36-350: Data Mining

Handout 2
August 27, 2003

Similarity searching procedure:

1. The user provides a document Q. Convert it into a bag of words.
2. For each document in the collection, measure the distance to Q.
3. Return the documents with smallest distance.

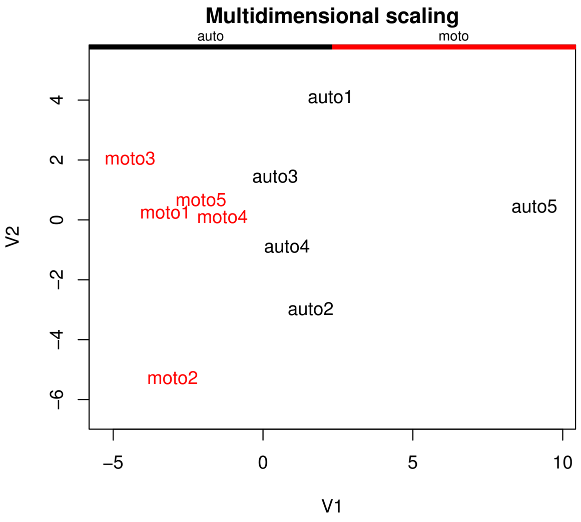
Multidimensional scaling—An approximate representation of a distance matrix. Points are placed in the plane so that the distances between them mimic the distances between the original objects. In general, the distances will not match exactly, and the rotation of the points is arbitrary. We will use it to visualize the quality of different distance measures between documents.

Document frequency (DF) of a word—The fraction of documents in the collection in which that word appears.

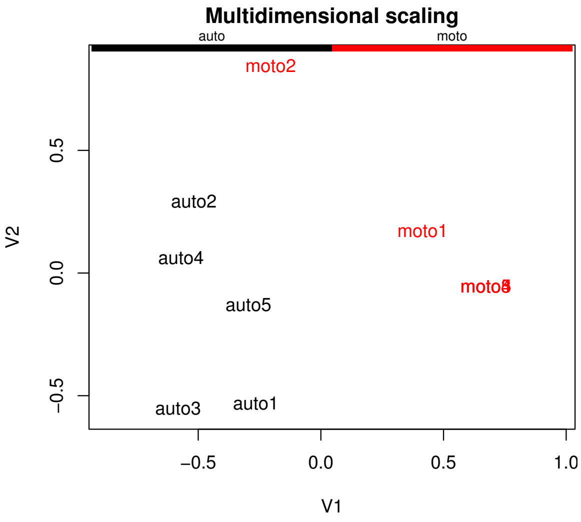
Inverse document frequency—A method for emphasizing "important" words by deemphasizing common words. The count for word w in each document is multiplied by

$$IDF(w) = \log\left(\frac{1}{DF(w)}\right)$$

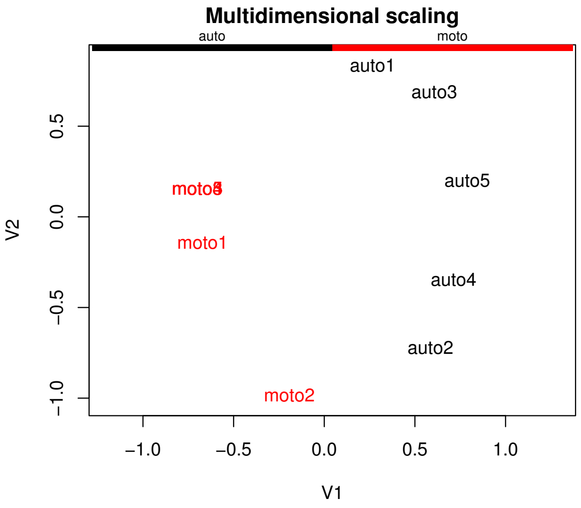
Words which occur frequently, like "the", will have a small $IDF(w)$. Then each document is normalized by Euclidean length, as before, and similarity is measured by Euclidean distance.



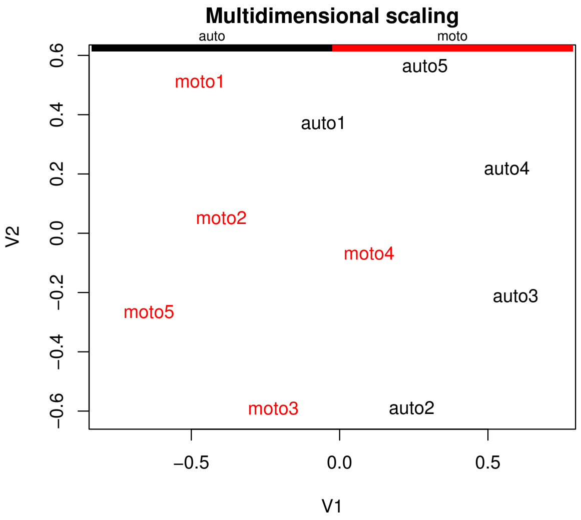
10 best words,
Un-normalized counts,
1 error (picks `moto4` for `auto3`)



