Information Retrieval

LIS 544 IMT 542 INSC 544

Welcome!

Your instructors

- Jeff Huang lazyjeff@uw.edu
- Shawn Walker stw3@uw.edu

Introductions

- Name
- Program, year
- Previous school(s)
- Most interesting thing you did this summer?
- What do you want to get from this class?

Goals for the Class

- Learn fundamentals of information retrieval
- Think about search engines like Larry and Sergey (or Jerry and David)
- Learn skills to apply elsewhere

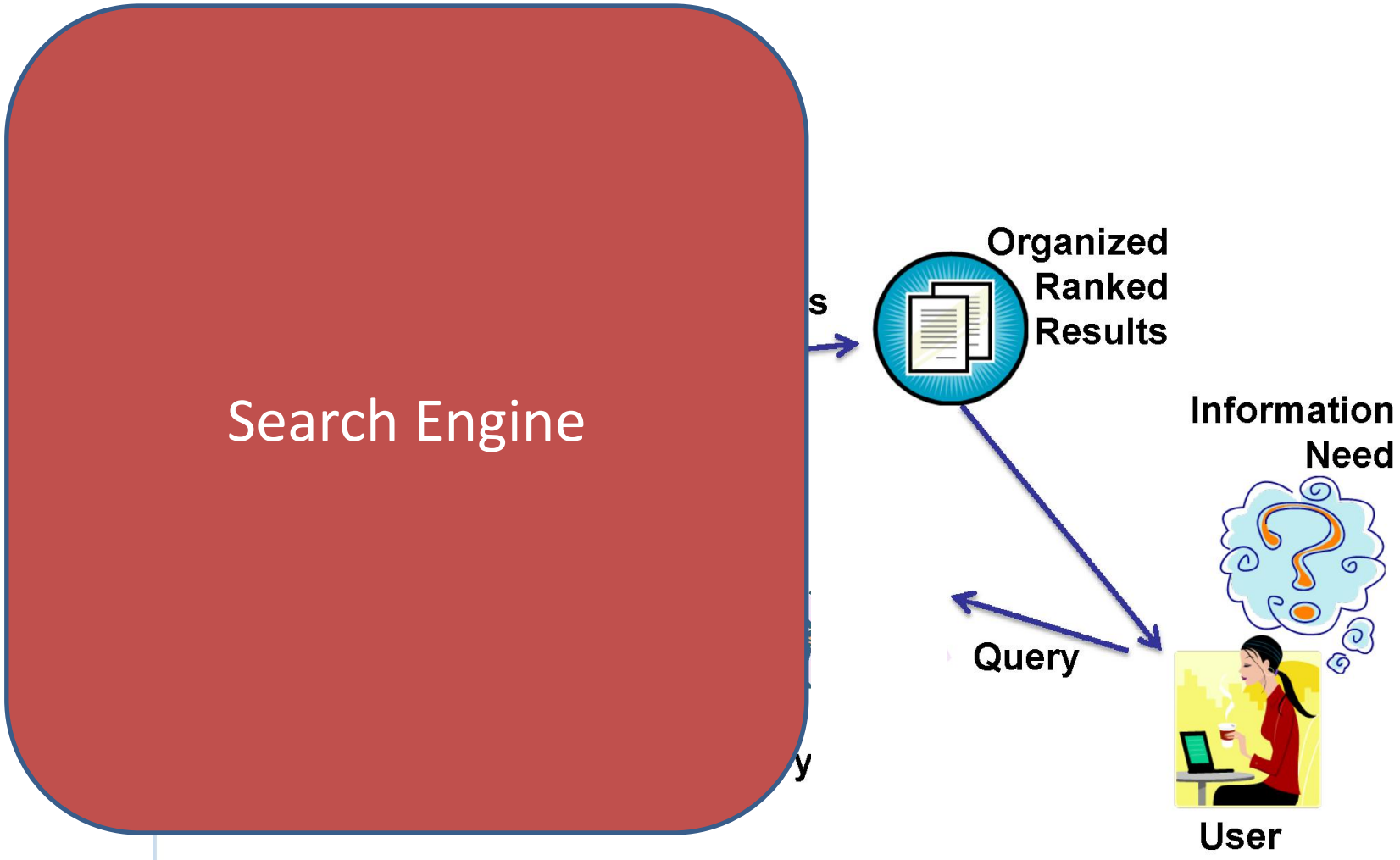
What kind of skills?

