Introduction to methods for digital humanities

Eetu Mäkelä, D.Sc.
Professor (tenure track) in Humanities–Computing Interaction / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University
http://presemo.helsinki.fi/meth4dh
Backgrounds

The level of studies you are currently doing

18 responses

- Bachelor's: 16.7%
- Master's: 77.8%
- Doctoral: 1.1%
- Post doctoral: 0.0%
Disciplinary backgrounds
Disciplinary backgrounds
Assignment: history of humanities computing

1. Did you manage to dig up any history of computational approaches applied to your own field?
2. Why didn’t the 60’s excitement ever turn into a true revolution?
Recap: What to learn if you’re a humanist?

1. Knowledge of easy to use end-user data processing and visualization tools
   ○ Easy to use for their intended purpose, but limited
2. Knowledge of the fundamentals concepts of programming
   ○ Frees you to process your data more efficiently
   ○ Allows you to more freely apply visualizations etc based on ready libraries and tutorials on the Internet
3. High-level understanding of what types of things can be accomplished with advanced CS methods
   ○ To be able to communicate in collaborative projects
Recap: final projects
Different types of data, data quality, available open datasets

Eetu Mäkelä, D.Sc.
Professor (tenure track) in Humanities–Computing Interaction / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University
Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!

“Any sufficiently advanced technology is indistinguishable from magic.”

- Arthur C. Clarke
Research process - Magic (?)

- Hedge magic (spreadsheets, Excel graphs)
- Common ritual magic (statistics: correlation, regression, ANOVA, PCA)
  - Relatively simple, commonly understood formulae you could mostly go through with pen and paper if you wanted to
- Higher ritual magic (SVM, LSA, LDA, SnE)
  - More complex, harder to follow formulae, impossible to work through manually
- Black magic (most machine learning, neural networks)
  - You feed the machine both an input and a desired output, it derives a mostly unintelligible black box that links the two
- Flashy magic (proper visualizations)
Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!
Digital humanities research process

1. Raw data
2. Iterative exploration of data
3. Analysis tools
4. Results
5. Research articles
6. Processing tools
Open data in the digital humanities - the good

- Great aggregators pushing for CC0 licenses, publishing participating data: Europeana, Digital Public Library of America & The European Library
- Influential national libraries moving to co-operative open (linked) data
  - Library of Congress, Deutsche Nationalbibliothek, British Library, Bibliothèque nationale de France
- Museums, Galleries and Archives catching up: British Museum, Finnish National Gallery, …
- Glue available: VIAF, CIDOC-CRM, Getty AAT, TGN, ULAN, CONA, Pleiades, …
Open data in the digital humanities - the bad

- Academic libraries have a long tradition of collaborating with library service companies (primarily EBSCO Information Services, ProQuest LLC and Gale Cengage Learning) to produce services.
- Often, they also participate in content creation projects, and then hold the rights for that content.
  - e.g. Early English Books Online (ProQuest), Nineteenth Century Collections Online (Gale), State Papers Online (Gale).
- But, this is also a wider culture inside humanities, e.g. Electronic Enlightenment.
Archetypes of data

- Structured data - Excel sheets, databases
- Unstructured data - text, sounds, images
Types of data

- Structured (databases) vs unstructured (text, image, video, audio)
- Clean vs messy
Types of data

- Structured (databases) vs unstructured (text, image, video, audio)
- Clean vs messy
- Biased? ← incomplete, messy, badly sampled
Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!
Research process

1. Have data ← 0. Get data, understand magic that went into data
2. Magic (?)
3. Something interesting shows up
4. Profit!
Data production pipelines: Early English Books Online

- Physical books
- Microfilms
- Early English Books Online
- OCR
But what’s in there?

20-30,000 novels published
6,000 novels still extant
3,500 novels available in full text

(Image not of EEBO, but of British Nineteenth-Century Novels)
Ist not Saint George we Sing of here,  
Nor George, the fatal Duke Villier;  
Nor George a Green, nor Castriot;  
Nor Buchanan the learned Scot;  
But us of George the Valiant Monck,  
That made Van-Trump in' S Blood deod  
And in theseus his Navy snuck. (drunk,  
Oh! this is our brave George!  