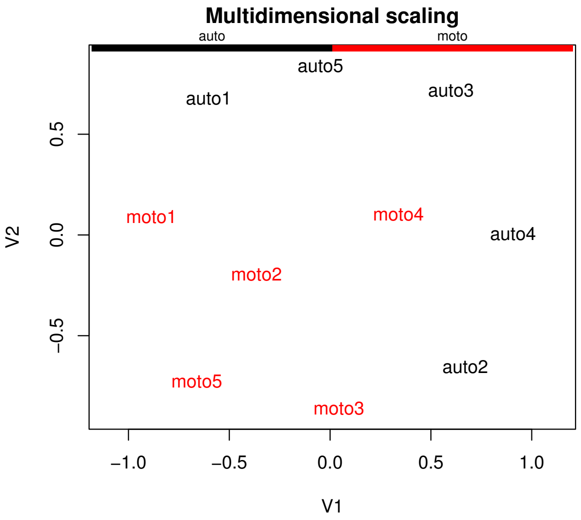
Normalized by document length,
1 error (picks `auto5` for `moto2`)



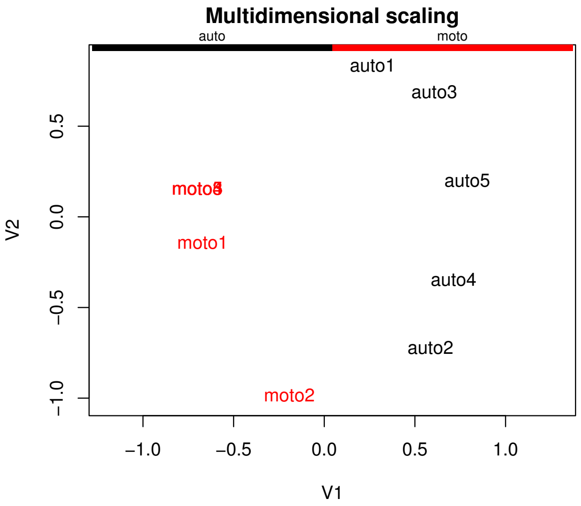
Normalized by Euclidean length,
No errors



182 words, equal weighting
5 errors (`auto1,2,4`, `moto2,4`)
(as bad as guessing)



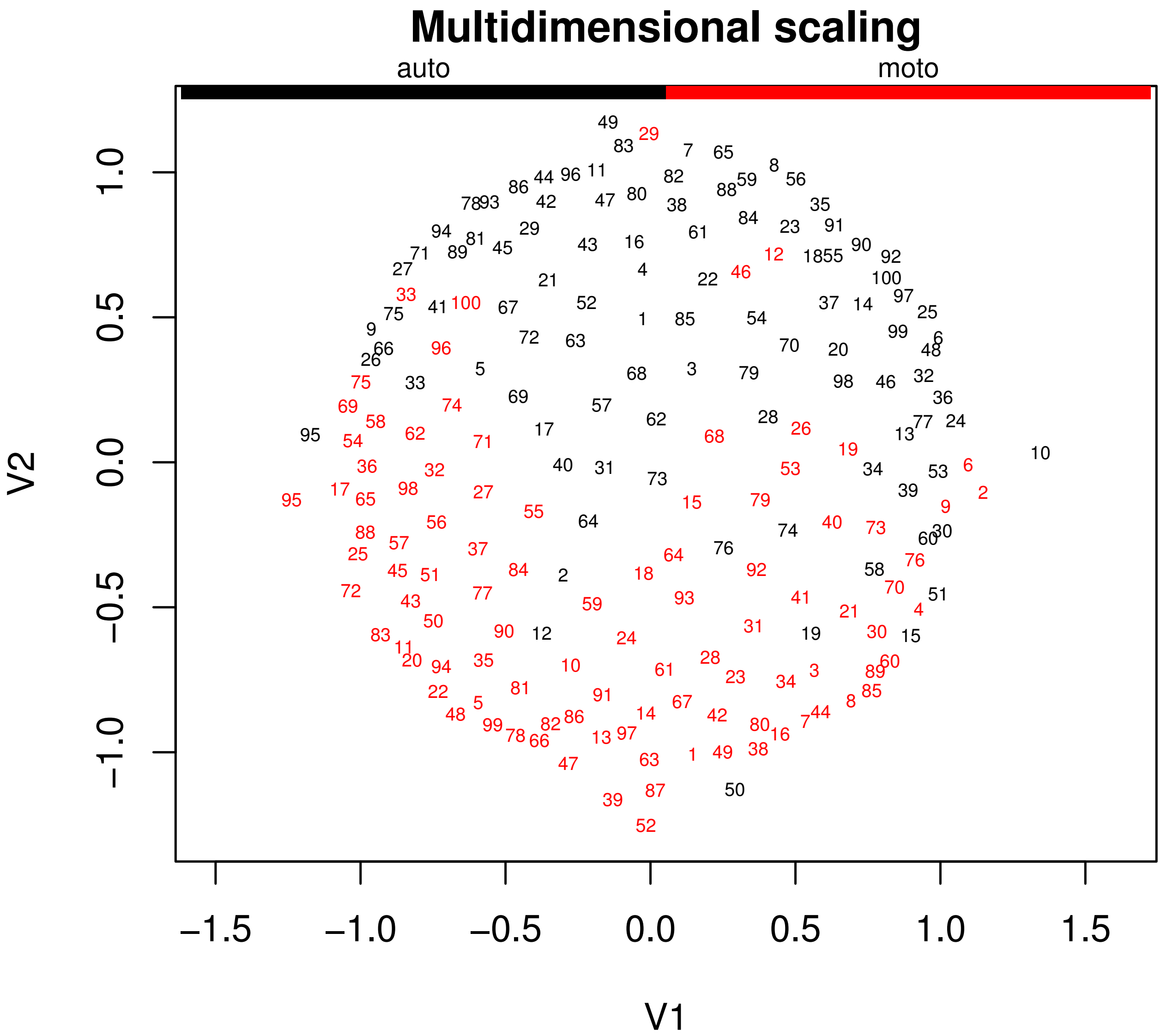
182 words, IDF weighting
3 errors (`auto4`, `moto1,4`)



