Vector-space models of meaning

Christopher Potts

CS 244U: Natural language understanding
Jan 19

A corpus in matrix form

Upper left corner of a matrix derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
!	3	0	0	1	0	0	11	0	1	0
):	0	0	0	0	0	0	0	0	1	0
);	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	1	0
1/10	0	0	0	0	0	0	0	0	0	0
1/2	0	0	0	0	0	0	0	0	0	0
10	2	0	1	0	0	0	0	0	0	0
10/10	0	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0

Guiding hypotheses (Turney and Pantel 2010:153)

Statistical semantics hypothesis: Statistical patterns of human word usage can be used to figure out what people mean (Weaver, 1955; Furnas et al., 1983). – If units of text have similar vectors in a text frequency matrix,[13] then they tend to have similar meanings. (We take this to be a general hypothesis that subsumes the four more specific hypotheses that follow.)

Bag of words hypothesis: The frequencies of words in a document tend to indicate the relevance of the document to a query (Salton et al., 1975). – If documents and pseudo-documents (queries) have similar column vectors in a term–document matrix, then they tend to have similar meanings.

Distributional hypothesis: Words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1957; Deerwester et al., 1990). – If words have similar row vectors in a word–context matrix, then they tend to have similar meanings.

Extended distributional hypothesis: Patterns that co-occur with similar pairs tend to have similar meanings (Lin & Pantel, 2001). – If patterns have similar column vectors in a pair–pattern matrix, then they tend to express similar semantic relations.

Latent relation hypothesis: Pairs of words that co-occur in similar patterns tend to have similar semantic relations (Turney et al., 2003). – If word pairs have similar row vectors in a pair–pattern matrix, then they tend to have similar semantic relations.

Overview: great power, a great many design choices

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

(Nearly the full cross-product to explore; only a handful of the combinations are ruled out mathematically, and the literature contains relatively little guidance.)

Overview: great power, a great many design choices

tokenization
annotation
tagging
parsing
feature selection
⋮ cluster texts by date/author/discourse context/...
⇓ ⇙

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

(Nearly the full cross-product to explore; only a handful of the combinations are ruled out mathematically, and the literature contains relatively little guidance.)

General questions for vector-space modelers

- How do the rows (words, phrase-types, . . .) relate to each other?
- How do the columns (contexts, documents, . . .) relate to each other?
- For a given group of documents D, which words epitomize D?
- For a given a group of words W, which documents epitomize W (IR)?

Matrix designs

- I'm going to set aside pre-processing issues like tokenization — the best approach there will be tailored to your application.
- I'm going to assume that we would prefer not to do feature selection based on counts, stopword dictionaries, etc. — our VSMs should sort these things out for us!
- For more designs: Turney and Pantel 2010:§2.1–2.5, §6

Word × document

Upper left corner of a matrix derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
!	3	0	0	1	0	0	11	0	1	0
):	0	0	0	0	0	0	0	0	1	0
);	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	1	0
1/10	0	0	0	0	0	0	0	0	0	0
1/2	0	0	0	0	0	0	0	0	0	0
10	2	0	1	0	0	0	0	0	0	0
10/10	0	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0

Word × word

Upper left corner of a matrix derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`.

	!	):	);	1	1/10	1/2	10	10/10	100	11
!	343744	225	441	2582	264	254	3211	307	683	179
):	143	218	9	17	4	0	36	5	2	2
);	291	5	472	39	2	6	37	4	3	0
1	1871	14	30	1833	17	63	523	20	74	41
1/10	195	2	1	8	107	0	20	10	5	5
1/2	174	0	1	41	0	161	26	3	5	1
10	2212	16	29	319	13	18	2238	27	56	65
10/10	208	4	2	13	5	3	15	166	2	4
100	482	1	3	52	3	2	38	2	523	11
11	116	1	0	13	3	1	46	3	9	172

Word × discourse context

Upper left corner of an interjection × dialog-act tag matrix derived from the Switchboard Dialog Act Corpus (Stolcke et al. 2000):
`http://compprag.christopherpotts.net/swda-clustering.html`

	%	+	^2	^g	^h	^q	aa
absolutely	0	2	0	0	0	0	95
actually	17	12	0	0	1	0	4
anyway	23	14	0	0	0	0	0
boy	5	3	1	0	5	2	1
bye	0	1	0	0	0	0	0
bye-bye	0	0	0	0	0	0	0
dear	0	0	0	0	1	0	0
definitely	0	2	0	0	0	0	56
exactly	2	6	1	0	0	0	294
gee	0	3	0	0	2	1	1
goodness	1	0	0	0	2	0	0

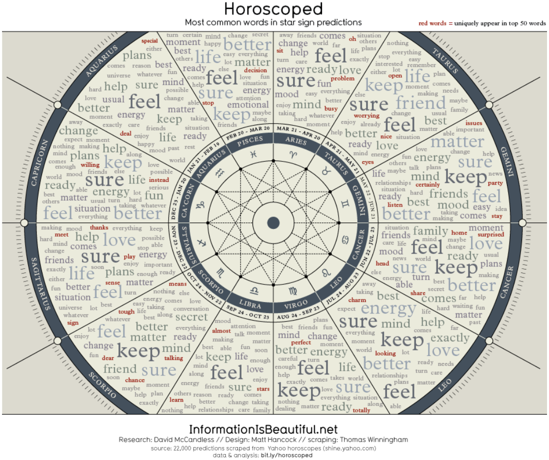
Other designs

- word × search query
- word × syntactic context
- pair × pattern (e.g., *mason* : *stone*, *cuts*)
- adj. × modified noun
- word × dependency rel.
- person × product
- word × person
- word × word × pattern
- verb × subject × object

$\vdots$

Challenge problem: Horoscoped

"Do horoscopes really all just say the same thing?"



http://www.informationisbeautiful.net/2011/horoscoped/

Challenge problem: Horoscoped

"Do horoscopes really all just say the same thing?"

