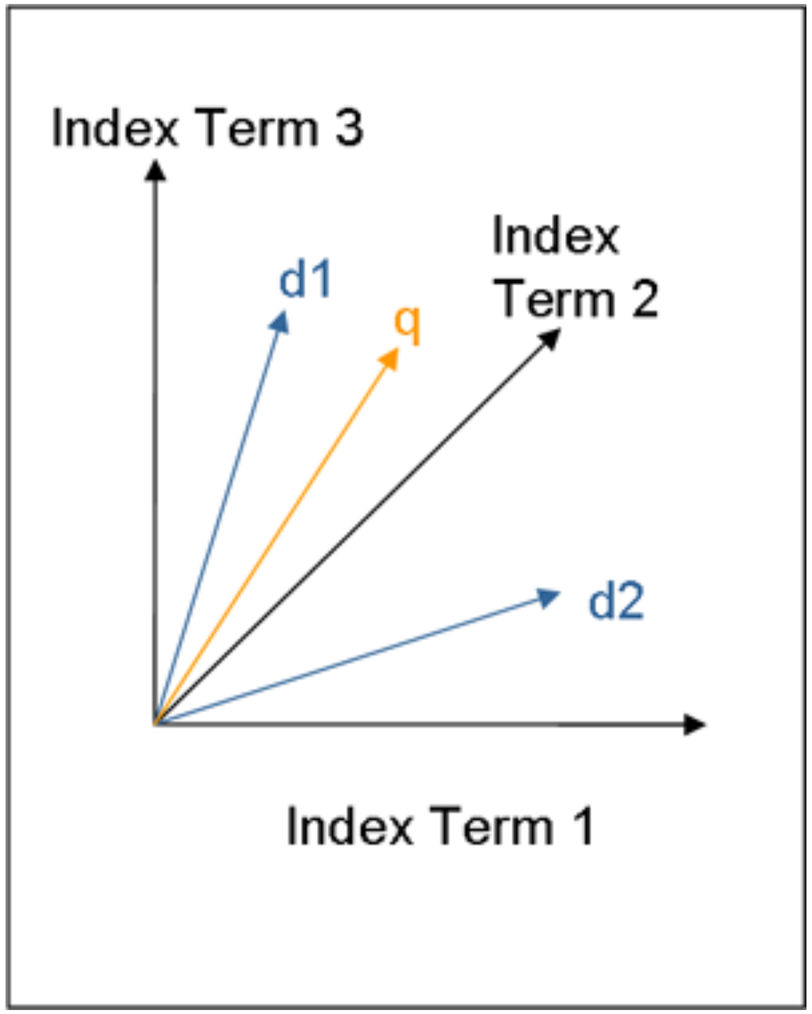
Machine Learning for Language Processing
ACS 2015/16
Stephen Clark
L6: Vector Space Models of Semantics

UNIVERSITY OF CAMBRIDGE

VSMs in Document Retrieval



UNIVERSITY OF CAMBRIDGE

Term-Frequency Model

Term vocabulary: $\langle$*England, Australia, Pietersen, Hoggard, run, wicket, catch, century, collapse*$\rangle$

Document $d1$: *Australia collapsed as Hoggard took 6 wickets . Flintoff praised Hoggard for his excellent line and length .*

Document $d2$: *Flintoff took the wicket of Australia 's Ponting , to give him 2 wickets for the innings and 5 wickets for the match .*

Query q: { *Hoggard, Australia, wickets* }

$$\vec{q1} \,.\, \vec{d1} = \langle 0,1,0,1,0,1,0,0,0 \rangle \,.\, \langle 0,1,0,2,0,1,0,0,1 \rangle = 4$$

$$\vec{q1} \,.\, \vec{d2} = \langle 0,1,0,1,0,1,0,0,0 \rangle \,.\, \langle 0,1,0,0,0,3,0,0,0 \rangle = 4$$

Figure 1. Simple example of document and query similarity using the dot product, with term-frequency providing the vector coefficients. The documents have been tokenised, and word matching is performed between lemmas (so *wickets* matches *wicket*).

