Distributed word representations: matrix reweighting

Chris Potts
Stanford Linguistics

CS 244U: Natural language understanding

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.
- So we should ask of each weighting scheme:

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.
- So we should ask of each weighting scheme:
  - How does it compare to the raw count values?

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.
- So we should ask of each weighting scheme:
    - How does it compare to the raw count values?
    - How does it compare to the word frequencies?

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.
- So we should ask of each weighting scheme:
    - How does it compare to the raw count values?
    - How does it compare to the word frequencies?
    - What overall distribution of values does it deliver?

Goal of reweighting and related questions

- Amplify the important, the trustworthy, the unusual; deemphasize the mundane and the quirky.
- Absent a defined objective function, this will remain fuzzy.
- The intuition behind moving away from raw counts is that frequency is a poor proxy for the above values.
- So we should ask of each weighting scheme:
    - How does it compare to the raw count values?
    - How does it compare to the word frequencies?
    - What overall distribution of values does it deliver?
- No feature selection based on counts, stopword dictionaries, etc.

Normalization

Definition (L2 norming)

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

Definition (Probability distribution)

Given a vector u of dimension n, the probability distribution of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\sum_{i=1}^{n} u_i$.

Vector L2 (length) normalization

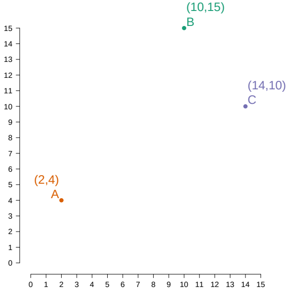
Definition

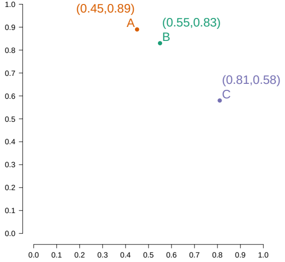
Given a vector u of dimension n, the normalization of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

