Master of Technology in Knowledge Engineering

Text Mining

Text Mining Process

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
email: zhenzhen@nus.edu.sg

Objectives

At the end of this module, you can:

- Define the process to perform text mining based on the business requirements
- Describe the differences between tasks in text mining and tasks in data mining

Outline

- CRISP-DM in text mining
  1. Business Understanding
  2. Data Understanding
  3. Data Preparation
  4. Modeling
  5. Evaluation
  6. Deployment

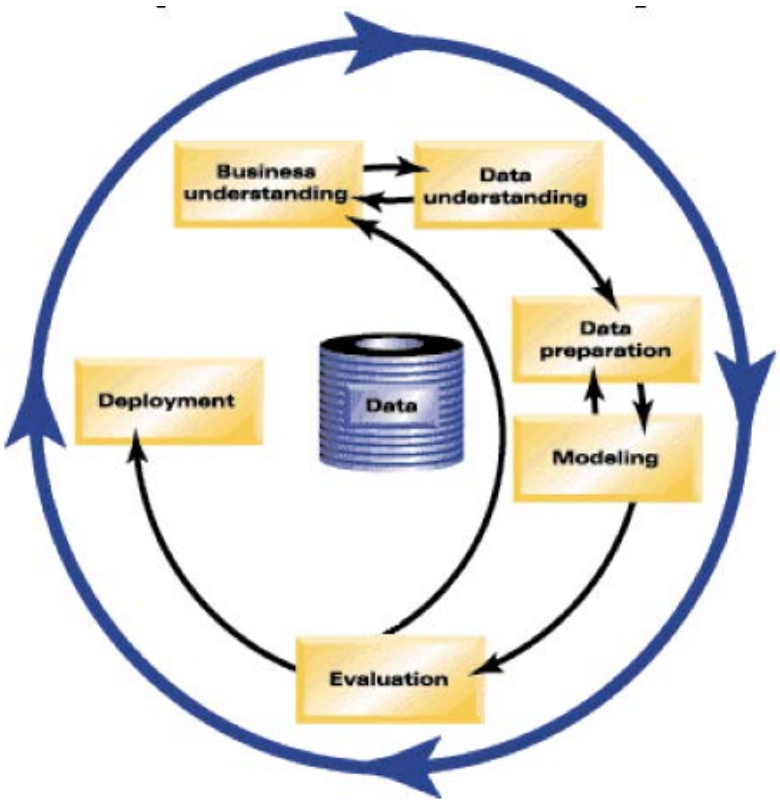
CRISP-DM

- **Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

- An industry-proven methodology tc guide data mining efforts and help project planning

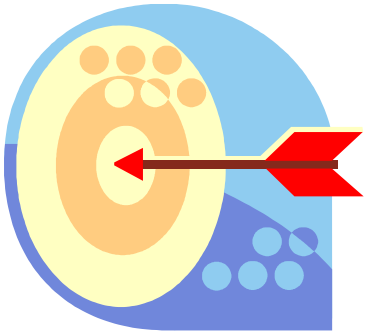
- A framework defining a lifecycle of 6 phases, including tasks typically done for each phase,

- Adaptable for Text Mining



1. Business Understanding

- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan



Don't find answers to the wrong questions!

- Understand the business case
- Determine the purpose of the study
- Inventory of available text data
- Text data alone or … ?

Project Plan

- A common experience for data mining projects is:
  - data preparation ~ 50-70 % of time
  - data understanding ~ 20-30 % of time
  - modeling, evaluation ~ 10-20 % of time
  - deployment ~ 5-10 % of time

- In text analytics, the data collection and processing phase is found to be more laborious, and therefore requires more time.

- Highly crucial to include business/domain expert in the project team.

2. Data Understanding

- Collect Initial data
- Describe data
- Explore data
- Verify data quality

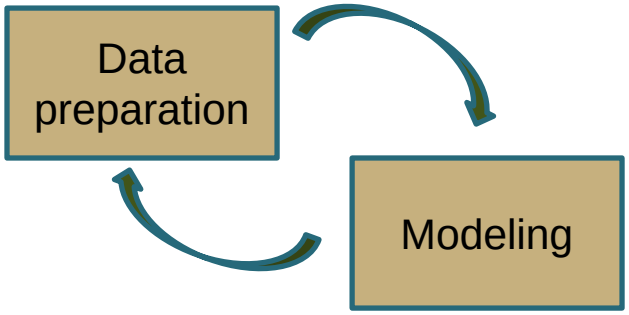
- Identify the text data sources (digitized or paper-based; internal or external to the organization)
- Assess the accessibility and usability of the data
- Collect an initial set of data
- Explore the richness of the data (e.g., does it have the information content needed?)
- Assess the quantity and quality of the data (any errors?)

3. Data Preparation

- Select data
- Clean data
- Construct new data
- Integrate data
- Format data

- Establish the text corpus
- Clean the text data
  - formatting, removal of irrelevant sections, combine text, etc.
- Preprocess the data
  - build stopword/include-word list (and other linguistic resources), identify candidate terms, create TDM, simplify TDM, etc.

4. Modeling



- Select modeling techniques
- Generate test design
- Build model
- Assess model

- Develop categorization model that can be used to classify/score text
- Other techniques like topic modeling, clustering and association analysis may also be used here
- The output of categorization model may be input to other prediction models (using structured data)

5. Evaluation

- Evaluate results
- Review process
- Determine next steps

- Verify and validate the proper execution of all the activities
- Ensure that the models developed and verified are addressing the business problem and satisfying the defined objectives
- Anything left out?

6. Deployment

- Plan deployment
- Plan monitoring and maintenance
- Produce final report and presentation
- Review project

- Deployment ranges from writing a report detailing the findings for the decision makers, to integrating the model into BI system
- Models should be updated periodically with new data

Reference and Resources

- CRISP-DM 1.0 Step-by-step data mining guide (ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf)
- Gary Miner, John Elder IV et. al. Chapter 5 Text Mining Methodology, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, 2012