TF-IDF Model

Term vocabulary: $\langle$*England, Australia, Pietersen, Hoggard, run, wicket, catch, century, collapse*$\rangle$

Document $d1$: *Australia collapsed as Hoggard took 6 wickets . Flintoff praised Hoggard for his excellent line and length .*

Document $d2$: *Flintoff took the wicket of Australia 's Ponting , to give him 2 wickets for the innings and 5 wickets for the match .*

Query q: { *Hoggard, Australia, wickets* }

$$\vec{q1} \cdot \vec{d1} = \langle 0,1,0,1,0,1,0,0,0 \rangle \cdot \langle 0, 1/10, 0, 2/5, 0, 1/100, 0, 0, 1/3 \rangle = 0.41$$

$$\vec{q1} \cdot \vec{d2} = \langle 0,1,0,1,0,1,0,0,0 \rangle \cdot \langle 0, 1/10, 0, 0/5, 0, 3/100, 0, 0, 0/3 \rangle = 0.13$$

Figure 2. Simple example of document and query similarity using the dot product, with term-frequency, inverse-document frequency providing the coefficients for the documents, using the same query and documents as Figure 1.

TF-IDF Model

$$
\begin{aligned}
\mathrm{Sim}(\overrightarrow{d}, \overrightarrow{q}) &= \frac{\overrightarrow{d}.\overrightarrow{q}}{\|\overrightarrow{d}\|\|\overrightarrow{q}\|} \\
&= \frac{\overrightarrow{d}.\overrightarrow{q}}{\sqrt{\sum_i d_i^2}\sqrt{\sum_i q_i^2}} \\
&= \mathrm{Cosine}(\overrightarrow{d}, \overrightarrow{q})
\end{aligned}
$$

Document and query length normalisation in combination with the dot product gives the cosine similarity measure

UNIVERSITY OF CAMBRIDGE

Term-Document Matrix

$$\begin{array}{l} \\ \textit{England} \\ \textit{Australia} \\ \textit{Pietersen} \\ \textit{Hoggard} \\ \textit{run} \\ \textit{wicket} \\ \textit{catch} \\ \textit{century} \\ \textit{collapse} \end{array}
\begin{array}{c} \begin{array}{cc} d1 & d2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1/10 & 1/10 \\ 0 & 0 \\ 2/5 & 0/5 \\ 0 & 0 \\ 1/100 & 3/100 \\ 0 & 0 \\ 0 & 0 \\ 1/3 & 0/3 \end{pmatrix} \end{array}$$

Figure 3. Term-document matrix for the simple running example, using *tf-idf* weights but without length normalisation.

Term-Document Matrix

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Documents are similar if they tend to contain the same (informative) terms

Terms are similar if they tend to occur in the same documents

A Finer Notion of Context

An automobile is a wheeled motor vehicle used for transporting passengers .
A car is a form of transport, usually with four wheels and the capacity to carry around five passengers .
Transport for the London games is limited , with spectators strongly advised to avoid the use of cars .
The London 2012 soccer tournament began yesterday , with plenty of goals in the opening matches .
Giggs scored the first goal of the football tournament at Wembley , North London .
Bellamy was largely a passenger in the football match , playing no part in either goal .

Term vocab: $\langle$*wheel, transport, passenger, tournament, London, goal, match*$\rangle$

$$\begin{array}{l} \\ automobile \\ car \\ soccer \\ football \end{array} \begin{array}{c} \begin{array}{ccccccc} wheel & transport & passenger & tournament & London & goal & match \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 2 & 1 \end{pmatrix} \end{array}$$

$automobile \,.\, car = 4$
$automobile \,.\, soccer = 0$
$automobile \,.\, football = 1$
$car \,.\, soccer = 1$
$car \,.\, football = 2$
$soccer \,.\, football = 5$

UNIVERSITY OF CAMBRIDGE

Alternative Definitions of Context

Giggs|NNP *scored*|VBD *the*|DT *first*|JJ *goal*|NN *of*|IN *the*|DT *football*|NN *tournament*|NN *at*|IN *Wembley*|NNP ,|, *North*|NNP *London*|NNP .|.

