models and algorithms #1

MGMT|from text analysis to actionable knowledge

kristoffer l nielbo
kln@cas.au.dk

DTL|IMC|AARHUS UNIVERSITY-CAS

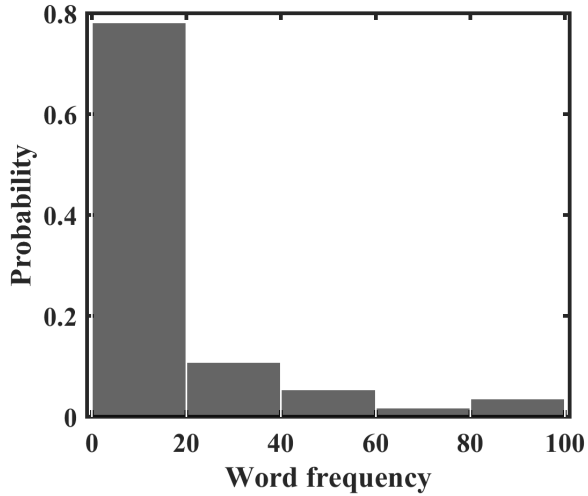
IMC
INTERACTING MINDS CENTRE

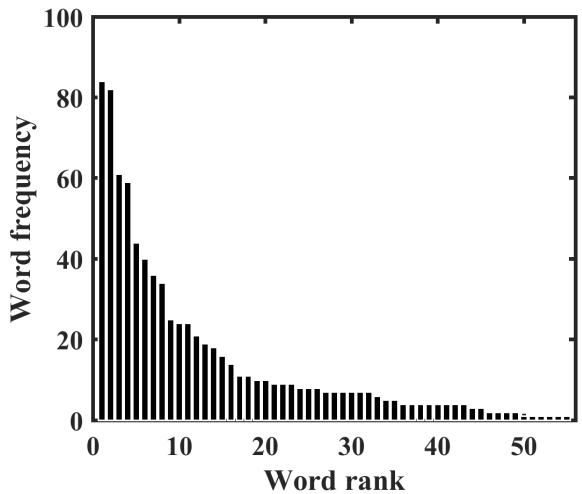
words are (one of) the basic units of meaning

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	
I am Sam	'I' 'am' 'Daniel' 'I' 'am'
Sam I am	'Sam' 'Sam' 'I' 'am'
That Sam-I-am	'That' 'Sam' 'I' 'am'
That Sam-I-am!	'That' 'Sam' 'I' 'am' 'I'
I do not like	'do' 'not' 'like' 'that'
that Sam-I-am	'Sam' 'I' 'am' ...
...	

a	1	59	0.073
am	1	16	0.02
and	1	24	0.03
anyhwhere	1	1	0.001
anywhere	1	7	0.009
...			
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]: $tfidf(t, d, D) = f_{t,d} \cdot log \frac{\mid D \mid}{\{d \in D : t \in d\}}$

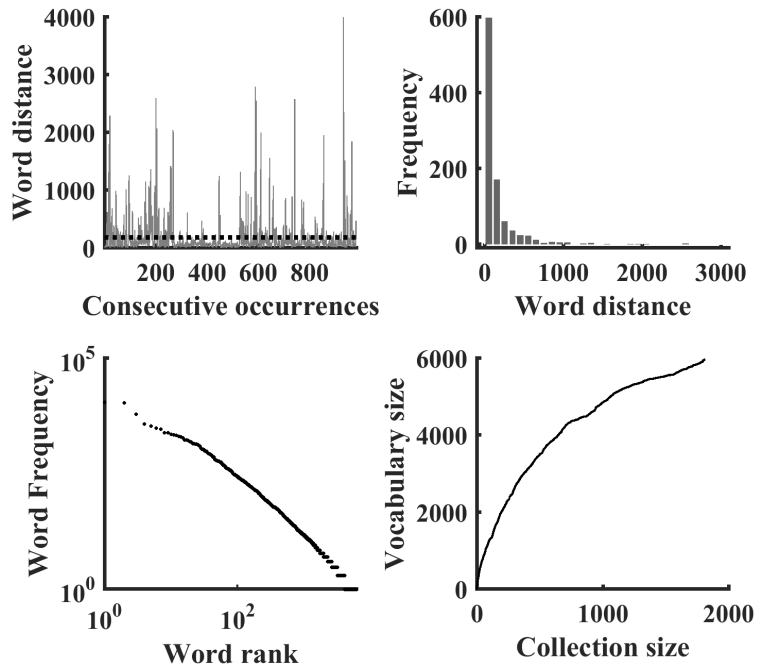
[1] prevents bias towards longer documents

