BLM5121: Introduction to Web Mining (3 CR)

Software and data resources

Note: some links may be out of date; please flag those to the instructor.

- GiveALink: donate your bookmarks to science -- could be a great source of project ideas and data
- Truthy: could be a great source of social media data for projects, or you could contribute to the project
- Scholarometer provides an API and linked data about authors and their citation impact
- Microsoft Academic Search provides an API to access rich data about authors, publications, and citations
- We have a huge Web click database collected at IU, available upon request
- Jure Leskovec's SNAP provides a C++ network analysis and graph mining library and a collection of about 50 large datasets including social networks, web graphs, internet networks, citation networks, communication networks, etc.
- Common Crawl provides open Web crawl data
- ICWSM 2011 provides links to various datasets including two from Spinn3r.com (a newer dataset of 386 million elements and an older one with 44 million blog posts), a Sentiment Corpus, and a Wikipedia User Contribution Dataset
- infochimps provides lots of APIs and datasets, many for free
- The UMBC Splog Blog Dataset can be used for research on blog spam
- Wikimedia provides downloads of all Wikipedia projects
- Data Science Toolkit: Various APIs for geolocation, text/HTML analysis and extraction
- JavaCrawlers: A Java library for topical crawlers
- Nutch: an open-source web search engine
- Jakarta Lucene: a high-performance, full-featured text search engine written in Java
- Lemur: a Toolkit for Language Modeling and Information Retrieval
- Clair Library: intended to simplify a number of generic tasks in Natural Language Processing (NLP) and Information Retrieval (IR)
- WIRE: Web IR Environment including a simple format for storing a collection of web documents, a crawler, and tools for generating stats and reports
- Terrier: modular software platform for the rapid development of large-scale Web IR applications, providing indexing and retrieval functionalities; Labrador is a distributed web crawler designed to be integrated with Terrier
- Network analysis tools: consider Gephi, iGraph, NWB, and many more
- Google Code offers a large number of APIs and tools
- Yahoo! Developer Network offers many APIs and tools
- Bing Developer offers Search and Maps APIs
- LETOR: Benchmark Datasets for Learning to Rank from Microsoft Research Asia
- The Boost Graph Library (BGL): a generic C++ library of graph algorithms developed at the Open Systems Lab in the IU CS department. It handles large graphs nicely and integrates (fairly) easily with existing code.
- WebGraph: a Java framework to study the web graph; WebGraph++ is a C++ port that bypasses some limitations imposed by the JVM
- Weka: Data Mining Software in Java

- WebBase: The Stanford WebBase project investigates various issues in crawling, storage, indexing, and querying of large collections of Web pages
- LWP: The World-Wide Web library for Perl
- Libwww: the W3C Protocol Library
- Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering
- MG: an open-source indexing and retrieval system for text, images, and textual images
- WebGlimpse: search engine software including a web administration interface, remote link spider, and the powerful Glimpse file indexing and query system
- ht://Dig: a complete world wide web indexing and searching system for a domain or intranet
- SWISH-E: a fast, powerful, flexible, free, and easy to use system for indexing collections of Web pages or other files
- Internet Archive: a digital library of Internet sites and other cultural artifacts in digital form, providing free access to researchers and scholars (see also Heritrix, the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project)
- Classifier Code (download): a collection of example classifier code written in Matlab, donated by Mark Meiss
- Software and utilities from Soumen Chakrabarti
- Kevin Chai's Homepage links to lots of datasets
- ... and of course many other data APIs are available from hundreds of services such as Twitter, Last.fm, Flickr, NYTimes, YouTube, etc.