(`ncmod` _ *goal first*)
(`det` *goal the*)
(`ncmod` _ *tournament football*)
(`det` *tournament the*)
(`ncmod` _ *London North*)
(`dobj` *at Wembley*)
(`ncmod` _ *scored at*)
(`dobj` *of tournament*)
(`ncmod` _ *goal of*)
(`dobj` *scored goal*)
(`ncsubj` *scored Giggs* _)

Contextual elements for target word *goal* using a 7-word window method:
{*scored, the, first, of, football*}

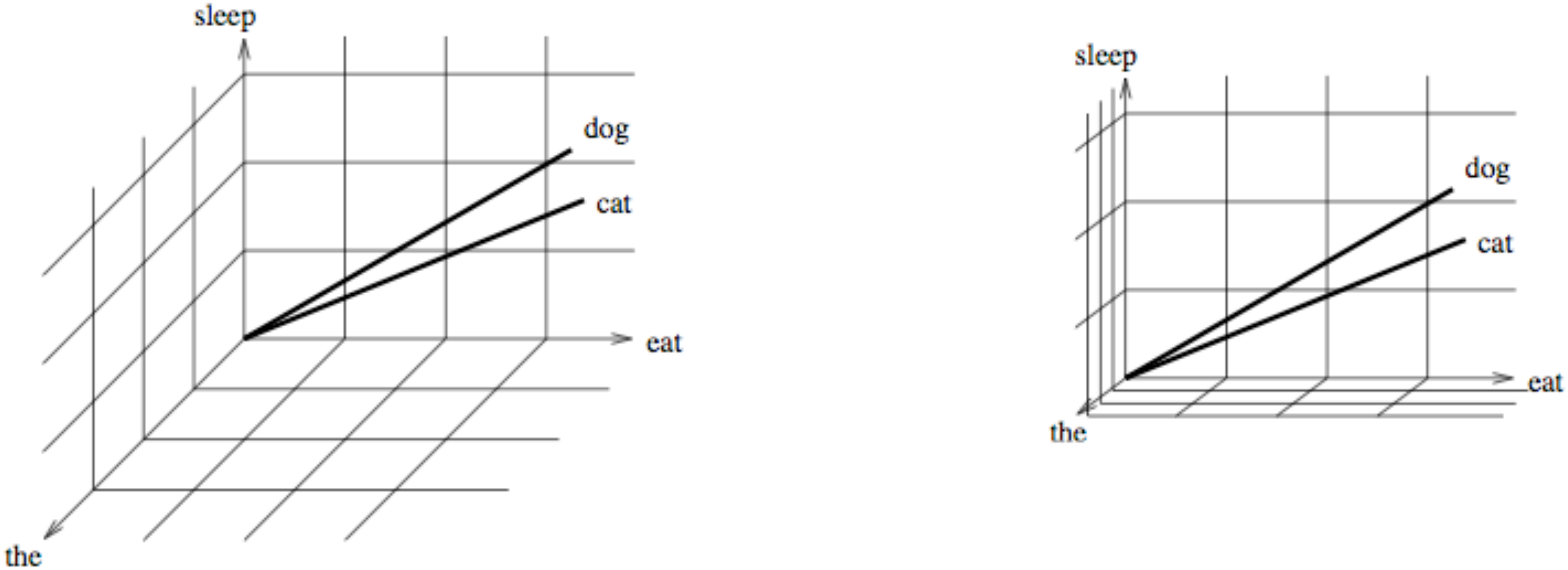
Contextual elements with parts-of-speech:
{*scored*|VBD, *the*|DET, *first*|JJ, *of*|IN, *football*|NN}

Contextual elements with direction (L for left, R for right):
{*scored*|L, *the*|L, *first*|L, *of*|R, *the*|R, *football*|R}

Contextual elements with position (e.g. 1L is 1 word to the left):
{*scored*|3L, *the*|2L, *first*|1L, *of*|1R, *the*|2R, *football*|3R}

Contextual elements as grammatical relations:
{*first*|`ncmod`, *the*|`det`, *scored*|`dobj`}

Weighting



The effect of IDF on a simple example vector space.

Similarity and Relatedness Datasets

love	sex	6.77
tiger	cat	7.35
tiger	tiger	10.00
computer	internet	7.58
plane	car	5.77
doctor	nurse	7.00
professor	doctor	6.62
smart	stupid	5.81
stock	phone	1.62

Some example human similarity/relatedness judgements from wordsim 353

UNIVERSITY OF CAMBRIDGE