[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

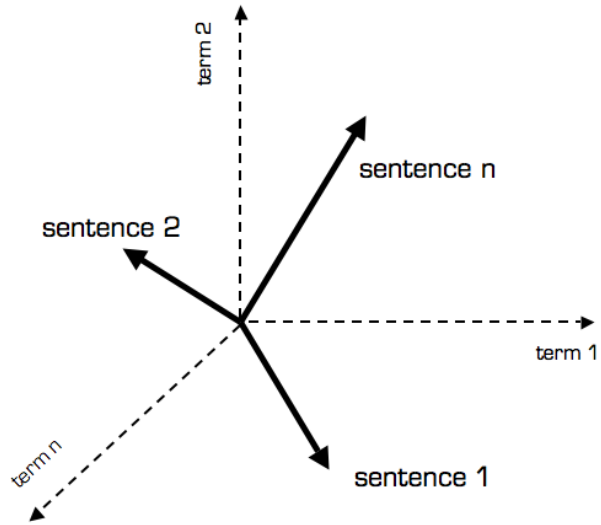
the vocabulary increases as a function of the number of texts, but the increase diminishes as more texts are included



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

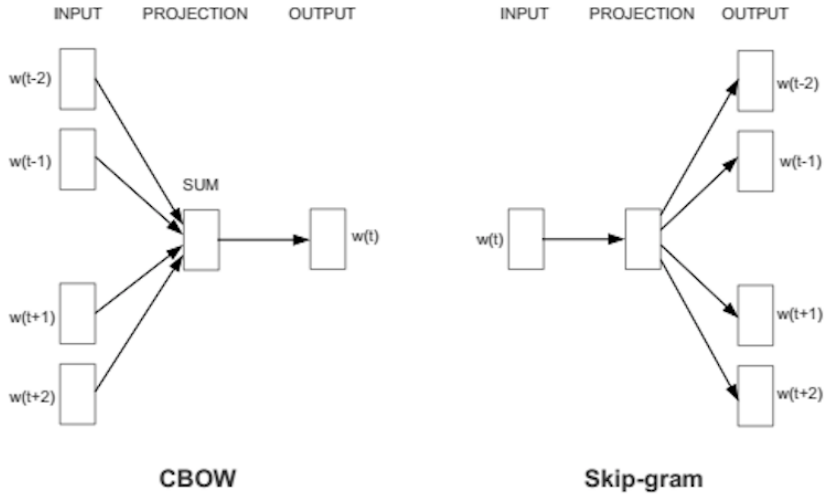


Document space	t_1	t_2	t_3	...	t_n ← Term vector space
D_1	a_{11}	a_{12}	a_{13}	...	a_{1n}
D_2	a_{21}	a_{22}	a_{23}	...	a_{2n}
D_3	a_{31}	a_{32}	a_{33}	...	a_{3n}
...					
D_m	a_{m1}	a_{m2}	a_{m3}	...	a_{mn}
Q	b_1	b_2	b_3	...	b_n

sample = {'I enjoy flying.', 'I like NLP.' 'I like deep learning.'} at *word window* = 1.[3]

