how to *do* computationally assisted research

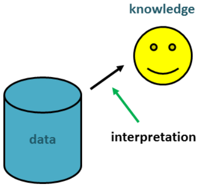
digital literacy @ comwell

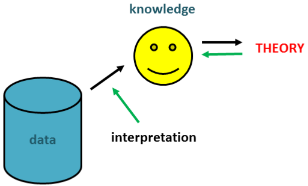
Kristoffer L Nielbo
knielbo@sdu.dk
knielbo.github.io/

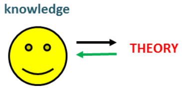
March 22, 2018

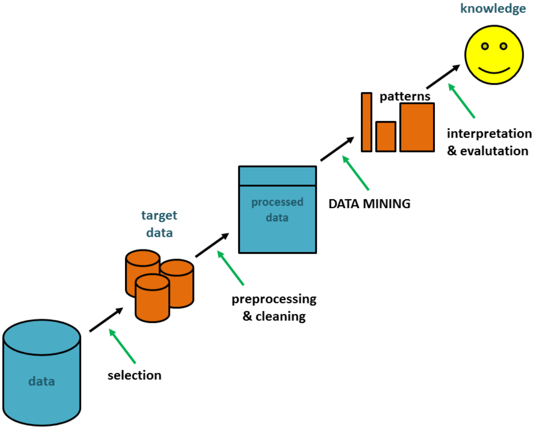
```
class Person(object):
    def __init__(self, name):
        self.name = name
    def says_hello(self):
        print ‘Hello, my name is’, self.name

class Researcher(Person):
    def __init__(self, title=None, areas=None, **kwargs):
        super(Researcher, self).__init__(**kwargs)
        self.title = title
        self.areas = areas

KLN = Researcher(name = ‘Kristoffer L Nielbo’,\
        title = ‘Associate professor’,\
        areas = [‘Humanities Computing’, ‘Culture Analytics’, ‘eScience’])

KLN.says_hello()
```

