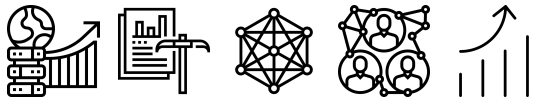


CSC 5741
Lecture 4: Data Pre-processing and Transformation

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia

Announcements—April 16, 2019 (1/2)

- **Paper reading suggestions**
  - Accounts towards class participation
  - HINT: Suggest papers you will include in the background section of the Technical Report
- **Grading of assessments**
  - Grading will be finalised before end of this week

No.	First Name	Lastname
1	Chola	Paul Modest
2	Daka	John Chrispin
3	Lamaswala	Inonge
4	Mubanga	Mubanga
5	Mukuma	Nonde
6	Mulenga	David
7	Mumbi	Memory
8	Mutende	Kaumba
9	Nongola	Justin
10	Nyambe	Teddy
11	Phiri	Jonathan
12	Sampa	Anthny Wila
13	Shamane	Tasha

https://groups.google.com/a/unza.zm/forum/?hl=en#!forum/csc5741

Announcements—April 16, 2019 (2/2)

- **Mini Project progress**
  - Ensure you draw up a plan, with specific details of tasks and activities
  - Get the easy portions of the project out of the way
- **Mini Project data collection**
  - Jupyter Notebook walkthrough

Implementation [8%]
30%: Data collection
30%: Code/scripts works correctly
20%: Novelty of key insights provided
10%: Relevance of implementation
10%: Demonstration

Presentation [4%]
20%: Contents of presentation
20%: Quality of presentation
20%: Visualisations
20%: Comprehensiveness of presentation
20%: Response to questions

Technical Report [8%]
10%: Abstract
10%: Aim/Problem Formulation and Background Work
10%: Implementation
10%: Dataset Description

https://groups.google.com/a/unza.zm/forum/?hl=en#!forum/csc5741

Lecture Series Outline

- **Part I: Academic Talk**
- **Part II: Paper Reading Discussion**
- **Part III: Data Pre-processing**
- **Part IV: Data Transformation**

Lecture Series Outline

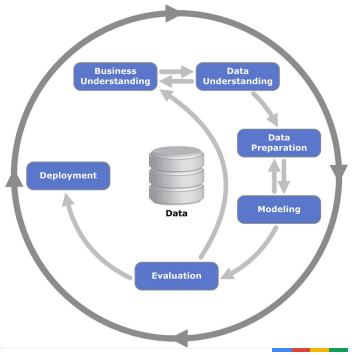
- Part I: Academic Talk
  - Friday Chazanga, University of Zambia
  - Title: "Development of a Two-Factor Authentication for Vehicle Parking Space Control Based on Automatic Number Plate Recognition and Radio Frequency Identification"
- Part II: Paper Reading Discussion
- Part III: Data Pre-processing
- Part IV: Data Transformation

Lecture Series Outline

- Part I: Academic Talk
- Part II: Paper Reading Discussion
- Part III: Data Pre-processing
  - Introduction
  - Text Preprocessing
  - Tokenization
  - Jupyter Notebook Walkthrough
- Part IV: Data Transformation

Introduction (1/3)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
  - CRISP-DM segments a data mining project into six phases with no strict order of execution
  - Surveys conducted suggest CRISP-DM is the most widely used methodology



Introduction (2/3)

- Select data required for modeling process/phase
- Clean the data
- Reconstruct the data and derive necessary attributes
- Merge different data sources
- Reformat the data

Introduction (3/3)

- **Terminologies**
  - Document—Set of terms such as a file
  - Term—Individual word contained in a document
  - Corpus—Collection of documents

Data Cleaning (1/2)

- **Data preprocessing typically involves data cleaning**
  - Removing duplicate entries
  - Dealing with null values: removing vs replacing null values
  - Dealing with outliers

Data Cleaning (2/2)

- **Textual content by far involves the most pre-processing steps**
- **Common text pre-processing techniques generally involve several iterations of cleanup steps**
  - Removing duplicate entries
  - Dealing with null values: removing vs replacing null values
  - Dealing with outliers

Text Processing (1/10)

- **Text processing techniques include**
  - Case folding
  - Stemming
  - Stopping
  - Removing Punctuations
  - Deduplication
  - Missing Values
  - Tokenization

Text Processing (2/10)

- **Case folding**
  - Textual content is generally case sensitive: e.g. RDBMS
    - Zambia vs ZAMBIA vs ZaMbia
    - **var_x = {"Zambia", "ZAMBIA", "ZaMbia", Zambia}**
    - **len(var_x)**
  - Case folding involves changing all document terms to a standard case, e.g. lower case

Text Processing (3/10)

- **Stemming**
  - Changing document terms into canonical versions
  - Stemming should avoid mapping words with different roots to the same stem
  - Poster's Stemming Algorithm applies rules based on patterns of vowel-consonant transition

Text Processing (4/10)

- **Stemming**
  - Changing document terms into canonical versions
    - Country vs Countries
  - Stemming should avoid mapping words with different roots to the same stem
  - Poster's Stemming Algorithm applies rules based on patterns of vowel-consonant transition

Text Processing (5/10)

- **Stopping**
  - Stopping involves the removal of stopwords
  - Stopwords are common words that do not discriminate in terms of relevance
  - Stopwords are not standard and depend on domain and language
    - Chemestry vs Engineering
    - English vs Lozi

Text Processing (6/10)

- **Removing Punctuations**
  - Open text typically contains punctuation marks that need to be removed

Text Processing (7/10)

- **Deduplication**
  - Duplicate data entries are a common occurance and careful attention must be placed in ensure that entries are unique

