COSC-589
Web Search and Sense-making
Information Retrieval In the Big Data Era

Spring 2016

Instructor: Grace Hui Yang

The Web provides abundant information which allows us to live more conveniently and make quicker decisions. At the same time, the growth of the Web and the improvements in data creation, collection, and use have lead to tremendous increase in the amount and complexity of the data that a search engine needs to handle. The increase of the magnitude and complexity of the data has become a major drive for new data analytics algorithms and technologies that are scalable, highly interactive, and able to handle complex and dynamic information seeking tasks in the big data era. How to effectively and efficiently search for the documents relevant to our information needs and how to extract the valuable information and make sense out from "**big data**" are the subjects of this course.

The course will cover Web search theory and techniques, including basic probabilistic theory, representations of documents and information needs, various retrieval models, link analysis, classification and recommender systems. The course will also cover programming models that allow us to easily distribute computations across large computer clusters. In particular, we will teach Apache Spark, which is an open-source cluster computing framework that has soon become the state-of-the-art for big data programming. The course is featured in step-by-step weekly/bi-weekly small assignments which composes a large big data project, such as building Google's PageRank on the entire Wikipedia. Students will be provided knowledge to **Spark, Scala, Web search engines, and Web recommender systems** with a focus on search engine design and "thinking at scale".

Prerequisites: None.

Course Website: https://piazza.com/georgetown/spring2016/cosc589/home

Class Time: Monday and Wednesday 1100-1215
Venue: Intercultural Center 205
Blackboard: Homework submission

Textbooks:

Information Retrieval: Implementing and Evaluating Search Engines
By Stefan Büttcher, Charles L. A. Clarke and Gordon V. Cormack
Publisher: The MIT Press
July 2010

Learning Spark: Lightning-Fast Big Data Analysis
By Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia
Publisher: O'Reilly Media

January 2015

Programming in Scala, Second Edition. A Comprehensive Step-by-Step Guide
By Martin Odersky, Lex Spoon, and Bill Venners
Publisher: Artima
January 2011

Lecturer:

Prof. Grace Hui Yang
Email: gh243@georgetown.edu
Office Phone: 202 687 6355
Website: http://www.cs.georgetown.edu/~huiyang

Teaching Assistants:

Jiyun Luo (jl1749@georgetown.edu)
Website: http://cs-class.uis.georgetown.edu/~jl1749/homepage/

Office Hours:

Prof. Grace Hui Yang
St Mary's Hall 338
Tuesday 1-2pm

TA hours
St Mary's Hall TA room 330
Wednesday evening 6-7:30pm

Grading:

- Midterm Exam 15%
- Assignments 80%
- Pop Quizzes 5%
- Bonus - Optional Assignment Problems 20%

Homework policy:

All homework should be submitted through Blackboard. Homework is due by 11:59pm on the due date. Three late days in total are allowed without penalty for the entire semester. For instance, you may be late by 1 day for homework 1 and be late by 2 days for homework 2. Once the three-late-dates are used, you will be penalized according to the policy below:

- a penalty of 50% will be applied for homework submitted within the next 24 hours.
- zero credit will be assigned after that.

All homework (even with zero credit) must be turned in order to pass the course.

Integrity policy:

All experimental results turned in must be true. No copying/cheating is allowed. Please check Georgetown's Honor system.

	1/13/16 Lecture: Introduction Spark Installation Readings: Chapter 1, 2 Assignment 1: Due 1/20/16 Install Spark
1/18/16 Break: Martin Luther King Day	1/20/16 Lecture: Scala Crash Course Readings: Scala Book Ch 2,3,7,8 Assignment 2: Due 1/27/16 Scala practice Due: Assignment 1
1/25/16 No class	1/27/16 Lecture: Scala Crash Course SBT Passing functions Readings: Scala Book Ch 2,3,7,8 Chapter 2 Assignment 3: Due 2/3/16 Scala practice Due: Assignment 2

2/1/16 Lecture: RDD File Operations Readings: Chapter 3 Chapter 5	2/3/16 Lecture: RDD Transformation Actions Spark UI Readings: Chapter 3 Chapter 7 Assignment 4: Due 2/10 Word Count Due: Assignment 3
2/8/16 Lecture: RDD Transformation Actions Write SBT Build File Readings: Chapter 3 Chapter 2	2/10/16 Lecture: RDD Persistence Regular Expression Readings: Chapter 3 Scala Book Section 26.7 Assignment 5: Due 2/17 Extract Patterns Due: Assignment 4

