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