Lesson 1

Text Mining Basics

Text Mining Definitions

1. “Text mining refers to the process of deriving high-quality information from text.” (Wikipedia)

Text Mining Definitions

2. “Text mining is the process of discovering and extracting knowledge from unstructured data.” (National Center of Text Mining—The University of Manchester)

http://www.nactem.ac.uk/brochure/NaCTeM_Brochure.pdf

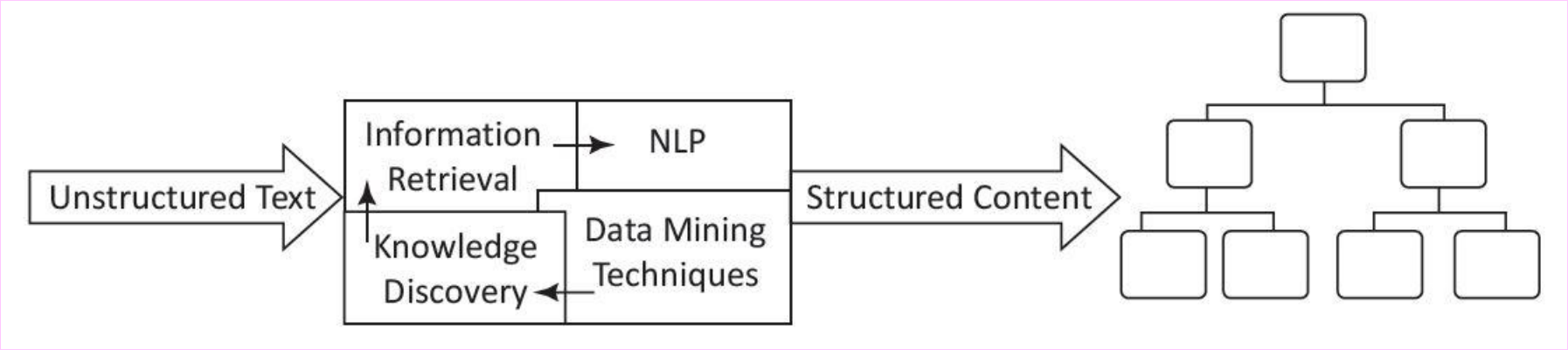
Text Mining Definitions

3. "Text mining is the process of analyzing collections of textual contents in order to capture key concepts themes, uncover hidden relationships, and discover the trends without requiring that you know the precise words or terms that authors have used to express those concepts." (IBM)

Text Mining Definitions

4. “Text mining is a technique which helps in revealing the patterns and relationships in large volumes of textual content that are not visible to the naked eye, leading to new business opportunities and improvements in processes.” (Amazon BigData Official Blog)

Figure 9.1 Text analytics process pipeline



Information Retrieval (IR) is a process of searching and retrieving a subset of documents from the abundant collection of documents. IR can also be defined as extraction of information required by a user.

Information Extraction (IE) is a process in which the software extracts structured information from unstructured and/or semi-structured documents. IE finds the relationship within text or desired contents from text.

Applications

(i) mail filtering (spam), (ii) drug action reports (iii) fraud detection (iv) knowledge management, and (iv) social media data analysis.

- .

Areas of Applications

- Natural Language Processing (NLP) is a technique for analyzing, understanding and deriving meaning from human language.
- Document Clustering is an application which groups text documents into clusters.

Areas of Applications

- Automating document organization, topic extraction and fast information retrieval or filtering use the document clustering method
- Document Classification is an application to classify text documents into classes or categories

Areas of Applications

- Concept Extraction─ an application that deals with the extraction of concept from textual data
- Business Documents Analysis

Applications in Business Domain

- Predicting stock movements from analysis of company results
- Decision making for product and innovations developed at the company and contextual advertising.
- Prediction of information or categories
- Information linkage with another information . .

Text Analytics Open source tool

- Python library *nltk*
- Refer Online contents accompanying book for using *nltk* in the solution of Practice Exercise 9.2

Summary

We learnt:

- Text Mining Definitions
- Text Mining Process Pipeline
- Text Mining Applications

End of Lesson 1 on

Text Mining Basics