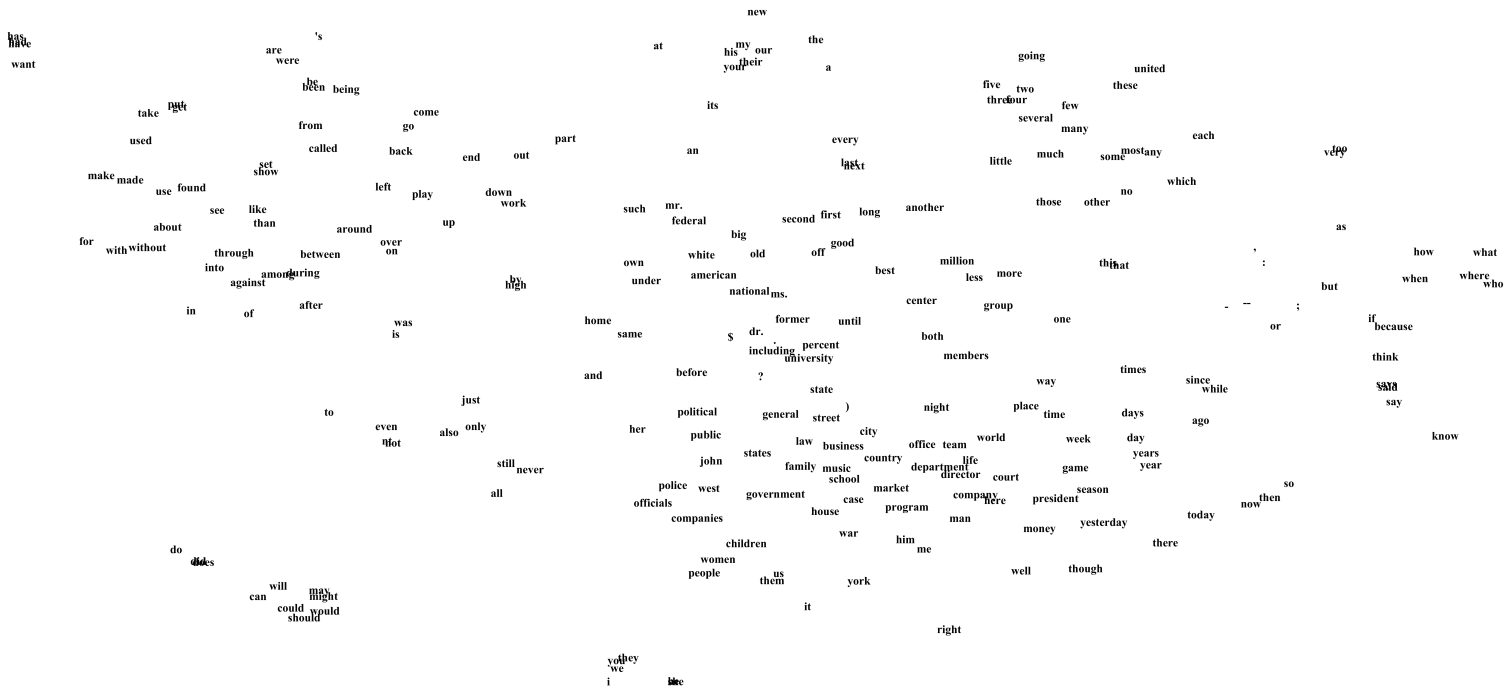
$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

