Hons. Degree in Computing

H4016 Text Analysis

The enrolment key for moodle is **miners**

Lecturer: Geraldine Gray

Contact details: geraldine.gray@itb.ie

office: E020

mail box: E6

Unit 1: Introduction to Text Mining

Objective for today

1) Overview of the course.
2) Understand what text analytics is.
3) How text analytics fits into CRISP-DM.
4) Review a publication on text analytics.

Course aims and objectives

AIM: The aim of this module is to give learners an understanding of how natural language processing (NLP) techniques and machine learning techniques can be applied to text analytics.

On successful completion of this module the learner will be able to

- Understand the application of NLP techniques to create a structured dataset for subsequent analysis.
- Discuss how prediction and clustering techniques can be applied to text mining. Exam
- Explain limitations of current information extraction techniques and the vision for the future.
- Work through the stages of a text mining project using a recognised methodology.
- Extract key concepts and relationships from textual or "unstructured" data. CA
- Apply prediction and clustering techniques to the prepared data.
- Evaluate the performance of text analytics tools.

Overview of the course

Lectures are organised as follows:

- Introduction to Text Mining methodology; recap on CRISP – DM
- Pre-processing for text mining
- Mining algorithms
  - Clustering
  - Classification
- Big Data Analytics

Plan for labs

1) Work sheets on text analytics with Rapid Miner (4-5 weeks)

2) Text Analytics assessment

Assessment

- Exam – 60%
- Continuous assessment – 40%
  - Lab work on Rapid Miner - 10%
    - Labsheets, to be completed by the end of each lab
  - Text Mining assessment – 30%

Reading Material

- Course notes
- Hofmann, Chisholm, Text Mining and Visualisation: Case studies using open-source tools, 2016
- Berry, Kogan (2010), Text Mining Applications and Theory. ISBN: 0470749822
- Manu Konchady, Text Mining Application Programming, Thomson, 2006. ISBN:1-58450-460-9
- Weiss et al (2004), Sholom Weiss, Nitin Indurkhya, Frederick Damerau, Tong Zhang , Text Mining: Methods for Analyzing Unstructured Information, Springer, 2004. ISBN: 0-387-95433-3
- Online journals

Overview of text analytics

Text Analytics. . . Definitions

- Many definitions in the literature:
- The non trivial extraction of **implicit**, **previously unknown**, and **potentially useful** information from (large amounts of) **textual data**"
- An exploration and analysis of textual (natural-language) data by automatic and semi automatic means to discover new knowledge

