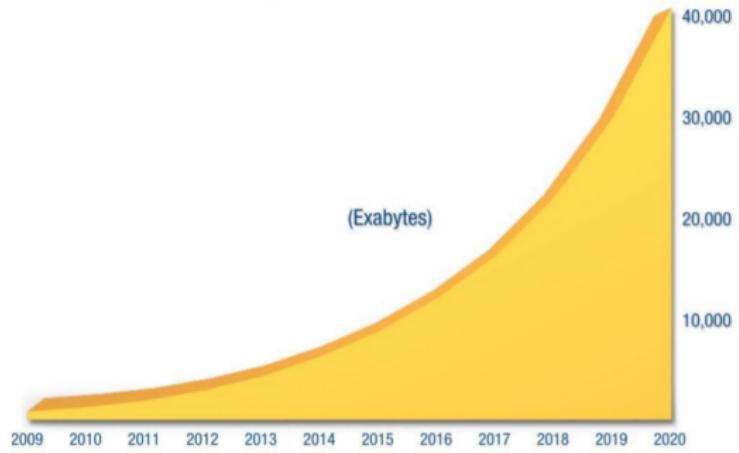
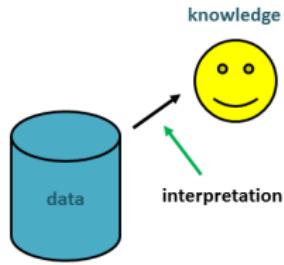