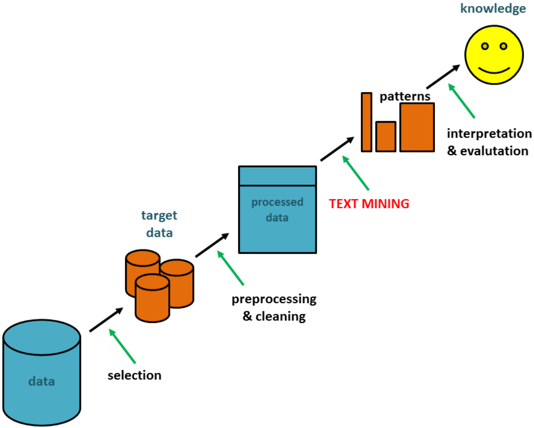
evolution of workflows

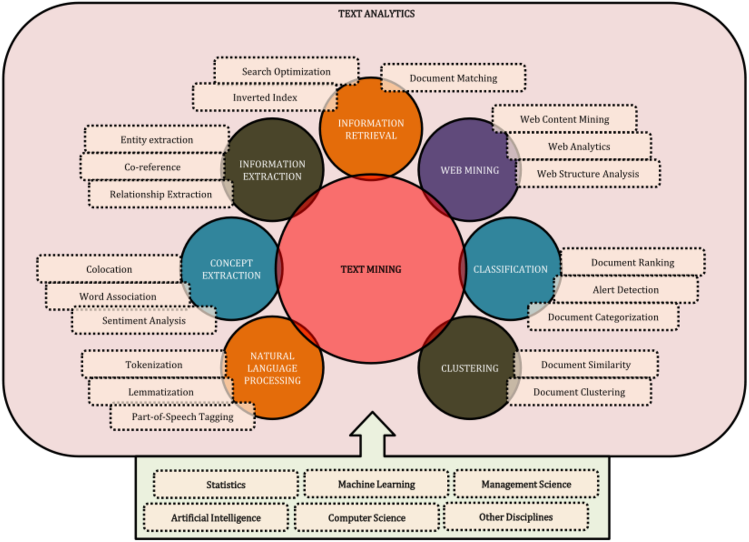

knowledge
data
interpretation


knowledge
THEORY
data
interpretation

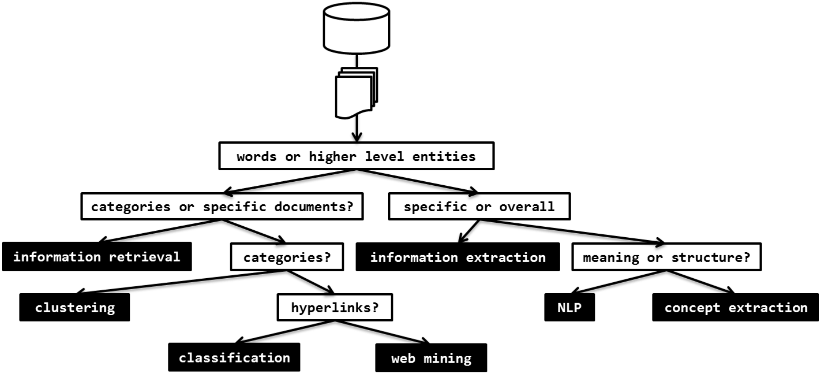

knowledge
THEORY


knowledge
patterns
interpretation
& evalutation
processed
data
DATA MINING
target
data
preprocessing
& cleaning
data
selection


knowledge
patterns
interpretation
& evalutation
TEXT MINING
processed
data
target
data
preprocessing
& cleaning
data
selection

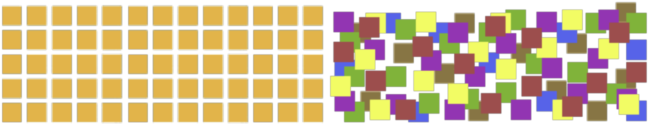


TEXT ANALYTICS
Search Optimization
Inverted Index
Document Matching
INFORMATION RETRIEVAL
Web Content Mining
Web Analytics
Web Structure Analysis
WEB MINING
Entity extraction
Co-reference
Relationship Extraction
INFORMATION EXTRACTION
CONCEPT EXTRACTION
Colocation
Word Association
Sentiment Analysis
TEXT MINING
CLASSIFICATION
Document Ranking
Alert Detection
Document Categorization
NATURAL LANGUAGE PROCESSING
Tokenization
Lemmatization
Part-of-Speech Tagging
CLUSTERING
Document Similarity
Document Clustering
Statistics
Machine Learning
Management Science
Artificial Intelligence
Computer Science
Other Disciplines


words or higher level entities
categories or specific documents?
specific or overall
information retrieval
categories?
information extraction
meaning or structure?
clustering
hyperlinks?
NLP
concept extraction
classification
web mining

`data`

data objects that are described over a set of (qualitative or quantitative) features



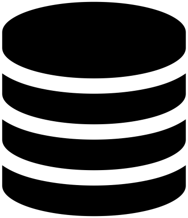
- fundamental difference between structured data and **unstructured* data**
- word processing files, pdfs, emails, social media posts, digital images, video, and audio
- today $>$ 80% of all data are unstructured
- unstructured data require expertise in culture, media, linguistic ...

data|access and sampling

select (sample*) a set of documents (target data) relevant to your research question from a data collection

>> online databases and research libraries are excellent resources

- proprietary issues
- data protection acts
- ethical concerns
- availability (e.g., historical sources)



>> sample requirements

- "all the data"
- balancing and stratification
- bias reduction

we will focus on documents stored locally in a *plain text* without markup

```
"""The First Book of Moses, called Genesis

   {1:1} In the beginning God created the heaven and the earth. {1:2}
And the earth was without form, and void; and darkness was upon the
face of the deep. And the Spirit of God moved upon the face of the
waters.

   {1:3} And God said, Let there be light: and there was light. {1:4}
And God saw the light, that it was good: and God divided the light"""
```

BUT with a bit of code everything is possible

```
import urllib2
from HTMLParser import HTMLParser

class html_parser(HTMLParser):
    def handle_starttag(self, tag, attrs):
        print "start tag:", tag
    def handle_endtag(self, tag):
        print "end tag :", tag
    def handle_data(self, data):
        print "data  :", data

url = "https://knielbo.github.io//"
response = urllib2.urlopen(url)
webpage = response.read()
parser = html_parser()
parser.feed(webpage)
```

`preprocessing`

to prepare a document we need to parse, slice and split it at the relevant level(s).

