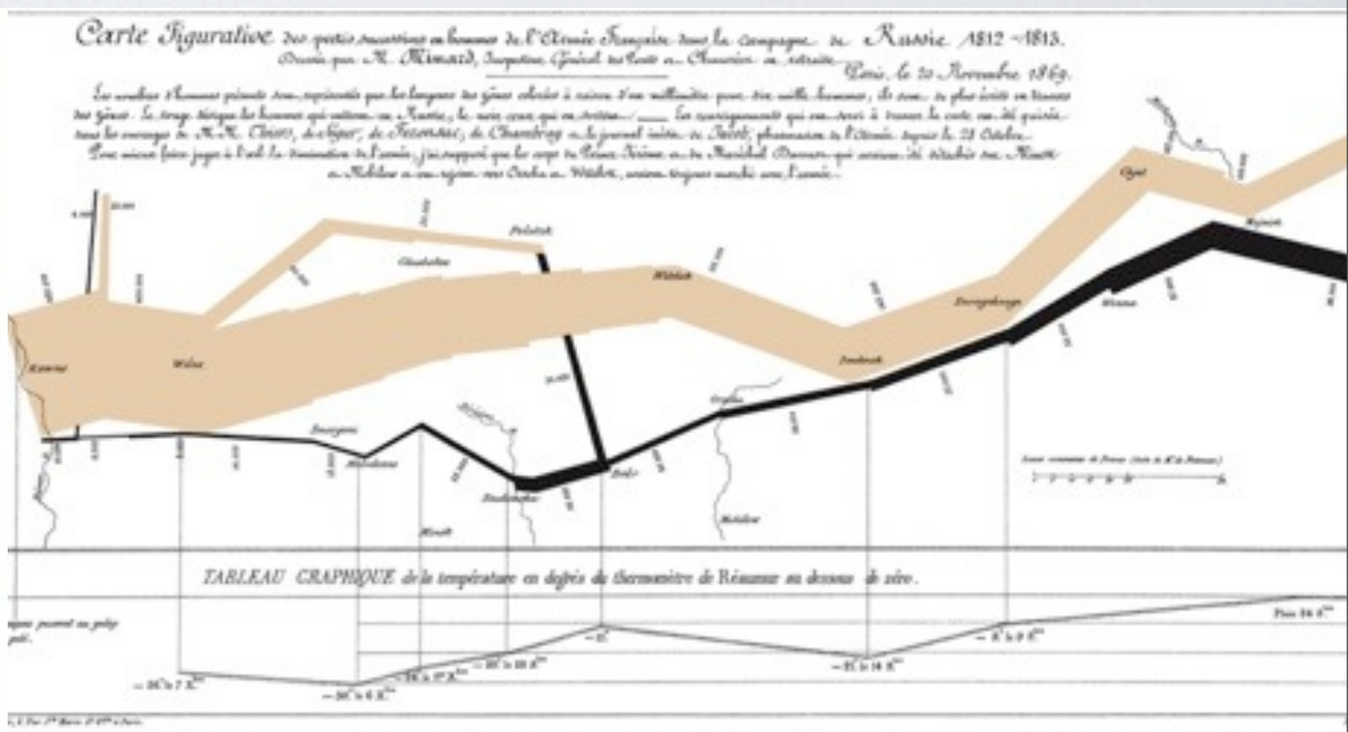
LAST WEEK ON IO LAB



We thought you should see these one more time.

INFORMATION ORGANIZATION LAB

RETRIEVAL & SEARCH

Answers.com
infospace
bing
Ask
Inktomi
overture
WolframAlpha
Westlaw
Google
altavista
yebol
LYCOS
YAHOO!
LexisNexis
excite
cuil
Technorati

Different search domains, different ideas of the importance of recall and and precision. Different sources of documents. Not all of these search services are in direct competition with Google (yes, some because they are defunct).

http://labs.ideeinc.com/multicolr/

CORPORA

Many options: text documents, multimedia documents, photos, music.
For any kind of corpus, different challenges.