Horoscoped

Unique words from the top 50 of each star sign

(scraped from daily predictions, common words tightly filtered)

2/3

	star sign	unique words	our interpretation
♒	aquarius	special, deal	bargain hunters?
♈	aries	busy, sit, problem	hard workers
♋	cancer	head, home, share, surprised	house cats
♑	capricorn	willing, instead	up for it?
♊	gemini	party, stay, issues, listen certainly	emotionally disturbed party animals who never say no
♌	leo	charm, looking	ever seductive
♎	libra	learn, stars, almost	nerds?
♓	pisces	stop, decision	just can't make up their minds
♐	sagittarius	thanks, sign, sense play, meet	they sound like fun!
♏	scorpio	chance, dear, means talking, tough	almost a sentence there
♉	taurus	nice, open, eyes worrying	naive?
♍	virgo	totally, perfect	hah! can it be true?

David McCandless – InformationIsBeautiful.net – data: bit.ly/horoscoped

http://www.informationisbeautiful.net/2011/horoscoped/

Challenge problem: Horoscoped

"Do horoscopes really all just say the same thing?"

"Ready? Sure?
Whatever the situation or secret moment, enjoy everything a lot.
Feel able to absolutely care. Expect nothing else. Keep making love.
Family and friends matter. The world is life, fun and energy.
Maybe hard. Or easy. Taking exactly enough is best.
Help and talk to others. Change your mind
and a better mood comes along…"

Meta-horoscope made from most common words in 4,000 star sign predictions

David McCandless – InformationIsBeautiful.net – data: bit.ly/horoscoped

http://www.informationisbeautiful.net/2011/horoscoped/

Challenge problem: Horoscoped

"Do horoscopes really all just say the same thing?"

Get my version of the data (restricted link):

`https://stanford.edu/class/cs224u/restricted/data/horoscoped.csv.zip`

Or: `/afs/ir/class/cs224u/restricted/data/horoscoped.csv.zip`

Sign	Texts
aquarius	2,744
aries	2,746
cancer	2,745
capricorn	2,744
gemini	2,745
leo	2,745
libra	2,745
pisces	2,746
sagittarius	2,740
scorpio	2,736
taurus	2,746
virgo	2,744
Total	32,926

80-texts per day	80-156
mean text length	54 words (median 43, std: 30)
token count	1,768,010
vocab size	23,091

Type	Texts
daily	30,634
monthly	432
weekly	1,860
Total	32,926

Category	Texts
career	5,129
extended	4,378
love	768
love-couples	4,375
love-flirt	4,375
love-singles	4,375
overview	5,147
teen	4,379
Total	32,926

Weighting and normalization

- This section focusses on methods for adjusting the counts in a matrix to better capture the underlying reationships.
- The examples are given in terms of word $\times$ document matrices, focussing on row-wise comparisons in places.
- The methods can also be applied column-wise, and to other kinds of matrices, though some (design, weighting) combos are better than others, as we will see.
- Further reading:
    - Manning and Schütze 1999:§15.2
    - Bullinaria and Levy 2007
    - Turney and Pantel 2010:§4.2

Relative frequencies

	d_1	d_2	d_3	d_4	d_5
A	10	15	0	9	10
B	5	8	1	2	5
C	14	11	0	10	9
D	13	14	10	11	12

Rows to $P(d|w)$ $\Rightarrow$

	d_1	d_2	d_3	d_4	d_5
A	0.23	0.34	0.00	0.20	0.23
B	0.24	0.38	0.05	0.10	0.24
C	0.32	0.25	0.00	0.23	0.20
D	0.22	0.23	0.17	0.18	0.20

Columns to $P(w|d)$ $\Downarrow$

	d_1	d_2	d_3	d_4	d_5
A	0.24	0.31	0.00	0.28	0.28
B	0.12	0.17	0.09	0.06	0.14
C	0.33	0.23	0.00	0.31	0.25
D	0.31	0.29	0.91	0.34	0.33

Dangers of prob. values: exaggerated estimates for small counts; comparisons that ignore differences in magnitude

Length (L2) normalization

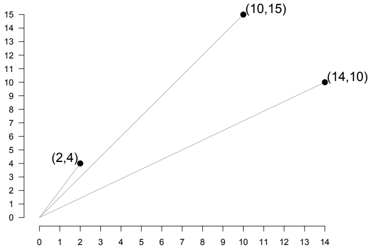
Definition (L2 normalization)

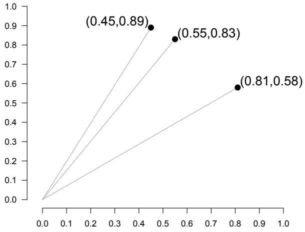
Given a vector x of dimension n, the normalization of x is a vector $\hat{x}$ also of dimension n obtained by dividing each element of x by $\sqrt{\sum_{i=1}^{n} x_i^2}$.

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm the rows $\Rightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58





Term Frequency–Inverse Document Frequency (TF-IDF)

Definition (TF-IDF)

For a corpus of documents D:

- Term frequency (TF): $P(w|d)$
- Inverse document frequency (IDF): $\log\left(\frac{|D|}{|\{d \in D | w \in d\}|}\right)$ (assume $\log(0) = 0$)
- TF-IDF: TF $\times$ IDF

