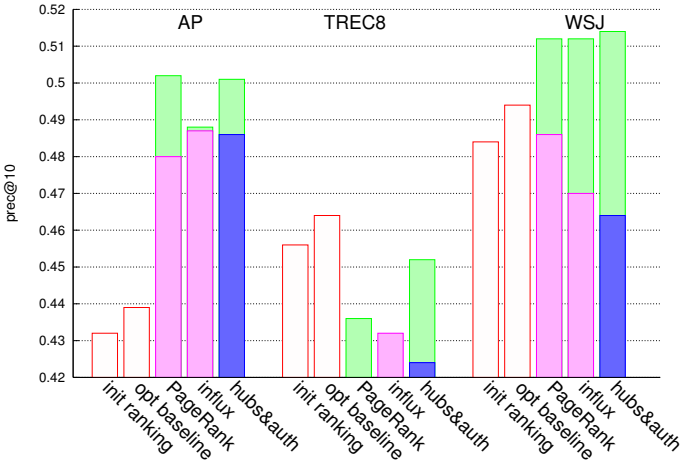
Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
3/16/07: **Lecture 24 aid — Introduction to natural language processing**

Topics: moving from "classic" "bag-of-words"-based information retrieval and non-content-based analysis to natural language processing (NLP); general applications and challenges.

Announcements: There will be no office hours between Sunday March 18 and Saturday March 24. Rafael Frongillo's Sunday March 25th hours (8-9pm) will be held in the Dickson computer lab, rather than the usual location. The usual office hour schedule (with the usual locations) resumes on Monday March 26th. All this information is also available at the course webpage, www.cs.cornell.edu/courses/cs172/2007sp (click on the link for the course calendar).

I. Link analysis on non-hyperlinked corpora Figure from O. Kurland and L. Lee, "Respect my authority! HITS without hyperlinks, utilizing cluster-based language models", SIGIR 2006. Links between documents are induced and weighted by their content-analysis-based similarities. Influx corresponds to in-degree in this link-weighed scenario. The "upper portions" of the bars show the performance improvement when document clusters are used as entities in the network (as opposed to just documents).



II. Some instructive examples

1. (a) "This document is about jaguars — the car, not the cat."
   (b) "This document is about jaguars — the cat, not the car."
2. "This document is about jaguars."
3. Query: "trucks". document: "Lorries galore..."
4. Top Google hit for query "Lilian Lee" is my home page (!)
5. "List all flights on Tuesday."
6. "List all flights on the double."
7. "Copy the local patient files to disk."
8. "I saw her duck with a telescope."