Masters of Technology

Text Mining Project Instructions (2017)

**Project Introduction**

This project is for you to practice the concepts and techniques you’ve learned in Text Mining elective. It will contribute 30% to the final marks for this elective. You should form a team of up to 6 students to undertake this project and submit a project report by **Oct 30, 2017**. (See section Submission Details below)

# Project Objective

Despite improvement in recent years, the construction industry remains the top contributor for workplace fatalities in Singapore. Similarly, poor construction safety performance can be observed in other countries. 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.

Many have proposed the use of leading indicators such as inspection findings, audit score and safety climate surveys to help construction-related organisations, e.g. large developers and contractors, forecast safety performance and improve safety risk controls proactively. However, these leading indicators were not created using rigorous methods and were not known to be reliable.

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.

Now assume you are engaged by a client to perform text mining on Fatality and Catastrophe Investigation Summaries to help find answers to the following questions:

1. Which type of accidents (in terms of main causes) are more common in fatal or catastrophic accidents?
2. What are the more risky occupations in such accidents?
3. Which parts of human body are more prone to be injured in such accidents
4. What are the common activities that the victims were engaged in prior to the accident?

You are given two files, *MsiaAccidentCases.xlsx* and *osha.xlsx*.

*MsiaAccidentCases.xlsx* is a small sample of accident summaries collected from Malaysia, which have been manually labelled with 9 main causes of accidents: caught in/between objects, collapse of object, drowning, electrocution, exposure to chemical substances, exposure to extreme temperatures, falls, fires and explosions, struck by moving objects, suffocation, and others.

The other document, *osha.xlsx*, is much larger. Due to resource constraint, it’s not labeled with the main causes.

You will use text mining techniques learned in class to analyze the data, extract useful information, build necessary models, and produce a report on your findings on the four questions.

# Tool and Method

You may use python and relevant packages. Your investigations and report should follow the text mining methodology introduced in class if possible (but additional steps and ideas are also welcome). Remember that text mining is a highly iterative process and it may be necessary to iterate around the “loop” many times to get the best results. Be sure to include in your report detailed results of data analysis performed, e.g. graphs, plots, tables, rules, performance figures etc.

**Submission Details**

In your project folder of IVLE, submit the following:

* a soft copy of your report
* cleaned datasets, *osha* dataset with causes predicted
* source codes

Make sure your submitted items state names and matriculation numbers of your team members.

# Project Hints

The following hints are for guidance only and your analysis is not restricted to this list. You will be assessed on your text mining process, the interpretation of your results, as well as any discovered knowledge.

Before you start

What are the objectives for this text mining project?

What’s the data?

How’s the data quality? Any errors in the data? Are the labels correct? Is the information in the data sufficient?

Text Mining

Which text mining techniques are required here? Which questions are more challenging?

Do I need to build classification models? What are the categories? One category per document, or a number of most likely categories? How to achieve that? Is the amount of labeled data enough? Any alternative approach?

Are there any terms that should be excluded? Any synonyms?

Which terms are important for my analysis? Shall I create different lists to keep them?

What’s next after I get scored documents using category model?

Is validation of the results applicable here? How?

What kind of information or entities need to be extracted? Any ready-made package to do that? What clues from the text can be used for extraction?

What’s next after extraction?

**Report Guidelines**

* List of contents
* Executive Summary (1 page)
  + Business goals
  + Findings
* Introduction
  + Purpose of Analysis, business & text mining goals
* Main Body
  + Description of each mining step and the results
    - Use Suitable diagrams & summary tables
    - State interpretation & interim conclusions clearly
* Conclusions
  + What was found
  + What was not found out
  + What actions should be done or taken out
  + How the results can be used (describe an implementation plan)
  + Further recommended research
* List of references
* Appendices
  + Details, listings, figures
  + Proper indexing & referencing