Text Mining Process

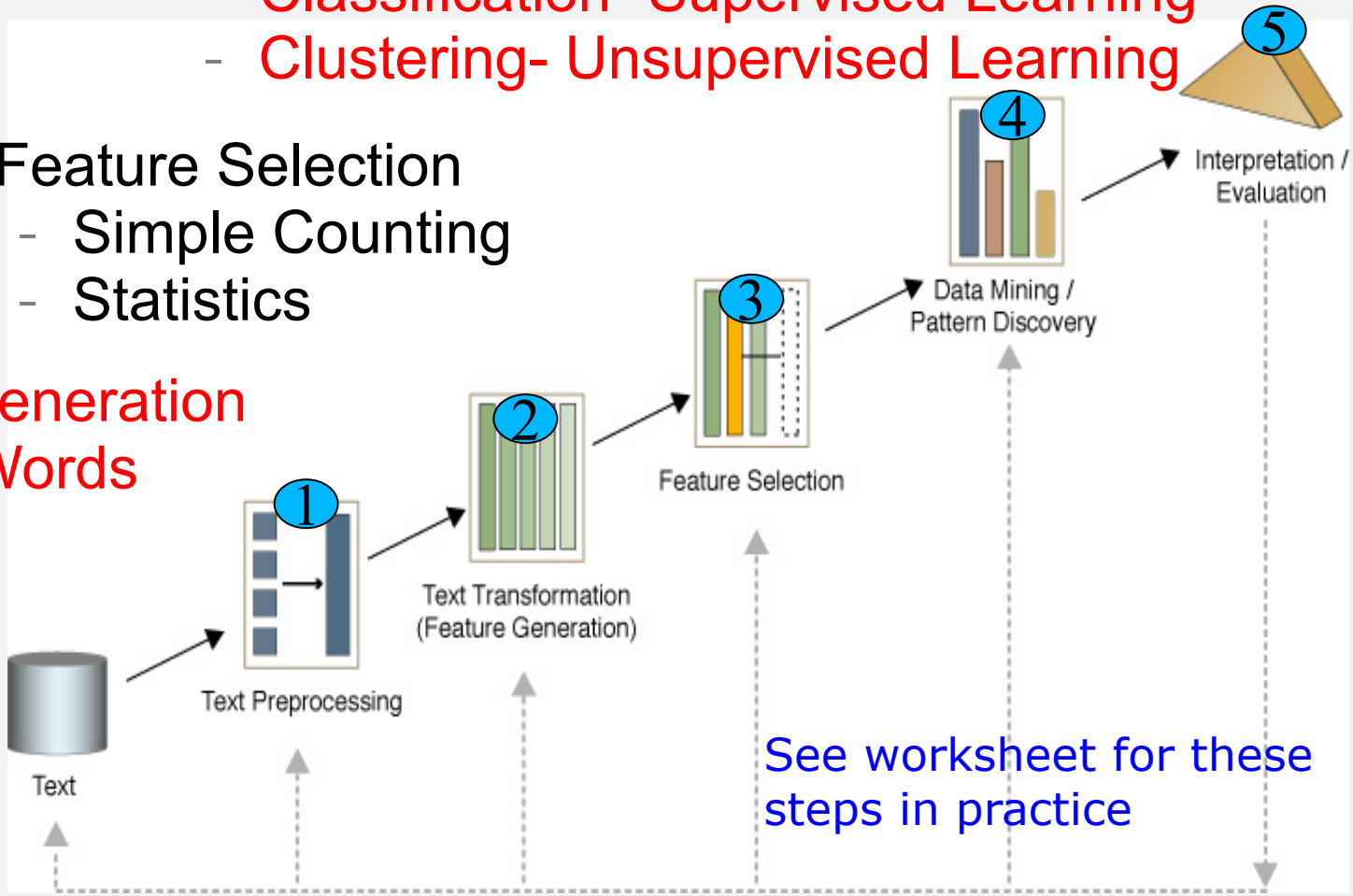
- 5. Analyzing Results

4. Text/Data Mining
   - Classification- Supervised Learning
   - Clustering- Unsupervised Learning

3. Feature Selection
   - Simple Counting
   - Statistics

2. Features Generation
   - Bag of Words

1. Collect data and identify the business objective



See worksheet for these steps in practice

WORK SHEET . . . .

Text Characteristics

1. Large textual database
   - Web is growing
   - Publications are electronic
   - Hugh volumes of content generated every day

2. High dimensionality
   - Each word/phrase in a document is a potential attribute

3. Dependency
   - Relevant information can be a complex conjunction of words and phrases.

4. Poorly structured text
   - Email/Chat rooms
     - "r u available ?"
     - "Hey whazzzzzz up"
   - Speech

5. Noisy data
   - Spelling mistakes
   - Abbreviations
   - Acronyms

Text Characteristics

6. Ambiguity
   - Word ambiguity
     - Pronouns (he, she …)
     - **Synonyms**: multiple words with the same meaning, e.g. buy, purchase
     - **Homonyms**: one word with multiple meanings, e.g. **bat** – is it related to baseball or a mammal.
   - Semantic ambiguity
     - The king saw the rabbit with his glasses. (multiple meanings)
   - Order of words
     - hot dog stand in the amusement park
     - hot amusement stand in the dog park

7. Authority of the source
   - Is IBM a more reliable source then your second cousin?
   - Fake news

Common text mining application areas

- Analysis of social media
  - There are vast stores in user opinion propagated daily on social media sites that are a powerful way to disseminate information from the ground up. Text analytics can automate the processing of this content into actionable information

- E-Mail and Call Center Analysis
  - Need some way to categorise customer communications so that all customer concerns are addressed appropriately.
  - E-mail Filter: using automatic classification of texts as 'junk mail' / channel e-mail queries to the correct person.

