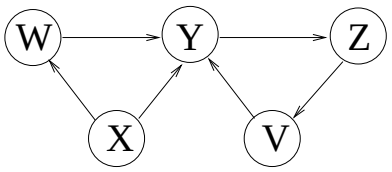
Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
10/19/05: **Lecture 23 aid — Introduction to NLP**

Agenda: move from standard information retrieval to natural language processing (NLP); discuss general applications and challenges

Follow-up to last time: Recall that in the example document set we considered,



page Z, which is pointed to by a page with relatively high in-degree, eventually gets relatively large PageRank but an authority score of 0. For either outcome, there is a sensible interpretation:

- If Y were a highly-trusted site disseminating high-quality scientific information and Z were a recent research paper, then it makes sense that Z should be considered an important document worth looking at.
- On the other hand, for the same Y, if Z were simply the "privacy policy" page for the site, it is probably not a worthwhile page despite being pointed to by an informative page.

This dichotomy can be related to the different contexts that the two algorithms were originally developed for: PageRank was originally conceived of as a *query-independent* ranking algorithm, whereas hubs and authorities was intended to be run on the set of pages retrieved in response to a particular query.

I. (a) "This document is about jaguars — the car, not the cat."
(b) "This document is about jaguars — the cat, not the car."

II. "This document is about jaguars."

III. query: "trucks". document: "Lorries galore..."

IV. Top Google hit for query "Lilian Lee" is my home page (!)

V. "List all flights on Tuesday."

VI. "List all flights on the double."

VII. "Copy the local patient files to disk."

VIII. "I saw her duck with a telescope."