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.

An example of such accident report is shown below:

Employee Suffers Burns When Caught In Flash Fire

At approximately 9:30 a.m. on September 3 2013 Employee # 1 was heating hot asphalt oil spray wand using a propane torch. During the heating of spray wand a coworker attempted to reposition the wand and accidentally pressed the nozzle control handle that blow out hot oil and diesel. A flash fire occurred and caught Employee #1's clothing on fire. The coworker responded to the accident and drove the employee to an area hospital where he was treated for second degree burns to his neck and arms. Employee #1 remained hospitalized.

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?**
2. **Which type of accidents has the largest number of occurrences?**
3. **What are the more risky occupations involved in such accidents?**
4. **Which part of the human body is injured most?**

The dataset is in file “osha.txt“.

## Data understanding and exploration

Load the data file and explore your data.

* How many records are there? 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 long are the titles?
* How’s the data quality?
* What are the contents of the reports roughly? [ Hint: create a word frequency distribution for the dataset ]

[ Note: use your results to answer the questions in “Day 1 – Quiz 3”. ]

## Q1. What are the major types of accidents reflected in the reports?

* Are there labels available?
* Is the task supervised or unsupervised?
* Classification or clustering? Or topic modelling?
* Which column to use? Title or details?

Q1 answers:

- No

- unsupervised

- topic modelling

- title

## Q2. Which type of accidents has more occurrences?

* Once you have the model from Q1, it’s straight-forward to get the number of reports wrt topic

## Q3. Which part of the human body is injured most?

* Information Extraction: extract the mentioning of injured body parts in the text, then count (in what way?)
* Which column to use?
* How to identify “body part” words? : still need case lowering, lemmatization, pos
* Any ambiguities? (e.g. is “hand” always referring to hand?)

1. Preprocess report text: pos tagging, lemmatization => keep the nouns (lemma)

[ [ lem1, lem2, lem3, lem1,…] , [ ], [], ]

1. Get document frequency : turn the lists into sets: set(list\_1) , the FreqDist
2. Get the list of body terms from wordnet
3. Look up the FreqDist to get the document frequencies for body terms, sort, check the results!

## Q4. What are the more risky occupations involved in such accidents?

* Information Extraction, the mentioning of occupations in the text.
* How to identify “occupations” words?