

models and algorithms #1

MGMT|from text analysis to actionable knowledge

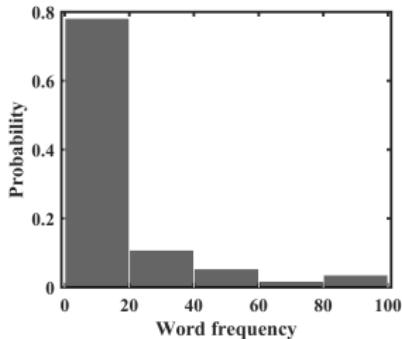
kristoffer l nielbo
kln@cas.au.dk

DTL|IMC|AARHUS UNIVERSITY-CAS

words are (one of) the basic units of meaning

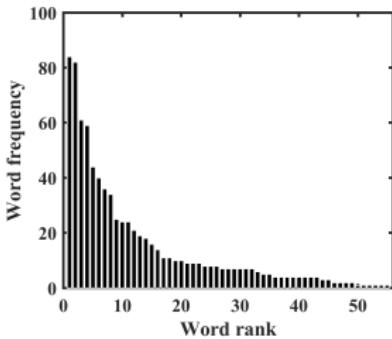
most **text analytics techniques rely on word frequencies**, that is, we tokenize a text at the word level and count the number of tokens for each type.

I am Daniel		a	1	59	0.073
I am Sam	'I' 'am' 'Daniel' 'I' 'am'	am	1	16	0.02
Sam I am	'Sam' 'Sam' 'I' 'am'	and	1	24	0.03
That Sam-I-am	'That' 'Sam' 'I' 'am'	anyhwhere	1	1	0.001
That Sam-I-am!	'That' 'Sam' 'I' 'am' 'I'	anywhere	1	7	0.009
I do not like	'do' 'not' 'like' 'that'	...			
that Sam-I-am	'Sam' 'I' 'am' ...	you	1	34	0.042
...		total	55	804	1.0



most words are infrequent, but a few words are very frequent

'i' 'not' 'them' 'a' 'like' 'in' 'do' 'you'
'would'



model a text as a distribution over words.
Some words are more likely than other.

often times we are looking at the mid-range (not too likely and not too unlikely).

binary term frequency: $f_{t,d} = 0, 1$

raw term frequency: $f_{t,d} = N(t, D)$

normalized term frequency ¹ $f_{t,d} = \frac{N(t, D)}{N(D)}$

IDF weighted term frequency ²: $tfidf(t, d, D) = f_{t,d} \cdot \log \frac{|D|}{\{d \in D : t \in d\}}$

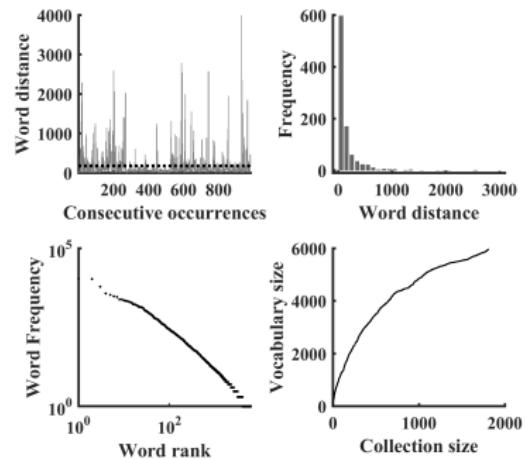
¹ prevents bias towards longer documents

² removes non-informative words

words occur in *bursts* - if word occurs it is likely to occur again in close proximity

a word's frequency is *inversely proportional* to its rank

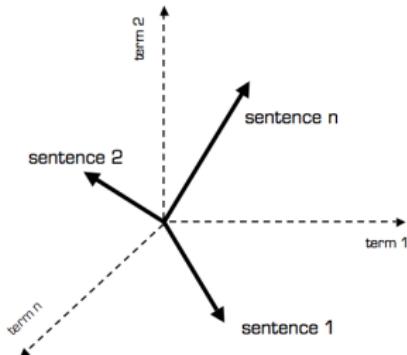
the vocabulary increases as a function of the number of texts, but the increase diminishes as more texts are included



any corpus (i.e., a collection of n documents) can be represented in the vector space model by a **document-term matrix**