	d_1	d_2	d_3	d_4
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

$\Rightarrow$

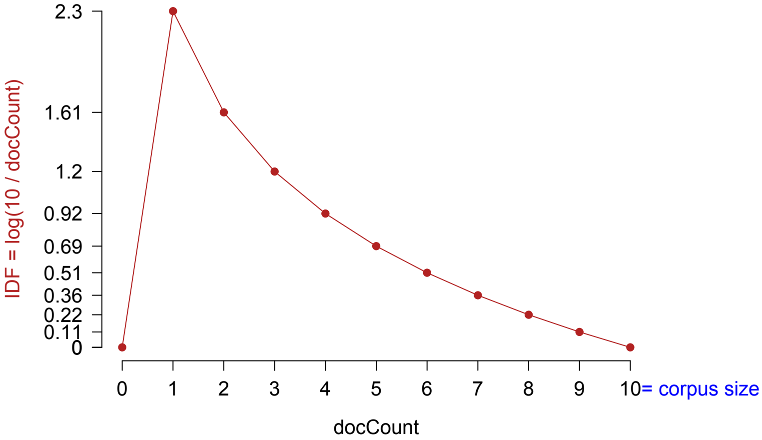
	IDF
A	0.00
B	0.29
C	0.69
D	1.39

$\Downarrow$

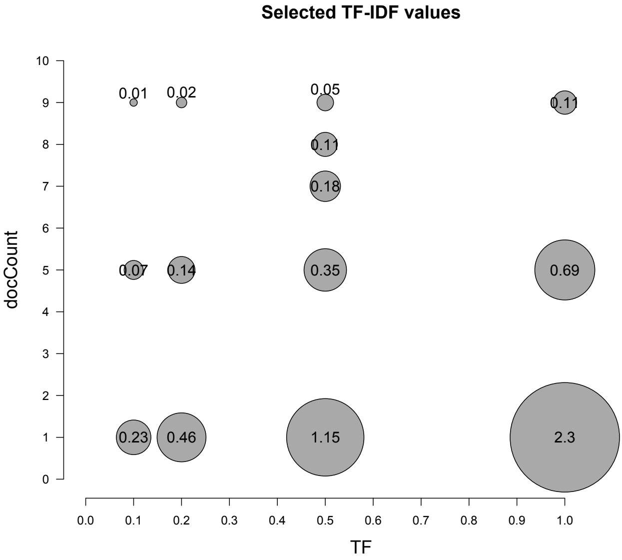
TF	d_1	d_2	d_3	d_4
A	0.33	0.33	0.50	0.91
B	0.33	0.33	0.50	0.00
C	0.33	0.33	0.00	0.00
D	0.00	0.00	0.00	0.09

TF-IDF	d_1	d_2	d_3	d_4
A	0.00	0.00	0.00	0.00
B	0.10	0.10	0.14	0.00
C	0.23	0.23	0.00	0.00
D	0.00	0.00	0.00	0.13

Term Frequency–Inverse Document Frequency (TF-IDF)



Term Frequency–Inverse Document Frequency (TF-IDF)



Pointwise Mutual Information (PMI)

Definition (PMI)

$$\log\left(\frac{P(w,d)}{P(w)P(d)}\right) \qquad \text{(assume } \log(0) = 0\text{)}$$

	d_1	d_2	d_3	d_4
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

$\Rightarrow$

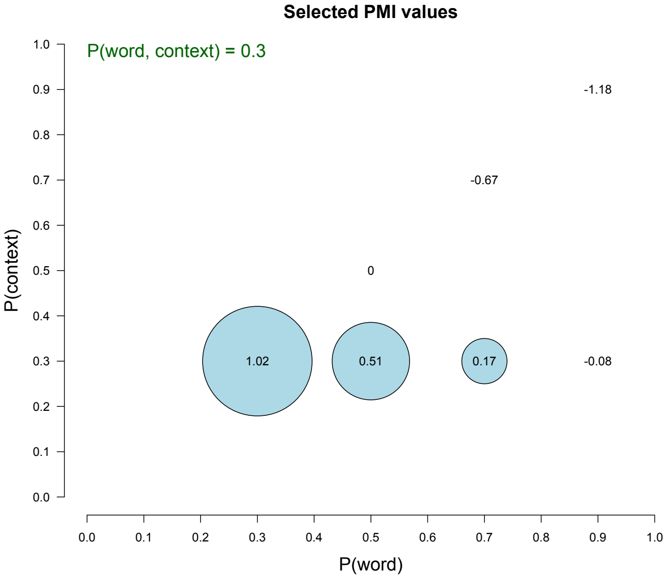
	$P(w,d)$				$P(w)$
A	0.11	0.11	0.11	0.11	0.44
B	0.11	0.11	0.11	0.00	0.33
C	0.11	0.11	0.00	0.00	0.22
D	0.00	0.00	0.00	0.01	0.01
$P(d)$	0.33	0.33	0.22	0.12	

PMI

$\Downarrow$

	d_1	d_2	d_3	d_4
A	−0.28	−0.28	0.13	0.73
B	0.01	0.01	0.42	0.00
C	0.42	0.42	0.00	0.00
D	0.00	0.00	0.00	2.11

Pointwise Mutual Information (PMI)



PMI with Lapacian smoothing

Definition (Lapacian smoothing)

Add a constant amount to all the counts.

	d_1	d_2	d_3	d_4
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

PMI $\Rightarrow$

	d_1	d_2	d_3	d_4
A	−0.28	−0.28	0.13	0.73
B	0.01	0.01	0.42	0.00
C	0.42	0.42	0.00	0.00
D	0.00	0.00	0.00	2.11

$\Downarrow +4$

	d_1	d_2	d_3	d_4
A	14	14	14	14
B	14	14	14	4
C	14	14	4	4
D	4	4	4	5

PMI $\Rightarrow$

	d_1	d_2	d_3	d_4
A	−0.17	−0.17	−0.17	−0.17
B	0.03	0.03	0.03	−1.23
C	0.52	0.52	−0.74	−0.74
D	0.30	0.30	0.30	0.52

PMI with contextual discounting

Definition (Contextual rescaling)

For a matrix with m rows and n columns:

$$\text{newpmi}_{ij} = \text{pmi}_{ij} \times \frac{f_{ij}}{f_{ij}+1} \times \frac{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})}{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})+1}$$

Count matrix

	d_1	d_2	d_3	d_4
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

PMI

	d_1	d_2	d_3	d_4
A	−0.28	−0.28	0.13	0.73
B	0.01	0.01	0.42	0.00
C	0.42	0.42	0.00	0.00
D	0.00	0.00	0.00	2.11

$f_{ij}/(f_{ij}+1)$

	d_1	d_2	d_3	d_4
A	0.91	0.91	0.91	0.91
B	0.91	0.91	0.91	0.00
C	0.91	0.91	0.00	0.00
D	0.00	0.00	0.00	0.50

$\frac{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})}{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})+1}$

	d_1	d_2	d_3	d_4	Sum
A	$\frac{30}{30+1}$	$\frac{30}{30+1}$	$\frac{20}{20+1}$	$\frac{11}{11+1}$	40
B	$\frac{30}{30+1}$	$\frac{30}{30+1}$	$\frac{20}{20+1}$	$\frac{11}{11+1}$	30
C	$\frac{30}{30+1}$	$\frac{30}{30+1}$	$\frac{20}{20+1}$	$\frac{11}{11+1}$	20
D	$\frac{1}{1+1}$	$\frac{1}{1+1}$	$\frac{1}{1+1}$	$\frac{1}{1+1}$	1
Sum	30	30	20	11	

