CS3245

Information Retrieval

0

Lecture 0: Course Organization

Why should you care about

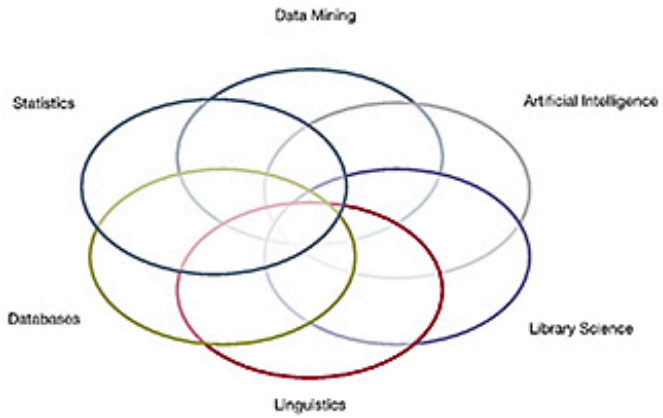
INFORMATION RETRIEVAL?

Text Mining: The Next Data Frontier

Mon, 01/06/2014 - 2:04pm

by Mark Anawis

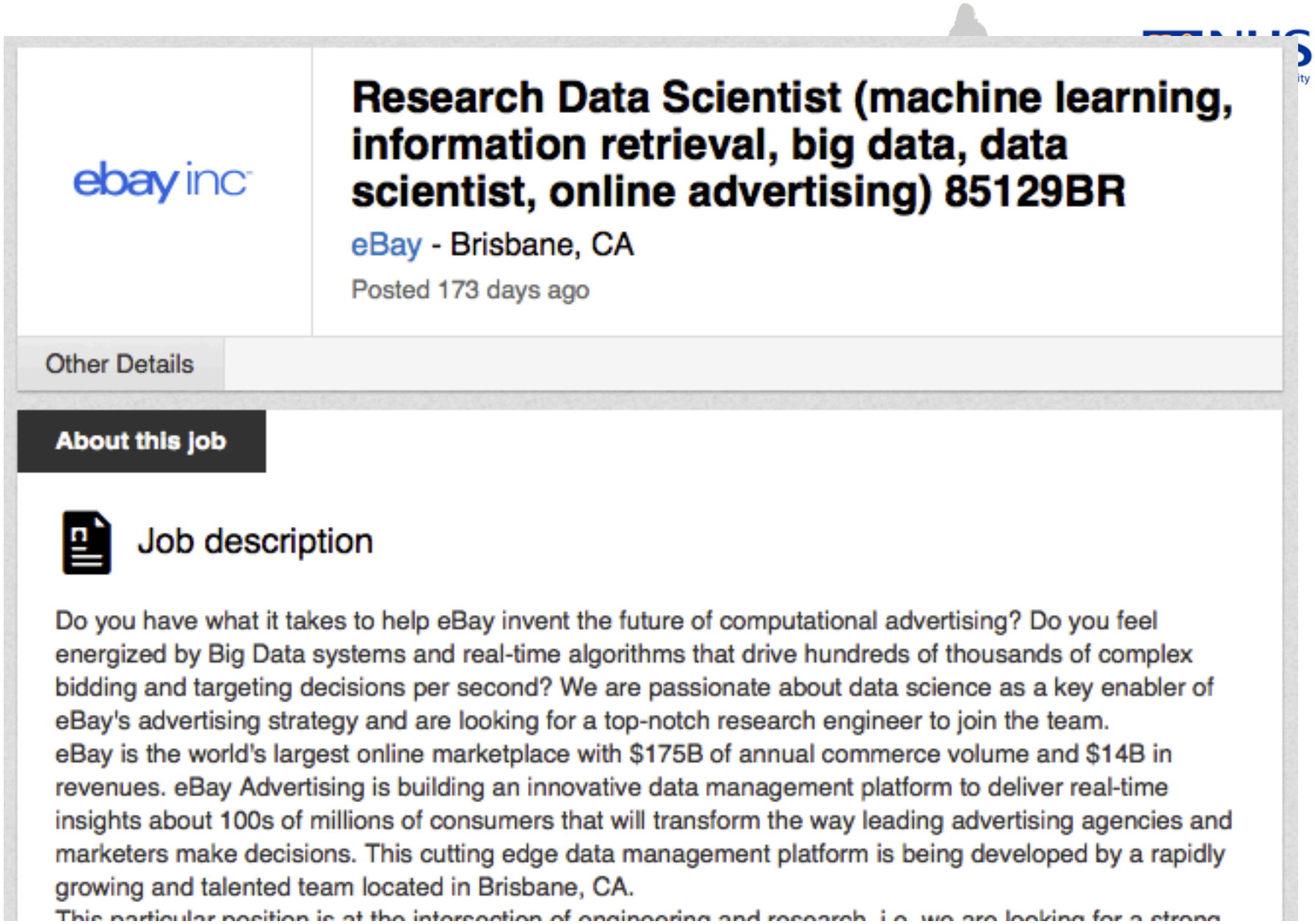
By some estimates, 80 percent of available information occurs as free-form text



