

## CS 5831: Information & Text Processing (4)

### Catalog Description:

The properties that underlie text processing and their application in terms of compression and encryption. Retrieval models. Digital libraries. Web applications.

**Textbook:** R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

**References:** Salton, G., *Automatic Text Processing*, Addison Wesley, 1989.  
Korfhage, R., *Information Storage and Retrieval*, Wiley, 1997.  
*Journal of the American Society for Information Science (JASIS)*  
*Information Processing and Management (IP&M)*  
Annual TREC conferences (e.g., *The 9<sup>th</sup> Text Retrieval Conference (TREC-9)*), published by NIST.  
Annual International ACM-SIGIR conferences (e.g., Proceedings of the 23<sup>rd</sup> Annual International Conference on Research & Development in Information Retrieval)

### Course Goals:

The objective of this course is to provide insight into the topics associated with the processing of textual material. It covers the theory and application of non-numeric processing with special emphasis on current research issues. You will learn of the statistical characteristics which underlie much of the processing of natural language text. We investigate the methods by which textual material may be automatically processed, stored, accessed, and retrieved. We examine current retrieval applications.

### Prerequisites by Course & Topic

CS 2511: Software Analysis & Design – proficiency in object-oriented design and coding, a systematic approach to testing and debugging

CS 2521: Computer Organization & Architecture – understanding how programs and data are stored and represented in a computer system

Math 3355: Discrete Mathematics – functions

### Major Topics Covered in the Course

- Statistical Characteristics of Text and Entropy
- Compression & Encryption
- Standard Text Processing
- Retrieval Models
- Text Analysis and Automatic Indexing (Automatic Classification, Relevance Feedback)
- Applications—XML and the Web, Digital Libraries, Distributed Retrieval

**Class/Laboratory Schedule:** Lecture: 3 hours per week, Laboratory: 1

### Laboratory Projects

- Calculating entropy (1)
- Coding various compression algorithms and comparing resulting compression ratios for files with different data characteristics (2)
- Coding simple encryption algorithms designed to encode/decode a message (2)
- Using modern encryption techniques to encode/decode messages (2)
- Apply clustering techniques to reveal underlying structure of data (2)
- Analyze the results of searches run on various commercial search engines; speculate on the engine's design based on the analysis (1)
- Apply current web technology to improve an XML-based retrieval system (4)

### Course Contribution to Program Objectives and Outcomes:

1. Understand the statistical characteristics of text that underlie all text processing and the concept of entropy and its relationship to the binary representation of information. (*d*)
2. Apply the basic methods of compression. (*d*)
3. Apply the basic building blocks of encryption. (*d*)
4. Understand the differences between symmetric and asymmetric encryption and the mathematical foundations and methods of public key encryption. (*d*)
5. Understand the basic retrieval models (vector space, Boolean) and their instantiations in modern systems. (*d*)
6. Understand how basic retrieval models are implemented with current technology: the web, distributed systems. (*d*)

### Estimate CSAB Category Content

	CORE	ADVANCED		CORE	ADVANCED
Data Structures			Computer Organization and Architecture		
Algorithms		1	Concept of Programming Languages		
Software Design		1			

### Oral and Written Communications

Every student is required to submit at least    written reports (not including exams, tests, quizzes, or commented programs) of typically    pages and to make    oral presentations of typically 20 minutes duration. Include only material that is graded for grammar, spelling, style, and so forth, as well as for technical content, completeness, and accuracy.

Each student selects a paper from the literature and presents it orally to the class with appropriate visual aids.

### Theoretical Content

- Applications of mathematics (entropy and its relationship to binary representation) (3)
- Applications of group theory, combinatorics and number theory in cryptography (3)

### Problem Analysis

Students decompose problems into tasks and determine appropriate approaches for the solution of text processing problems.

### Solution Design

Students design algorithms and write code for the solution of problems dealing with the processing of text.

**Coordinator/Prepared by:** C. Crouch