text analytics

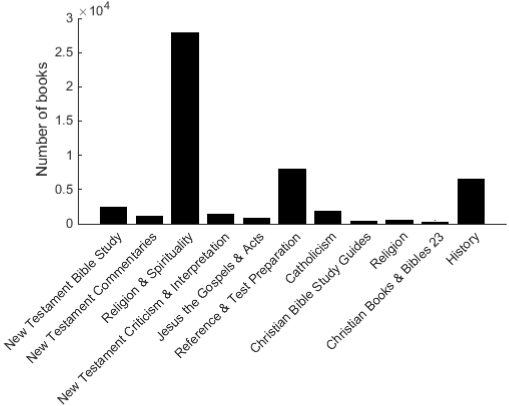
DTL SU @ AU

kristoffer l nielbo
kln@cas.au.dk
github.com/kln-courses/tmgu17
tmgu17.slack.com

DAI|IMC|AARHUS UNIVERSITY

- domain knowledge in history, language, literature &c combined with microscopic and (predominantly) qualitative analysis of human cultural manifestations

Gospel of Marc (KJV) ∼ 16500 words in 16 chp. on 11 p.

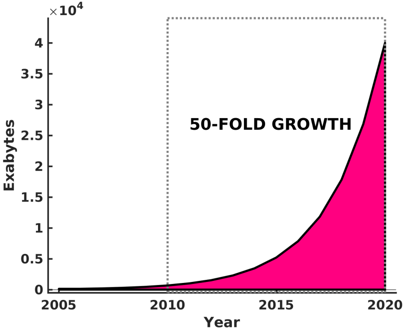


‘from the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days ... and the pace is accelerating’

Eric Smith (Google)

‘increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets’

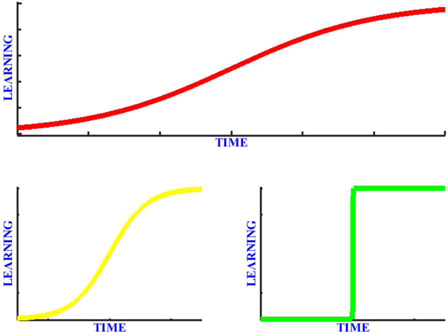
Jim Gray (Fourth Paradigm)



computational sciences are entering the exa-scale era

+

digital technologies are disruptive on a new scale

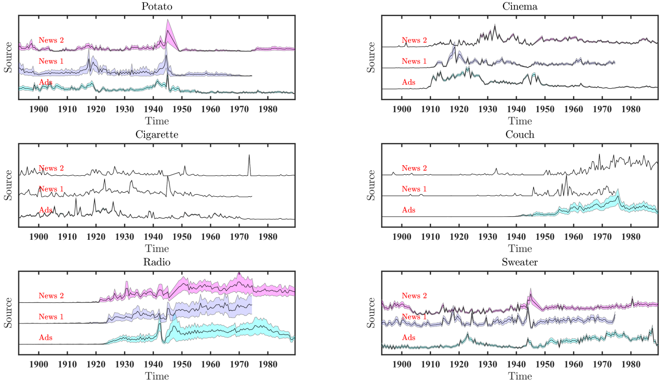


every knowledge-intensive industry have to “break” the learning curve

INTERVENTION|from the console

GUI → CLI

- novice-friendly visual approach to computer interaction w. a fast learning curve ERROR
- expert-friendly text-based approach to computer interaction w. ++freedom VALID
- CONFLICT break the learning curve through training intensive, non-intuitive, and specialized tools



Potato

Cinema

Cigarette

Couch

Radio

Sweater

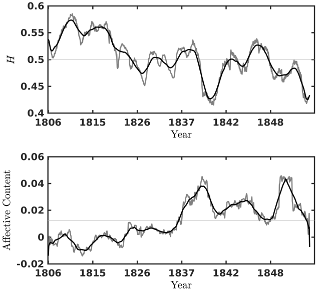
Digital history and media studies

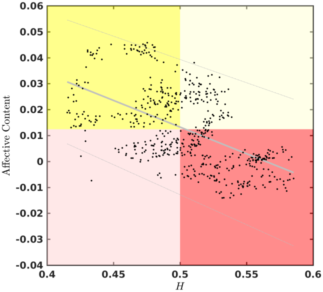
– prerequisite: humanistic domain experts that use content analysis

– source digitization (newspapers) og super computing change resolution and scale

– technologies create new standards for the domains involved

– share technology, but not data!