- recommendation systems
- measure how much a document changed
- automatic classification
- trend detection
- evaluate how good search results are
- text processing

Administrivia

- Website – http://courses.washington.edu/ir2010
- Calendar – http://courses.washington.edu/ir2010/calendar.html
- Syllabus – http://courses.washington.edu/ir2010/syllabus.pdf

What is Information Retrieval?



Lecture 1: Documents

Document analysis and indexing

History

- The decline of structured data
- Boolean retrieval

You are the search engine

plato Search

DOCUMENT_A

- **Plato** (pronounced /ˈpleɪtoʊ/) (Greek: Πλάτων, Plátōn, "broad")[1] (428/427 BC[a] – 348/347 BC), was a Classical Greek philosopher, mathematician, writer of philosophical dialogues, and founder of the Academy in Athens, the first institution of higher learning in the Western world. Along with his mentor, Socrates, and his student, Aristotle, **Plato** helped to lay the foundations of natural philosophy, science, and Western philosophy.[2] **Plato** was originally a student of Socrates, and was as much influenced by his thinking as by what he saw as his teacher's unjust death.

DOCUMENT_B

- Neuroplasticity (also referred to as brain plasticity, cortical plasticity or cortical re-mapping) is the changing of neurons and the organization of their networks and so their function by experience. This idea was first proposed in 1892 by Santiago Ramón y Cajal the proposer of the neuron doctrine though the idea was largely neglected for the next fifty years.[1] The first person to use the term neural plasticity appears to have been the Polish neuroscientist Jerzy Konorski.[2]

Which document should rank higher?

- The document that contains the query

Both documents have the term

plato Search

DOCUMENT_A

- **Plato** (pronounced /ˈpleɪtoʊ/) (Greek: Πλάτων, Plátōn, "broad")[1] (428/427 BC[a] – 348/347 BC), was a Classical Greek philosopher, mathematician, writer of philosophical dialogues, and founder of the Academy in Athens, the first institution of higher learning in the Western world. Along with his mentor, Socrates, and his student, Aristotle, **Plato** helped to lay the foundations of natural philosophy, science, and Western philosophy.[2] **Plato** was originally a student of Socrates, and was as much influenced by his thinking as by what he saw as his teacher's unjust death.

DOCUMENT_B

- Socrates (pronounced /ˈsɒkrətiːz/; Greek: Σωκράτης, Sōkrátēs; c. 469 BC–399 BC[1]) was a Classical Greek philosopher. Credited as one of the founders of Western philosophy, he is an enigmatic figure known only through the classical accounts of his students. **Plato**'s dialogues are the most comprehensive accounts of Socrates to survive from antiquity.[2]

Which document should rank higher?

- *t* = how often a query appears in a document
- Pick the document with the highest *t*

Both documents have the term the same number of times

plato Search

DOCUMENT_A

- **Plato** (pronounced /ˈpleɪtoʊ/) (Greek: Πλάτων, Plátōn, "broad")[1] (428/427 BC[a] – 348/347 BC), was a Classical Greek philosopher, mathematician, writer of philosophical dialogues, and founder of the Academy in Athens, the first institution of higher learning in the Western world. Along with his mentor, Socrates, and his student, Aristotle, **Plato** helped to lay the foundations of natural philosophy, science, and Western philosophy.[2] **Plato** was originally a student of Socrates, and was as much influenced by his thinking as by what he saw as his teacher's unjust death.