Isw-not- Saint George we Sing of here,  
Nor George, the fatal Duke Villier;  
Nor George a Green, nor Castriot;  
Nor Buchanan the learned Scot;  
But us of George the Valiant Monck,  
That made Van-Trump in' S Blood deod  
and in theseus his Navy snuck. (drunk,  
Ok I this is our brave George!
THE
First volume of the Chronicles of England, Scotland, and Irelande.

Conteyning,
The description and Chronicles of England, from the first inhabiting unto the conquest of that nation, till the yeare of our Lorde 1571.
The description and Chronicles of Scotland, from the first original of the Scotch nation, till the yeare of our Lorde 1547.

Faithfully gathered and set forth, by Raphael Holinshed.

AT LONDON,
Imprinted for John Harrison.
Data production pipelines: EEBO-TCP (Text Creation Partnership)

Images + OCR

2x manual keying

Transcribed by hand. Owned by libraries. Made for everyone.
Data production pipelines: EEBO-TCP (Text Creation Partnership)

TCP I and TCP II are now available on EEBO, adding transcriptions of approximately 50% of the texts on EEBO.
1. All material should be recorded in the form in which it appears in the book: do not attempt to correct spelling or typographic error.

2. Illegible text, missing and damaged text, or clear but unrecognized symbols all will require some attention from us. Two extremes should be avoided as far as possible: (1) using the illegibility markers promiscuously to avoid capturing text about which there is some difficulty; and (2) "creative" capture of text that really cannot be read (from the EEBO TCP keying instructions)

---

**FLINTSHIRE.**

FLINT-shire (named from Flint, formerly an eminent place therein) hath the Sea on the North, Shropshire on the South, Cheshire on the East, and Denbighshire on the West. The smallest Shire in Wales. Flint the Shire Town is no Market Town, nor nor St. Asaph (a City and Bishops See) till made to go very late; but it is near to Chester, the Market General of those parts, and beneficial, every Village hath a Market in it all, as affording all necessary Commodities. This County was Parcel of the County Palatines, paying 2000 Marks (called a Mone) at the Charge of every Earl of Chester, until the year 1658. For then upon the occasion of one Thomas Radford, committed to Prison by the Chamberlain of Chester, this County disowned it self from the Earlom of Chester, and united it self to the Principality, though I cannot see how the one or the other could be done, without an Act of Parliament.

**Notes.**

1. Mary nag on basa ysgob ysgob? That is, Mary there or High-loy in Chester. Chester men have been very famous for Archery. It is applied to such who take other folk's Goods for their own, being mistaken with the semblable and reemblemance.

2. A. Mary nag on basa ysgob ysgob? That is, More then High-loy in Chester. Chester men have been very famous for Archery. It is applied to such who take other folk's Goods for their own, being mistaken with the semblable and reemblemance.
Amongst the rest (who attended Divine Service) St. Aſaph was eminently conspicuous for Piety and Learning, infomuch that Mungo, (in Latine Quentigernus) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that Llan-Elvy was after his death, called from him St. Aſaph. He was an affidious Preacher, having this Speech in his Mouth, Such who are against the Preaching of Gods word, envy Mans Salvation. He is thought by some to have dyed about 569. After which, his See was Vacant above 500 years.
honour should be done unto him whilst living, who was so solemnly disgraced after his death; his Books being then publicly prohibited by the Court of Rome. See Writers in this Shire.

John of Marmouth, D. D. and Canon of Lincoln, was chosen Bishop of Llandaff, 1256, after that he had been 7 years vacant. He was a Learned and Pious Divine. Besides other Benefactions to his See, he procured the Rectory of Newland, to the Farrell of Deans, to be appropriated to it; but Bishop Kerkeo afterwards impoverished the same, more than all his Predecessors had endow'd it in 400 years. This John dyed April 1252, and was buried in St. Mary's Chapel.

Walter Castlesie, Son to William the elder, Lord Castlesie (whose prime residence was at Aberavon in this County) was made (by Henry 5.) Bishop of Wrexham. He would not yield to the Popes Legate, who complained of many clergy murders keeping their Living's against the Canons, intending to make room for the Popes Favourites, or force such irregular incumbents to a Compromise. He was one of a keen nature, whose two-edged spirit did cut on both sides, against the King and Popes. Against the former, he fled with the Barons, to whom he promised Heaven for the reward of their Rebellion against their Prince, though it cost him an Excommunication from the Pope, who was the more forward in denouncing that fatal Sentence against him, because he had told Paulus his Legate, coming hither 1255, that he would prefer him to be hang'd on the Gallows, rather than ever consent to such excommunication of the Church, as aforesaid. Lying on his death bed, he was told with true remorse for his disobedience, and was advised not to resist. The Bishop, who was in attendance, was immediately informed. He dyed February 1169, when I beheld as Uncle to Thomas, the Sainted Bishop of Hereford.