- Efficient and Reliable analysis of Open Ended Responses in Surveys
  - In many surveys, open ended questions are asked. These responses need to be analysed. Text mining can help automate that task.

Common text mining application areas

- Discovery search / drug discovery
  - Need some way to quickly review and analyze the information contained in thousands of scientific articles and patents to better understand the relationships existing between a large amount of genes.

- Competitive intelligence
  - Need to monitor newsfeeds, trade press, patents and the Internet to watch for new product announcements, mergers, acquisitions, etc.

- Investigating competitors by crawling their web sites: automatically process the contents of Web pages in a particular domain. For example, you could go to a Web page, and begin "crawling" the links you find there to process all Web pages that are referenced.

Common text mining application areas

- Analysis of interviews/conversations, e.g. patient diagnosis, report problems with cars etc. Notes of such interactions are generally recorded electronically. Analysis can give insight into common complaints and solutions.

- Government Intelligence analysis

  Need to monitor huge volumes of unstructured information e.g. Web pages, NewsGroups, e-mails.

Examples from student projects . . .

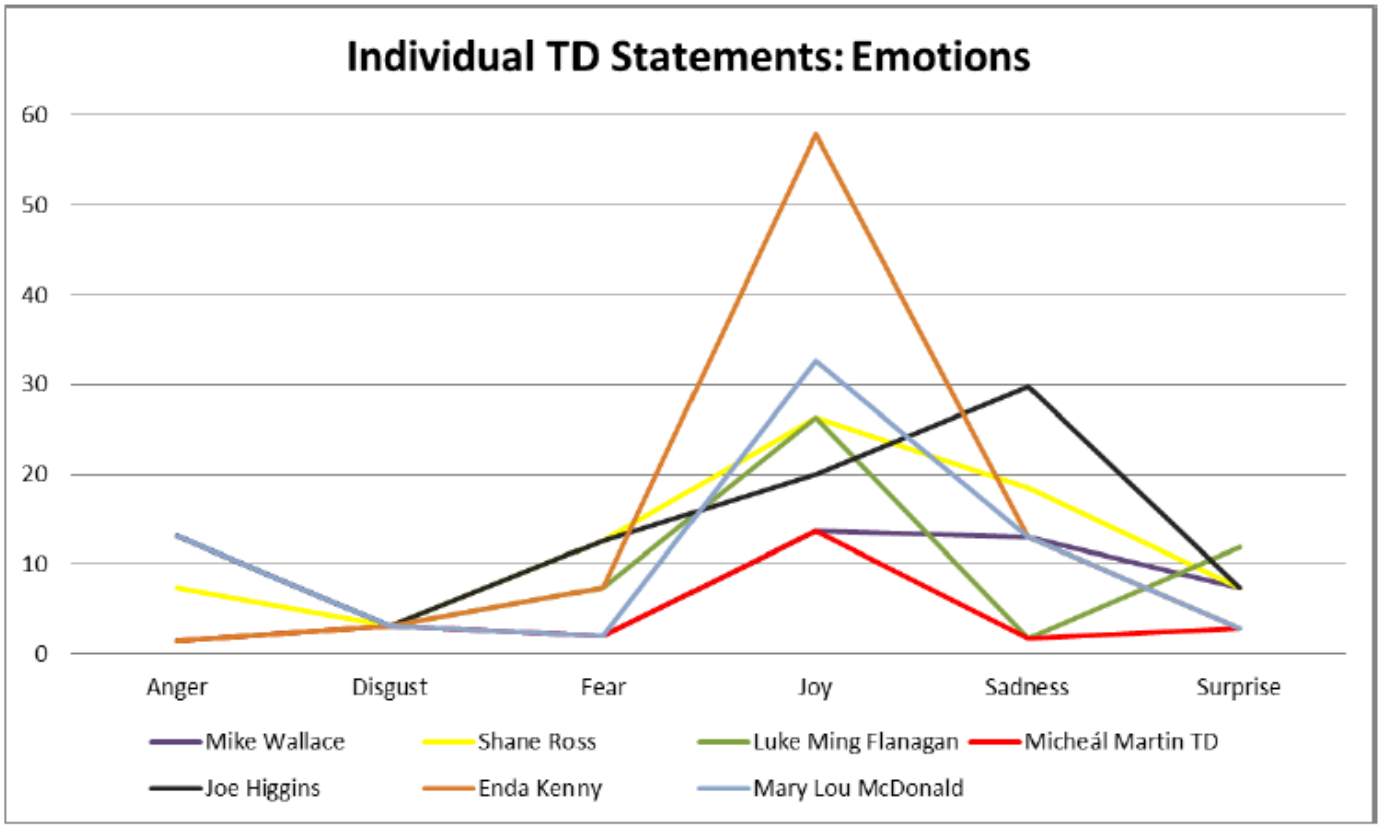


Figure 6.25: TD's Individual Statements: Emotions

Figure 5.3: FF 2011 General Election Manifesto

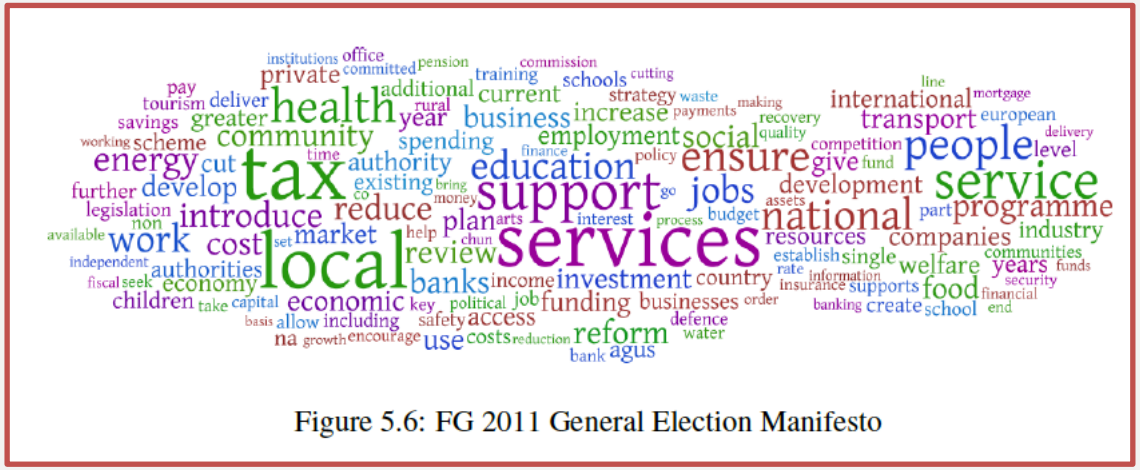
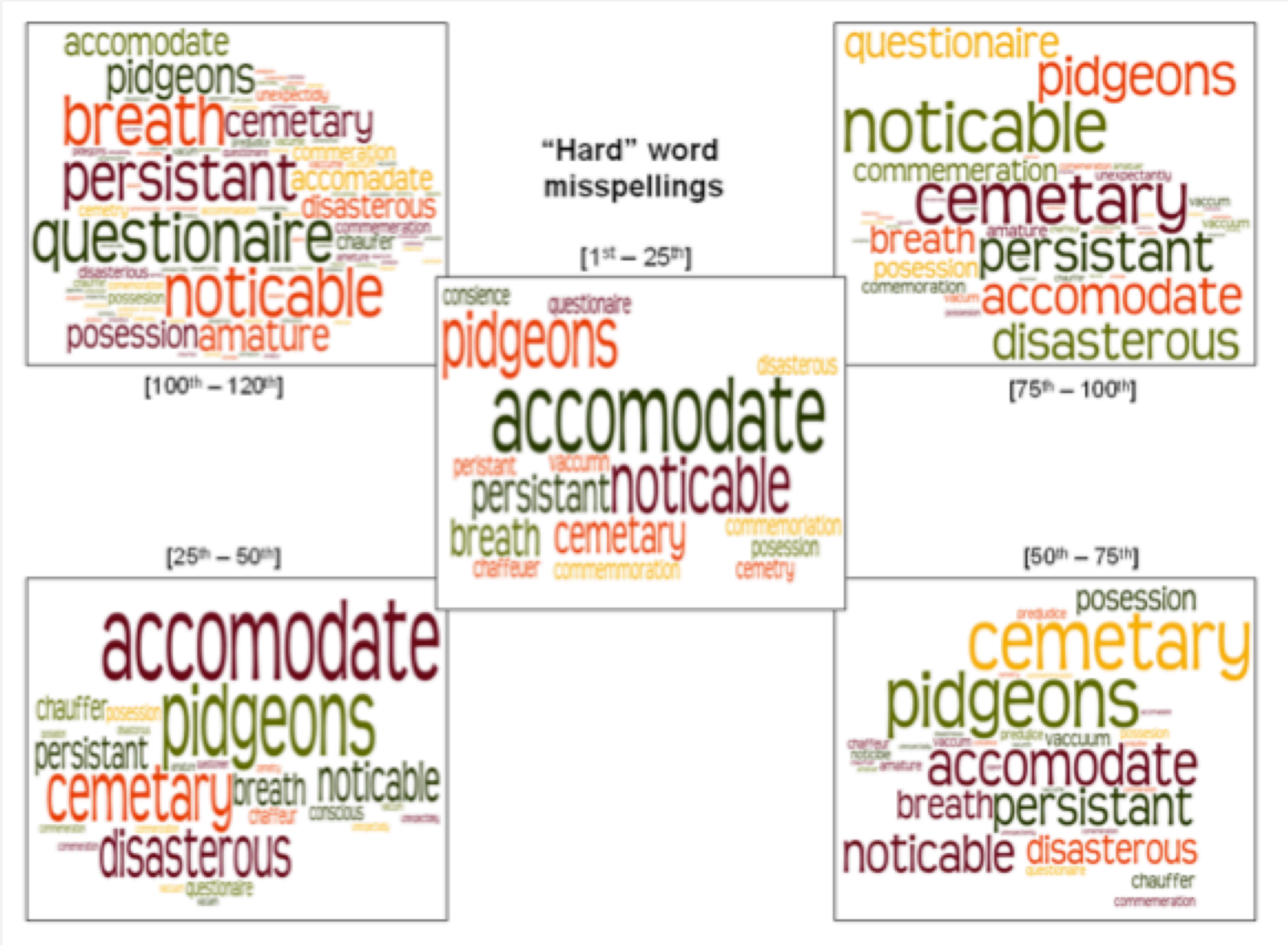