2/15/16 Break: President Day	2/17/16 Lecture: Text Processing Bag of Words Query and Documents Vector Space Model Readings: IR Book Chapter 2 IR Book Chapter 3 Assignment 6: Due 3/2 Download Wikipedia Dump Clean the dataset Due: Assignment 5
2/22/16 Break: WSDM Week	2/24/16 Break: WSDM Week
2/29/16 Lecture: Accumulators Broadcast Variables Readings: Chapter 6	3/2/16 Lecture: XML Parser Readings: Scala Book Chapter 28 Assignment 7: Due 3/14 Parse the Wikipedia data using Scala.xml Word Count Due: Assignment 6
3/7/16 Spring Break	3/9/16 Spring Break
3/14/16 Lecture: Midterm Review Due: Assignment 7	3/16/16 Midterm Exam (in class, closed book)

3/21/16 Break: ECIR	3/23/16 Easter Break
3/28/16 Easter Break	3/30/16 Lecture: Web Search Link Graph Readings: IR Book Chapter 15 Assignment 8: Due 4/6 Build Wikipedia Link graph Write to a file
4/4/16 Lecture: PageRank Key/Value Pairs Readings: Chapter 4 IR Book Chapter 15	4/6/16 Lecture: PageRank Readings: IR Book Chapter 15 PageRank paper Assignment 9: Due 4/20 Compute PageRank for Wikipedia pages Due: Assignment 8
4/11/16 Break: WWW Conference	4/13/16 Break: WWW Conference

4/18/16 Lecture: Web Knowledge Base Topic-Oriented PageRank Readings: IR Book Chapter 15 Papers	4/20/16 Lecture: Recommender Systems Readings: Chapter 11 IR Book Section 10.1 Assignment 10: Due 5/2 Compute Topic-oriented PageRank for Wikipedia Due: Assignment 9
4/25/16 Lecture: Regression & Classification Readings: Chapter 11 Papers	4/27/16 Lecture: Advanced Search Readings: Papers Assignment 11 (optional): Due 5/14 Secret
5/2/16 (Class End) Lecture: Conclusion Due: Assignment 10	5/7/16 Study Week
5/9/16 Exam Week	5/14/16 Exam Week Due: Assignment 11

See Sample Lecture Notes on the Next Page.

It is also available at http://people.cs.georgetown.edu/~huiyang/cosc-589/intro.pdf

COSC589
WEB SEARCH AND SENSE-MAKING
INFORMATION RETRIEVAL IN THE BIG DATA ERA

SPRING 2016

1

WHAT IS SEARCH (TO YOU)?

2

LET'S PRACTICE SOMETHING YOU DO EVERY DAY...

3

EXERCISE

- Find what city and state Dulles airport is in, what shuttles ride-sharing vans and taxi cabs connect the airport to other cities, what hotels are close to the airport, what are some cheap off-airport parking, and what are the metro stops close to the Dulles airport.

4

HOW DO YOU SEARCH?

5

- Recognizing a need for information,
- Accepting the challenge to take action to fulfill the need,
- Formulating the problem,
- Expressing the information need in a search system,
- Examination of the results,
- Reformulation of the problem and its expression, and
- Use of the results.

6

SEARCH IS ALSO KNOWN AS INFORMATION SEEKING

7

"Information-seeking is a special case of problem solving. It includes recognizing and interpreting the information problem, establishing a plan of search, conducting the search, evaluating the results, and if necessary, iterating through the process again."

–GARY MARCHIONINI, 1989

8

TWO MODES OF SEARCH

9

1. SEARCH BY QUERY

10

SEARCH BY QUERY

- The user issues a textual query
  - types the query in a search box

- After examining the search results, the user
  - either stops the search
  - or modifies the query to start another query

11

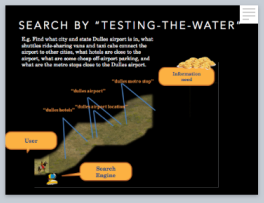
- … it is like testing the water …

12

13

14

SEARCH BY "TESTING-THE-WATER"



15

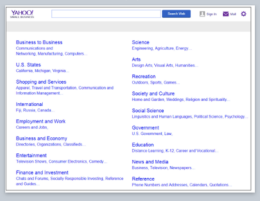
2. SEARCH BY BROWSING

16

SEARCH BY BROWSING

- The system can provide the searcher with structure that characterizes the available information
- The searcher (user) browses or navigate the organized information

17

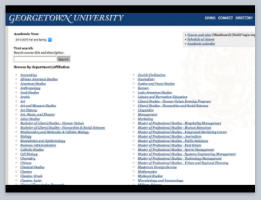


18

19

- … but you are still using browsing …

20



21

QUERYING VS. BROWSING

- Search queries tend to produce new, ad hoc collections of information that have not been gathered together before,
- Navigation/browsing refers to selecting links or categories that produce pre-defined groups of information items.
- Browsing also involves following a chain of links, switching from one view to another, in a sequence of scan and select operations.
- Browsing can also refer to the casual, mainly undirected exploration of navigation structures.

