36-350: Data Mining

Handout 1
August 25, 2003

Similarity searching and information retrieval: better ways to search the web.

Similarity searching—“Find me a document like this one.” Does not require picking a set of keywords. Similarity searching can also be used in a database of images or a database of sounds. E.g. “Find a patient with a similar X-ray.”

Today’s data: `rec.autos` and `rec.motorcycles`—discussion lists on the Internet.

Bag of words representation—For each word, the number of times it appears in the document (including zeros). Every document has the same size representation.

	Word									
Document	car	bike	cars	his	tires	she	ive	her	#k	are
auto1	5	0	0	0	0	1	0	2	1	0
auto2	0	0	3	0	3	0	1	0	0	1
auto3	2	0	0	0	0	0	1	0	0	0
auto4	1	0	1	0	2	0	0	0	0	1
auto5	5	0	2	0	0	4	2	2	3	7
moto1	0	3	0	1	0	0	0	0	0	0
moto2	0	0	0	6	0	0	0	0	0	1
moto3	0	5	0	0	0	0	0	0	0	0
moto4	0	1	0	0	0	0	0	0	0	0
moto5	0	2	0	0	0	0	0	0	0	0

Measuring similarity—a fundamental operation in data mining. All other tasks (clustering, modeling, etc.) are based on it. Sometimes it is more convenient to work with **dissimiliarity** or **distance**. There are many ways to define distance. One that has proven useful for text is Euclidean distance, after the normalizing the document vectors by Euclidean length.

Euclidean distance—A measure of distance between two vectors (points in space).

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_k (x_k - y_k)^2}$$

For document vectors, this becomes

$$\sqrt{\sum_{\text{words } w} (doc_1(w) - doc_2(w))^2}$$

We want documents to be matched based on the relative proportion of different words, not on the document's length. Thus we **normalize** the word counts before computing the distance.

Document length normalization—Divide the word counts by the total number of words in the document. This turns the word counts into word fractions. This treats a word which occurs once in a 100-word document the same as a word which occurs ten times in a 1000-word document.

Euclidean length normalization—Divide the word counts by the Euclidean length of the count vector. This tends to perform better, since it de-emphasizes words that have occurred only once.

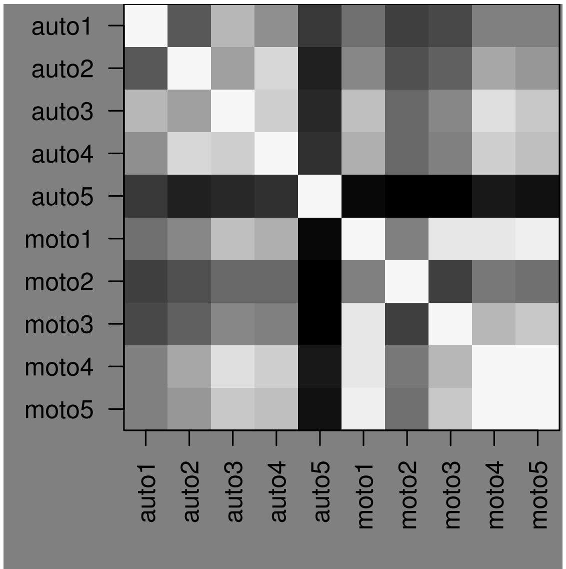
Euclidean length of a vector—The distance from the vector to the origin.

$$||\mathbf{x}|| = \sqrt{\sum_k x_k^2}$$

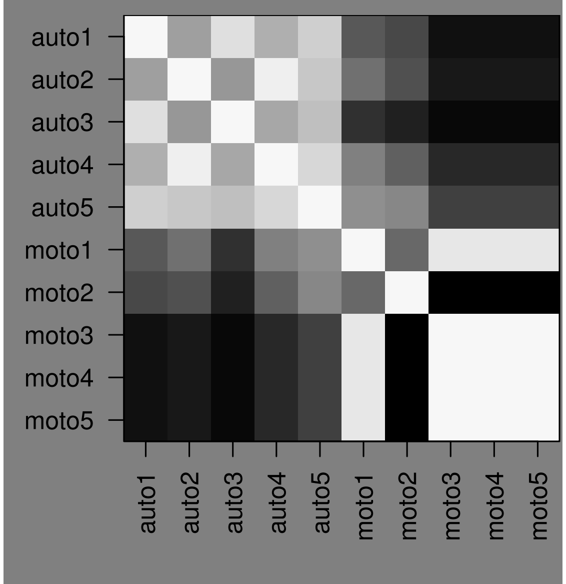
Similarity measures can be compared by **error rate**—the number of documents for which the closest match is in the wrong category.

References

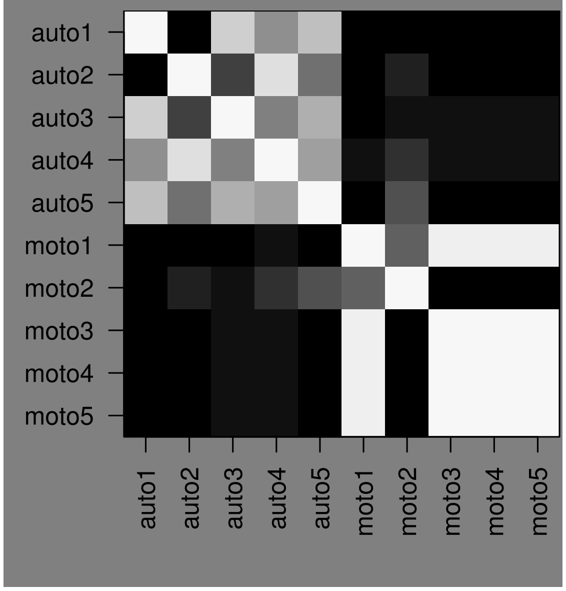
[1] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*, Section 14.3. MIT Press, 2001.



Distance matrix for un-normalized counts
Lighter = Closer
1 error (picks `moto4` for `auto3`)



Normalized by document length
1 error (picks `auto5` for `moto2`)



Normalized by Euclidean length
No errors