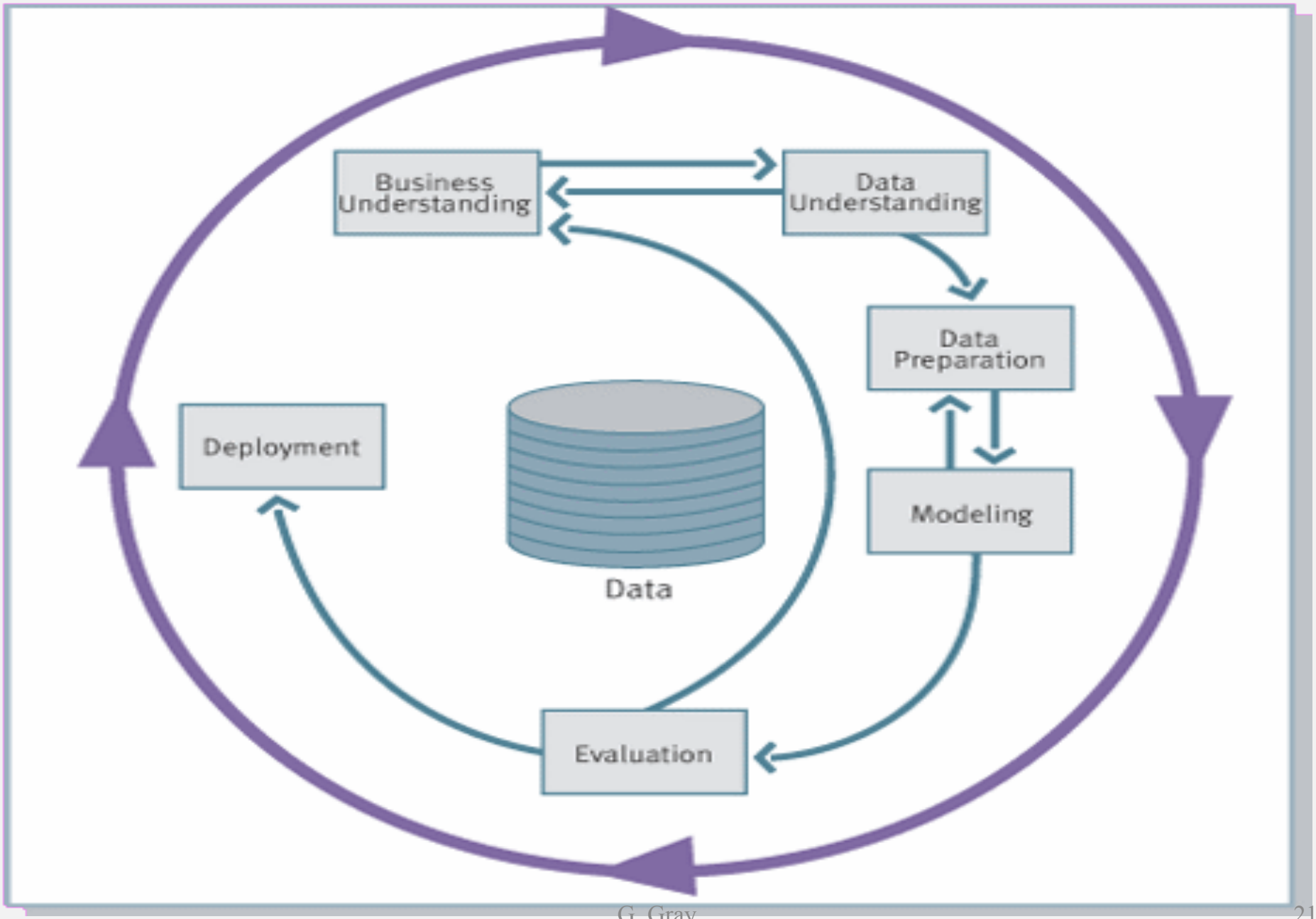
Figure 5.6: FG 2011 General Election Manifesto

Figure 5.10: Labour 2011 General Election Manifesto


"Hard" word misspellings
[1st – 25th]
consience
questionaire
pidgeons
disasterous
accomodate
peristant
vaccum
persistant
noticable
breath
cemetary
commemoriation
posession
chaffeuer
commemmoration
cemetry
[25th – 50th]
accomodate
chauffer
posession
pidgeons
persistant
cemetary
breath
noticable
chaffeur
conscious
disasterous
questionaire
[50th – 75th]
posession
cemetary
prejudice
pidgeons
vaccum
accomodate
breath
persistant
noticable
disasterous
chauffer
commemeration
[75th – 100th]
questionaire
pidgeons
noticable
commemeration
unexpectantly
cemetary
vaccum
breath
amature
persistant
posession
comemoration
accomodate
disasterous
[100th – 120th]
accomodate
pidgeons
breath
cemetary
commeration
persistant
accomadate
disasterous
questionaire
chauffer
noticable
posession
amature

CRISP-DM for text analytics

CRISP – DM for text analytics



Business Understanding

- As with data mining, this section includes:
  - Business objectives
  - Mining objectives
  - Project analysis (cost benifit analysis; risk assessment; resouces needed; assumptions)
  - Project plan

Example:

- Business objective: Improve SPAM detection by 10%
- Mining objective: Generate a predictive model that will classify an email as either SPAM or not SPAM.

