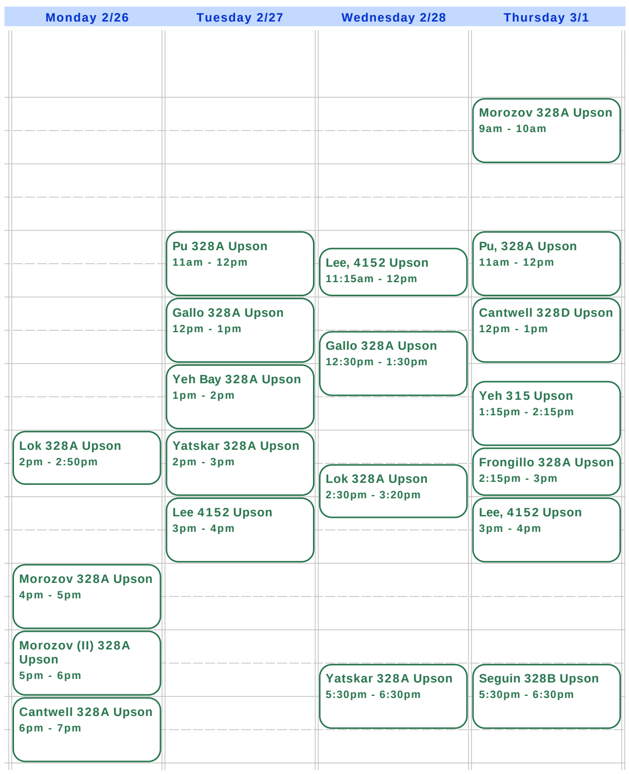
Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
2/26/07: **Lecture 16 aid — Completion of tf-idf weighting; beginning of link-analytic approaches**

Topics: tf-idf weighting example; potential utility of link-analytic approaches; in-degree models.

Announcements: The office hours schedule for this week, altered because of Friday's in-class prelim, is as follows. This information is also available at www.cs.cornell.edu/courses/cs172/2007sp/calendar.htm . Graded HW2's can be picked up at the Monday and Tuesday hours as well as Wednesday's lecture.



I. Altered example data[1] W: w_1: "the"; w_2: "wolf"; w_3: "lady"; w_4: "of"; w_5: "shalott".
Corpus:

d': the wolf the wolf
d'': lady lady lady, the lady of shalott
d''': the the
d'''': of the lady

Query: "the shalott painting"

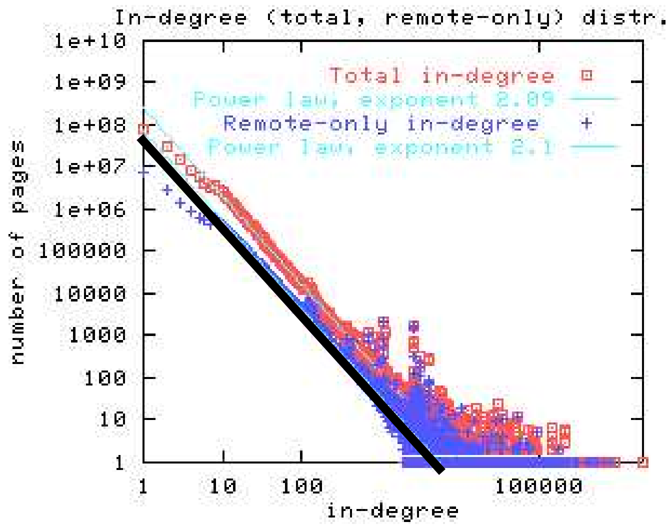
Self-check: Under tf weighting (and no normalization of the query vector), $\overrightarrow{d'} \cdot \overrightarrow{q} = 2/\sqrt{8}$ and $\overrightarrow{d''} \cdot \overrightarrow{q} = 2/\sqrt{19}$.

[1] The example provided on last lecture's aid also shows a case where tf.idf-weighting yields a different ranking than tf weighting, but the numerical results given aren't correct (I forgot to update all my calculations after changing the example documents a bit).

(OVER)

II. A Web power law A "power law" is a relationship of the form $y = x^{-\alpha}$, where α is a constant. Observe that if we take the log of both sides, we get the linear relationship $\log(y) = -\alpha \log(x)$.

The *(in-)degree distribution* of a given collection of linked documents gives, for each possible in-degree x, the number (or fraction) of documents that have in-degree equal to x. Here is Figure 1 of Broder et al. (2000), which shows the in-degree distribution, on a log-log scale, for their 200M-document Web crawl. The line corresponding to $\alpha = 2.1$ is highlighted.



III. Web-growth models: Template, conventions and notation We use the integer-valued variable $t \geq 0$ to stand for time.

1. The constant $n_0 \geq 1$ is the number of documents that exist at time $t = 0$; call them $d_{-1}, d_{-2}, \ldots, d_{-n_0}$, and assume they have no links between them.

2. At the j^{th} time step, we add a new document, named d_j; hence, a positive subscript indicates when a document was added.

3. We then grant to d_j a constant number ℓ of links, where $1 \leq \ell$, to some of the $n_0 + j - 1$ pre-existing documents, allowing repeated links to the same document.

4. We are interested in computing $\mathrm{In}(\mathrm{doc} = j, \mathrm{time} = t)$, which is our estimate of d_j's in-degree at time $t \geq \max(j, 1)$ (there is no point computing the in-degree of a document at a time before it existed), where $j \geq 1$.