	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





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

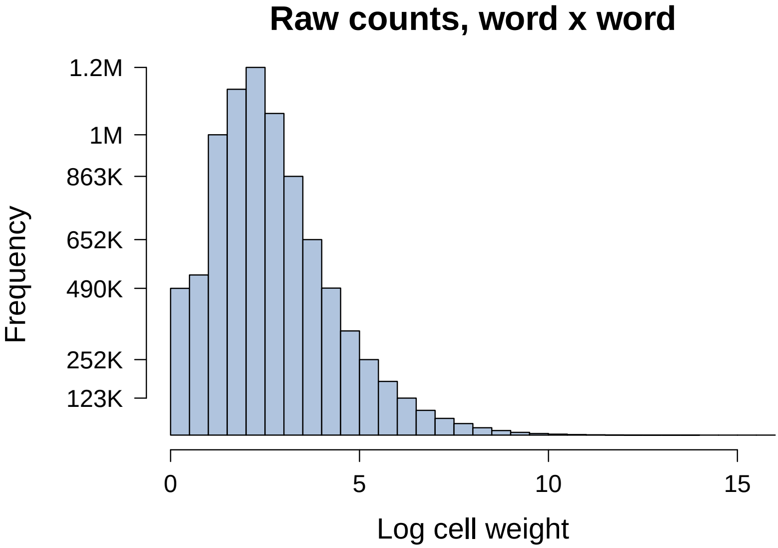
Columns to $P(w|d)$

$\Downarrow$

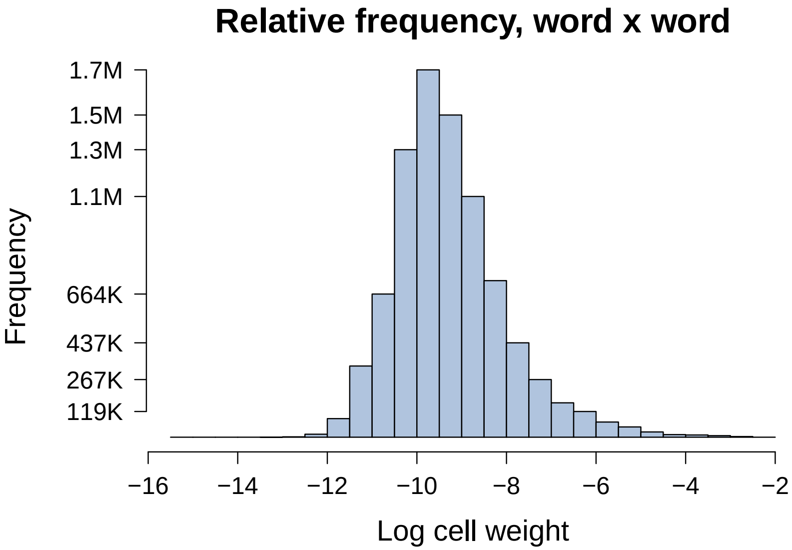
	d_1	d_2	d_3	d_4	d_5
A	0.24	0.31	0.00	0.28	0.28
B	0.12	0.17	0.09	0.06	0.14
C	0.33	0.23	0.00	0.31	0.25
D	0.31	0.29	0.91	0.34	0.33

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

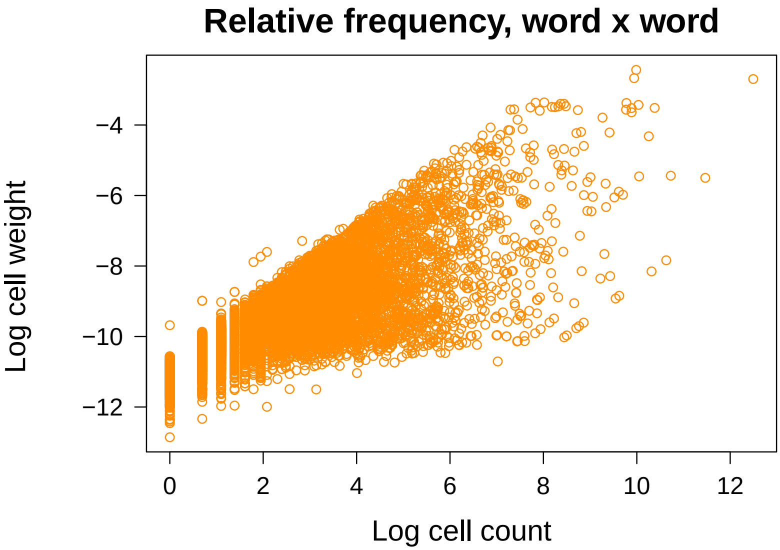
Relative frequencies compared to counts



Relative frequencies compared to counts



Relative frequencies compared to counts



Term Frequency–Inverse Document Frequency (TF-IDF)

Definition

For a corpus of documents D:

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

Term Frequency–Inverse Document Frequency (TF-IDF)

Definition

For a corpus of documents D:

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

Term Frequency–Inverse Document Frequency (TF-IDF)

Definition

For a corpus of documents D:

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

$\Longrightarrow$

	IDF
A	0.00
B	0.29
C	0.69
D	1.39

Term Frequency–Inverse Document Frequency (TF-IDF)

Definition

For a corpus of documents D:

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

Term Frequency–Inverse Document Frequency (TF-IDF)

Definition

For a corpus of documents D:

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

$\Rightarrow$

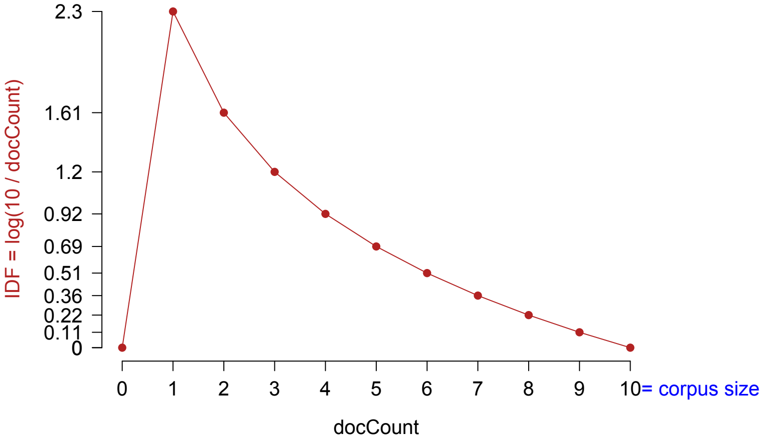
	IDF
A	0.00
B	0.29
C	0.69
D	1.39

$\Downarrow$

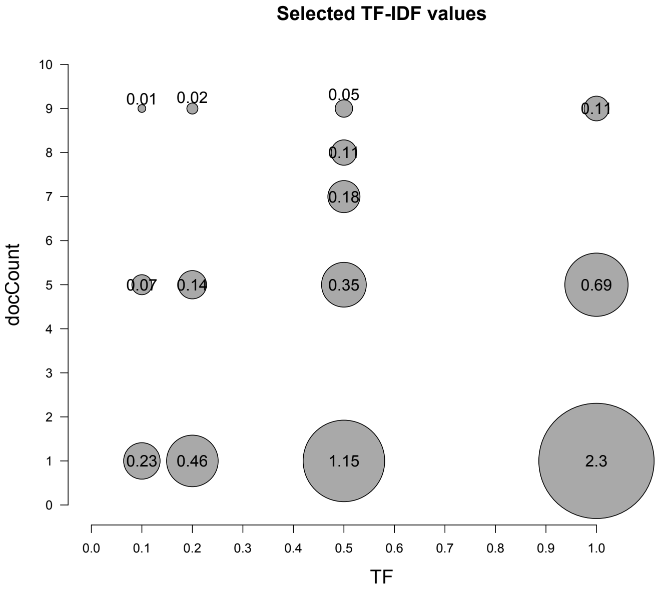
	TF			
	d_1	d_2	d_3	d_4
A	0.33	0.33	0.50	0.91
B	0.33	0.33	0.50	0.00
C	0.33	0.33	0.00	0.00
D	0.00	0.00	0.00	0.09

	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

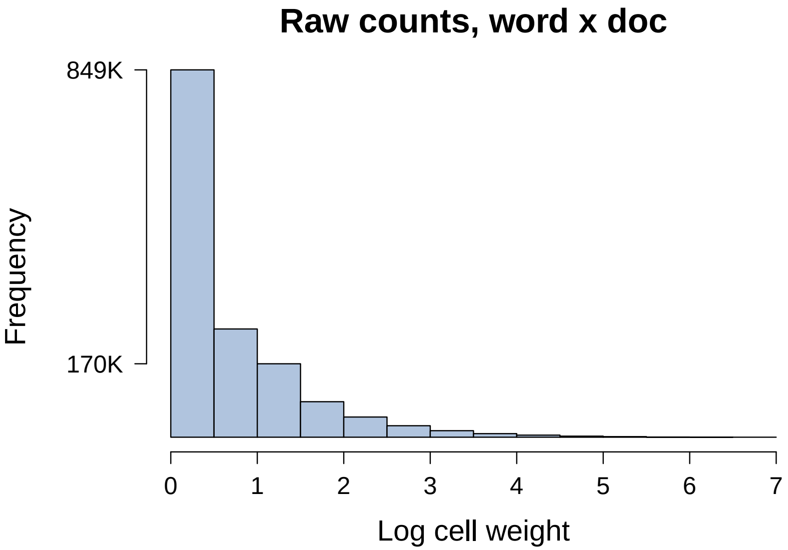
IDF values


2.3
1.61
1.2
0.92
0.69
0.51
0.36
0.22
0.11
0
IDF = log(10 / docCount)
0
1
2
3
4
5
6
7
8
9
10
= corpus size
docCount