Discounted PMI

	d_1	d_2	d_3	d_4
A	−0.24	−0.24	0.11	0.61
B	0.01	0.01	0.36	0.00
C	0.36	0.36	0.00	0.00
D	0.00	0.00	0.00	0.53

PMI with contextual discounting

Definition (Contextual rescaling)

For a matrix with m rows and n columns:

$$\text{newpmi}_{ij} = \text{pmi}_{ij} \times \frac{f_{ij}}{f_{ij}+1} \times \frac{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})}{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})+1}$$

Count matrix

	d_1	d_2	d_3	d_4
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

PMI

	d_1	d_2	d_3	d_4
A	−0.28	−0.28	0.13	0.73
B	0.01	0.01	0.42	0.00
C	0.42	0.42	0.00	0.00
D	0.00	0.00	0.00	2.11

$f_{ij}/(f_{ij}+1)$

	d_1	d_2	d_3	d_4
A	0.91	0.91	0.91	0.91
B	0.91	0.91	0.91	0.00
C	0.91	0.91	0.00	0.00
D	0.00	0.00	0.00	0.50

$\frac{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})}{\min(\sum_{k=1}^{m} f_{kj}, \sum_{k=1}^{n} f_{ik})+1}$

	d_1	d_2	d_3	d_4	Sum
A	0.97	0.97	0.95	0.92	40
B	0.97	0.97	0.95	0.92	30
C	0.95	0.95	0.95	0.92	20
D	0.50	0.50	0.50	0.50	1
Sum	30	30	20	11	

Discounted PMI

	d_1	d_2	d_3	d_4
A	−0.24	−0.24	0.11	0.61
B	0.01	0.01	0.36	0.00
C	0.36	0.36	0.00	0.00
D	0.00	0.00	0.00	0.53

Expected and observed/expected values

Definition (Expected values)

$$\text{expected}_{ij} = \sum_r \text{observed}_{ir} \times \left(\frac{\sum_k \text{observed}_{kj}}{\sum_{kr} \text{observed}_{kr}} \right)$$

Observed

	d_1	d_2	d_3	d_4	Sum
A	10	10	10	10	40
B	10	10	10	0	30
C	10	10	0	0	20
D	0	0	0	1	1
Sum	30	30	20	11	91

Expected

	d_1	d_2	d_3	d_4	Sum
A	13.19	13.19	8.79	4.84	40
B	9.89	9.89	6.59	3.63	30
C	6.59	6.59	4.40	2.42	20
D	0.33	0.33	0.22	0.12	1
Sum	30	30	20	11	91

Observed/Expected

	d_1	d_2	d_3	d_4
A	0.76	0.76	1.14	2.07
B	1.01	1.01	1.52	0.00
C	1.52	1.52	0.00	0.00
D	0.00	0.00	0.00	8.27

Other weighting/normalization schemes

- t-test: $\frac{p(w,d)-p(w)p(d)}{\sqrt{p(w)p(d)}}$
- Positive PMI: set all PMI values < 0 to 0
- TF-IDF variants that seek to be sensitive to the empirical distribution of words (Church and Gale 1995; Manning and Schütze 1999:553; Baayen 2001)

Relationships and generalizations

- Many weighting schemes end up favoring rare events that may not be trustworthy. Discounting procedures seek to combat this.
- The magnitude of counts can be important; $[1, 10]$ and $[1000, 10000]$ might represent very different situations; creating probability distributions or length normalizing will obscure this.
- TF-IDF severely punishes words that appear in many documents — it fails for dense matrices, which can include word $\times$ word matrices

Back to the Horoscoped challenge problem

Get my version of the data (restricted link):

`https://stanford.edu/class/cs224u/restricted/data/horoscoped.csv.zip`

Or: `/afs/ir/class/cs224u/restricted/data/horoscoped.csv.zip`

Sign	Texts
aquarius	2,744
aries	2,746
cancer	2,745
capricorn	2,744
gemini	2,745
leo	2,745
libra	2,745
pisces	2,746
sagittarius	2,740
scorpio	2,736
taurus	2,746
virgo	2,744
Total	32,926

80-texts per day	80-156
mean text length	54 words (median 43, std: 30)
token count	1,768,010
vocab size	23,091

Type	Texts
daily	30,634
monthly	432
weekly	1,860
Total	32,926

Category	Texts
career	5,129
extended	4,378
love	768
love-couples	4,375
love-flirt	4,375
love-singles	4,375
overview	5,147
teen	4,379
Total	32,926

Vector distance measures

- All the definitions are in terms of *distance* measures. They can be turned into similarity measures by subtracting appropriate constants.
- Examples focus on row vectors; the definitions and assessments hold for column-wise comparisons as well.
- Further reading:
  - van Rijsbergen 1979:§3
  - Manning and Schütze 1999:§8.5
  - Lee 1999
  - Bullinaria and Levy 2007
  - Turney and Pantel 2010:§4.4–4.5

Euclidean distance

Definition (Euclidean distance)

Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

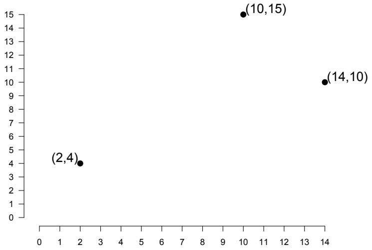
	d_x	d_y
A	2	4
B	10	15
C	14	10

Euclidean distance

Definition (Euclidean distance)

Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

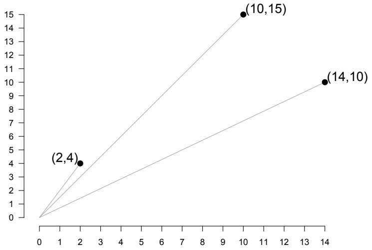


Euclidean distance

Definition (Euclidean distance)

Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

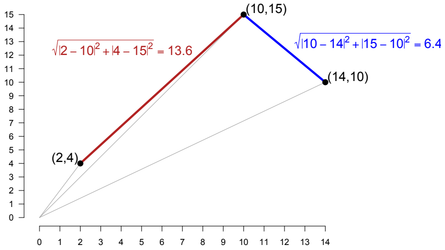


Euclidean distance

Definition (Euclidean distance)

Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

	d_x	d_y
A	2	4
B	10	15
C	14	10



Euclidean distance

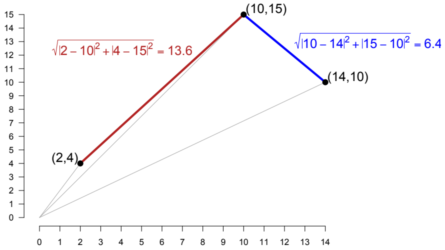
Definition (Euclidean distance)

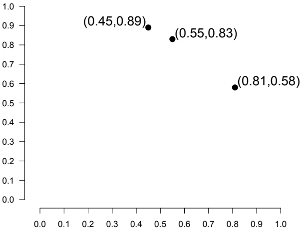
Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm the rows $\Longrightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58





Euclidean distance

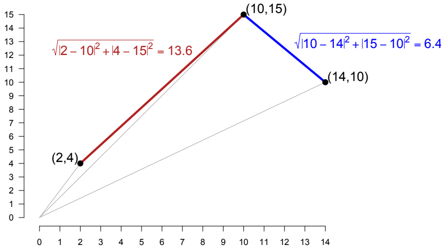
Definition (Euclidean distance)

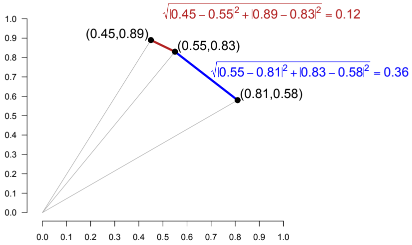
Between vectors x and y of dimension n: $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm the rows $\Rightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58



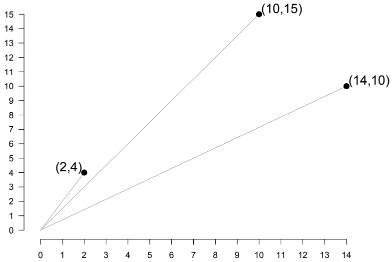


Cosine distance

Definition (Cosine distance)

Between vectors x and y of dimension n: $1 - \frac{\sum_{i=1}^{n} x_i \times y_i}{\|x\| \times \|y\|}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

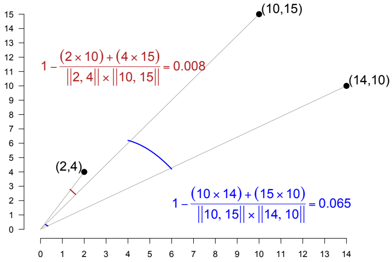


Cosine distance

Definition (Cosine distance)

Between vectors x and y of dimension n: $1 - \frac{\sum_{i=1}^{n} x_i \times y_i}{\|x\| \times \|y\|}$

	d_x	d_y
A	2	4
B	10	15
C	14	10



Cosine distance

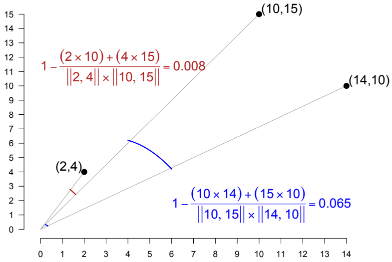
Definition (Cosine distance)

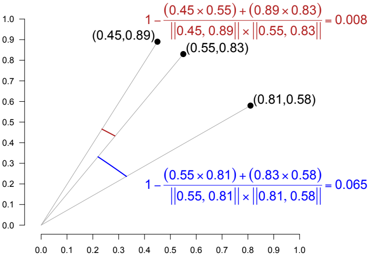
Between vectors x and y of dimension n: $1 - \frac{\sum_{i=1}^{n} x_i \times y_i}{\|x\| \times \|y\|}$

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm has no effect $\Rightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58





Dice and Jaccard distances

Definition (Dice distance; Dice 1945)

Between vectors x and y of dimension n:

$$1 - \frac{2 \times \sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} x_i + y_i}$$

Alternatively, define a mapping S_n from vectors to sets such that $S_n(v) = \{v_i > n\}$ for $n \geqslant 0$, and use $1 - \frac{2 \times |S_n(x) \cap S_n(y)|}{|S_n(x)| + |S_n(y)|}$

Definition (Jaccard distance)

Between vectors x and y of dimension n:

$$\frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)}$$

Alternatively, with S_n from above, use $\frac{|S_n(x) \cap S_n(y)|}{|S_n(x) \cup S_n(y)|}$

- Jaccard and Dice give different numerical values, with Jaccard penalizing large numerical differences more, but the two deliver identical rankings (van Rijsbergen 1979:§3; Lee 1999).
- Cosine distance penalizes large numerical differences less than both (Manning and Schütze 1999:299).
- Dice is not a true distance metric: it fails the triangle inequality.

KL divergence

Definition (KL divergence)

Between probability distributions p and q: $$D(p\|q) = \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i})$$

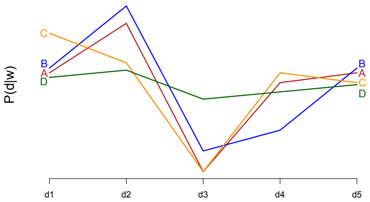
p is the reference distribution.

Before calculation, map all 0s to ϵ.

	d_1	d_2	d_3	d_4	d_5
A	10	15	0	9	10
B	5	8	1	2	5
C	14	11	0	10	9
D	13	14	10	11	12

Rows to prob. dists. $\Rightarrow$

	d_1	d_2	d_3	d_4	d_5
A	0.23	0.34	0.00	0.20	0.23
B	0.24	0.38	0.05	0.10	0.24
C	0.32	0.25	0.00	0.23	0.20
D	0.22	0.23	0.17	0.18	0.20



Word	KL distance from *A*	Rank
A	0.00	1
C	0.03	2
B	0.10	3
D	0.19	4

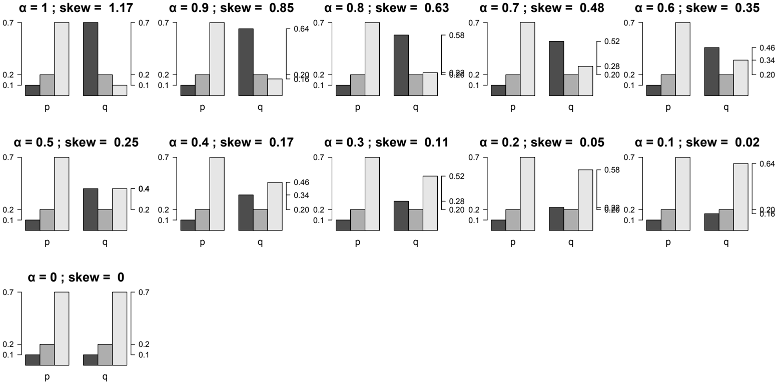
KL divergence with skew

Definition (α skew; Lee 1999)