10 best words (from last time)

Retrieval errors on a larger collection (199 articles from `rec.auto` and `rec.motorcycles`):

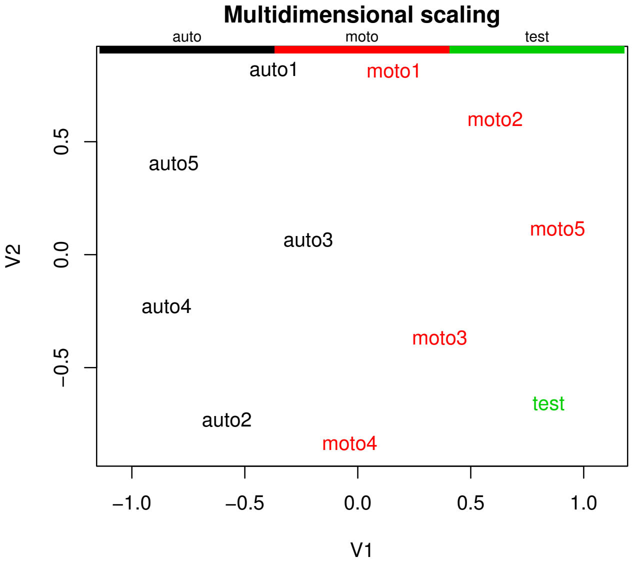
Normalization	Equal weight	IDF weight
None	83	79
Document length	63	60
Euclidean length	59	21



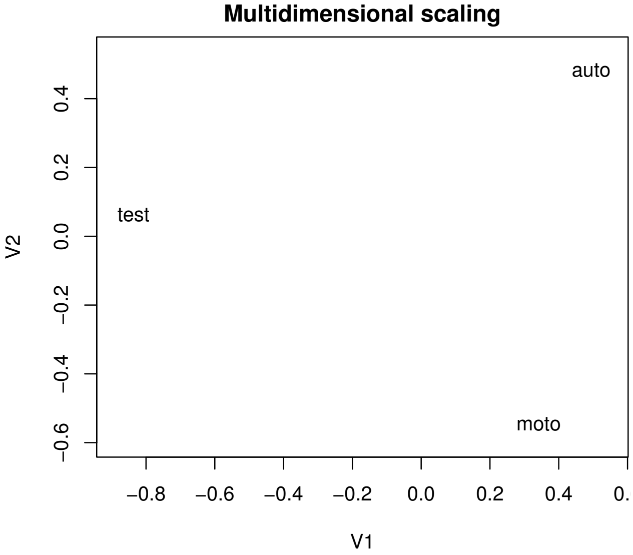
Q: Why does it look like a disk?

Similarity searching can also be used to **classify** a new document into known categories.

Nearest-neighbor method—Label the new document according to the nearest labeled document.



Prototype method—Average documents in the same category to obtain a single "prototypical" document. Label the new document according to the nearest prototype.


Multidimensional scaling
auto
test
moto
V2
V1