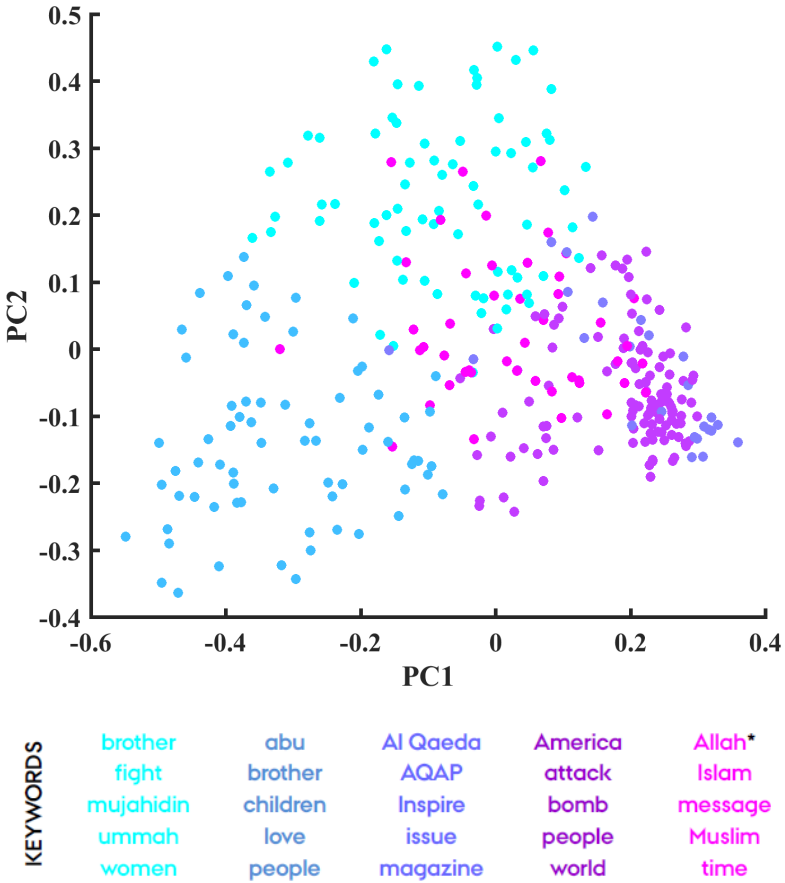
[3] example from Franck Dernoncourt



$$king - man + woman = queen$$ [4]

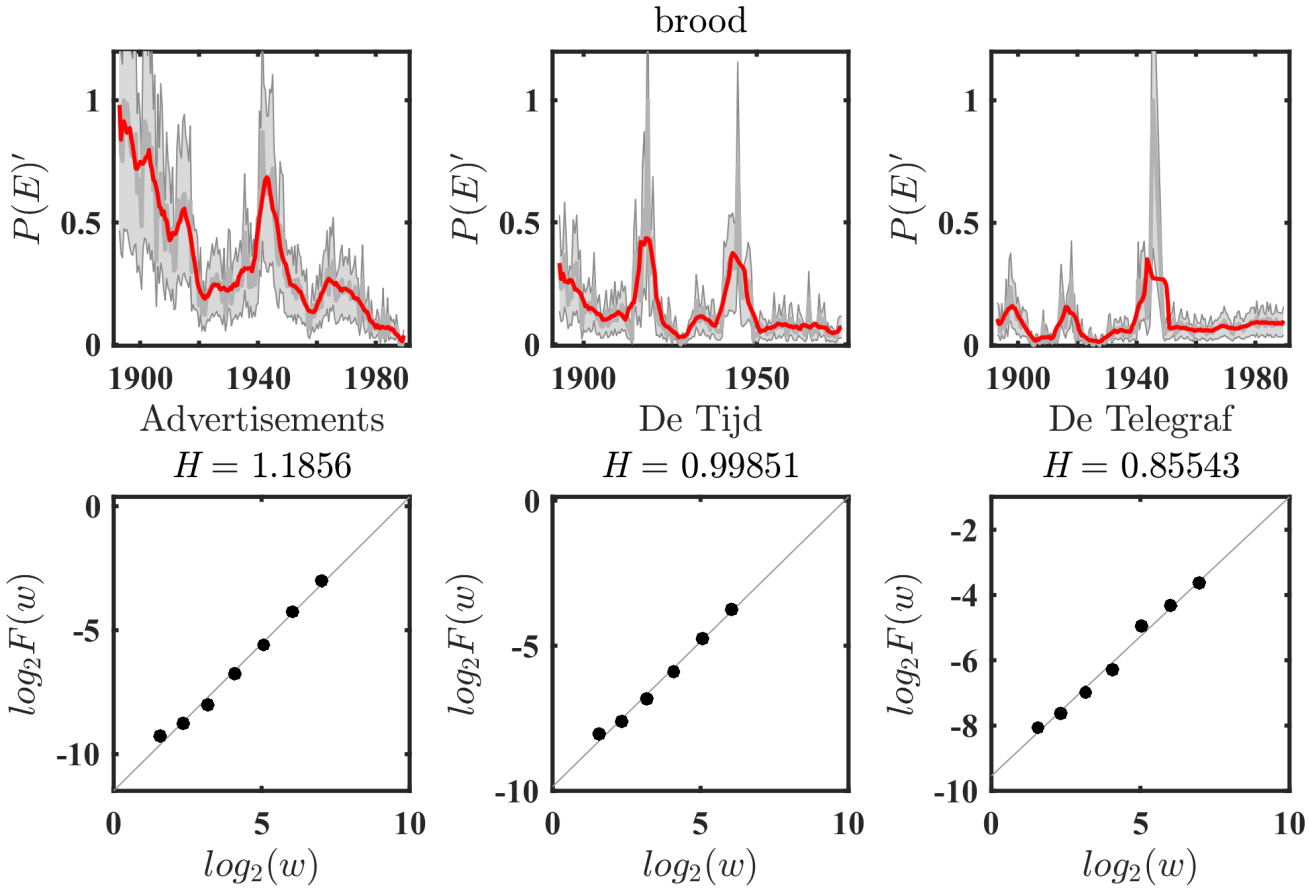
[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119)