Computational literary history

– prerequisite: humanistic domain experts that study writers and literary periods

– high quality digitization of writers, annotation and NLP changes perspective and scale

– technologies that are creating new standards

– sharing of tehcnology and data

Data

Information

Presentation

Knowledge

Data

Information

Presentation

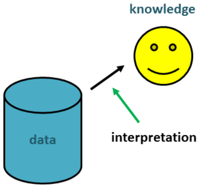
Knowledge

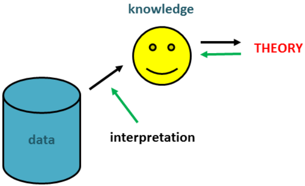
Data

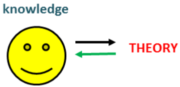
Information

Presentation

Knowledge

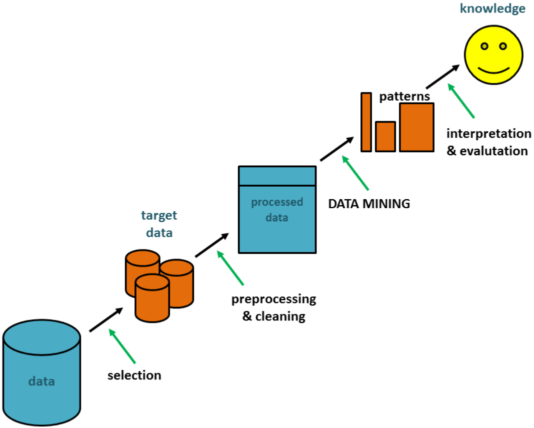

knowledge
data
interpretation


knowledge
THEORY
data
interpretation



knowledge

THEORY



data

selection

target data

preprocessing & cleaning

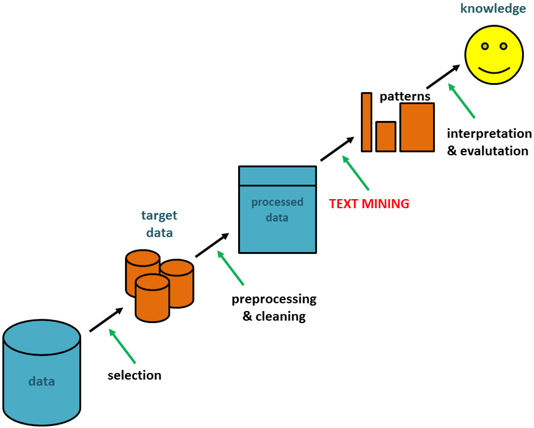
processed data

DATA MINING

patterns

interpretation & evalutation

knowledge


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

text analytics $\sim$ text mining $\sim$ automated text analysis

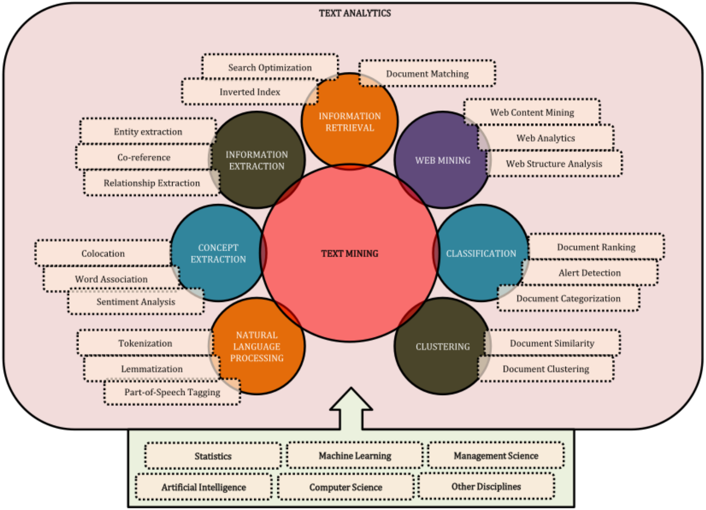
set of data mining[1] techniques for extracting high quality information from large scale text-heavy (unstructured) data sets

($\sim$ Miner et al 2012)

a tool for discovery and measurement in textual data of prevalent attitudes, concepts, or events

