet injuries are the most common.

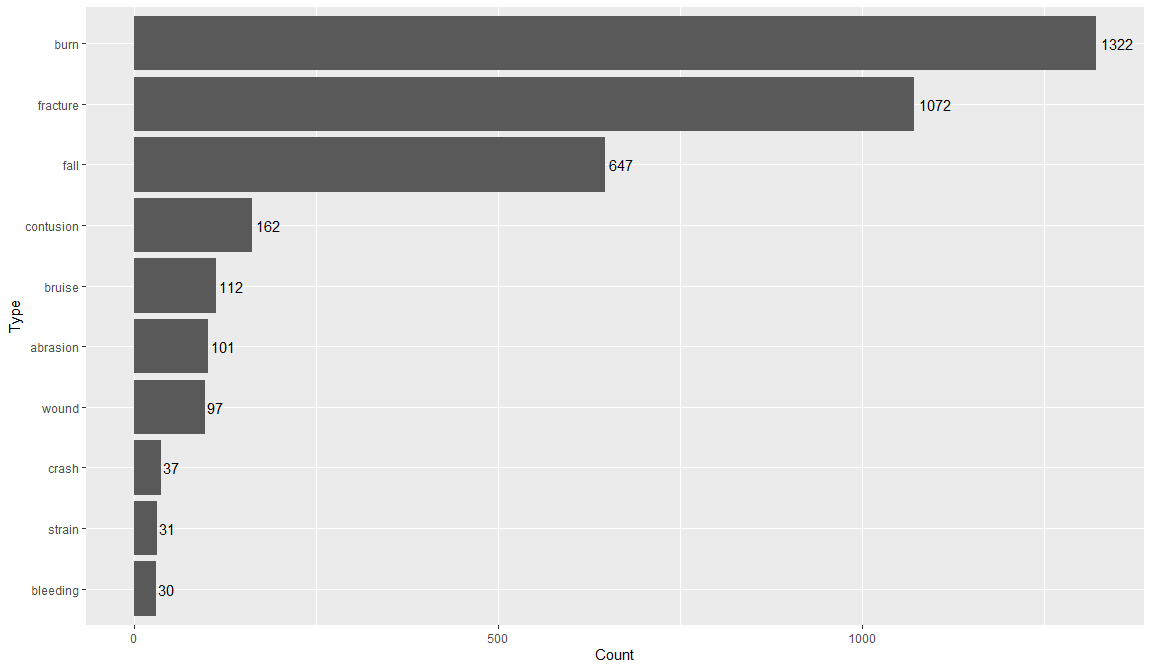
However, as the calculation was based on how many times the text appears in the report, this set of results may not be conclusive. Calculation based on whether the text appears once in the report would be a better indicator (See next graph on Binary indexing).

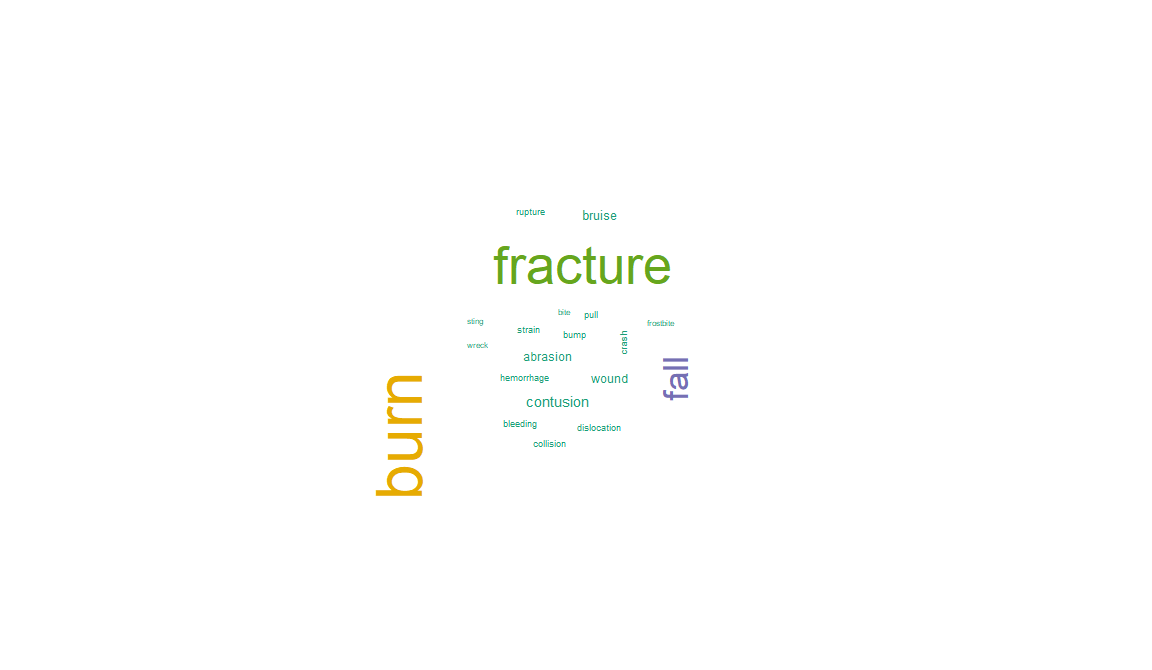


**Common accidents by DTM (Binary indexing):**

Burn is the most common accident, followed by fracture and fall.