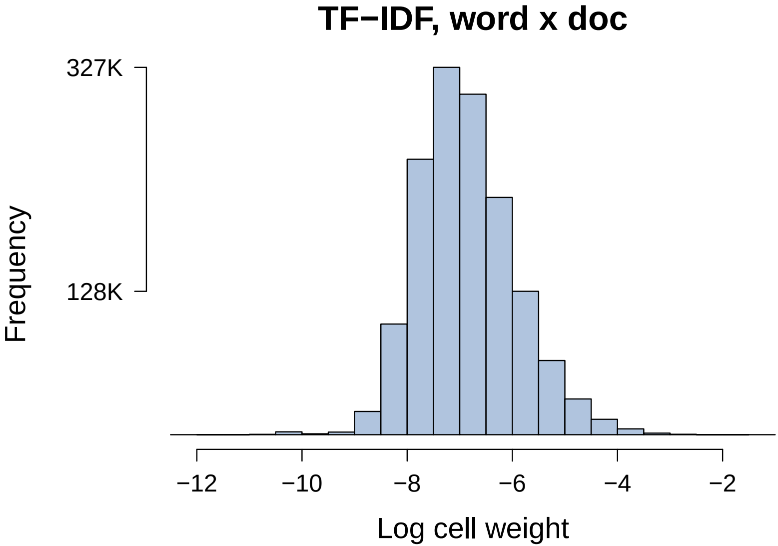
TF-IDF values



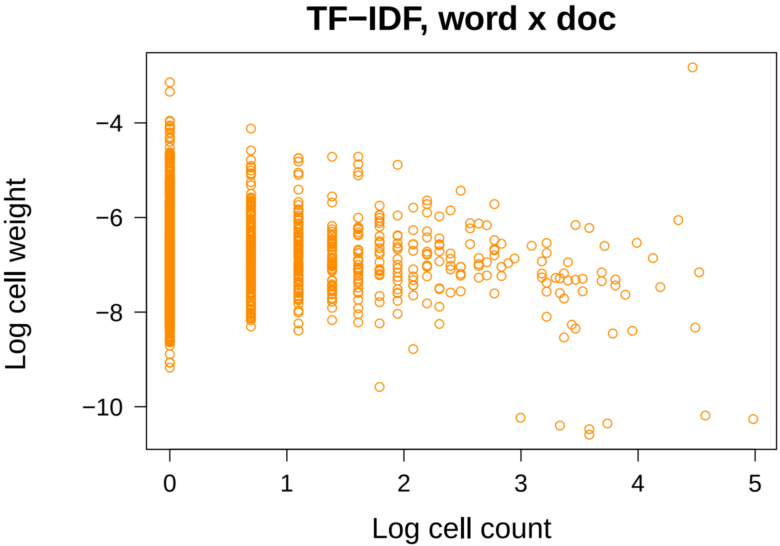
TF-IDF compared to counts



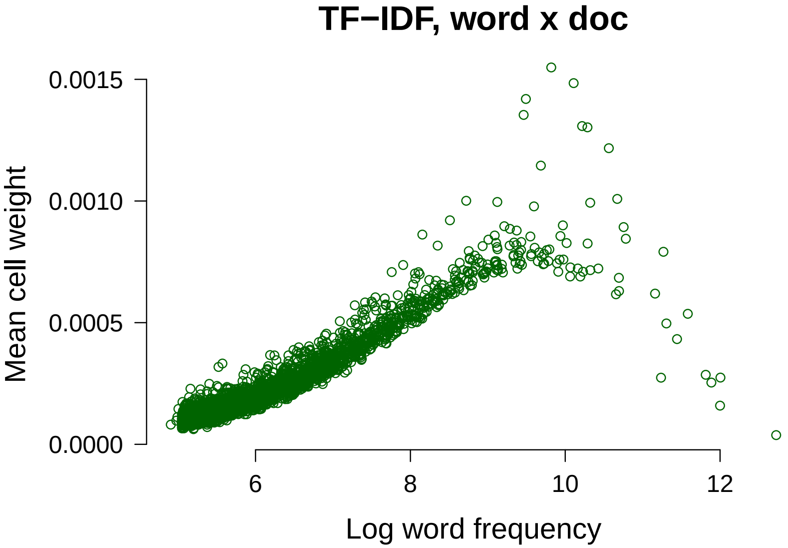
TF-IDF compared to counts



TF-IDF compared to counts



TF-IDF compared to counts



Pointwise Mutual Information (PMI)

Definition (PMI)