Text Mining: The Next Data Frontier
Figure 1: Text Mining and Related Fields

Josiah Stamp said: "The individual source of the statistics may easily be the weakest link." Nowhere is this more true than in the new field of text mining, given the wide variety of textual information. By some estimates, 80 percent of the information available occurs as free-form text which, prior to the development of text mining, needed to be read in its entirety in order for information to be obtained from it. It has been applied to spam filters, fraud detection, sentiment analysis, identification of trends and authorship.

Text mining can be defined as the analysis of semi-structured or unstructured text data. The goal is to turn text information into numbers so that data mining algorithms can be applied. It arose from the related fields of data mining, artificial intelligence, statistics, databases, library science, and linguistics (Figure 1).



ebay inc

Research Data Scientist (machine learning, information retrieval, big data, data scientist, online advertising) 85129BR

eBay - Brisbane, CA

Posted 173 days ago

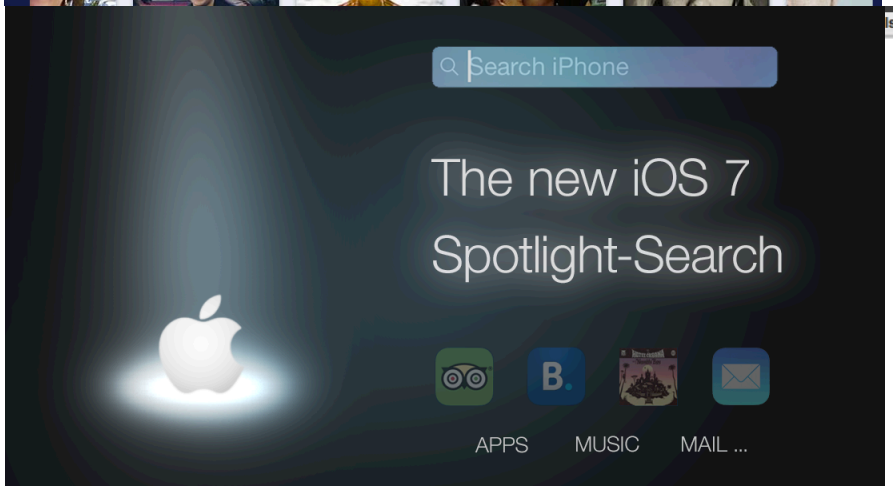
Other Details

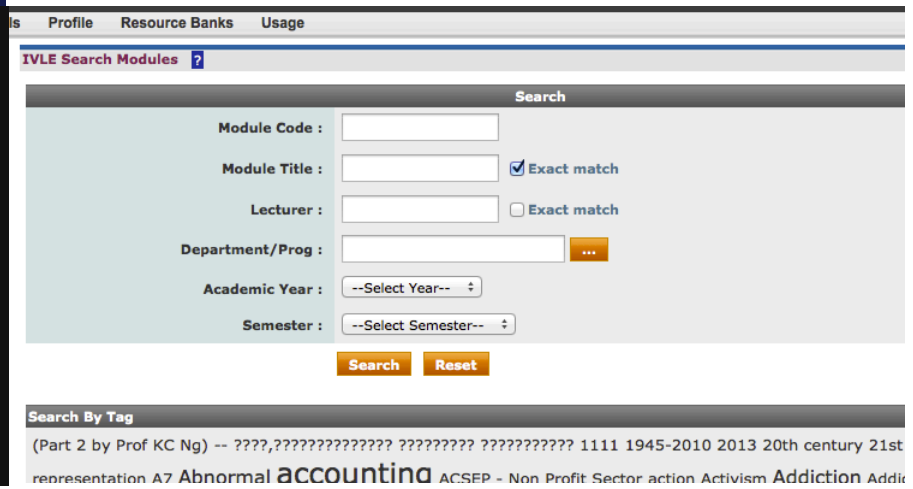
About this job

Job description

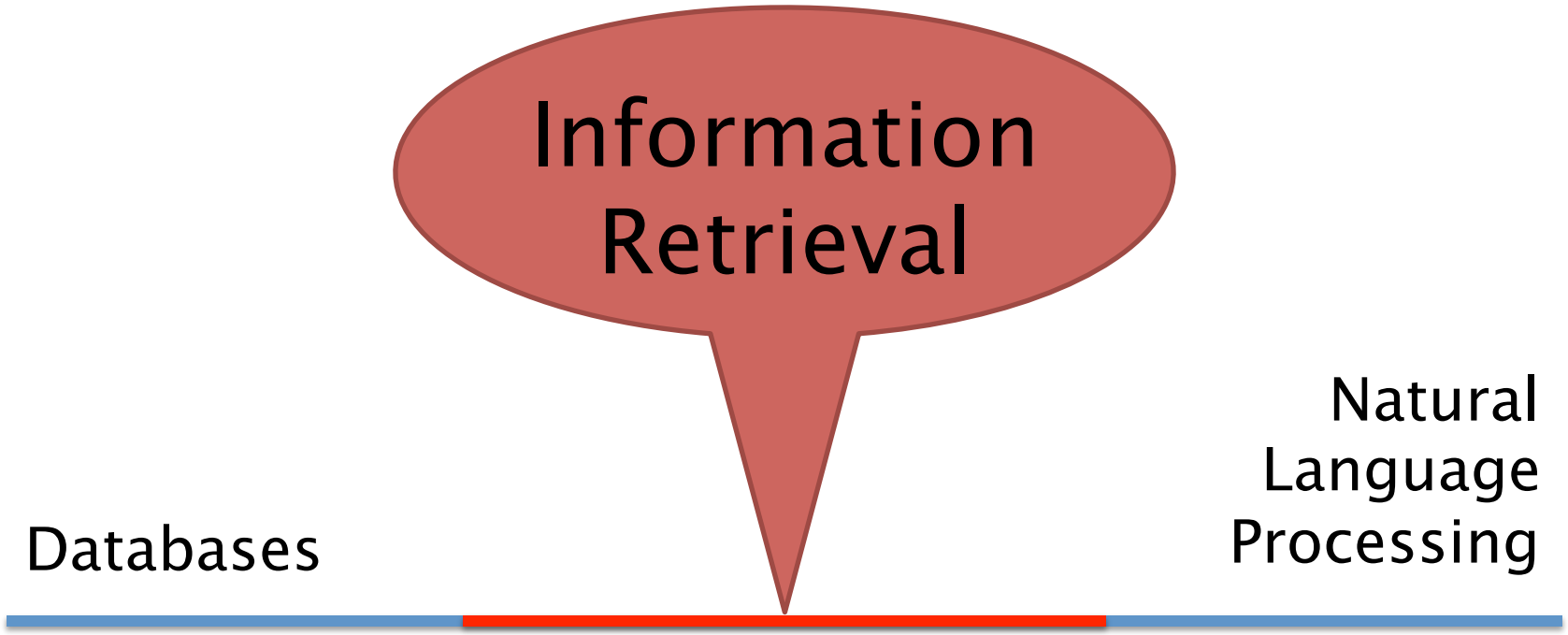
Do you have what it takes to help eBay invent the future of computational advertising? Do you feel energized by Big Data systems and real-time algorithms that drive hundreds of thousands of complex bidding and targeting decisions per second? We are passionate about data science as a key enabler of eBay's advertising strategy and are looking for a top-notch research engineer to join the team.
eBay is the world's largest online marketplace with $175B of annual commerce volume and $14B in revenues. eBay Advertising is building an innovative data management platform to deliver real-time insights about 100s of millions of consumers that will transform the way leading advertising agencies and marketers make decisions. This cutting edge data management platform is being developed by a rapidly growing and talented team located in Brisbane, CA.
This particular position is at the intersection of engineering and research, i.e. we are looking for a strong