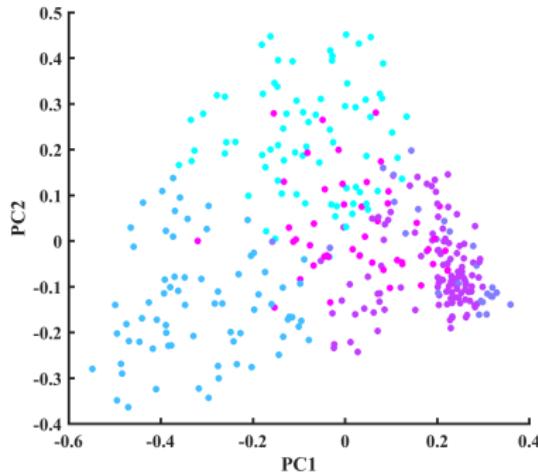
a vector space model is a basic modeling mechanism for a word- or document-space (whether we look at rows or columns)

- a document vector with only one word is collinear to the vocabulary word axis
- a document vector that does not contain a specific word is orthogonal/perpendicular to the word axis
- two documents are identical if they contain the same words in a different order



Document space $t_1 \quad t_2 \quad t_3 \quad \dots \quad t_n$ ← Term vector space

D_1	a_{11}	a_{12}	a_{13}	\dots	a_{1n}
D_2	a_{21}	a_{22}	a_{23}	\dots	a_{2n}
D_3	a_{31}	a_{32}	a_{33}	\dots	a_{3n}
\dots					
D_m	a_{m1}	a_{m2}	a_{m3}	\dots	a_{mn}
Q	b_1	b_2	b_3	\dots	b_n

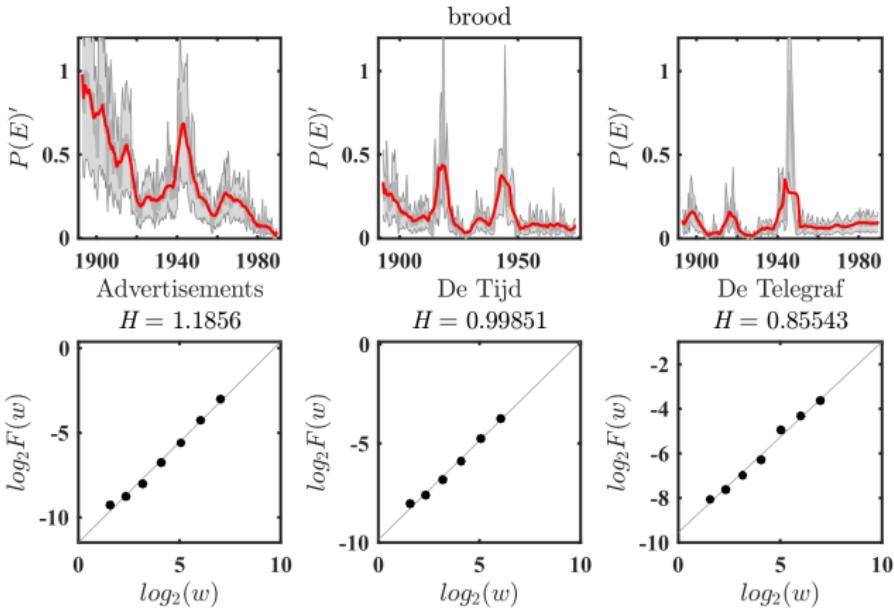


KEYWORDS

brother	abu	Al Qaeda	America	Allah*
fight	brother	AQAP	attack	Islam
mujahidin	children	Inspire	bomb	message
ummah	love	issue	people	Muslim
women	people	magazine	world	time

word counts can capture groups of semantic features in the word frequency mid-range (minimize computation & maximize discrimination)

quick and dirty 'thematic' analysis when prior knowledge and metadata are lacking