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

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

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

$\Longrightarrow$

	$P(w,d)$				$P(w)$
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	

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

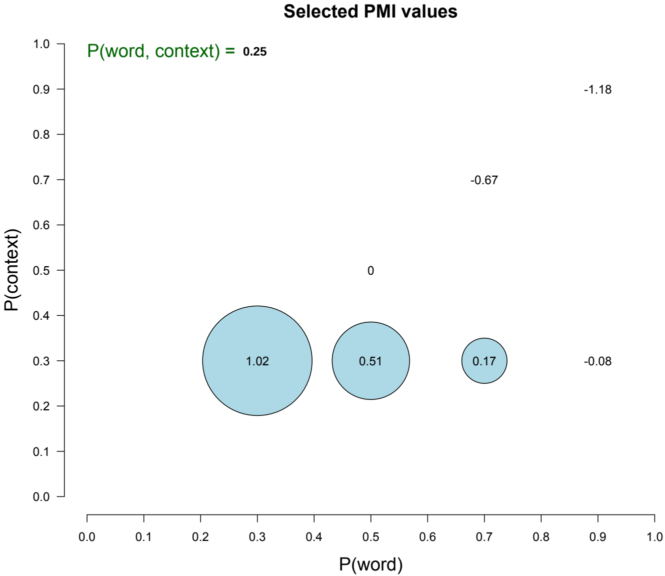
$\Rightarrow$

	$P(w,d)$				$P(w)$
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	

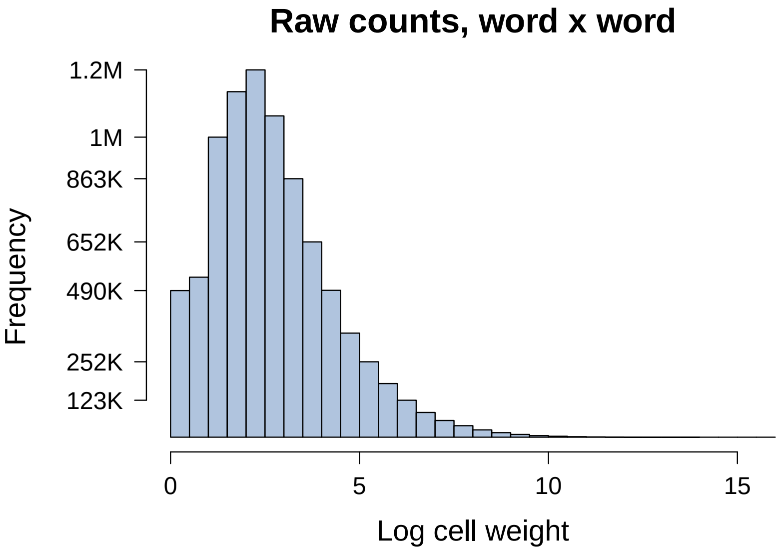
PMI $\Downarrow$

	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

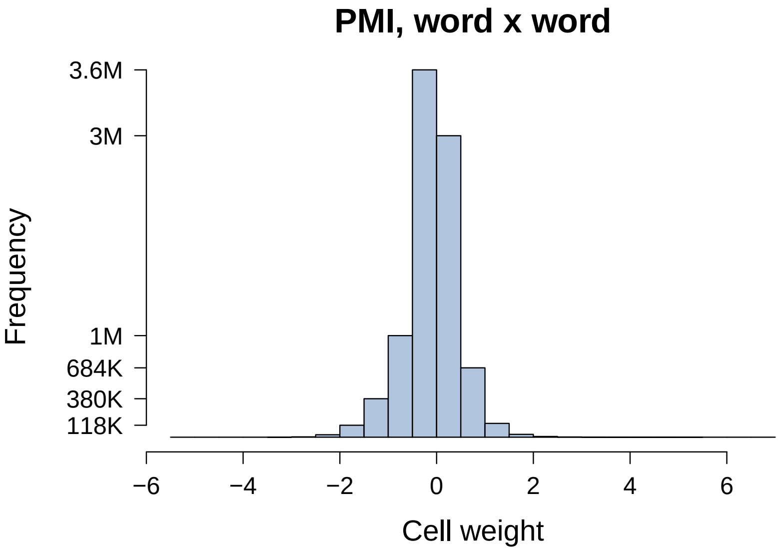
PMI values


Selected PMI values
P(word, context) = 0.25
-1.18
-0.67
0
1.02
0.51
0.17
-0.08
P(context)
P(word)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

