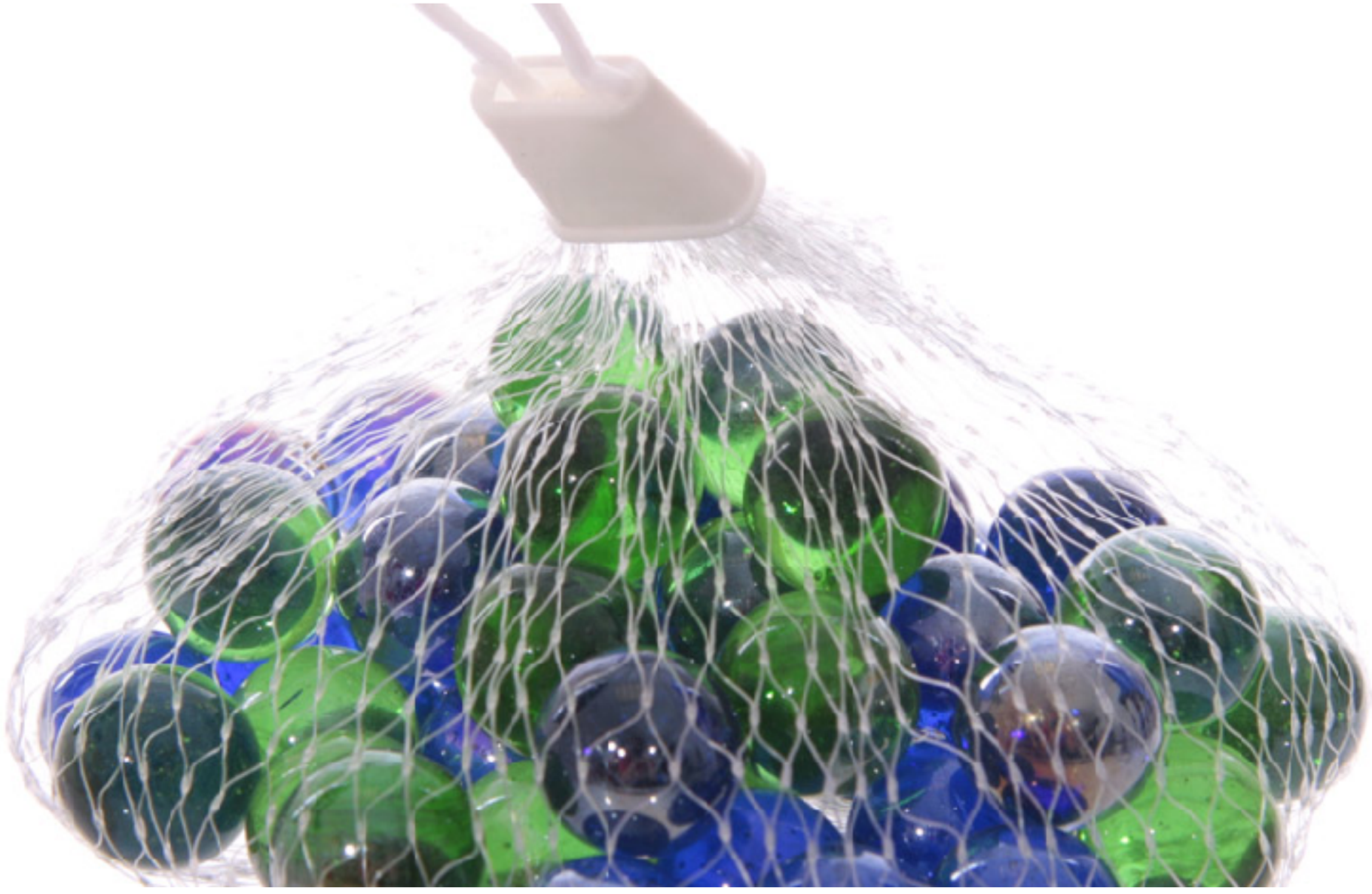


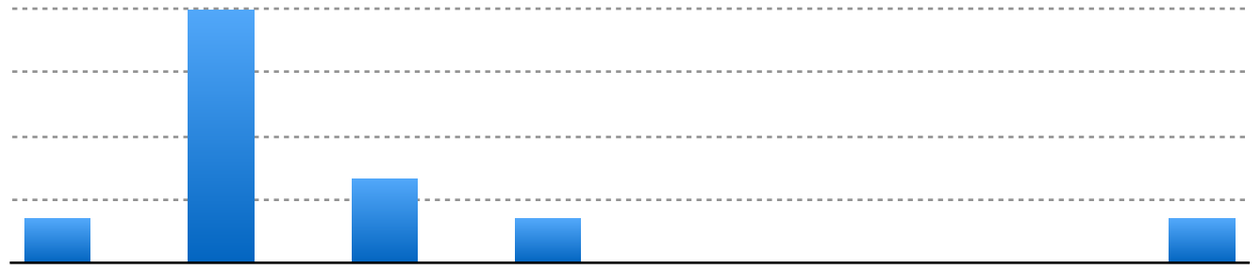
TF-IDF

Computer Vision

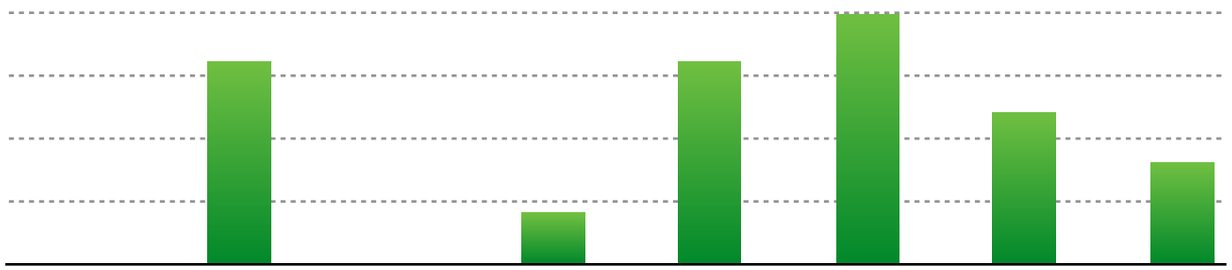
Carnegie Mellon University (Kris Kitani)

Vector Space Model

(aka Bag-of-Words)



1	6	2	1	0	0	0	1
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor



0	4	0	1	4	5	3	2
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

G. Salton. ‘Mathematics and Information Retrieval’ Journal of Documentation,1979

http://www.fodey.com/generators/newspaper/snippet.asp

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

just a histogram over words

$n(\cdot)$ counts the number of occurrences

What is the similarity between two documents?

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

just a histogram over words

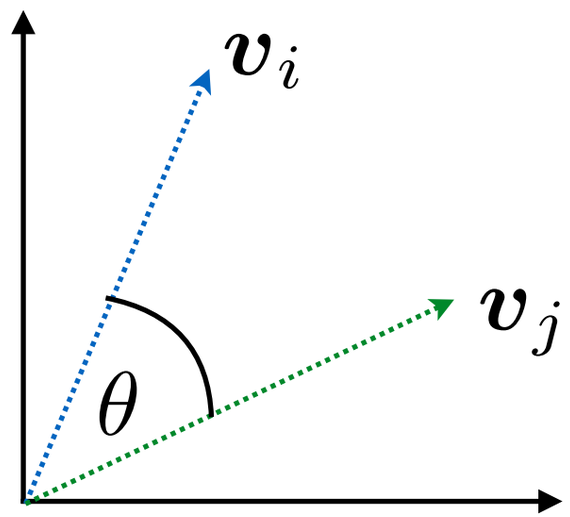
$n(\cdot)$ counts the number of occurrences

What is the similarity between two documents?