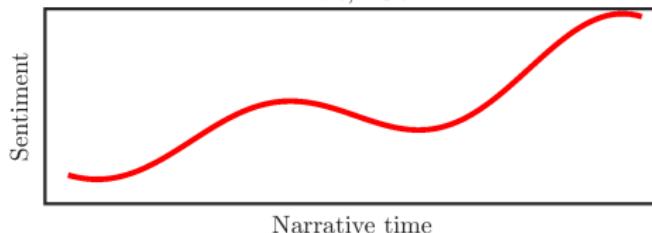
a dictionary can be used to extract the frequency of cognitive and affective keywords from a collection of documents and perform a **sentiment analysis**

```
1 'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S. citizen
2 so she could use her in the debate?'
3
4 Positive sex, citizen
5 Negative crooked, hillary, disgusting, out
6 Sentiment Score (2+1) + (-2-1-3-1) = -4
7 Sentiment Polarity Negative
8 Overall Score Sum of all sentence scores
```

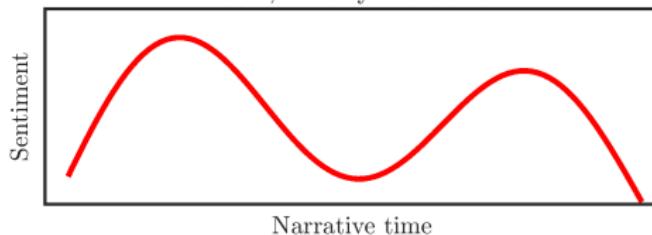
word ratings for dictionaries (sentiment scores) are based on more or less principled procedures

techniques originate in psychological and sociological scale studies and span a range of approaches (dictionary, machine learning)

Bible, KJV



Koran, Arberry Translation

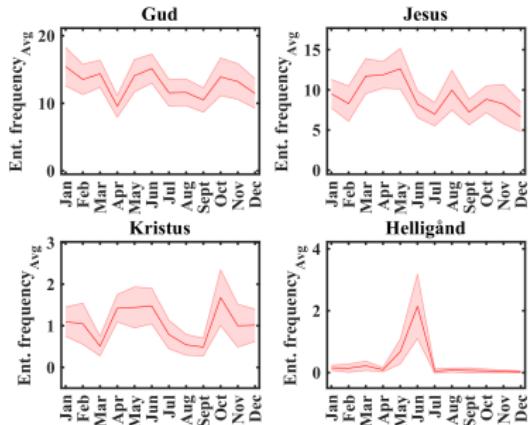


words acquire meaning from their mutual dependencies to or **co-occurrence** with other words

```
1 [1] "it hath been said thou shalt love thy neighbour and hate
2 thine enemy but i say unto you love your enemies"
3 [2] "no man can serve two masters for either he will hate the
4 one and love the other or else he will"
```

identify words that collocate with a specific probe or node words beyond chance level

used for tracking consumer preferences, text retrieval, recommender systems, OCR ...

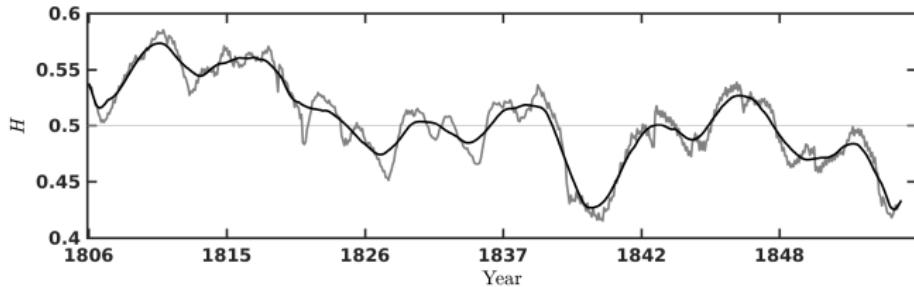


named entities are unique identifiers for entities (a proper noun serving as a name for someone or something)

detection of proper names and
classification into categories: person,
localization, organization

recognition of other data/time,
percentage, money, email addresses ...

more general, given concrete semantics,
locate elements in text that fit those
semantics



measures of **lexical density** or variation can capture global text features that are indicative of age, education, and cognitive impairment: $L_d = \frac{N_{lex}}{N} \times 100$

“a rose is a rose is a rose” is less lexically dense than “a rose is red and thorny”

lexical density \sim text predictability: $Entropy = \sum_{i=1}^n p_i \log_2 p_i$