Between probability distributions *p* and *q*:

$$\text{Skew}_{\alpha}(p, q) = D(p \| \alpha q + (1 - \alpha)p)$$

$$p = [0.1, 0.2, 0.7] \qquad q = [0.7, 0.2, 0.1] \qquad D(p\|q) = 1.17$$



Relationships and generalizations

1. Euclidean, Jaccard, and Dice with raw count vectors will tend to favor raw frequency over distributional patterns.
2. Euclidean with L2-normed vectors is equivalent to cosine w.r.t. ranking (Manning and Schütze 1999:301).
3. Jaccard and Dice are equivalent w.r.t. ranking.
4. Both L2-norms and probability distributions can obscure differences in the amount/strength of evidence, which can in turn have an effect on the reliability of cosine, normed-euclidean, and KL divergence. These shortcoming might be addressed through weighting schemes.
5. Skew is KL but with a preliminary step that gives special credence to the reference distribution.

Other vector distance measures

For vectors x and y of dimension n

Let $X = S_n(x)$ and $Y = S_n(y)$, where $S_n(v) = \{v_i > n\}$ for $n \geqslant 0$.

- Matching coefficient (counts): $\sum_{i=1}^{n} \min(x_i, y_i)$
- Matching coefficient (binary): $\left|X \cap Y\right|$
- Overlap (counts): $\frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\min\left(\sum_{i=1}^{n} x_i \ , \ \sum_{i=1}^{n} y_i\right)}$
- Overlap (binary): $\frac{\left|X \cap Y\right|}{\min\left(|X| \ , \ |Y|\right)}$
- Manhattan distance: $\sum_{i=1}^{n} |x_i - y_y|$

For probability distributions p and q

- Symmetric KL: $D(p\|q) + D(q\|p)$
- Jensen-Shannon: $\frac{1}{2}D(p\|\frac{p+q}{2}) + \frac{1}{2}D(q\|\frac{p+q}{2})$

Back to the Horoscoped challenge problem

Get my version of the data (restricted link):

`https://stanford.edu/class/cs224u/restricted/data/horoscoped.csv.zip`

Or: `/afs/ir/class/cs224u/restricted/data/horoscoped.csv.zip`

Sign	Texts
aquarius	2,744
aries	2,746
cancer	2,745
capricorn	2,744
gemini	2,745
leo	2,745
libra	2,745
pisces	2,746
sagittarius	2,740
scorpio	2,736
taurus	2,746
virgo	2,744
Total	32,926

80-texts per day	80-156
mean text length	54 words (median 43, std: 30)
token count	1,768,010
vocab size	23,091

Type	Texts
daily	30,634
monthly	432
weekly	1,860
Total	32,926

Category	Texts
career	5,129
extended	4,378
love	768
love-couples	4,375
love-flirt	4,375
love-singles	4,375
overview	5,147
teen	4,379
Total	32,926

Some experimental comparisons

- Matrices derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`:
  - word $\times$ document matrices: 3000 $\times$ 3456
  - word $\times$ word matrices: 3000 $\times$ 3000
- For word $\times$ document, all the reviews for each movie were pooled into a single document. (These matrices are sparse but not absurdly so.)
- For word $\times$ word, two words co-occur if they appear in the same document as defined above. (This gives really dense matrices.)
- For the sake of computational efficiency, the matrices contain only the top 3,000 words ordered by frequency. I did no additional filtering.
- Available:
  - `http://www.stanford.edu/class/cs224u/data/imdb-worddoc.csv.zip` (From your Stanford account: `/afs/ir/class/cs224u/WWW/data/imdb-worddoc.csv.zip`)
  - `http://www.stanford.edu/class/cs224u/data/imdb-wordword.csv.zip` (From your Stanford account: `/afs/ir/class/cs224u/WWW/data/imdb-wordword.csv.zip`)

outstanding (417 tokens): raw counts

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
delight	and	superb	and	great	excellent
successfully	as	supporting	as	as	performances
extraordinary	in	powerful	in	and	performance
fortunately	of	moving	is	best	wonderful
nonetheless	great	today	of	in	great
nowadays	who	perfectly	the	well	best
poignant	is	emotional	a	of	perfect
viewed	the	roles	to	very	as
marvelous	performance	tells	this	is	well

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
intense	performances	stunning	performances	performances	performances
stunning	excellent	recommended	performance	excellent	excellent
lovely	superb	intense	excellent	best	best
thoroughly	beautifully	lovely	best	performance	performance
delivers	brilliant	delivers	brilliant	as	as
fascinating	cinematography	fascinating	wonderful	brilliant	brilliant
tragic	strong	thoroughly	as	wonderful	wonderful
fresh	memorable	fresh	role	great	story
recommended	and	includes	great	role	great

good (14,841 tokens): raw counts

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
really	a	some	a	a	a
some	but	if	the	the	the
very	and	has	and	and	and
can	the	out	of	it	but
when	it	just	to	this	it
time	this	there	this	but	is
up	is	very	is	is	this
more	to	like	in	to	to
only	for	when	it	of	of

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
very	pretty	even	but	but	but
even	better	very	it	it	it
no	but	it's	this	this	this
it's	acting	no	really	really	really
up	worth	up	some	some	some
only	actually	only	like	like	like
time	basically	which	better	better	all
which	like	can	not	not	not
can	decent	time	all	all	better

outstanding (417 tokens): TF-IDF

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
a	viewed	superb	and	great	superb
of	remain	excellent	as	as	excellent
the	kim	supporting	is	excellent	wonderful
and	superb	wonderfully	of	very	performance
to	aware	wonderful	in	and	great
this	remarkable	perfect	the	time	best
in	adds	performances	a	best	perfect
viewed	existence	powerful	this	has	performances
remain	color	today	to	story	supporting