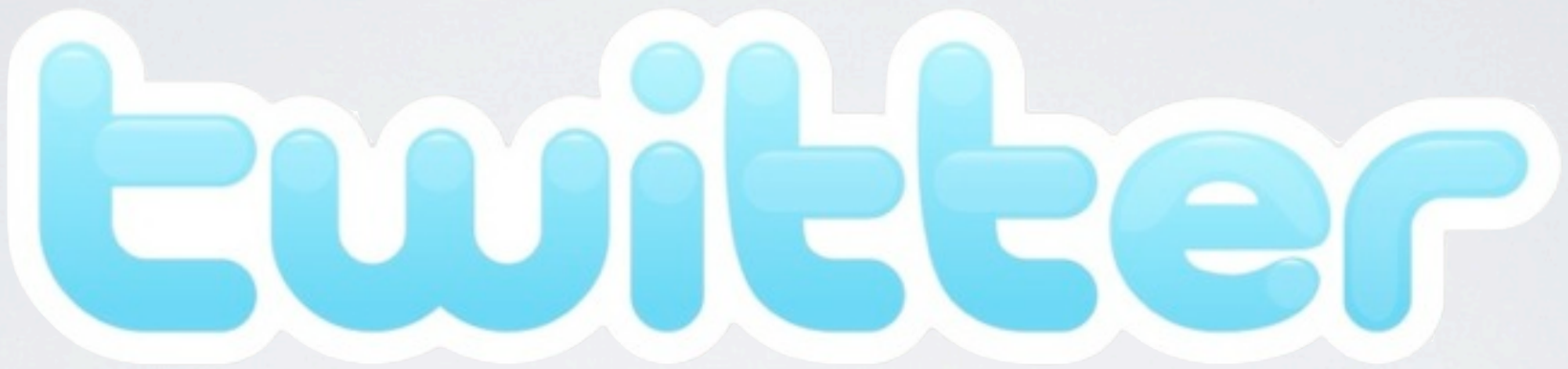
80legs

DISTRIBUTED SPIDERING

One corpus you might be interested in is all the documents on the Web. But writing software to crawl all that is time-consuming and requires extensive expensive infrastructure.

80legs lets you write your own Java software to process a document and have it run on the Web for $2 per million URLs. It's not designed for downloading the Web (where would you put it all anyway?), but is good for searching for specific things you're interested in, data mining, measuring usage, etc.

Example: W3C Geolocation API (Javascript), but we could also use it to measure rel="license" usage, or find hCards to compile into a big rolodex, or create some other specific corpus you'd like to make available.