DOCUMENT_B

Greek philosophy focused on the role of reason and inquiry. Many philosophers today concede that Greek philosophy has shaped the entire Western thought since its inception. As Alfred Whitehead once noted, with some exaggeration, "Western philosophy is just a series of footnotes to **Plato**."[1] Clear unbroken lines of influence lead from ancient Greek and Hellenistic philosophers, to medieval Muslim philosophers, to the European Renaissance and Enlightenment.

Early Greek philosophy, in turn, was influenced by the older wisdom literature and mythological cosmogonies of the Near East. As **Plato** points out: "[...]contact with oriental cosmology and theology helped to liberate their [the early Greek philosophers'] imagination; it certainly gave them many suggestive ideas. But they taught themselves to reason. Philosophy as we understand it is a Greek creation."[2]

Neither reason nor inquiry began with the Ancient Greeks, but the Socratic method, along with the idea of Forms, allowed great advances in geometry, logic, and the natural sciences. Defining the difference between the Ancient Greek quest for knowledge and the quests of the elder civilizations, such as the ancient Egyptians and Babylonians, has long been a topic of study by theorists of civilization. Benjamin Farrington, former Professor of Classics at Swansea University wrote:

"Men were weighing for thousands of years before Archimedes worked out the laws of equilibrium; they must have had practical and intuitional knowledge of the principles involved. What Archimedes did was to sort out the theoretical implications of this practical knowledge and present the resulting body of knowledge as a logically coherent system."

and again:

"With astonishment we find ourselves on the threshold of modern science. Nor should it be supposed that by some trick of translation the extracts have been given an air of modernity. Far from it. The vocabulary of these writings and their style are the source from which our own vocabulary and style have been derived."[3]

[edit] Pre-Socratic philosophy

Main article: Pre-Socratic philosophy

The presocratics were primarily ontologists who rejected mythological explanations for reasoned discourse. **Plato**, for example, gave one of the first documented logical arguments: How could what is perish? How could it have come to be? For if it came into being, it is not; nor is it if ever it is going to be. Thus coming into being is extinguished, and destruction unknown.

Which document should rank higher?

- *t* = how often the query appears in a document
- T = total number of terms in a document
- Pick the document with the highest

$$\frac{t}{T}$$

- Basically, the document with the highest proportion of terms which are the query

What about for multi-term queries?

- *t* = how often the query appears in a document
- T = total number of terms in a document
- Pick the document with the highest

$$\sum_{terms} \frac{t}{T}$$

- Basically, the document with the highest proportion of terms which are the query

Equal number of terms, but…

princeton university Search

DOCUMENT_A

- **Princeton** is a private research university located in **Princeton**, New Jersey, United States. The school is one of the eight universities of the Ivy League and is considered one of the Colonial Colleges.
- **Princeton** has traditionally focused on undergraduate education, although it has almost 2,500 graduate students enrolled in the **university**.[6] A unique blend of research and liberal arts, **Princeton** does not offer professional schooling generally, but it does offer professional master's degrees (mostly through the Woodrow Wilson School of Public and International Affairs) and doctoral programs in the sciences, humanities, and social sciences, as well as engineering.

