

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 9th Text Retrieval Conference (TREC-9)*), published by NIST.

Annual International ACM-SIGIR conferences (e.g., *Proceedings of the 23rd 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 2521: Computer Organization & Architecture – understanding how programs and data are stored and represented in a computer system

CS 3512, Computer Science Theory: functions and via 3512 prerequisite 2511: proficiency in object-oriented design and coding, a systematic approach to testing and debugging

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

Course Outcomes

1. Understanding the statistical characteristics of text that underlie all text processing and the concept of entropy and its relationship to the binary representation of information.
 - a. Given a body of text, compute the average entropy of the message in terms of bits per character.
 - b. Given probability distributions, compute the average entropy of a message when probabilities are both known, and unknown, and compare the results.
2. Apply the basic methods of compression.
 - a. Implement basic compression techniques for files of characteristic data and compare compression ratios.
 - b. Demonstrate an understanding of the importance of redundancy in compression.
3. Apply the basic building blocks of encryption (substitution, permutation, polyalphabetic substitution, stream and block ciphers).
 - a. Write programs to encode and/or decode messages using these basic methods.
 - b. Demonstrate an understanding of the importance of redundancy in encryption.

4. Understand the differences between symmetric and asymmetric encryption and the mathematical foundations and methods of public key encryption.
 - a. Solve problems using the most frequently used methods of asymmetric encryption (e.g., RSA).
 - b. Correctly identify the components with respect to both public and private key data.
 - c. Utilize current methods (RSA, PGP) to send and receive coded messages.
5. Understand the basic retrieval models (vector space, Boolean) and their instantiations in modern systems.
 - a. Understand the advantages of each model.
 - b. Solve small-scale problems demonstrating the utility of these techniques (homework, tests).
 - c. Understand the importance of redundancy in IR.
6. Understand how basic retrieval models are implemented with current technology: the web, distributed systems.
 - a. Study and compare existing systems by analyzing their methods and results.
 - b. Understand the specific problems associated with distributed retrieval.

Relationship to Program Outcomes

In order to take CS 5831, students must have completed discrete math, systems design and analysis, and computer organization and architecture. This course contributes to meeting the following program outcomes:

2. *Students can design, develop, and analyze significant software systems.*
 Students gain experience in the design, implementation and analysis of algorithms dealing with the manipulation of textual data--in particular, entropy, compression, encryption, information retrieval, and web applications. Course outcomes 1-6 map to this program outcome.
3. *Students understand the fundamentals of computer organization and architecture, data structures and related algorithms, and programming languages..*
 Students use appropriate data structures and programming languages (C++ and Java) in solving problems related to the processing of textual data. A typical problem might be the design and implementation of a set of compression routines applied to very large data files with differing characteristics, with an analysis of results based on compression ratios. . Course outcomes 1-6 map to this program outcome.
4. *Students can apply computer science principles and practices to a variety of problems.*
 Students have the opportunity to design and develop software that uses the characteristics of the data in processing it – for example, uses the redundancy inherent in natural language to facilitate data compression. . Course outcomes 1-6 map to this program outcome.
6. *Students can communicate effectively both orally and in writing.*
 Students make two oral presentations (on issues related to cryptography and security, respectively) during the course of the semester with the objective of increasing the breadth of the material covered. Appropriate visual aids are created and made available to the class. . Course outcomes 1-6 map to this program outcome.

Assessment Plan for Course:

This course is assessed every third year by the instructor and a course assessment document covering all of the course outcomes and their effect on the program outcomes is prepared.

Estimate CSAB Category Content

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

Coordinator/Prepared by: C. Crouch