

Search Engines Architecture

*“So, You Want to Build your Own Search Engine?
If so, this Course is for You.”-Dr. E. Garcia.*



**Crawlers, Parsers, Meta Data, Keywords, Regular Expressions,
Stemmers, Search Engine,s Look-Ups Directories,
Inverted Index, SVD, Collection, Vector Space Models**

About Dr. Garcia

Dr. Garcia research interests include Web Mining, Search Engine Architectures, and Information Retrieval at the intersection of Information Security and Intelligence. He is a program committee member of W3C's Adversarial Information Retrieval on the Web Workshops (AIRWeb) and has served as reviewer for *JASIST*, *IBM's Computer and Graphics*, and has co-chaired several local conferences on search engine technologies. At Polytechnic University, he is a visiting lecturer, having taught the graduate course *Web Mining & Business Intelligence*. He is the founder of <http://www.miislita.com>, an online resource on information retrieval and search engine technologies.

Title: Search Engines Architecture

Description: This is a hands-on, one-semester course on search engines architecture and their algorithms. Each class consists of lecture and lab sessions. This course is laboratory oriented, wherein students are expected to build and test their own search engines and related components on a server dedicated for this purpose.

Target: Students in Business, Engineering, and Computer Sciences and from other disciplines are encouraged to register for this special course.

Requirements: Permission from advisor or department and knowledge of matrix algebra.

Grading: Weekly Lab Reports, Project Presentations, and a Final Exam.

Topics: Although not necessarily in this order, some of the topics to be covered, include, but are not limited to the followings:

Linear Algebra Fast Track Tutorial: Brief tutorial on matrix operations with emphasis on vector theory and Singular Value Decomposition

Parser Building: Use of regular expressions to build, test, and use a parser.

Crawler Building: Use of AJAX to build a client-side crawler and a dedicated server-side crawler.

Look-Up Directories: Implementing a look-up directory and pseudo site search tool.

Search Interfaces: Developing and testing search interfaces, their search modes, and advanced search features.

Index and Database Building: Data fragmentation and storing.

Sanitizing and Ranking Answer Sets: Filtering, De-duplicating, and ranking answer set results.

Textbook: There is no official textbook. Open source components will be used or developed by the students. However, the following reference books are recommended for research. Additional references and extended syllabus will be provided in class.

1. *Modern Information Retrieval* (Baeza-Yates and Ribeiro-Neto; Addison Wesley).
2. *Information Retrieval – Algorithms and Heuristics* (Grossman and Frieder; Springer).