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
it's	performances	beautifully	performances	performances	performances
mother	excellent	stunning	excellent	excellent	excellent
complex	although	finest	wonderful	wonderful	wonderful
portrayal	wonderful	fascinating	brilliant	brilliant	brilliant
fantastic	gives	tragic	perfect	!	!
innocent	actor	provides	roles	10	10
convincing	perfect	surprising	although	?	?
superb	brilliant	terrific	!	a	a
minor	it's	physical	10	able	able

good (14,841 tokens): TF-IDF

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
but	a	i	the	a	a
is	the	but	a	the	the
it	is	not	of	and	and
that	and	as	and	of	is
for	of	was	this	is	of
in	this	are	to	this	to
with	to	for	is	to	but
i	but	movie	in	it	this
not	in	with	it	in	it

word × word

Fail! *good* co-occurs with every other word (document-level)!

outstanding (417 tokens): PPMI

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
and	superb	superb	and	superb	superb
the	excellent	terrific	of	excellent	wonderful
of	wonderful	date	is	wonderful	excellent
in	performance	10/10	great	performances	powerful
a	performances	emotional	as	performance	emotional
to	supporting	incredible	an	perfect	terrific
is	finest	powerful	in	great	performances
as	emotional	compelling	well	supporting	10/10
that	10/10	supporting	film	brilliant	supporting

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
performances	performances	performances	as	performances	performances
performance	performance	finest	and	as	performance
excellent	excellent	performance	an	and	wonderful
best	wonderful	superb	of	performance	excellent
wonderful	finest	portrayal	by	wonderful	as
brilliant	brilliant	excellent	performances	excellent	and
role	superb	wonderful	in	finest	finest
great	as	terrific	youth	an	superb
as	and	stunning	performance	superb	brilliant

good (14,841 tokens): PPMI

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
a	movie	movie	movie	movie	movie
is	bad	acting	this	this	bad
the	acting	very	a	but	acting
but	but	not	but	bad	but
and	very	bad	was	acting	not
of	not	really	i	not	this
this	this	i	is	i	very
to	was	like	it	was	i
in	i	was	not	like	was

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
it	really	really	better	really	really
but	pretty	better	really	better	better
really	movie	movie	pretty	pretty	pretty
this	better	lot	acting	acting	movie
like	acting	acting	entertaining	movie	acting
some	ok	pretty	lot	lot	lot
all	liked	like	some	ok	ok
so	watch	some	decent	watch	watch
have	it	watch	average	liked	liked

outstanding (417 tokens): PPMI with discounting

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
the	superb	superb	and	performances	superb
and	performances	performances	of	excellent	wonderful
of	excellent	wonderful	great	wonderful	performances
in	wonderful	terrific	is	superb	excellent
to	performance	excellent	as	performance	performance
a	great	supporting	well	great	brilliant
is	actor	10/10	in	perfect	emotional
that	supporting	date	an	brilliant	supporting
victoria	perfect	performance	film	supporting	perfect

word × word

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
outstanding	outstanding	outstanding	outstanding	outstanding	outstanding
performances	performances	performances	as	performances	performances
performance	performance	performance	and	as	performance
excellent	excellent	finest	an	performance	wonderful
best	wonderful	excellent	performances	and	excellent
as	finest	superb	of	wonderful	as
great	brilliant	wonderful	by	excellent	and
wonderful	superb	portrayal	in	finest	finest
story	as	terrific	youth	an	superb
brilliant	and	brilliant	performance	superb	brilliant

good (14,841 tokens): PPMI with discounting

word × document

Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
a	movie	movie	movie	movie	movie
the	acting	acting	this	this	acting
is	bad	very	a	but	bad
and	but	not	but	acting	but
but	very	but	was	bad	very
to	not	i	is	i	not
of	this	really	it	not	this
in	pretty	bad	i	was	i
that	is	was	not	a	really

word × word

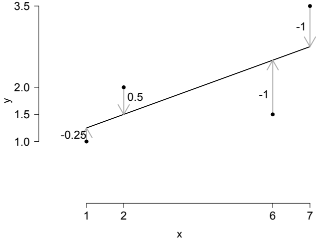
Euclidean	Cosine	Jaccard/Dice	KL	Skew95	Skew80
good	good	good	good	good	good
it	really	really	better	really	really
but	pretty	better	really	better	better
really	movie	movie	pretty	pretty	pretty
this	better	lot	acting	acting	movie
like	acting	acting	entertaining	movie	acting
some	ok	pretty	lot	lot	lot
all	liked	like	some	ok	ok
so	watch	some	decent	watch	watch
have	it	watch	average	liked	liked

Dimensionality reduction

- The goal of dimensionality reduction is eliminate rows/columns that are highly correlated while bringing similar things together and pushing dissimilar things apart.
- This section looks briefly at Latent Semantic Analysis (Deerwester et al. 1990), which seeks not only to find a reduced-sized matrix but also to capture similaries that come not just from direct co-occurrence, but also from second-order co-occurrence.
- Latent Semantic Analysis is an application of truncated singular value decomposition (SVD). SVD is a central matrix operation; ‘truncation’ here means looking only at submatrices of the full decomposition.
- For more:
  - Turney and Pantel 2010:§4.3
  - Manning and Schütze 1999:§15.4
  - Manning et al. 2009:§18

Latent Semantic Analysis (truncated singular value decomposition)

- I won't try to give a complete exposition of SVD. Instead, I'll try to convey the intuition in 2d and then work through an example.
- For the 2d case, SVD is closely related to fitting a least-squares regression, where the idea is to find a line that minimizes the errors (equivalently, whose vector of errors is orthogonal to the fitted line):



- The least-squares regression reduces the matrix to a line.
- Trunctated SVD, as applied in LSA, is the process of reducing a rectangular $m \times n$ matrix to an $i \times n$ matrix where $i \ll m$ or a $m \times j$ matrix where $j \ll n$.
- In the reduced dimension matrices, once-correlated variables are orthogonal and only the dimensions of greatest variation remain.