A keystone of today's tech





Right at the center of attention



How is it structured?

NUS National University of Singapore

And now for the

COURSE ORGANIZATION

Lecturer

Min-Yen KAN

(as in $min(x)$)

kanmy@comp.nus.edu.sg

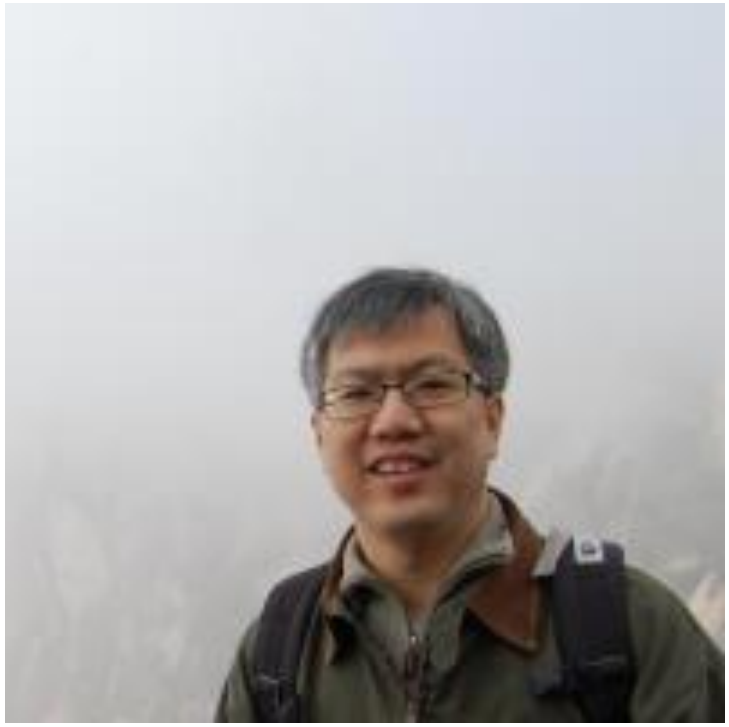
http://www.comp.nus.edu.sg/~kanmy

Office: AS6 05-12

6516-1885

Hours: 13:00-14:00 Fri,
or by appointment

Hobbies:
Taking care of a little one,
and soon to be two



Teaching Assistant

Xiangnan HE

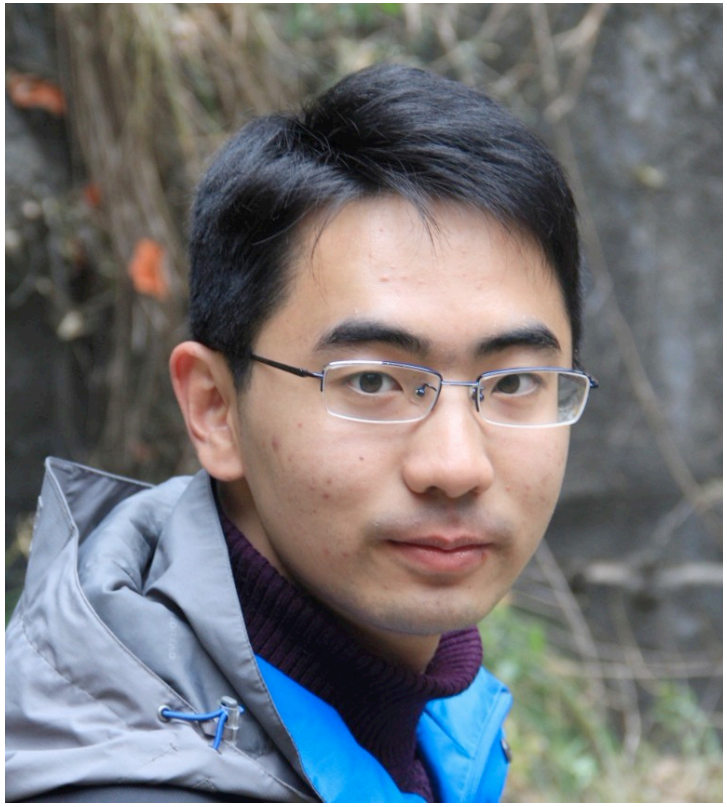
xiangnan@comp.nus.edu.sg

http://www.comp.nus.edu.sg/~xiangnan

Office: AS6 #04-13
(Media Research Lab 2)

Hours: TBA and by appointment

Hobbies: Table Tennis, among others



Course web sites

http://www.comp.nus.edu.sg/~cs3245

- Homework
- Other supplementary content

When using a search engine to find our course materials, make sure you find our site for 2013/14 Sem II.

http://ivle.nus.edu.sg/

- Lecture notes (via workbin)
- Discussion forum (participation counts!)
  - Any questions related to the course should be raised on this forum
  - Emails to me are considered public unless otherwise specified
- Announcements
- Homework submissions (via workbin)

Freedom of information rule

- Collaboration is acceptable and encouraged.
- To assure that all collaboration is on the level, **you must always fill in the name(s) of your collaborators on your assignment.**
- You will be assessed for the parts for which you claim is your own contribution.

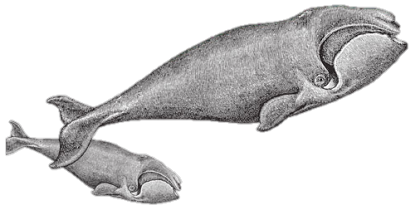
Facebook rule

- You are free to meet with fellow students(s) and discuss assignments with them.