Richard de Clare, alias Strongbow, born (probably) at Stringale Castle, was Earl of Stringale and Pembroke. A period of effectual performance. It happened, Sir Alan Marshall, in 1169, being expelled his Territories, for several Tyrannies, by the Lords of Mach and Conisbrey, repaired to King Henry 2. and invited him to Ireland. That Politic King fear'd over this R. Strongbow (with 1000 Men) who possessed himself of the Ports of Limerick and Waterford, with large Lands thereunto belonging; inasmuch that the King growing jealous of his greatness remanded him home, and commanded him to surrender his Acquaint with his hands, which done, he received them by a remon from the Kings love that Henry refered the City of Dublin for himself. This Strongbow is commonly called Dumnis Hervinia, the Tanner of Ireland. Yet some of the great Lords there did still retain the Power and Title of King; witness the Preface in the Commission, whereby King Henry 2. made William Fitz-Adams his Lieutenant of Ireland, Archbishop, Bishop, Regius, etc. Saluarem. This Earl dyed at Dublin, 1177.

Sir Roger Williams, born of an ancient Family at Pembrey, was first a Souldier of Fortune under the Duke of Ana, and afterwards served Queen Elizabeth. A man extremely forward to Fight. When a Spanish Captain challenged Sir John Norris to fight a single Combat (which was beneath him to accept, being a General) this Roger undertook the Don. And after they had fought some time (both Armies beholding them) without any hurt, they pledged each other a deep draught of Wine, and kindly departed. Another time at midnight, he assaulted the Camp of the Prince of Parma, near Pembrey; few from the banditti, and pierced Pp 2
Walter Cantilupe, Son to William the elder, Lord Cantilupe, (whose prime residence was at Abergavennie in this County) was made (by Henry 3.) Bishop of Worcester. He would not yield to the Popes Legate.
1 videocassette (10 min.) : sd., col. ; 1/2 in.
Cataloged from contributor's data.

Safety Frog teaches children to be fire safe, explaining that smart kids never play with matches. She shows how smoke detectors work and explains why they are necessary. She also describes how to avoid house hold accidents that lead to fires and how to stop, drop, and roll if clothing catches fire.

VHS.
Elementary grades.
Issued also as motion picture.
Fire prevention Juvenile films.
81 pages of documentation on the exact annotation practices used in the digital edition of the Potage Dyvers

Library cataloguing standards:
- 302 pages of ISBD
- 750 pages AACR, 1056 pages of RDA
  - Helmetin luettelointiohjeet
- 1020 pages of the SPECTRUM standard for museum cataloguing
- A single page of field descriptions in the Schoenberg database
“We changed our cataloguing standards once in the 80’s, and then a second time in 1998.”
“Most of our older entries have actually been copied from the national library that has different cataloguing standards”
“A lot of the Swedish language publications from the 19th century are simply missing, as they were never indexed.”
“This database was gathered based on the whimsies of what the participating researchers researched. It’s probably thus quite biased.”
Documentation?

https://pro.europeana.eu/data/linked-open-data-data-downloads
Open data in the digital humanities - the ugly

- Different forms of encoding, typos
  (Paris,) Paris [Paris,] [Paris]
  (Paris) A Paris À Paris (Paris)
  (Paris.) [A Paris]

Amsterdam. - et Paris Amsterdam ; et Paris
Amsterdam. - et à Paris Amsterdam [Paris]
(Paris. - Amsterdam A Amsterdam [i. e. Paris]. M. DCC. LXX.
Data woes: viaf.org

- Automatic conversions from “Lastname, Firstname” to “Firstname Lastname” does not always work due to bad data

http://viaf.org/viaf/41896578/
KLK Newspaper Pipeline: from archives to a researcher

1. Actual newspapers
2. Digitisation in National Library
3. Optical Character Recognition (OCR) in Russian & Germany
4. Some metadata work in National Library (XML)
5. Kielipankki (XML)
6. Automatic grammatical and morphological analysis plus lemmatization in Turku