This is by far the best method to identify common accidents as this was based whether the word appeared in the report. The words were matched against a common list of accidents (provided the list is comprehensive).

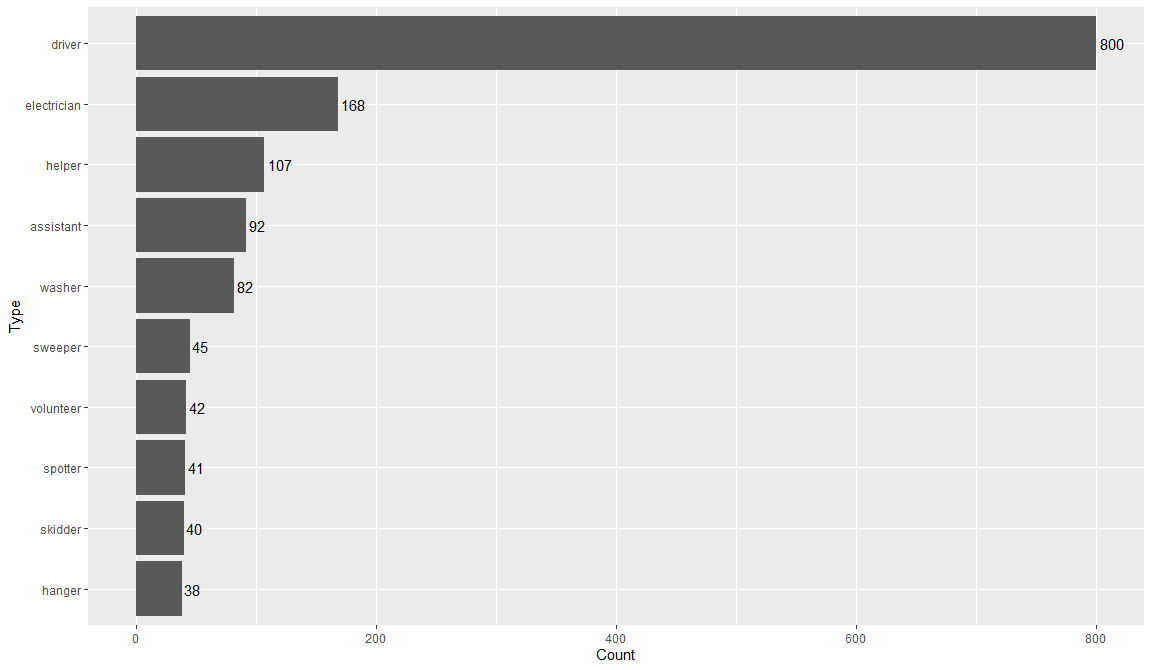


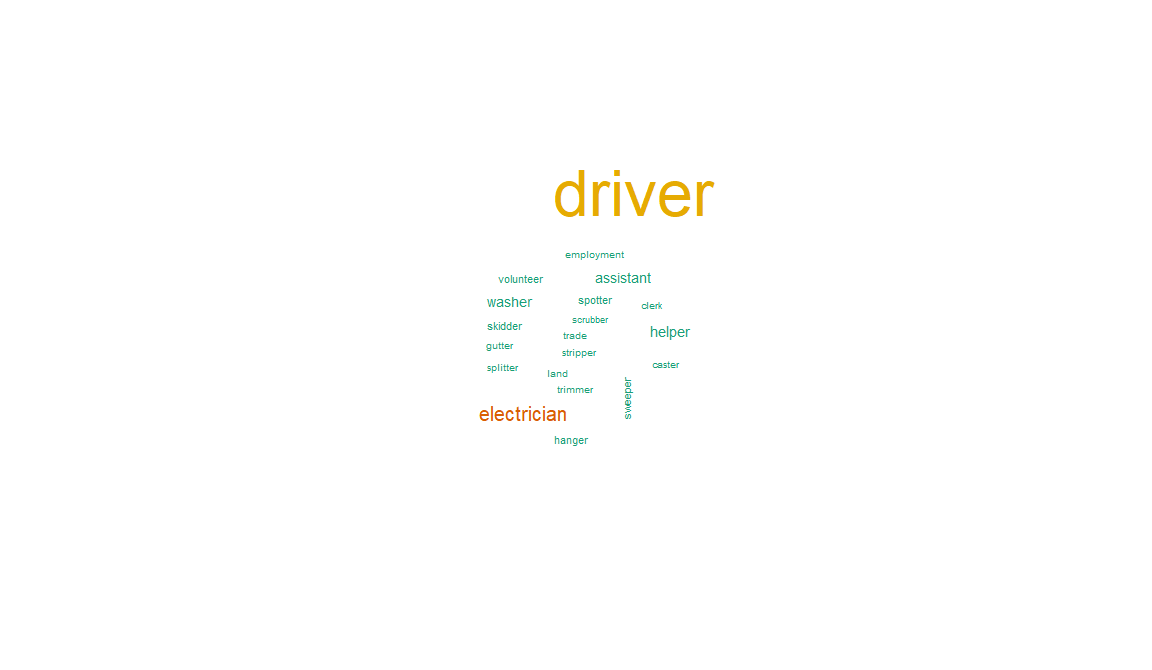


1. Can we find out the more risky occupations in such accidents?

**Risky occupations:**

Driver is the riskiest occupation, followed by electrician.





1. Which part of the body is injured most? (Optional)

**Body parts that are injured most:**

Foot injury is the most common, followed by head and arm.