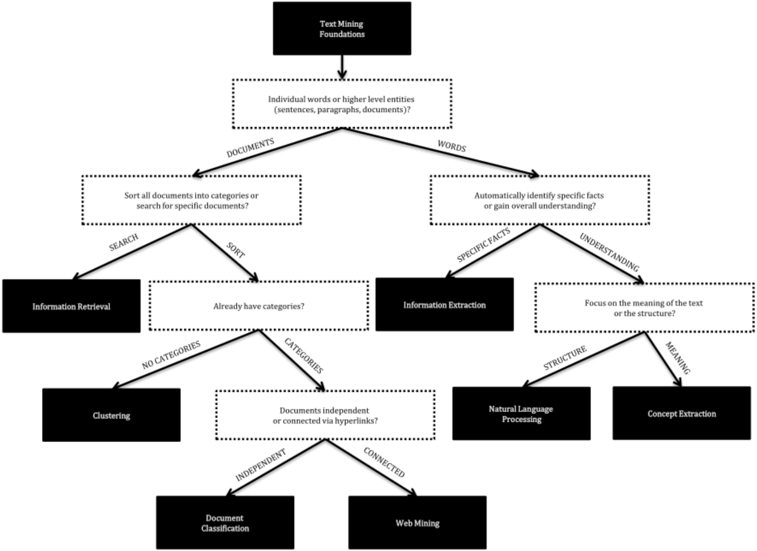
($\sim$ O'Connor, Bamman & Smith 2011)

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37.

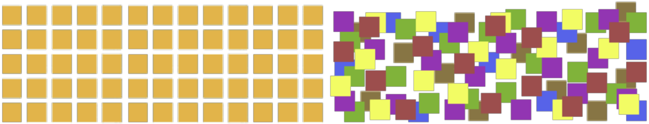


AARHUS UNIVERSITET

IMC INTERACTING MINDS CENTRE


Text Mining Foundations
Individual words or higher level entities (sentences, paragraphs, documents)?
DOCUMENTS
WORDS
Sort all documents into categories or search for specific documents?
Automatically identify specific facts or gain overall understanding?
SEARCH
SORT
SPECIFIC FACTS
UNDERSTANDING
Information Retrieval
Already have categories?
Information Extraction
Focus on the meaning of the text or the structure?
NO CATEGORIES
CATEGORIES
STRUCTURE
MEANING
Clustering
Documents independent or connected via hyperlinks?
Natural Language Processing
Concept Extraction
INDEPENDENT
CONNECTED
Document Classification
Web Mining

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

- increased demand for expertise from culture, media and linguistic domains the goal of **statistical learning** is to build a machine that can learn from data and automatically make the right decisions

supervised learning infer mapping between data & class-information → ‘ground truth’

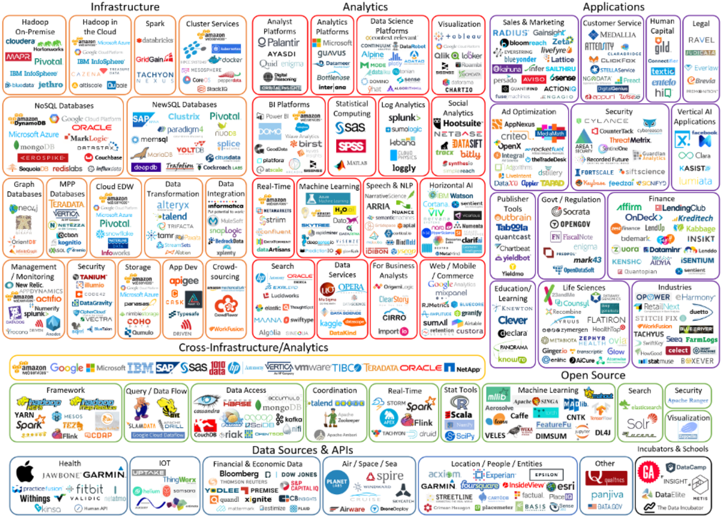
unsupervised learning identify latent classes in the data → lack ‘ground truth’ adequate problem solution requires that we test a range of approaches (algorithms, (hyper-)parameter estimation) - the validation of an approach is an **experiment**

experiment input: code, data sets, hyperparameter values

experiment output: model definition (weights), metric values (experiment comparison), execution logs

a complex and error-prone process

⇒ systematically comment your work and process and use version control and source code management


Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt / Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Education/ Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Incubators & Schools
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other

