Introduction to IR and IE

Vasudeva Varma

Goal of this course

- To introduce and provide *hands-on exposure* in the areas related to *Information Access*.

- To provide necessary background for potential *research* students in the areas of Information Access technologies.

Course Topics and Roadmap

- **Introduction (4)**
- **IR Fundamentals (9)**
  - **Models,**
  - **Scoring functions**
  - **Index design**
  - **Crawling**
  - **IR Evaluation**
- **NLP/Text Mining for IR (4)**
- **Machine Learning & IR (9)**
- **Information Extraction (3)**
  - **IE Fundamentals**
  - **Named Entity Recognition**
- **Information Access and IR Applications (9)**
  - **Summarization**
  - **Social Computing**

Instructors

- Vasudeva Varma
- Manish Gupta
- Niyati Chaya

- Tutorials:
  - Senior Research (PhD&MS) students of IREL
  - Formal TAs

Course Administration

- Teaching Assistants:
  - Teaching Associate: Himanshu Maheshwari
  - Mentors: All PhD/Senior MS Students of IREL
  - Teaching Assistants: Sayar Ghosh, Tanmay Sachan, …
- Tutorial: Every Thu/Fri ??
- TA Office Hours: TBA

Grading

- Quiz/In Class Activities: 10%
- Assignments 15%
- Project 60% (20+40)
- Term Paper 15%

Text/Reference Books/material

- Introduction Information Retrieval – Chris Manning et al (the Stanford IR Book)
- Search Engines IR in Practice – Bruce Craft et al

Projects

- Mini Project – Individual Project (4 Weeks)
  - Two deliverables
- Major Project – Teams with 3-5 members each (10 weeks)
  - Three deliverables

Mini Project – 4 Weeks

- Objective: Design and develop a scalable and efficient search engine using the Wikipedia data.
- Features:
  - Dump of Wikipedia as document repository
  - Results obtained in less than a sec (even for long queries)
  - Supports field queries (ex: title)
  - Index size should be less than 1/4 of the data size.
  - You have to build your own indexing mechanism
    - i.e. you cannot use Nutch or Lucene to index the Wikipedia data.
- Platform:
  - OS: Preferably Linux
  - Languages: Java/C++/Python

Mini Project evaluation

- The evaluation will be done on 4 parameters:
  - Search time,
  - Search efficiency
  - Indexing time
  - Index Size
- You can use compression techniques
- Explore several ranking functions (tf,tf-idf, normalized tf, normalized idf etc) and
- Create a secondary index if required.

Mini Project Deadlines

- First evaluation: 3rd September
  - Indexing time and efficiency will be evaluated.
- Second (Final) evaluation: 17th September
  - Dummy queries will be provided before September 10th.
  - All four evaluation parameters will be considered

Major Project (10 Weeks)

- Team project (4 members) – Constrained choice
- Advanced topics
- Well defined project
- Major implementation component
- Three deliverables
  - Scope document Oct 1st
  - End-to-end system – MVP Oct 29th
  - Complete system - Demo/presentation Video, Code, Report… Nov 21st

Follow the course on…

Web: http://moodle.iiit.ac.in

Information-Retrieval-Course-at-IIIT-Hyderabad

thank you

Vasudeva Varma

@devvarma

vv@iiit.ac.in www.iiit.ac.in/~vv