School of Informatics Teaching Course Proposal Form

This version was generated **February 14, 2017**. User 'wmagdy@inf.ed.ac.uk' verified.

Proposal

Course Name: TEXT TECHNOLOGIES FOR DATA SCIENCE
Proposer's Name: Walid Magdy
Email Address: wmagdy@inf.ed.ac.uk
Course Year: 4
Names of any courses that this new course replaces :
Text Technologies for Data Science

Course Outline

Course Level: 11
Course Points: 20
Subject area: Informatics
Programme Collections:
Computer Science, Artificial Intelligence, Computational Linguistics.

Teaching / Assessment

Number of Lectures: 18
Number of Tutorials or Lab Sessions: 10
Identified Pre-requisite Courses: 1. Probability theory. 2. Vectors and matrices. 3. Calculus. 4. Python (c
Identified Co-requisite Courses: none
Identified Prohibited Combinations: none

Assessment Weightings:
Written Examination: 60%
Assessed Coursework: 40%
Oral Presentations: 0%

Description of Nature of Assessment:
Written examination will evaluate students' understanding to fundamentals of text technologies and IR. In adiditon, coursework will include three practical assignments to show the depth of understanding to the basics when applied to real-life problems.

Assignments will be designed as follows: 1) Two assignements for student to work individually (10% each). 2) One course project assignment for group of students, 2-4 students per group (20%).

Course Details

Brief Course Description:
The course is an upgrade to the "TEXT TECHNOLOGIES FOR DATA SCIENCE" course, where some content to be updated with recent science and some new content to be added.

The course will continue to focus on text technologies, especially information retrieval, as before, however, it will expand to include new technologies such as information filtering, text classification, and applications on domains such as social media and user generated content in general.

Detailed list of Learning Objectives:

On completion of this course, the student should be able to:

1: Describe the main algorithms for processing, storing and retrieving text.

2: Show familiarity with theoretical aspects of IR, including the major retrieval models.

3: Discuss the range of issues involved in building a real search engine

4: Evaluate the effectiveness of a retrieval algorithm

5: Build social media applications using text processing techniques

Syllabus Information:

1. Introduction: search applications, search tasks, users information need 2. Definitions: documents, queries, bag-of-words trick 3. Laws of text: Zipf, Heaps, clumpting, index size. 4. Vector space: term weighting, similarity functions. 5. Vocabulary mismatch: tokenization, stemming, synonyms. 6. Indexing: inverted lists, compression, query execution. 7. Web crawling: XML feeds, crawling, expected age. 8. Content Extraction: XML tags, DOM, Finn's method. 9. Locality Sensitive Hashing: duplicates, Simhash. 10. Evalaution: recall, precision, F1, MAP, nDCG, query logs. 11. Web search: PageRank, hubs and authorities, link spam. 12. Probabilistic model: probability ranking principle, BM25. 13. Relevance models: exchangeability, cross-language search. 14. Language models for IR: query likelihood, smoothing. 15. Machine learning in IR: PA, SVM, SMO algorithms, LeToR. 16. Social media search, nature, challenges, tasks 17. Information filtering, topic drift 18. Text classification

Recommended Reading List:

Text books: "Introduction to Information Retrieval", C.D. Manning, P. Raghavan and H. Schutze "Search Engines: Information Retrieval in Practice", W. Bruce Croft, Donald Metzler, Trevor Strohman

Readings: "Machine Learning in Automated Text Categorization". F Sebastiani "The Zipf Mystery", Youtube video: https://www.youtube.com/watch?v=fCn8zs912OE "Information Retrieval", C.J. van Rijsbergen "Recommended Reading for IR Research Students", A. Moffat, J. Zobel, D. Hawking

Any additional case for support information:

Adding new content to the TTDS course + changing its credit to 20 points