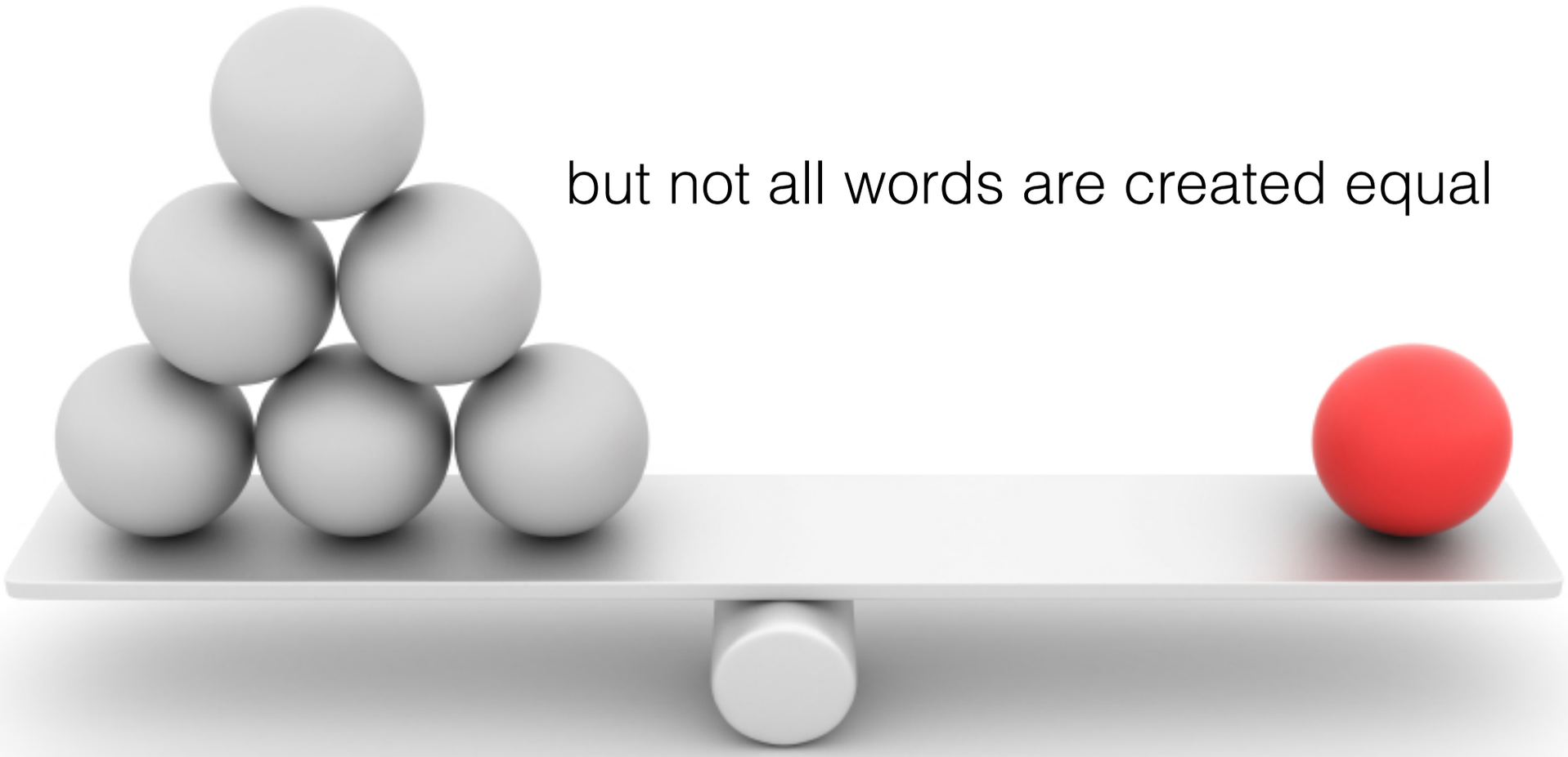
Use any distance you want but the cosine distance is fast.

$$d(\boldsymbol{v}_i, \boldsymbol{v}_j) = \cos\theta$$

$$= \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\| \|\boldsymbol{v}_j\|}$$



but not all words are created equal



TF-IDF

Term **F**requency **I**nverse **D**ocument **F**requency

$$\boldsymbol{v}_d = [n(w_{1,d}) \;\; n(w_{2,d}) \;\; \cdots \;\; n(w_{T,d})]$$

weigh each word by a heuristic

$$\boldsymbol{v}_d = [n(w_{1,d})\alpha_1 \;\; n(w_{2,d})\alpha_2 \;\; \cdots \;\; n(w_{T,d})\alpha_T]$$

$$n(w_{i,d})\alpha_i = \overbrace{n(w_{i,d})}^{\text{term frequency}} \log \left\{ \frac{D}{\overbrace{\sum_{d'} \mathbf{1}[w_i \in d']}^{\text{inverse document frequency}}} \right\}$$

(down-weights **common** terms)