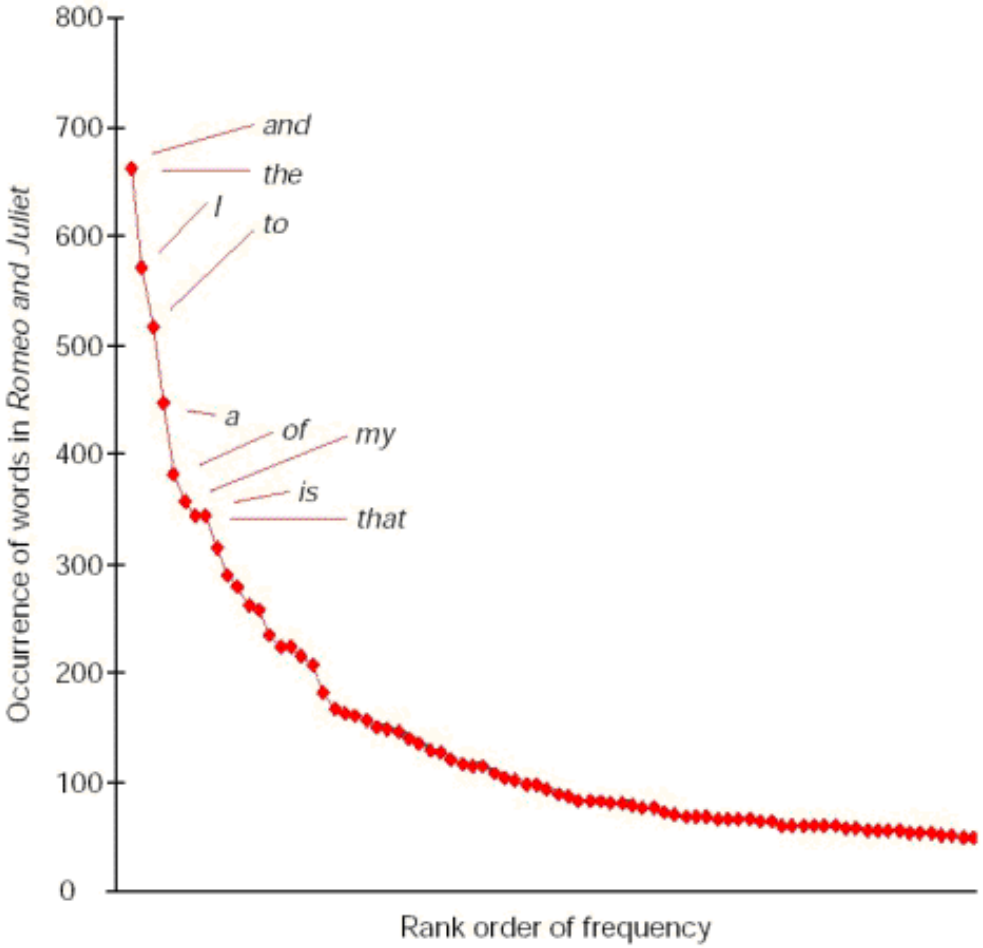
DOCUMENT_B

So what's a **university** admissions director to do? You can be sure that one of the most closely watched trends in every admissions office is the yield, the fraction of successful applicants who accept your **university**'s offer of admission. **Universities** like Harvard, **Princeton**, Stanford, and MIT typically have low acceptance rates. There are many factors contributing to that yield. For example, contrary to most expectations, the school with the highest yield in the country isn't Harvard, it's Brigham Young **University**, a school that admits 70% of its applicants. Presumably, the appearance of Yeshiva at 8th is also affected by similar factors, i.e., students applying to BYU or Yeshiva do so for specific cultural reasons. If admitted, they'll probably attend.

Term Distribution

- What is a distribution?

Zipf for Languages



- Rare terms should be weighted higher
- Detecting power law

tf-idf

- *t* = how many times the query appears in a document
- T = total number of terms in a document
- *D* = set of all documents
- *d* = number of documents with that term
- Pick the document with the highest

$$\sum_{terms} \frac{t}{T} \times \log \frac{D}{d}$$

- Basically, the document with the highest proportion of terms which are the query

tf-idf Intuition

- Documents containing more of the term(s) scored higher
- Longer documents discounted
- Rare terms weighted higher

More Document Analysis

- Vector-Space Model
  - Cosine similarity
- Kl-divergence
- Statistical information retrieval

Question and Break

- You have 1,000 text documents. How would you instruct your intern to quickly find the ones with the word “acquisition” in them?
- What could you do beforehand so the intern could find them quickly in the future?

Indexing

- Documents
- Matrix
- Inverted Index

Optimizing the Inverted Index

- Ordering by position
- Ordering the key

Tokenizing

- Multiple words
- Punctuation
- Different languages
- We'll talk more about query processing / stemming next time

The SEO Game

(handout)

Reading Response for next class