Data Understanding

- Find the key concepts in the collection of document
  - Eliminate common words & stop words
  - Eliminate rare words
  - Identify phrases

Data Preparation

- Generate a document vector based on selected concepts (terms and phrases)

Mining

- Select the model
  - As per data mining; Primarily focus on
    - Classification
    - Clustering
    - Association analysis
- Generate test design
  - e.g. split documents into training set and test set
- Build model
- Asses model
  - Evaluate model accuracy

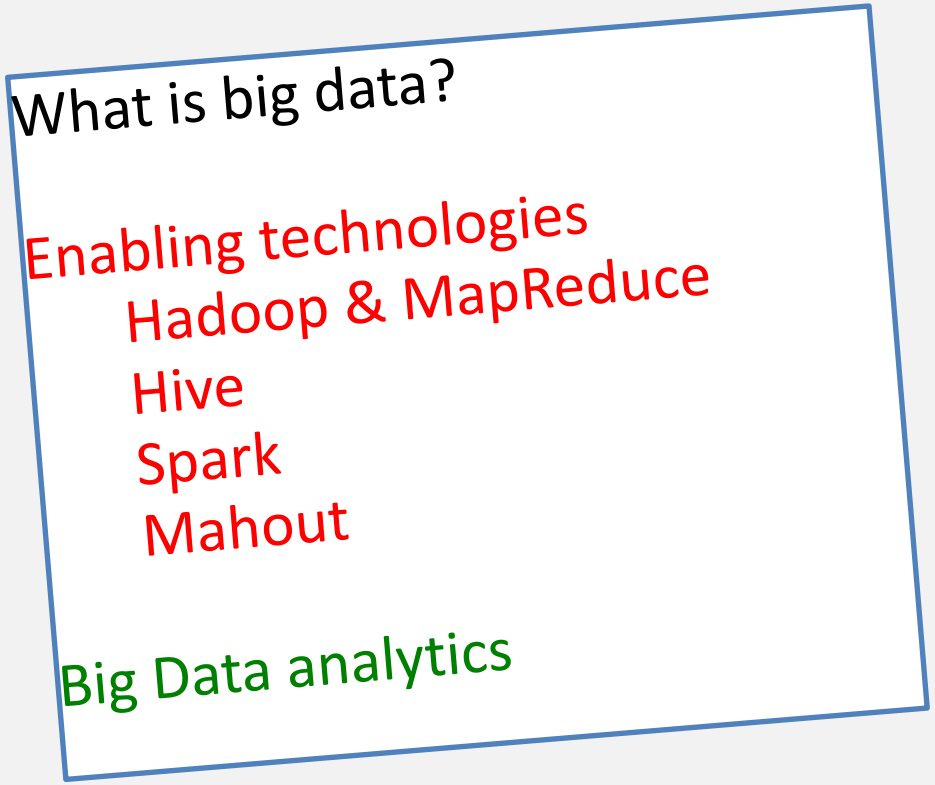
Evaluation & Deployment

- Assessment of data mining reults with respect to original business objectives
- Project review
- Project deployment
  - Generate document vectors for new documents, and run the model.

Other topics on the course. . .

Big Data Analytics

- Introduction to the challenges of working with Big Data and Large scale file systems.
- Adapting standard analytics techniques to run in a distributed environment.



Summary

- Recap . . . .
  - What is text mining?
  - What are the stages involved in processing the data?
  - What will be covered in this module?