unstructured data are very noisy, so to increase the signal, we therefore remove irrelevant data through `preprocessing`

range of text normalization techniques to `preprocess` the data:

- casefolding
- removal of non-alphanumeric characters (punctuation, blanks) and numerals
- vocablary pruning
- identification of parts of speech
- reduction of inflectional forms through stemming and lemmatization
- disambiguation
- synonym substitution

  ...

one man's rubbish may be another's treasure

example

- normalization by reducing inflected words to their stem, base or root form

- the stem need *not* be identical to the morphological root
- sufficient that related words map to the same stem (stem $\neq$ valid root)

- search engines treat words with the same stem as synonyms (`conflation`)

Porter stemming algorithm - step 1a

```
SSES -> SS    caresses -> caress
IES  -> I     ponies   -> poni
              ties     -> ti
SS   ->  SS   caress   -> caress
S    ->       cats     -> cat
```

proprocessing|structuring words

– selecting the right **formalism** for representing a problem over a data set
– many techniques rely on basic probabilistic or geometric properties of the data set

example

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

'I' 'am' 'Daniel' 'I' 'am'
'Sam' 'Sam' 'I' 'am'
'That' 'Sam' 'I' 'am'
'That' 'Sam' 'I' 'am' 'I'
'do' 'not' 'like' 'that'
'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

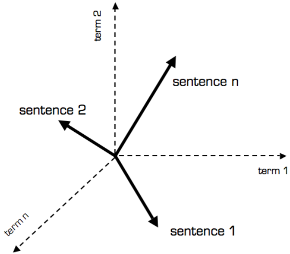
SDU

example

any collection of *m* documents can be represented in the vector space model by a document-term matrix of *m* documents and *n* terms

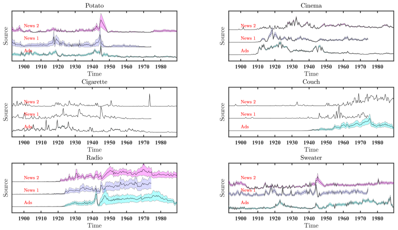
a vector space model is a basic modeling mechanism for a word- or document-space (whether we look at rows or columns)

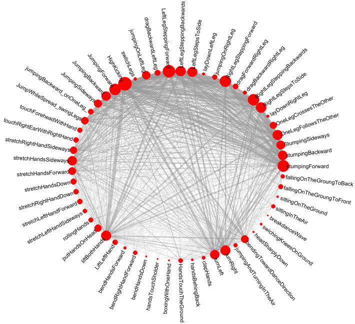
- a document vector with only one word is collinear to the vocabulary word axis
- a document vector that does not contain a specific word is orthogonal/perpendicular to the word axis
- two documents are identical if they contain the same words in a different order (BOW assumption)



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	

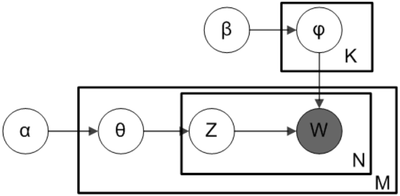
`analysis`





– describe basic properties of the data, e.g., simple distributions and relations
– result in themselves or input to more advanced analysis
– the value depends critically on domain knowledge

beauty lies in simplicity



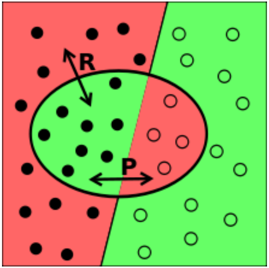
α	Dirichlet prior for per-doc topic dist - proportions parameter
β	Dirichlet prior for per-topic word dist - topic parameter
θ_i	word dist for topic - per-document topic proportions
ϕ_k	word dist for topic k - topics
Z_{ij}	topic for j^{th} word in doc i - per-word topic assignment
W_{ij}	the observed word

Procedure 1 Generative Model

1: **choose** $\theta_i \sim Dir(\alpha)$, i is a document
2: **choose** $\phi_k \sim Dir(\beta)$, k is a topic
3: **for** each word position **do**
4: **choose** a topic $z_{ij} \sim Multinomial(\theta_i)$
5: **choose** a word $w_{ij} \sim Multinomial(\phi_{z_{ij}})$
6: **end for**

