

National University of Singapore

Institute of System Science

AY 2017/2018

KE5205

TEXT MINING

Construction Industry  
Determining Pain Points

TEAM PT 03

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| Anusuya Manickavasagam | A0163300Y |
| Chinnasubbareddygari Mohan Reddy | A0163433L |
| Gello Mark Vito | A0163448Y |
| Sridharan Kesavan | A0163207M |
| Nandha Kumar Murugesan | A0163346E |
|  |  |

Table of Contents

[Executive Summary 3](#_Toc496346956)

[Introduction 3](#_Toc496346957)

[Purpose of Analysis 3](#_Toc496346958)

[Business and Text Mining Goals 3](#_Toc496346959)

[Model Building and Testing 3](#_Toc496346960)

[Conclusion 3](#_Toc496346961)

[Reference 3](#_Toc496346962)

[Appendix 3](#_Toc496346963)

# Executive Summary

**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

# Introduction

## Purpose of Analysis

This study aims to identify the most common pain points in the construction industry. The construction industry is a booming field which serves as a foundation of a growing economy. However, accidents are prone and usually leads to deaths as well. To help mitigating these risks, the proponents of this study try to identify the factors contributing to these risks so safeguards can be put in place.

## Business and Text Mining Goals

The data used in this study are from 2 files: MsiaAccidentCases.xslx and osha.xslx. The former contains a more structured dataset with entries categorized between these groups:

* Exposure

The latter contains more entries which are not classified into groups.

# Model Building and Testing

# Conclusion

# Reference

# Appendix

STEPS

Verify data integrity of MsiaAccidentCases.xslx

Build a model for osha.xlsx: preprocess x[train], make sure y[train] categorization is correct

Categorize the cause from osha.xslx using the built model

Check for model accuracy

Additionally verify the osha.xlsx cause, normalize the final training data to contain at least 100 per category

8 categories = 800 dataset (if 100/category), train a model and apply to osha.xslx

Q1: highest count(cause)

Q2: highest(extract occupation from Q1 entries)

Q3: highest(extract body part from Q1 entries)

Q4: highest(extract activities from Q1 entries)