R

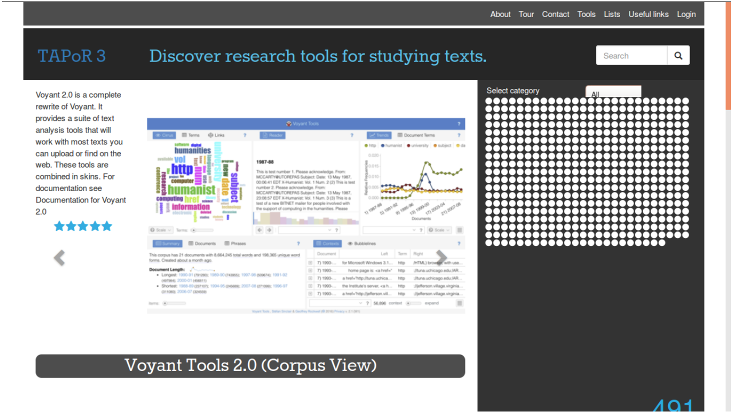
python™

MATLAB

rapidminer

tableau®

VOYANT
see through your text



Discover research tools for studying texts.

Voyant 2.0 is a complete rewrite of Voyant. It provides a suite of text analysis tools that will work with most texts you can upload or find on the web. These tools are combined in skins. For documentation see Documentation for Voyant 2.0

★★★★★

1987-88

This is test number 1. Please acknowledge. From: MCCARTY@UTOREPAS Subject: Date: 13 May 1987, 00:06:41 EDT X-Humanist: Vol. 1 Num. 2 (2) This is test number 2. Please acknowledge. From: MCCARTY@UTOREPAS Subject: Date: 13 May 1987, 23:08:57 EDT X-Humanist: Vol. 1 Num. 3 (3) This is a test of a new BITNET mailer for people involved with the support of computing in the humanities. Please

This corpus has 21 documents with 8,664,245 total words and 198,365 unique word forms. Created about a month ago.

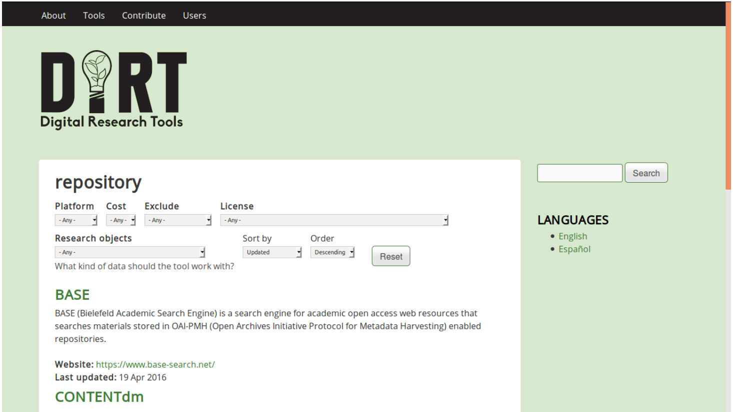
Document Length:
- Longest: 1990-91 (791280); 1989-90 (743955); 1997-98 (509674); 1991-92 (497984); 2000-01 (456811)
- Shortest: 1988-89 (237107); 1994-95 (245689); 2007-08 (271099); 1996-97 (311080); 2006-07 (324559)

Document	Left	Term	Right
7) 1993-...	for Microsoft Windows 3.1...	http	/HTML) browser, with use...
7) 1993-...	home page is: <a href="	http	://tuna.uchicago.edu/AR...
7) 1993-...	a href="http://tuna.uchica...	http	://tuna.uchicago.edu/AR...
7) 1993-...	the Institute's server, <a h...	http	://jefferson.village.virginia...
7) 1993-...	a href="http://jefferson.vill...	http	://jefferson.village.virginia...

56,896

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2016) Privacy v. 2.1 (M1)

Voyant Tools 2.0 (Corpus View)



repository

Platform: - Any -
Cost: - Any -
Exclude: - Any -
License: - Any -
Research objects: - Any -
Sort by: Updated
Order: Descending

What kind of data should the tool work with?

BASE

BASE (Bielefeld Academic Search Engine) is a search engine for academic open access web resources that searches materials stored in OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) enabled repositories.

Website: https://www.base-search.net/
Last updated: 19 Apr 2016

CONTENTdm

LANGUAGES
- English
- Español

“There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea”

Andreas Buja