new
's
at
the
my
our
his
their
your
a
going
united
five
two
these
three
four
few
several
many
each
its
every
part
an
last
next
little
much
some
most
any
too
very
which
no
those
other
as
such
mr.
federal
second
first
long
another
big
white
old
off
good
million
best
less
more
this
that
how
what
when
where
who
but
own
under
american
national
ms.
center
group
one
or
if
because
think
says
said
say
former
until
dr.
$
including
percent
university
members
both
home
same
and
before
?
times
since
while
way
state
night
place
time
days
ago
know
political
general
street
city
her
public
law
business
office
team
world
week
day
years
year
john
states
country
family
music
department
director
court
game
life
school
police
west
government
market
season
president
company
here
case
program
house
officials
companies
man
money
yesterday
today
now
then
so
war
him
me
there
children
women
people
us
them
york
well
though
it
right
they
you
we
i
do
will
may
can
might
could
would
should
was
is
to
just
even
also
only
not
still
never
all
by
high
in
of
after
against
during
among
into
through
between
with
without
for
about
than
like
see
found
use
made
make
left
play
down
work
up
around
over
on
back
end
out
called
from
go
come
take
used
been
being
are
were
have
has
had
want



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


brood
P(E)'
1
0.5
0
1900
1940
1980
Advertisements
H = 1.1856
1900
1950
De Tijd
H = 0.99851
1900
1940
1980
De Telegraf
H = 0.85543
log2F(w)
0
-5
-10
-2
-4
-6
-8
0
5
10
log2(w)

