Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
2/23/07: **Lecture 15 aid — TF and TF-IDF weighting**

Topics: document ranking based on document vectors using term-frequency (tf) or tf-idf weighting.

Announcements: Due to the upcoming prelim, the office-hours schedule for next week is changing. The finalized schedule will be announced in Monday's lecture and posted to the online course calendar (www.cs.cornell.edu/courses/cs172/2007sp/calendar.htm). To facilitate your study planning, we are informing you now of the Sunday and Monday portions of next week's schedule:

Sunday (2/25/07)	8pm-9pm	3331 Balch (Tatkon Center)	Rafael Frongillo
Monday (2/26/07)	2-2:50pm	328A Upson	Selina Lok
	4-6pm	328A Upson	Anton Morozov
	6-7pm	328A Upson	Jared Cantwell

I. Reminder: normalized term-frequency vectors

$$\overrightarrow{d} = \left(\frac{\text{freq}(w_1 \in d)}{N(d)}, \frac{\text{freq}(w_2 \in d)}{N(d)}, \ldots, \frac{\text{freq}(w_m \in d)}{N(d)} \right)$$

where $N(d) = \sqrt{\sum_{i=1}^{m} \text{freq}(w_i \in d)^2}$ is the *(vector-)length-normalization factor*.
Self-check: verify that $\overrightarrow{d}$ is of unit length using the inner-product characterization of vector length.
Note: we typically do *not* normalize the query vectors.

II. Example data W: w_1: cat; w_2: dog; w_3: news. Query: "cat dog". Corpus:

d: "news news news cat dog"
d': "cat dog news dog news"

III. Tf-idf weighting We[1] define $\text{IDF}(w_i)$, the *inverse document frequency* of term w_i, as $n/\text{doccount}(w_i)$, where $\text{doccount}(w_i)$ is the number of documents in the n-document corpus that contain w_i.
The resulting alternative to term-frequency weighting converts a document d to the vector

$$\overrightarrow{d} = \left(\frac{\text{freq}(w_1 \in d) \times \text{IDF}(w_1)}{N_I(d)}, \frac{\text{freq}(w_2 \in d) \times \text{IDF}(w_2)}{N_I(d)}, \ldots, \frac{\text{freq}(w_m \in d) \times \text{IDF}(w_m)}{N_I(d)} \right)$$

where $N_I(d) = \sqrt{\sum_{i=1}^{m} \left(\text{freq}(w_i \in d) \times \text{IDF}(w_i)\right)^2}$. Where no confusion can result, we will often drop the "I" subscript.
Note: we typically use non-normalized tf (*not* tf-idf) weighting for the query vector.

IV. Example for tf-idf computations Vocabulary: 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, lady of shalott
d'': the the
d''': of the lady

Query: "the shalott painting"
We **don't**[2] have $\text{IDF}(w_1) = 1$, $\text{IDF}(w_2) = \text{IDF}(w_5) = 4$, etc. Note that the within-document repetitions of "the" don't matter. We **don't** find $\overrightarrow{d} \cdot \overrightarrow{q} = 2/\sqrt{68}$ and $\overrightarrow{d'} \cdot \overrightarrow{q} = 4/\sqrt{84}$. What if we use tf weighting instead?

[1] We are using a reduced form to simplify (your) calculations.
[2] I forgot to update some calculations when editing the example documents. We'll fix this next time.