- Writing on a board or shared piece of paper is acceptable during the meeting; however, you **may not take any written (electronic or otherwise) record away from the meeting**.
- After the meeting, do something else for at least a half-hour (Facebook, or doing an assignment for a different class), before working on the assignment.
  - This will assure that you are able to reconstruct what you learned from the meeting, by yourself, using your own brain.
  - You will be asked to certify that you meet this requirement per assignment.

Python

- We'll be using the Python programming language for our entire class homework assignments.
- You are expected to learn Python and use it for the assignments.
- Don't worry, Python is an easy language to transition to for most programmers and it is very useful for string manipulation (a key characteristic for IR).

NLTK



- We will also be using the Natural Language Tool Kit (NLTK) for helping us deal with some parts of the course.
- It's coded in Python (surprise!)
  - Yes, we know that this is a course on **information retrieval** and not **natural language processing**, but the two are forever intertwined.
- Please install both Python 2.7.x and NLTK on your own personal computer (they recommend 2.7.3)
  - I will try to get a Programming Lab installed with the appropriate Python.

NUS National University of Singapore

- NUS is also opening certain courses and sessions to the public. We may have some external guests in a few of our classes:
  - Today's class (17 Jan) – Wk 1
  - Next week (24 Jan) – Wk 2
    and
  - After Recess Week (7 Mar) – Wk 7

http://www.nus.edu.sg/oam/experiencesNUS/experiences.html