Example: toy dialect difference (*gnarly* for LA; *wicked* for Boston)

	d1	d2	d3	d4	d5	d6
gnarly	1	0	1	0	0	0
wicked	0	1	0	1	0	0
awesome	1	1	1	1	0	0
lame	0	0	0	0	1	1
terrible	0	0	0	0	0	1

$\Downarrow\Uparrow$

Distance from *gnarly*
1. gnarly
2. awesome
3. terrible
4. wicked
5. lame

Example: toy dialect difference (*gnarly* for LA; *wicked* for Boston)

	d1	d2	d3	d4	d5	d6
gnarly	1	0	1	0	0	0
wicked	0	1	0	1	0	0
awesome	1	1	1	1	0	0
lame	0	0	0	0	1	1
terrible	0	0	0	0	0	1

Distance from *gnarly*
1. gnarly
2. awesome
3. terrible
4. wicked
5. lame

$\Downarrow\Uparrow$

T(erm)					
gnarly	0.41	0.00	0.71	0.00	-0.58
wicked	0.41	0.00	-0.71	0.00	-0.58
awesome	0.82	-0.00	-0.00	-0.00	0.58
lame	0.00	0.85	0.00	-0.53	0.00
terrible	0.00	0.53	0.00	0.85	0.00

$\times$

S(ingular values)					
1	2.45	0.00	0.00	0.00	0.00
2	0.00	1.62	0.00	0.00	0.00
3	0.00	0.00	1.41	0.00	0.00
4	0.00	0.00	0.00	0.62	0.00
5	0.00	0.00	0.00	0.00	-0.00

$\times$

$\left(\text{D(ocument)} \right)^T$

D(ocument)					
d1	0.50	-0.00	0.50	0.00	-0.71
d2	0.50	0.00	-0.50	0.00	0.00
d3	0.50	-0.00	0.50	0.00	0.71
d4	0.50	-0.00	-0.50	-0.00	0.00
d5	-0.00	0.53	0.00	-0.85	0.00
d6	0.00	0.85	0.00	0.53	0.00

Example: toy dialect difference (*gnarly* for LA; *wicked* for Boston)

	d1	d2	d3	d4	d5	d6
gnarly	1	0	1	0	0	0
wicked	0	1	0	1	0	0
awesome	1	1	1	1	0	0
lame	0	0	0	0	1	1
terrible	0	0	0	0	0	1

$\Downarrow\Uparrow$

Distance from *gnarly*
1. gnarly
2. awesome
3. terrible
4. wicked
5. lame

T(erm)

gnarly	0.41	0.00	0.71	0.00	-0.58
wicked	0.41	0.00	-0.71	0.00	-0.58
awesome	0.82	-0.00	-0.00	-0.00	0.58
lame	0.00	0.85	0.00	-0.53	0.00
terrible	0.00	0.53	0.00	0.85	0.00

×

S(ingular values)

1	2.45	0.00	0.00	0.00	0.00
2	0.00	1.62	0.00	0.00	0.00
3	0.00	0.00	1.41	0.00	0.00
4	0.00	0.00	0.00	0.62	0.00
5	0.00	0.00	0.00	0.00	-0.00

×

D(ocument) ^{T}

d1	0.50	-0.00	0.50	0.00	-0.71
d2	0.50	0.00	-0.50	0.00	0.00
d3	0.50	-0.00	0.50	0.00	0.71
d4	0.50	-0.00	-0.50	-0.00	0.00
d5	-0.00	0.53	0.00	-0.85	0.00
d6	0.00	0.85	0.00	0.53	0.00

gnarly	0.41	0.00
wicked	0.41	0.00
awesome	0.82	-0.00
lame	0.00	0.85
terrible	0.00	0.53

×

2.45	0.00
0.00	1.62

=

gnarly	1.00	0.00
wicked	1.00	0.00
awesome	2.00	0.00
lame	0.00	1.38
terrible	0.00	0.85

Distance from *gnarly*
1. gnarly
2. wicked
3. awesome
4. terrible
5. lame

Other dimensionality reduction techniques

- Probabilistic LSA (PLSA; Hofmann 1999)
- Latent Dirichlet Allocation (LDA; Blei et al. 2003; Steyvers and Griffiths 2006)
- t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Geoffrey 2008)
- For even more: Turney and Pantel 2010:160

Tools

VSMs

- See Turney and Pantel 2010:§5 for lots of open-source projects
- Python NLTK's `text` and `cluster`: `http://www.nltk.org/`
- R's `topicmodels` package (mostly for LDA)

Visualization

- t-SNE implementations for dimensionality reduction and 2d visualization: `http://homepage.tudelft.nl/19j49/t-SNE.html`
- Gephi: `http://gephi.org/`

Looking ahead in the course

- VSMs and semantic composition (Socher et al. 2011)
- VSMs and sentiment analysis (Turney and Littman 2003)
- VSMS and relation extraction (see Turney and Pantel 2010:§2.3-2.4, §5.3)

References I

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Blei, David M.; Andrew Y. Ng; and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bullinaria, John A. and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3):510–526.

Church, Kenneth Ward and William Gale. 1995. Inverse dcument frequency (IDF): A measure of deviations from Poisson. In David Yarowsky and Kenneth Church, eds., *Proceedings of the Third ACL Workshop on Very Large Corpora*, 121–130. The Association for Computational Linguistics.

Deerwester, S.; S. T. Dumais; G. W. Furnas; T. K. Landauer; and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407. doi:\bibinfo{doi}{10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9}.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3):267–302.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. New York: ACM. doi:\bibinfo{doi}{http://doi.acm.org/10.1145/312624.312649}. URL `http://doi.acm.org/10.1145/312624.312649`.

Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 25–32. College Park, Maryland, USA: Association for Computational Linguistics. doi:\bibinfo{doi}{10.3115/1034678.1034693}. URL `http://www.aclweb.org/anthology/P99-1004`.

van der Maaten, Laurens and Hinton Geoffrey. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Manning, Christopher D.; Prabhakar Raghavan; and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

References II

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*. London: Buttersworth.

Socher, Richard; Jeffrey Pennington; Eric H. Huang; Andrew Y. Ng; and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 151–161. Edinburgh, Scotland, UK.: Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1014`.

Steyvers, Mark and Tom Griffiths. 2006. Probabilistic topic models. In Thomas K. Landauer; D McNamara; S Dennis; and W Kintsch, eds., *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum Associates.

Stolcke, Andreas; Klaus Ries; Noah Coccaro; Elizabeth Shriberg; Rebecca Bates; Daniel Jurafsky; Paul Taylor; Rachel Martin; Marie Meteer; and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26(3):339–371.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21:315–346. doi:\bibinfo{doi}{http://doi.acm.org/10.1145/944012.944013}. URL `http://doi.acm.org/10.1145/944012.944013`.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.