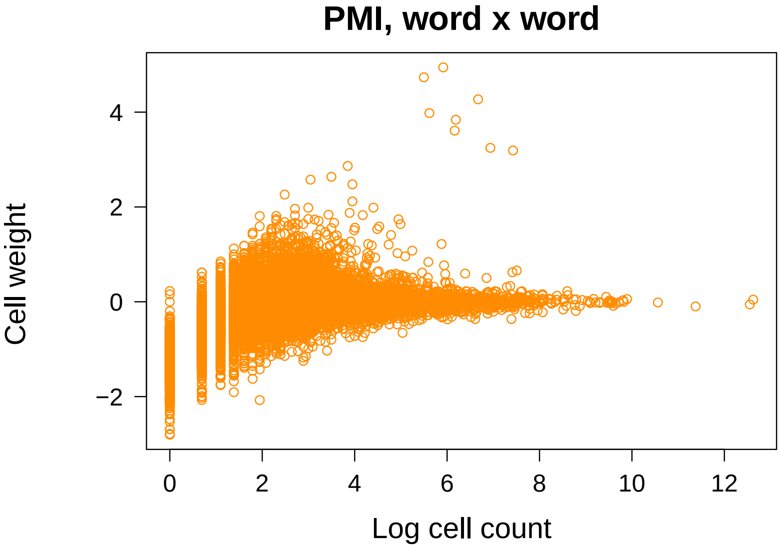
PMI compared to counts



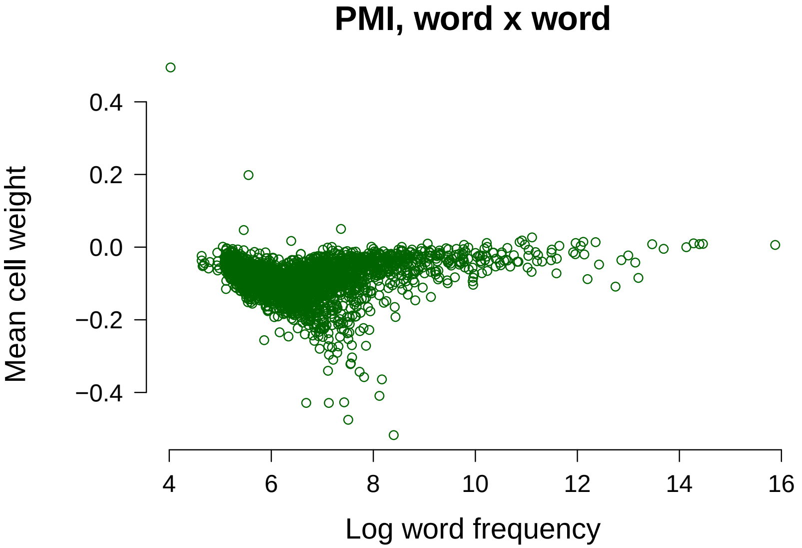
PMI compared to counts



PMI compared to counts



PMI compared to counts



PMI variants

PMI variants

Definition (Lapacian smoothing)

Add a constant amount to all the counts.

PMI variants

Definition (Lapacian smoothing)

Add a constant amount to all the counts.

Definition (Contextual discounting)

For a matrix with m rows and n columns:

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

PMI variants

Definition (Lapacian smoothing)

Add a constant amount to all the counts.

Definition (Contextual discounting)

For a matrix with m rows and n columns:

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

Definition (Positive PMI)

PPMI(w, d) = max(0, PMI(w, d))

Other weighting/normalization schemes

Other weighting/normalization schemes

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

Other weighting/normalization schemes

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

- t-test: $\frac{P(w,d) - P(w)P(d)}{\sqrt{P(w)P(d)}}$

Other weighting/normalization schemes

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

- t-test: $\frac{P(w,d) - P(w)P(d)}{\sqrt{P(w)P(d)}}$

- TF-IDF variants that seek to be sensitive to the empirical distribution of words (For discussion and references, see Manning and Schütze's textbook *Foundations of Statistical Natural Language Processing*, p. 553)

Other weighting/normalization schemes

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

- t-test: $\frac{P(w,d) - P(w)P(d)}{\sqrt{P(w)P(d)}}$

- TF-IDF variants that seek to be sensitive to the empirical distribution of words (For discussion and references, see Manning and Schütze's textbook *Foundations of Statistical Natural Language Processing*, p. 553)

- Pairwise distance matrices:

	d_x	d_y
A	2	4
B	10	15
C	14	10

cosine $\Rightarrow$

	A	B	C
A	0	0.008	0.116
B	0.008	0	0.065
C	0.116	0.065	0