The joint distribution defines a posterior probability: $P(\theta, z, \phi)$
use posterior to:
Train on a corpus: Bayesian inference on θ and ϕ
Train on a new documents d: fix $P(w \mid z)$ to infer $P(z \mid d)$
– Multiple inference algorithms available (expectation-maximization/VEM and Gibbs sampling/GIBBS)

interpretation and evaluation



← `relevant objects` (e.g., ham)

→ `irrelevant objects` (e.g., spam)

objects classified with relevant class label

ERROR

CORRECT

Precision: fraction of retrieved instances that are relevant

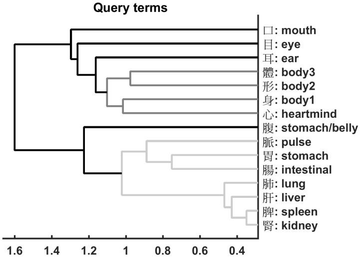
$$P = \frac{TP}{TP + FP}$$

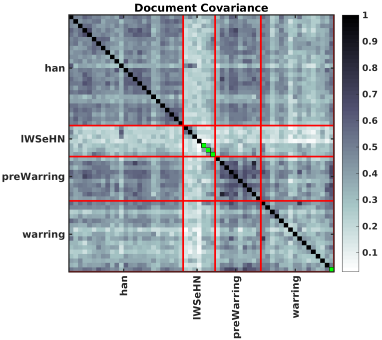
Recall: fraction of relevant instances that are retrieved

$$R = \frac{TP}{TP + FN}$$

P and R are inversely related. Identify balance through a Precision-Recall curve.

interpretation|what does our model mean?





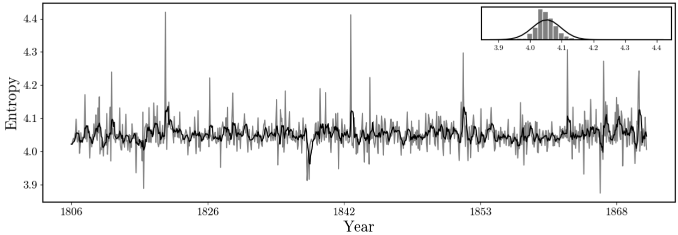
– philosphers and sinologists have been debating the existence of mind-body dualism in classical Chinese philosophy

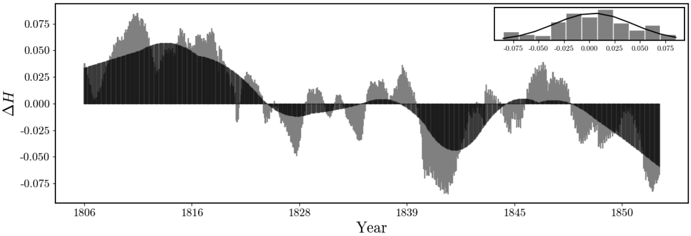
– with domain experts, unsupervised learning was used to identify a multi-level dualistic semantic space

– one model (LDA) was further utilized to predict class of origin for controversial texts slices knowledge

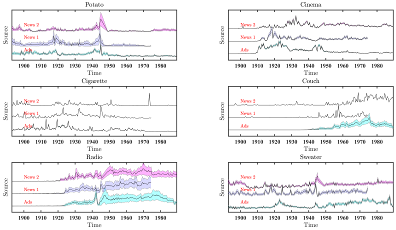

YOUR
GAME
HERE

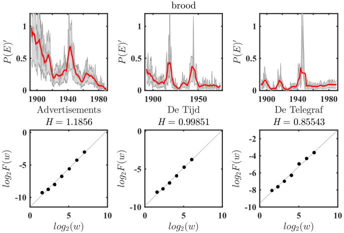
cases


Entropy
4.4
4.3
4.2
4.1
4.0
3.9
1806
1826
1842
1853
1868
Year
3.9 4.0 4.1 4.2 4.3 4.4


ΔH
0.075
0.050
0.025
0.000
-0.025
-0.050
-0.075
1806
1816
1828
1839
1845
1850
Year
-0.075 -0.050 -0.025 0.000 0.025 0.050 0.075

History|Predictive Causality & Slow Decay





– historians and media researchers theorize about the causal dependencies between public discourse and advertisement

– time series analysis of keyword frequencies (from seedlists) indicated that for some categories 'ads shape society', while other categories merely 'reflect'

– advertisements show a faster decay (on-off intermittant behavior) than public discourse (long-range dependencies)