- Numpy DataFrame
  - Data from the local copy (Python 3.4). Can include some "unique" NLP stuff.
  - Local copy through KORP API (JSON)
  - Kielipankki KORP server in CSC

- Analysis of the data (e.g. cosine similarities, MDS, hierarchical clustering, LDA)
  - Visualisation through Excel (Google Sheets) or Pyplot.
22(!) different versions of pipeline
Data woes: National Newspaper Collection (KLK)

- On the surface, the Korp API allows one to search by lemma.
- However, these lemmas have been automatically generated, and are only as good as the process that generated them.
- Examples:
  - Early modern Finnish allowed words with the letter “v” to be written as “w”. These are all passed through unprocessed by the analyzer.
  - Fraktur fonts, which are hard for OCR engines, appear in the early parts of the collection.
EXAMPLE:
KLK-1800-subcorpora contains the highly frequent lemma ‘niisi’, because of faulty disambiguation of a morphological analysis (should be *ne : niiden* instead of *niisi : niiden*).
Data woes: National Newspaper Collection (KLK)

• In fact, due to 1) OCR errors and 2) historical language, only a small fraction of KLK is accurately lemmatized
  • 1851-1910: 9.6% of distinct words (66.0% of tokens)
  • before 1851: 15.0% of distinct words (69.3% of tokens)
• Number of lemmas in data sets of comparable sizes:
  • KLK_1980-2000: 201
  • KLK_1820-1859: 528
Data woes: National Newspaper Collection (KLK)

The data is (for obvious historical reasons) temporally imbalanced, this causes the earlier part to be more fragile to metadata problems.

EXAMPLE: “Tähdenvälejä” is located in the year 1842 instead of 1942 where it belongs, and is the only paper in the 1842 corpus.
Data woes: National Newspaper Collection (KLK)

Finnish and Swedish Newspapers are classified under separate subcorpora. Sometimes individual papers contain both languages. These have not been consistently separated.

EXAMPLE: Helsingfors Tidningar
Data woes: National Newspaper Collection (KLK)
KLK Newspaper Pipeline: from archives to a researcher

Actual newspapers
Digitisation in National Library
Optical Character Recognition (OCR) in Russian & Germany
Some metadata work in National Library (XML)
Kielipankki (XML)
Automatic grammatical and morphological analysis plus lemmatization in Turku

Data from the local copy (Python3.4) can include some “unique” NLP stuff.
Local copy through KORP API (JSON)
Kielipankki KORP server in Turku

DataFrame
Analysis of the data (e.g., cosine similarities, MDS, hierarchical clustering, LDA)

Visualisation through Excel (Google Sheets) or Pyplot.

Valid results
Research articles

Bias handling
Bias handling
Bias handling
Bias handling
Bias handling
Bias handling

UNIVERSITY OF HELSINKI
FACULTY OF ARTS
Digital humanities research process

Iterative exploration of data

- raw data
- processing tools
- analysis tools
- results
- research articles
Digital humanities research process

Iterative cleanup, exploration of data

raw data → understanding data → clean data → processing tools → analysis tools → results → research articles
Digital humanities research process

80% of your time for data cleanup, another 80% for algorithms, …
Leverage collaboration, open science workflows to reduce individual workload

raw data → cleaning up data (80% of work) → exploratory tools → results → research articles

understanding data, 2

collaborate, share these, speed up research for everyone + reproducibility
Sample available datasets and APIs

- Korp API (hits in texts+metadata)
- Finnish national gallery API / dump (metadata)
- Schoenberg database (metadata)
- Cushman collection metadata (metadata)
- WW2 covert support networks (metadata)
- Europeana APIs (metadata)
- DPLA APIs (metadata)
- The European Library API (metadata)
- Sydney Powerhouse Museum (metadata)
- EEBO-TCP Phase I (full texts+metadata)
- ECCO-TCP (full texts+metadata)
Assignments for next week

1. Find a dataset that could be of interest to you in your final project. Post a message on #datasets on Slack giving a link to the dataset and a note on why you selected it.

2. Read some background on visualisation for next week:
   - What is Visualization Really for? sections 2.1-2.4 for a categorisation of different uses for visualisation
   - Perception deception & Common visualization mistakes for learning to not trust visualisations blind

3. Read your first DH research article on approaches to visualising relationships between different versions of texts.
eetu.makela@helsinki.fi
http://iki.fi/eetu.makela
http://presemo.helsinki.fi/meth4dh