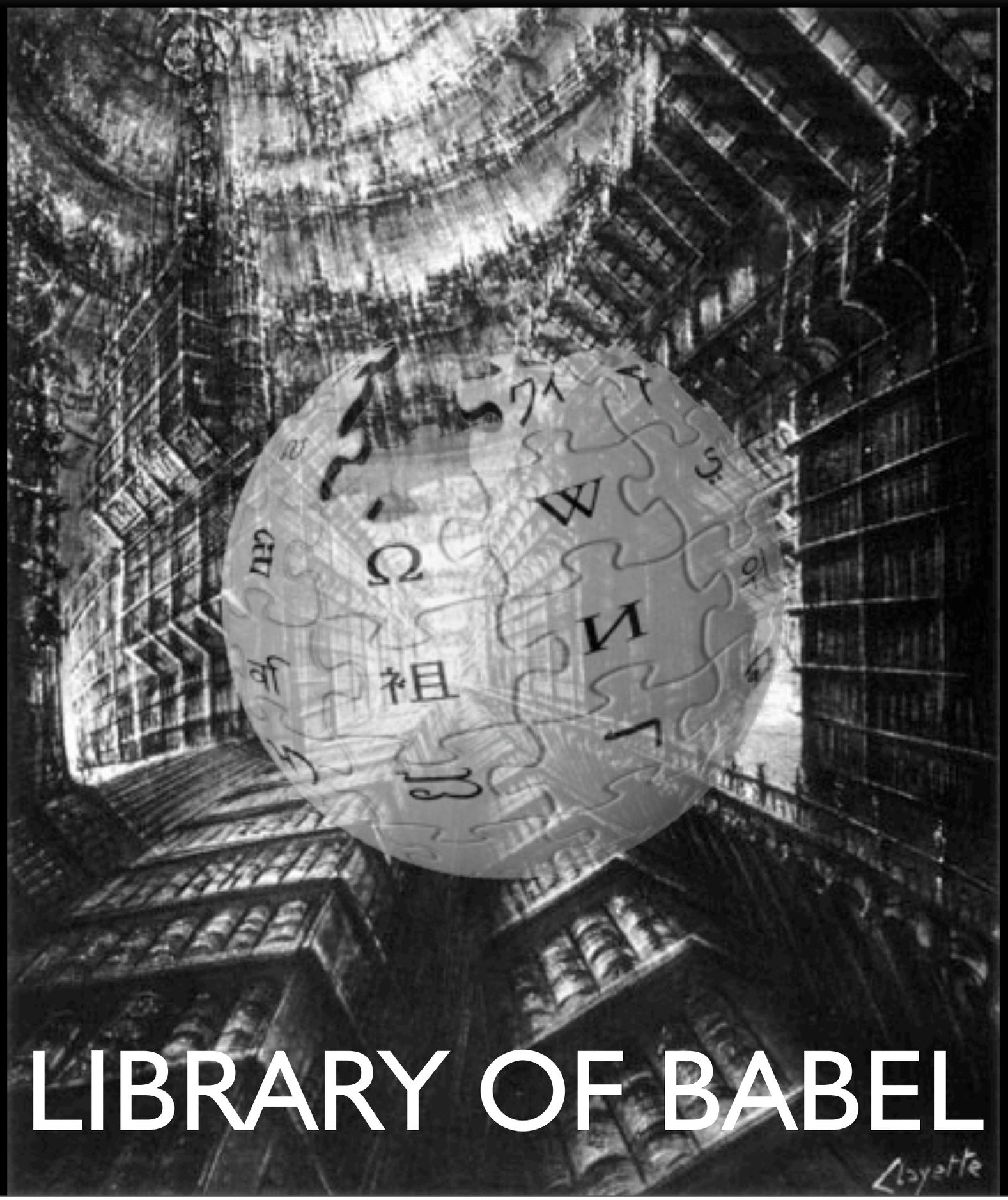


TWITTER FRAMEWORK

Document are short, lots of relevance issues. Demonstration of Twitter IR framework where you can write a short JavaScript module that indexes available documents, returns results to a query, and orders documents by relevance.

NOISE@ISCHOOL

Private list, so we can't have Google index it for us. Own set of problems. We have a cleaned up email corpus from another class, with message attachments, replies, signatures, etc. removed.



If we had a corpus based on Wikipedia, but with words jumbled in hundreds of iterations, could we find the real information? Imagine taking 10,000 Wikipedia articles, jumbling the words in each one, and creating 100,000 articles. How would you find the relevant information in the noise?

i202 ASSIGNMENT 7

$$w_{ik} = \frac{tf_{ik} \cdot idf_k}{\sqrt{\sum_{n=0}^{M-1} w_{ik} \cdot w_{jk}}} \qquad tf_{ik} = \frac{n_{ik}}{\sum_j n_{jk}} \qquad idf_k = log\frac{N}{n_k}$$

Illustrations and equations from i202 assignment 7. Portions removed from lecture because they contain answers to the currently posted assignment for another class.

GRADING & FEEDBACK

Grading scale: Maybe somewhat unintuitive. 10 reserved for above and beyond. We've assigned letter grades as if 9-9-9 were the highest score you could receive. We provide written feedback. Our written comments aren't intended to note deficiencies; highlight possibilities for expansion, improvements, etc.

ROADMAP

Next week we're going to look more in-depth demonstrations of some of the tools we previewed today, functional programming. Project 5 due date is the final end-of-semester presentation date.

FOR NEXT WEEK

This space intentionally left blank.