Text Processing (8/10)

- **Deduplication**
  - Duplicate data entries are a common occurance and careful attention must be placed in ensure that entries are unique

Text Processing (9/10)

- **Missing Values**
  - Careful emphasis must be placed on how to deal with missing and/or null values
  - Depending on the problem, this could involve excluding records with null values or replacing the null values with placeholder text

Text Processing (10/10)

- **Tokenization**
  - Splitting a document up into constituent words is referred to as tokenizing
  - There are a number of strategies for tokenising document
  - Simple strategy: create a vector of all possible words
    - Count number of times word appears in each document

Lecture Series Outline

- **Part I: Academic Talk**
- **Part II: Paper Reading Discussion**
- **Part I: Data Pre-processing**
- **Part II: Data Transformation**
  - Introduction
  - Bag-of-Words Model
  - Term Frequency
  - TF-IDF Vectorising
  - Jupyter Notebook Walkthrough

Bag-of-Words Model

- **Bag-of-Words**
  - Computers are generally not good at processing text, however, they are generally good at working with numbers
  - Each document, once tokenised can be thought of as a bag of words.

Term Document Frequency

- **Term Document Frequency**
  - Vector representation of document terms, with their corresponding frequency of occurrence
  - Note: Commonly used in Information Retrieval

TF-IDF

- **TF-IDF**
  - Frequency distribution of words in a document is not sufficient to rank important of worlds
  - TF-IDF provides a better way of scoring the relative relevant of document terms

TF-IDF

- **TF-IDF**
  - tf-idf = tf(w) * idf(w)
  - tf(w)— Number of times word appears in a document
  - idf(w)—log(number of documents/number of documents that contain word)

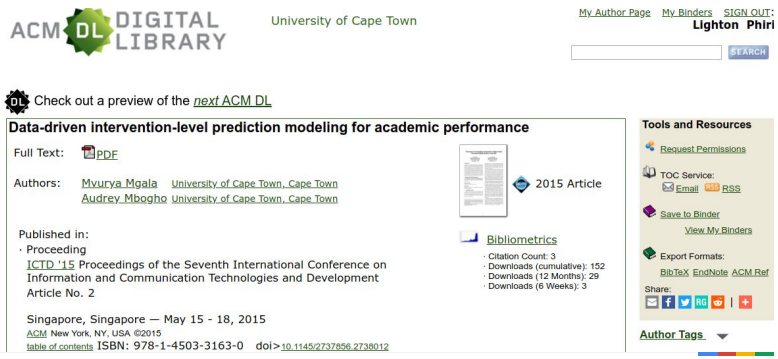
Q & A Session

- **Comments, concerns and complaints?**

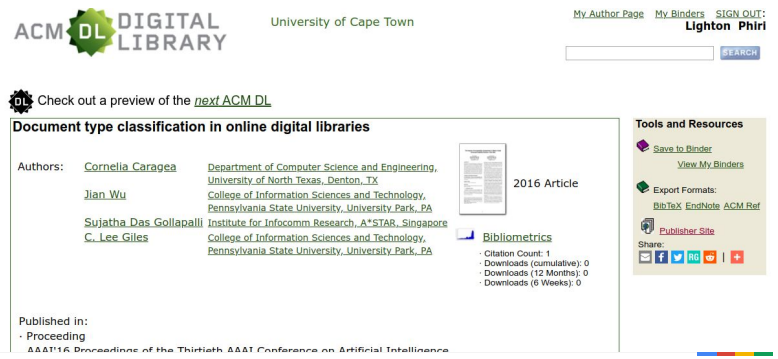
Lecture Series Outline

- **Part I: Academic Talk**
- **Part II: Paper Reading Discussion**
  - M. Mgala and A. Mbogho (2015). "Data-driven intervention-level prediction modeling for academic performance"
  - Caragea et al. (2016). "Document Type Classification in Online Digital Libraries"
  - Moro et al. (2011). "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology"
- **Part III: Data Pre-processing**
- **Part IV: Data Transformation**

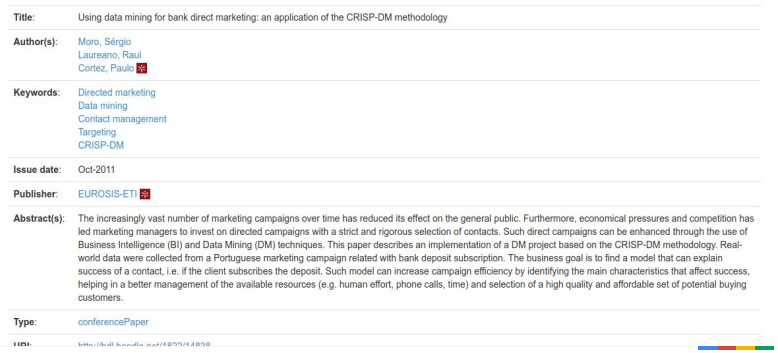
Paper Reading Session (1/3)



Paper Reading Session (2/3)



Paper Reading Session (3/3)

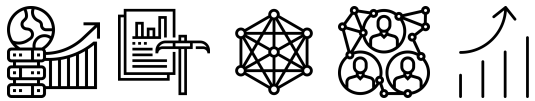


Bibliography

[1] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 2
https://www.cs.waikato.ac.nz/ml/weka/book.html

[2] Introduction to Information Retrieval. Chapter 2
https://nlp.stanford.edu/IR-book

[3] Regular Expressions Tutorial - Learn How to Use and Get The Most out of Regular Expressions
https://www.regular-expressions.info/tutorial.html



CSC 5741
Lecture 4: Data Pre-processing and Transformation

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia