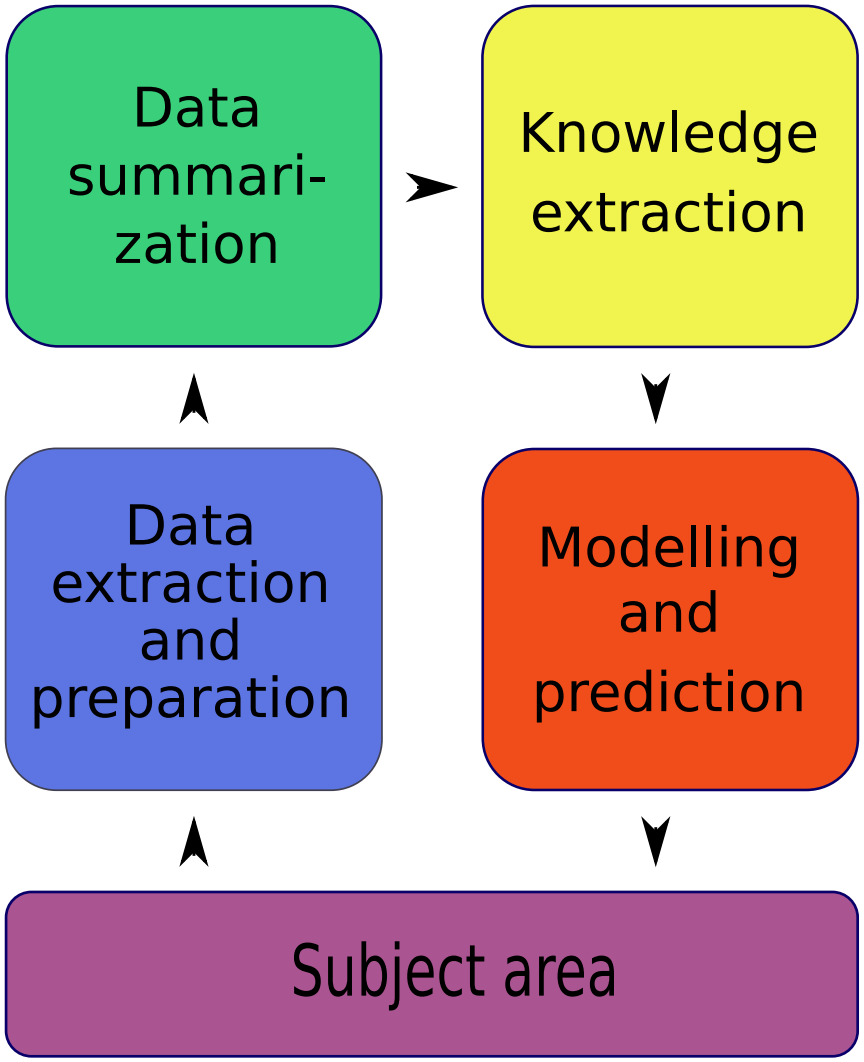
The data analysis cycle



Data Analysis: Special Topics

Institute of Technology Tallaght

Department of Computing

Introduction The topics covered in this last session are each somewhat different from data analysis as we have studied it so far in this module:

- Text mining - unstructured data, with data preparation an important part of the process
- Social network analysis - data in the form of network topology information, including nodes and connections
- Web mining - a combination of three aspects
  - network topology (www structure)
  - text and media (www content)
  - usage and user behaviour (www log files)

Text mining

- Many and varied sources of data:
  - by type: business reports, business communication, online news, personal messaging, standards documents, specialist document archives, scientific literature etc.
  - by domain: law, science, medicine, business, finance etc.
  - by format: PDF files, MS Word files, XML files, text messages, tweets etc.
- Applications:
  - **marketing:**
    * 'social personas' clustering technique from customer communication, social media sources, blogs etc.
    * 'listening platform' for real-time gathering of customer sentiment
    * analysis of call centre conversations to understand performance and potential problems
  - **business operations:**
    * analysis of sentiment among employees

    - general understanding of consumer behaviour and its applications
  - **legal:**
    - e-discovery platforms that help to minimize risk when sharing documents
    - analysis of case histories and risk assessment
    - archive searches for help with case handling based on precedences etc.
  - **government and politics:**
    - assessment of general population or constituent sentiment
    - geopolitical security
- Text is unstructured data but we have to start somewhere when analysing it
  - Bag of words - in the first instance, a body of text is viewed as a grouping of words, which may be repeated
  - Every word is a possible attribute, but what value can we give it for a document?
    - 1 vs 0 - depending on whether the word is present in the document or not
    - count - representing the number of times the word appears in the document - results in a term-document-table:

	why	where	how	the
doc1	3	2	0	7
doc2	1	0	2	10
doc3	0	1	1	10

* normalization (lowercase, only roots, stopwords removed, division by number of words if multiple documents)

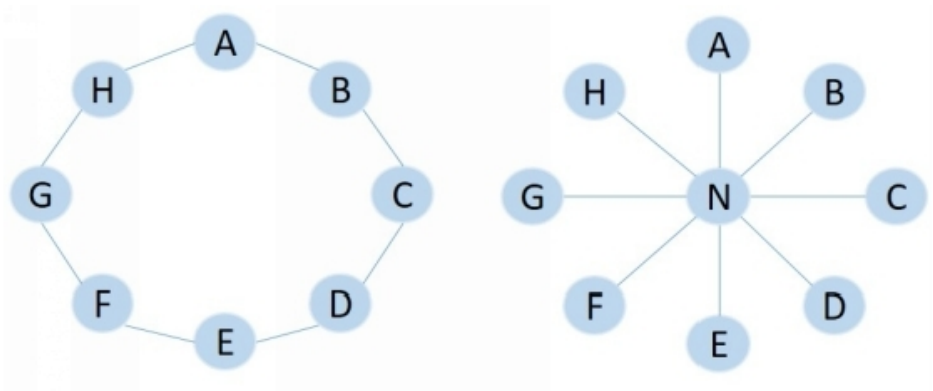
Table 10-3. Terms after normalization and stemming, ordered by frequency

Term	Count	Term	Count	Term	Count	Term	Count
skype	3	microsoft	3	agreement	2	global	1
approv	1	announc	1	acquir	1	lead	1
definit	1	lake	1	communic	1	internet	1
board	1	led	1	director	1	corp	1
compani	1	investor	1	silver	1	billion	1

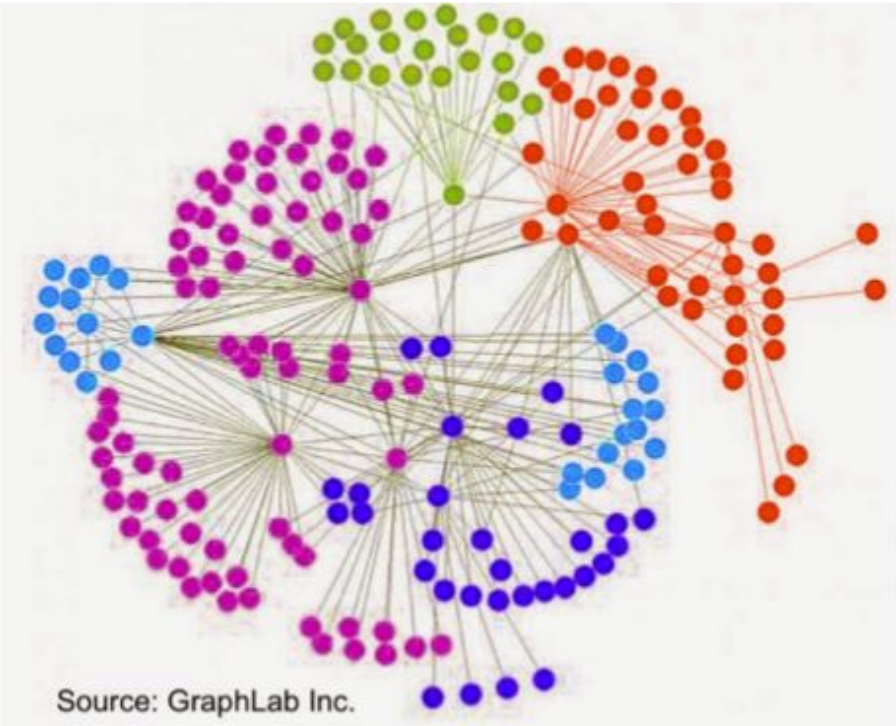
- What for?
  - incidence of words (e.g. word-cloud)
  - predicting chances of a document being liked based on comparison with previous documents (predictive modelling)
  - clustering for determining document type, sentiment etc.
  - association rule analysis e.g. word 'happy' occurs with what other words?

Social network analysis

- Applications:
  - study of communities
  - marketing
  - public health (disease spread)
  - human behaviour and self awareness
- Differs from data analysis in general because it is based around analysing network topologies and their properties:
  - ring / hub and spokes



- sparse / dense
- subnetworks



- node importance (+)

- Google page rank (85% importance and 15% Teleporting)

Web mining

- Content mining (HTML etc)
- Structure mining (hyperlinks resulting in network topology)
- Usage (site visits, clicks)