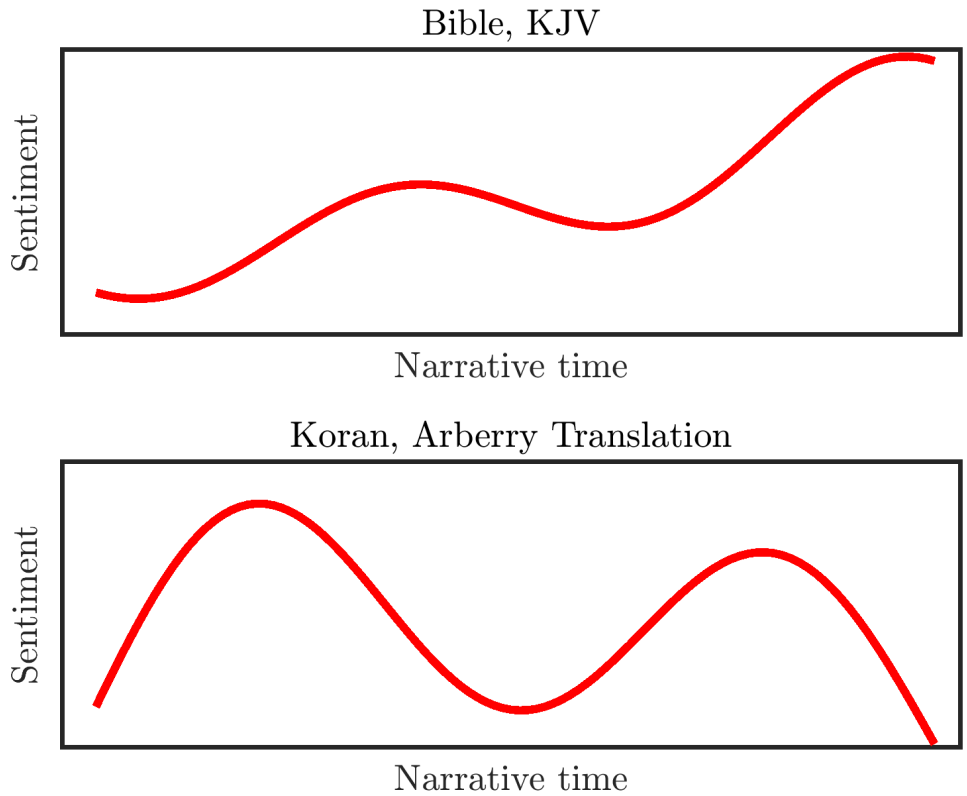
a dictionary can be used to to extract the frequency of cognitive and affective keywords from a collection of documents and perform a **sentiment analysis**

```
'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S. citizen
so she could use her in the debate?'

Positive sex, citizen
Negative crooked, hillary, disgusting, out
Sentiment Score (2+1) + (-2-1-3-1) = -4
Sentiment Polarity Negative
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)

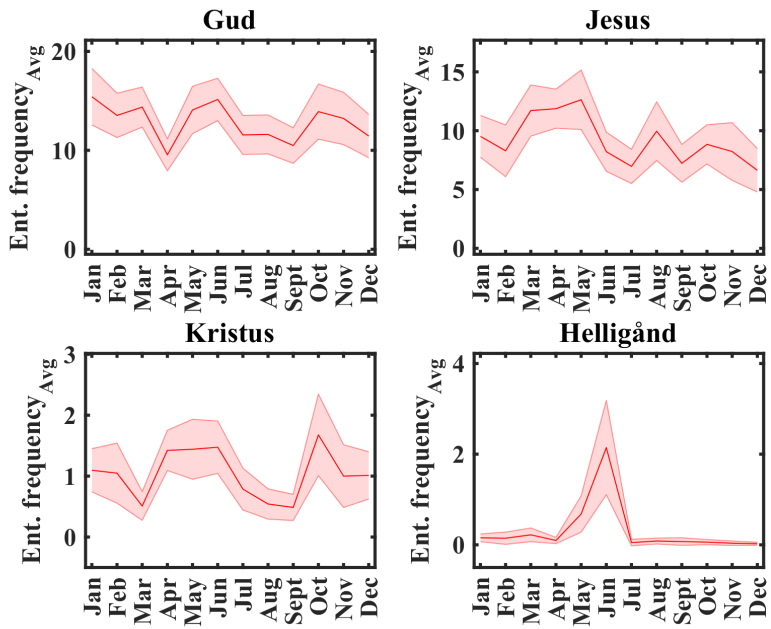

Bible, KJV
Sentiment
Narrative time
Koran, Arberry Translation
Sentiment
Narrative time

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

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

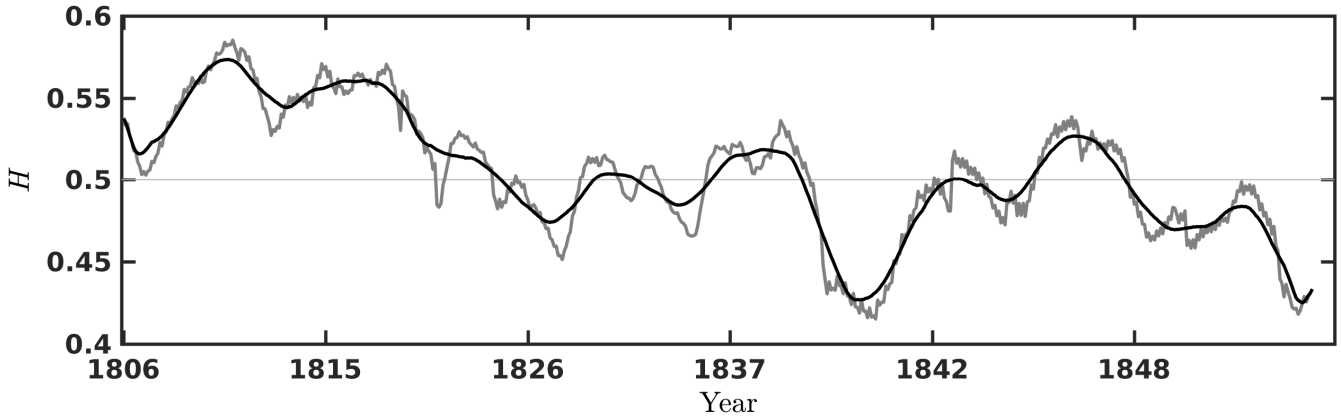


named entities are unique identifiers for entities (a proper noun serving as a name for someone or something)

