ADSA H6014: Text Analysis and Web Content Mining

Short Title:	Text Analysis and Web Content Mining APPROVED
Full Title:	Text Analysis and Web Content Mining

Module Code:	ADSA H6014

ECTS credits:	10

NFQ Level:	9

Module Delivered in	1 programme(s)

Module Contributor:	Markus Hofmann

Module Description:	Module aims include: • Investigate state of the art and research trends in text analysis and web content mining. • Critique and evaluate the performance of algorithms for both text analysis and web content mining

Learning Outcomes:

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

1. Demonstrate an awareness and critical understanding of ways to extract key concepts and relationships from semi-structured and unstructured text, and structure them for data mining
2. Discuss current research activities relating to text mining and web content mining
3. Understand limitations of current information extraction techniques and the vision for the future
4. Extract key concepts and relationships from semi-structured and unstructured data
5. Apply prediction and clustering techniques to the prepared data, and critically evaluate the results in particular with the aid of modern text visualisation techniques.
6. Independently research current trends and developments relating to the processing of semi-structured unstructured data

Module Content & Assessment

Indicative Content
Preparing text documents for mining § Extracting key concepts, sentiments, and relationships from semi-structured and unstructured data; § Structural representations for text documents (e.g. Vector Space Model). § Apply appropriate visualisation techniques pre-model, model and post-model.
Mining the data § Learning methods for sparse, high simensional data (e.g. support vector machines) and performance evaluation. § Clustering methods (e.g. k-means clustering; hierarchical clustering) and similarity measures for asymmetric data. § Visualisation techniques. § Case studies.
Web Crawling XPATH, web crawlers, regular expressions, crawling rules
Knowledge Extraction Concept extraction based on both syntactic and semantic natural language processing.

Indicative Assessment Breakdown	%
Course Work Assessment %	100.00%

Course Work Assessment %

Assessment Type	*Assessment Description*	*Outcome addressed*	*% of total*	*Assessment Date*
Reflective Journal	Students must prepare a portfolio of literary reviews and analysis covering a range of topics across all areas of the syllabus, and give an oral presentation of at least one of their research areas. For example, • An exploration of current trends in methods for structuring text in preparation for mining. • Investigation into at least one algorithm suitable for classifying text documents, and an analysis of current research in learning methods. • Investigation into at least one algorithm suitable for clustering text documents, and an analys	1,2,3,6	30.00	Week 6
Practical/Skills Evaluation	Work through all stages of a text mining project life cycle using an appropriate text mining tool. For examples students would be presented with raw text and business objectives from which they would mine the data in an ongoing project during the semester with the following deliverables: • Step 1. Business Understanding: Evaluate the appropriate text mining function to be used to achieve the business objectives. • Step 2. Data Preparation: Structure the data in a format suitable for the relevant text mining function. • Step 3. Data Modelling: Apply the appropriate mining algorithm(s). • Evaluation: Evaluate the results.	4	40.00	Sem 1 End
Practical/Skills Evaluation	Students are asked to compile a unique data set by applying web crawling strategies and implementations. Once the data have been obtained, an appropriate analysis technique such as visualisation, classification, association rules or clustering need to be applied.	4	30.00	Sem 1 End

No Final Exam Assessment %

Indicative Reassessment Requirement
Coursework Only *This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.*
Reassessment Description As per course work

ITB reserves the right to alter the nature and timings of assessment

Indicative Module Workload & Resources

Resources
Recommended Book Resources
Hofmann, Chisholm 2016, *Text Mining and Visualization: Case Studies Using Open-Source Tools*, 1 Ed., Chapman & Hall/CRC Data Mining and Knowledge Discovery Series [ISBN: 1482237571] **Ashok Srivastava (Editor), Mehran Sahami (Editor), *Text Mining: Classification, Clustering, and Applications* [ISBN: 1420059408]** **Sholom M. Weiss... [et al.] 2005, *Text mining*, Springer New York [ISBN: 0-387-95433-3]** **editor, Michael W. Berry 2003, *Survey of text mining*, Springer New York [ISBN: 0-387-95563-1]**
Supplementary Book Resources
Ian H. Witten, Alistair Moffat, Timothy C. Bell 1999, *Managing gigabytes*, Morgan Kaufmann Publishers San Francisco, Calif. [ISBN: 1558605703]
Recommended Article/Paper Resources
Madjid Tavana 2015, *International Journal of Knowledge Engineering and Data Mining* [ISSN: 1755-2087] http://www.inderscience.com/jhome.php?jcode=ijkedm
This module does not have any other resources

Module Delivered in

Programme Code	Programme	Semester	Delivery
BN_KADSA_R	Master of Science in Computing in Applied Data Science & Analytics	3	Elective