Introduction

• Fernando Diaz
  – Ph.D. in CS, 2007, University of Massachusetts Amherst
  – Research Scientist, Microsoft Research
  – Interests: Information Retrieval, Information Extraction, Time-Sensitive Information
  – http://ciir.cs.umass.edu/~fdiaz/

• Cong Yu
  – Ph.D. in CSE, 2007, University of Michigan
  – Research Scientist, Google Research
  – Interests: Information Extraction, Social Data Mining, Structured Data Management
  – http://www.eecs.umich.edu/~congy/
Outline

Website: http://cs.nyu.edu/courses/fall16/CSCI-GA.2580-001/

• Course Overview
  – Topics
  – Homework / Project
  – Exam
• Course Logistics
  – HW0
• Introduction to Information Retrieval
Course Topics

- **Divided into two parts**
  - Part 1: Fundamentals of search engines (and information retrieval)
  - Part 2: Advanced topics in modern day search engines

- **Fundamentals**
  - Textbook-based, we will cover the basic building blocks underneath all search engine implementations

- **Advanced topics**
  - Research-oriented (i.e., papers), we will cover some of the most interesting and cutting edge topics on going at academia and (more importantly) industry.
Fundamentals of Search Engines

- Server & UI
- Log
- Evaluator
- Miner
- Index / Corpus
- Crawler
- Indexer
- Doc Store
- Ranker

Dashed box: online
Advanced Topics

• **Common components:**
  – Advertising
  – Personalization

• **Advanced features:**
  – Real time
  – Big Data
  – Machine Learning
  – Knowledge / Structured Data

• **For each topic, we will study**
  – Why it is important
  – What are the technical challenges
  – What are the cutting edge solutions
  – What are the future directions
Homework / Project

• Similar to the course material, the project is divided into two stages
  – First: implementation of a basic search engine through a series of 4 homeworks (3 graded ones)
  – Second: design and implementation of any advanced component of your choosing on top of the basic search engine you build through the homeworks

• Involve heavy programming!
  – Requirements: comfortable with Java, some understanding of server, familiar with data structures and fundamental algorithms

• WARNING:
  – The course is designed based on similar ones taught at other schools and frequently getting complained as one of the most difficult courses, so be prepared!
Homeworks at a Glance

HW0 (not graded)
- Server & UI
- Log
- Ranker
- Doc Store

HW1
- Evaluator
- Index / Corpus
- Miner

HW2
- Crawler
- Indexer

HW3
Advanced Component

• Pick from one of the advanced topics covered in the second half of the lectures
  – Decide on the scope of your solution (i.e., to what extent you will be addressing the challenges, what assumptions you will be making, and what limitations you will allow)
  – Design and architect your system on top of your basic search engine
  – Implementation

• You may also choose another topics not covered in the lectures, but you must obtain our approval first
Exams

• Midterm will be take home and cover materials taught up to and including Lecture 6.
  – Must complete the exam *alone*: no discussion allowed with others, including those in the same project group.
  • We caught several cheaters last time, so don’t try
  – We will try a paperless process, stay tuned.

• **Final will be on December 19**\(^{th}\), **2 hours**, and cover materials taught through the *entire* course.
  – Open book, but not laptops.

• **No class on 12/05**
Logistics

• **Time and Location**
  – Mondays 5:10p – 7:00p, CIWW 312
  – We will usually have a 5 minute break mid-class

• **Office Hour**
  – Mondays 4:00p – 5:00p before each class, CIWW 328
  – No class or office hour on 12/05

• **Teaching Assistants**
  – Yanghui Zhuang: yanghui.zhuang
  – Qi Feng: qf264

• **Website**
  – [http://cs.nyu.edu/courses/fall16/CSCI-GA.2580-001/](http://cs.nyu.edu/courses/fall16/CSCI-GA.2580-001/)

• **Mailing List**
  – csci.ga.2580.001.fa16 [AT] cs dot nyu dot edu
Grading

• Not curve based
  – You do your job well, you will get a good grade

• Group work 50%
  – 30% for homeworks: 10% each, HW0 is not graded
  – 10% for the project report
  – 10% for a 15-min demo with us (to be scheduled)
  – Your contribution to the project will be judged by your group members and communicated to us.

• Exams 40%
  – Midterm (take home): 15%
  – Final (in class): 25%

• Class participation 10%
  – Mainly used to reward students who engage actively in the class, sometimes can be used to punish those with poor attendances 😊
Schedules (see course homepage)

• **HW1 will be assigned on 9/26 and due 10/17.**
  – A basic retrieval engine with a provided corpus index.
  – Extra week for group formation, spans fall recess, don’t procrastinate!!

• **HW2 will be assigned on 10/17 and due 10/31.**
  – A basic crawl and indexing pipeline over a small Web corpus.

• **Midterm will be assigned on 10/31 and due 11/7.**
  – Covers lectures 1-6.

• **HW3 will be assigned on 11/7 and due 11/21.**
  – A basic query miner to improve search quality.

• **Project due 12/14.**
  – 3 weeks to work on the advanced component: be wise with your time!
  – Report due on 12/14 at 9am and code due on 12/17 at 9am.
  – A 15-min demo should be arranged with us for one of the slots from 12/14 to 12/16.

• **Final on 12/19 in class.**
Homework 0

• **Due two weeks from today!**
  – Handout on the course homepage
  – Implement a simple Java server that echoes the user query.

• **Start thinking about your group**
  – Sign up your group at the online form:
    https://docs.google.com/forms/d/e/1FAIpQLSe6Li3cCme19XaaNk7rrWRX0JnGhLcnzPzualvurq5VYPrufw/viewform
Any Questions?