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

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

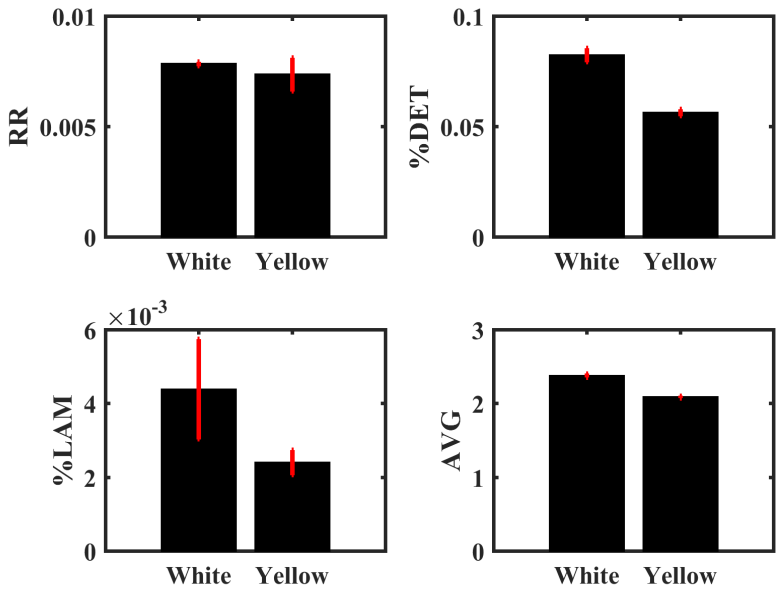
more general, given concrete semantics, locate elements in text that fit those semantics

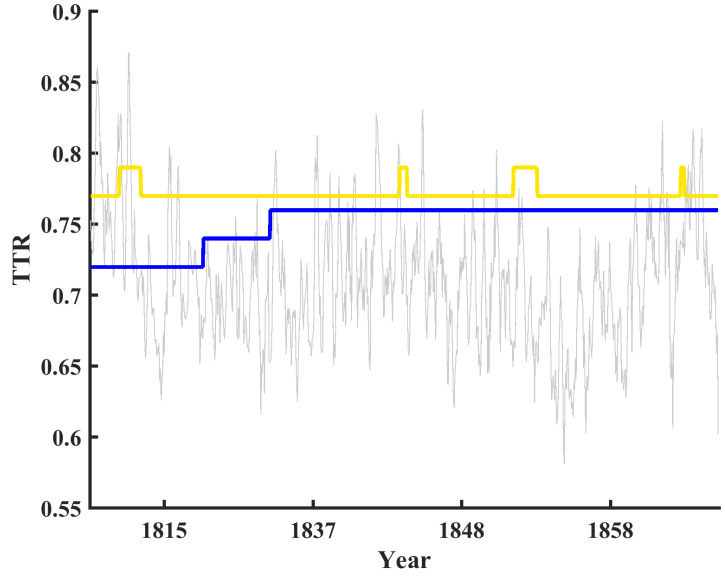


measures of **lexical density** or variation can capture global text features that are indicative of age, education, and cognitive impairment: $L_d = \frac{N_{lex}}{N} \times 100$

"a rose is a rose is a rose" is less lexically dense than "a rose is red and thorny"

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


RR
0.01
0.005
0
White
Yellow
%DET
0.1
0.05
0
White
Yellow
%LAM
×10^-3
6
4
2
0
White
Yellow
AVG
3
2
1
0
White
Yellow


TTR
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
1815
1837
1848
1858
Year