22

23

HTTP://INFOSENSE.CS.GEORGETOWN.EDU/

A QUICK INTRODUCTION OF ME

24

WHAT BRINGS ME HERE

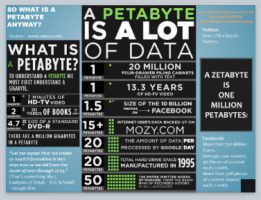
- To have a wonderful semester with you :-)
- As an educator, I want to teach a class that is timely and useful, helping you in the job market
- Search relates to everyone!!!

25

- … and there are new challenges and new opportunities in this big data era…

26

27



28

- For myself, to be honest, I am not a big fan of huge data volume. However, big data not just means bigger volume, it also means higher data variety and faster data change rate (velocity)
- I am a fan of complexity. ;-)

29

COURSE PURPOSE

- Able to code and design large scale data analytics tools
  - Master spark programming
  - Understand how web search engine works
- Focus on text analytics, but the techniques we will learn are generic
- Have fun!

30

31

IF YOU ARE THE GOD OF DATA

- What are the typical uses of your data?
  - understand trends and patterns
  - prediction
  - search

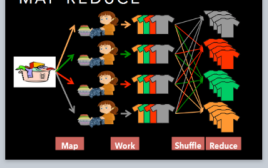
32

IF YOU ARE THE GOD OF DATA

- What will be the challenges/problems when your data is big?
- What is your solution?
  - divide-and-conquer
  - parallelization
  - compression

33

MAP REDUCE



34

MAPREDUCE IS A LITTLE BIT OUTDATED

- It is great at one-pass computation
- But not efficient enough for multiple-pass algorithms
  - Things that require repeatedly hashing or other operations
- States go to file systems
  - A lot of I/Os
  - Slow

35

SPARK

- Key idea:
  - Load things in the memory
  - Resilient Distributed Datasets (RDDs)
- Clean APIs in Java, Scala, Python, R
  - not for c++
  - We will learn Scala

36

COURSE PLAN

- Key topics
  - Spark
  - Web search engine
- January - Scala, Set You Up
- February - Spark Essentials
- March - Web Search
- April - PageRank and Other Apps (Recommender Systems, Dynamic Search, Regression, Classification)

37

"A homework is worth a thousand lectures."

–GRACE HUI YANG

38

ASSIGNMENTS

- Build Google's pagerank algorithm over Wikipedia
- A big project broken into small pieces
- (Nearly) weekly
- 10 + 1 of them
- (Almost) all due on Wednesday 11:59PM

39

40

HIGHLIGHT OF TODAY

- Install Spark
- First Spark Program

41

SPARK INSTALLATION

- Note: Please do not install/run Spark using:
  - Homebrew on MacOSX
  - Cygwin on Windows

42

STEP 1 - INSTALL JAVA JDK 6/7 ON MACOSX OR WINDOWS

- oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html
  - Follow the license agreement instructions
  - Then click the download for your OS
  - Need JDK instead of JRE (for Maven, etc.)

43

STEP 2: GET SPARK

- We will use Spark 1.1.0
- 1. copy from the USB sticks
- 2. connect into the newly created directory
- Or you could download from spark.apache.org/downloads.html

44

STEP 3: RUN SPARK SHELL

- We'll run Spark's interactive shell…
- Within the "spark" directory, run:
- ./bin/spark-shell
- Then from the "scala>" prompt,
- Let's create some data…
- val data = 1 to 10000

45

STEP 4: CREATE AN RDD

- Create an RDD based on that data…
- val distData = sc.parallelize(data)
- Then use a filter to select values less than 20…
- distData.filter(_ < 100).collect()

46

ASSIGNMENT 1

- Use a filter to select values less than your age
- Submit the screen captures of the results of above programs

47

SUMMARY

- Search engine
- Spark
- Syllabus
- Installation of spark
- Assignment 1
  - due next Wednesday
- No class on Monday

48