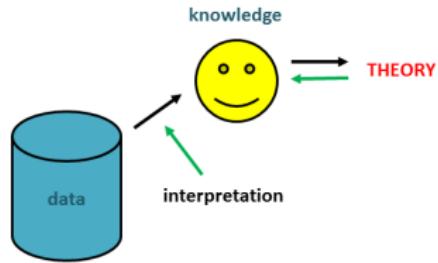
introduction to text analytics

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

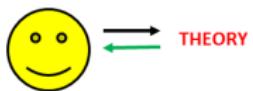


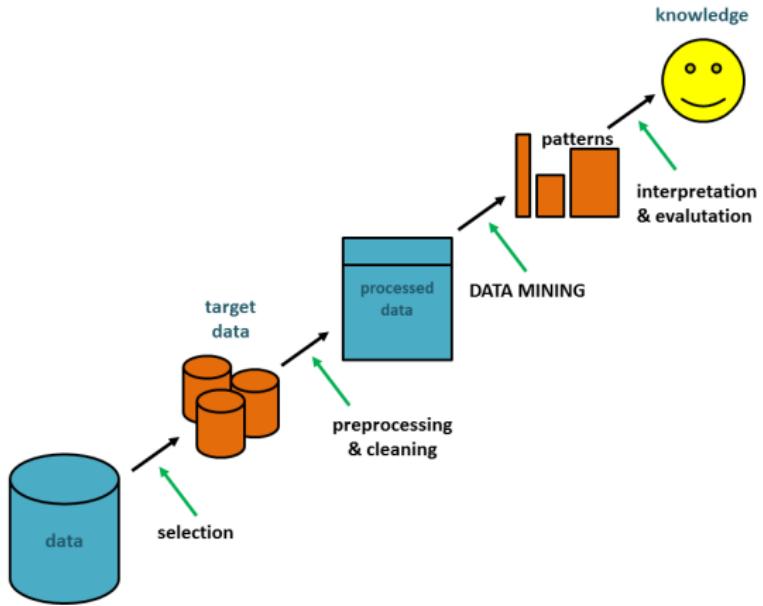


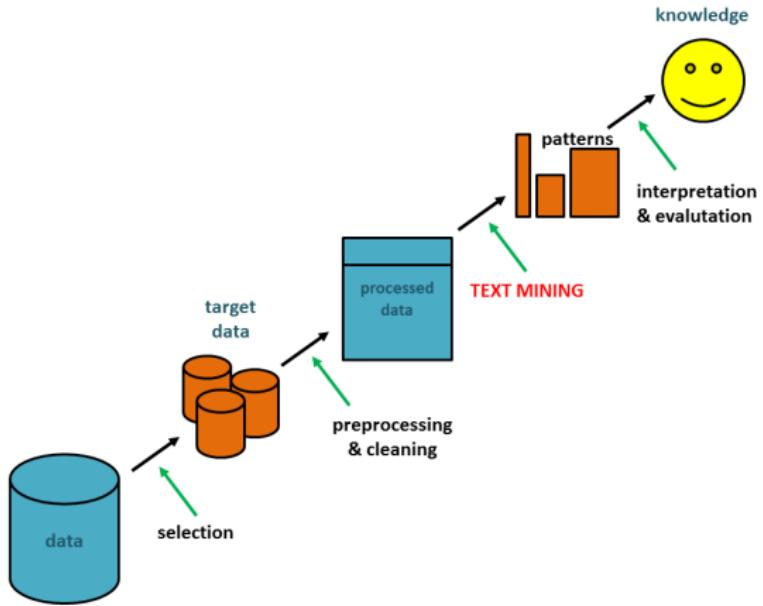




knowledge



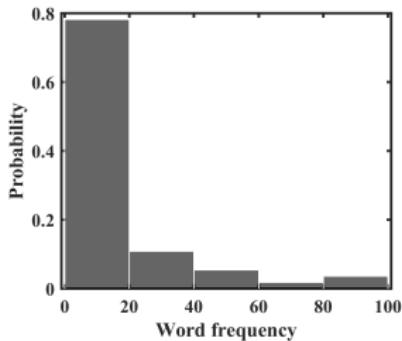




words are (one of) the basic units of meaning

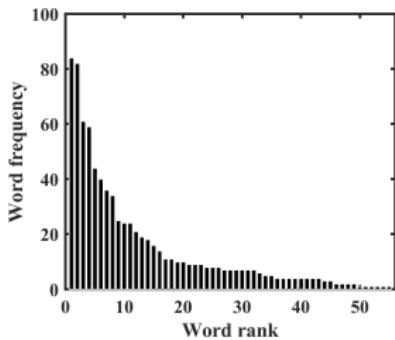
most **text analytics techniques rely on word frequencies**, that is, we tokenize a text at the word level and count the number of tokens for each type.

I am Daniel	'I' 'am' 'Daniel' 'I'	a	1	59	0.073
I am Sam	'am' 'Sam' 'Sam'	am	1	16	0.02
Sam I am	'I' 'am' 'That'	and	1	24	0.03
That Sam-I-am	'Sam' 'I' 'am'	anywhere	1	1	0.001
That Sam-I-am!	'That' 'Sam' 'I'	anywhere	1	7	0.009
I do not like	'am' 'I' 'do' 'not'	...			
that Sam-I-am	'like' 'that' 'Sam'	you	1	34	0.042
...	'I' 'am' ...	<i>total</i>	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).

before calculating word frequencies, we use a range of **text normalization techniques**

- ▶ casefolding
- ▶ removal of non-alphanumeric characters (punctuation, blanks)
- ▶ removal of numeral and stopwords*
- ▶ reduction of inflectional forms through stemming and lemmatization
- ▶ synonym substitution

remember that **one man's rubbish may be another's treasure**

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: $tfidf(t, d, D) = f_{t,d} \cdot \log \frac{|D|}{\{d \in D : t \in d\}}$

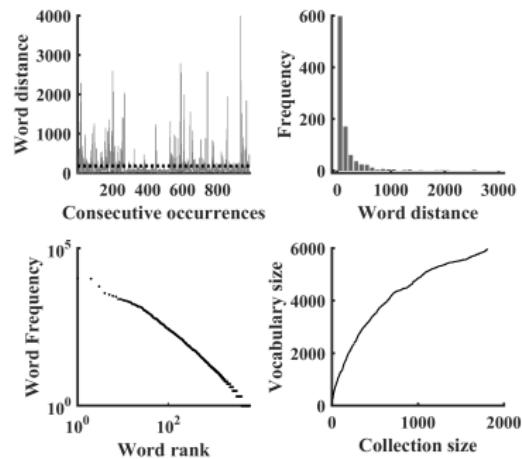
¹ prevents bias towards longer documents

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

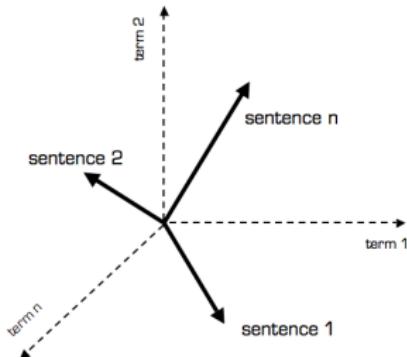
the vocabulary increases as a function of the number of texts, but the increase **diminishes as more texts are included**



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	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	