INF 141
COURSE SUMMARY

Crista Lopes

Lecture Objective

Know what you know

Problem Space of this course

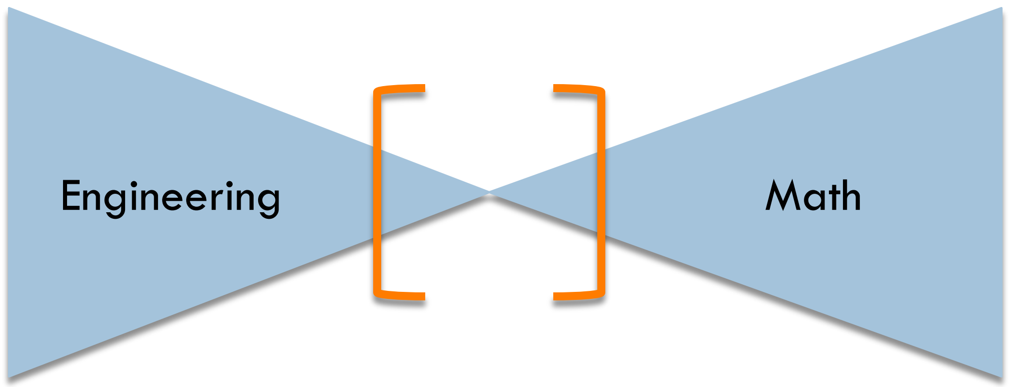
- “Big Data”
- How to
  - collect it
  - index it
  - search it for relevant information

Industry segment of this course

- Search engines
  - Google
  - MS Bing
  - nameless others

- Web information retrieval is big $$$

Technical content of this course



Lecture 2

- Search engines history
- Search & advertising on the Web
- Web corpus

Lecture 3

- Characteristics of the Web
  - duplication, linkage, spam
  - how big
  - rate of change
  - evolution
- Characteristics of Web search Users
  - reasons for searching
  - characteristics of queries used
  - behavior towards results
  - the need behind the query

Lecture 4

- Search Engine Optimization

Lectures 5 and 6

- Web crawling
  - architecture of a crawling infrastructure
  - algorithms
  - constraints

Lectures 7, 8 and 9

- Index construction
  - what index is
  - efficient data structures
  - efficient algorithms for constructing it

Lecture 10

- Map Reduce
- Index compression

Lecture 11

- Retrieval
  - boolean
  - zones
  - TF metrics

Lecture 12

- Ranked Retrieval
  - weighting fields

Lecture 13

- Better scoring
  - TF-IDF
  - Corpus-wide statistics

Lecture 14

- Vector Space model
- Score by magnitude (euclidian distance)
- Score by angle (cosine distance)

Lecture 15

- Language statistics
- Language processing
  - tokenizing
  - stemming
  - stopping
- Link analysis
  - PageRank

Lecture 16

- Hadoop

Lecture 17

- Latent Semantic Analysis
  - Singular Matrix Decomposition

Lecture 18

- Retrieval on LSI

- Use of Latent Dirichlet Allocation (LSA)

All together

- Search engines history
- Search & advertising on the Web
- Web corpus
- Characteristics of the Web
- Characteristics of Web search Users
- Search Engine Optimization
- Web crawling
- Index construction
- Index compression
- Map Reduce
- Boolean retrieval
- Parametric retrieval
- Scored retrieval
- TF-IDF and corpus-wide statistics
- Language statistics
- Language processing
- Link analysis (PageRank)
- Hadoop
- LSI (and LDA)

Big Data jobs

- plenty…
- not just traditional search
  - making sense of data
- search on google

Where to go from here

- Data mining
- Machine learning