Using Matlab for LSA

Introduction to Information Retrieval
CS 221
Donald J. Patterson

Learning Objective

“Be able to use MATLAB to conduct LSI analysis on your own data”

What is MATLAB?

- A numerical computing environment
- An interpreter for a specialized programming language
- Many libraries for complex mathematical operations
- Support for:
  - Matrix Operations
  - Graphing
  - User Interfaces
- Great for rapid prototyping complex algorithms
- Cross -platform

What MATLAB isn’t

- A production ready commercial software development tool
- Free
- Open-source

Where is MATLAB at UCI?

- CS 364 Lab - about 30 machines with one machine licenses
- NACS PC Labs - “mpc cluster”
- Remote access through openlab if you buy a license and tell ICS support.
- Student edition is about $100.00
- Open-source alternative called “octave” is available.

http://www.ics.uci.edu/~smyth/courses/matlab.html

Using MATLAB For LSA

Demo

- 6 documents
  - Wikipedia entry for “baseball bat”
  - Wikipedia entry for “bat”
  - Wikipedia entry for “coffee”
  - Wikipedia entry for “starbucks”
  - Starbucks’ home page
  - First page of a recent publication of mine

Demo

- I pulled out 14 words
  - BALL
  - BASEBALL
  - BAT
  - CALIFORNIA
  - COFFEE
  - COMPANY
  - ENCYCLOPEDIA
  - IRVINE
  - RUN
  - SPECIES
  - STARBUCKS
  - STORES
  - UNIVERSITY
  - USERS

Using MATLAB For LSA

Demo

- Calculate the TFIDF score
- Plot the documents on a two term axis
- Perform SVD decomposition
  - Validate decomposition
- Reduce rank of system
- Show “M”
  - Demonstrate what SVD is capturing
- Execute a query

LUCI