Data analysis method literacy

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!
Research process

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

raw data

understanding data

cleaning up data (80% of work)

exploratory tools

results

research articles
To do anything, there is always a long path. But! Many parts of that path may be shared between different projects.
Data analysis method literacy by example
The word most often used on the Suomi24 discussion board (91602 times) that a morphological analyzer couldn’t recognize: Ķ
Average lengths of Finnish and translated detective/suspense novels (in pages)

Simple counting by metadata facet with data from the Finnish public libraries
Coverage of different sports in Urheiluruutu in the 1980’s (in minutes)

Counting from mapped mentions inside program metadata
More metadata questions

- What are the most popular colors of underwear in Finnish museum collections?
- What are the most popular themes in Finnish fiction literature?
- On the other hand, what should one write about if one wants to win Finnish literary prizes or government grants?
- Does the Finnish public broadcaster YLE cover the whole of Finland equally?
"human nature" --- timeline for co-terms in paragraphs

simple co-occurrence counts
The discussion context of soul separates from the mind in Britain during the 18th Century. Human nature and dignity become associated.
Words used disproportionately much in socialist newspapers

Simple word frequency comparisons
counting smell words from geotagged photos and tweets
Protagonist presence in films

Simple counting of protagonist name appearance in scenes
Emotional valence as proxy for plot structure

Simple counting of words indexed in dictionary as positive/negative
Complexity of language used by German political parties
Layout choices by topic in in-flight magazines

Amount of snippets in two topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Text</th>
<th>Picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>12.5</td>
<td>25</td>
</tr>
<tr>
<td>Cultural Events</td>
<td>25</td>
<td>37.5</td>
</tr>
</tbody>
</table>

Share of area in two topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Text</th>
<th>Picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>600000</td>
<td>1500000</td>
</tr>
<tr>
<td>Cultural Events</td>
<td>900000</td>
<td>1800000</td>
</tr>
</tbody>
</table>
Inflection forms automatically analyzed from Wikipedia
Rhythm/theme visualization in poems
“For example, the two poems at the top clearly exhibit more dynamic changes in vowel positions than either of the prose pieces. This provides evidence to support the understanding of literary scholars that poets strive to enhance the density of their poems by creating dynamic sound through sonic recurrence and change in every word throughout the poem; it also leads to a hypothesis that language in prose texts or everyday use may have evolved for practical reasons to aid audience comprehension speed and accuracy by avoiding such sonic turbulence.”
Optimal alignments can be automatically derived.
Simulating the habitation of the Anasazi Valley
More stuff

- Topics talked about in the Corpus of Early English Correspondence
- True structure of Finnish politics
- Where are the manuscripts collected by Sir Thomas Phillipps from?
- How many hands have the manuscripts passed through?
- Co-citation graph of mythical and real authorities in ancient Greek scientific texts
Crowdsourcing

- **Chinese Text Project** ← Wiki-style correcting of (originally automatically OCR’d) transcriptions with original shown side by side. Suggested corrections increased tenfold after moving to this type of UI from old-style “send feedback” box
- **Zooniverse** ← the absolute best in designing crowdsourcing for research. See e.g. *Operation War Diary*
What can be done?

Generally:

● First, automatically extract/calculate some statistical features from the data (or, code by hand)

● Then, either:
  a. Compare (the variation of) these features (different types of visualizations), or
  b. Automatically categorize the data according to these features (e.g. topic modeling, clustering)
Remember, to accomplish anything is always a long path.

Thus, it makes sense not to try to do everything yourself (as a service), but to create a community of researchers with similar interests, who'll share the burden, and in the best case also feed back to the raw data provider ← the open science approach.
Assignments

• **Explore bootstrapping**
• Check *Explained Visually*, especially regression and PCA
• Read on some small, actual work:
  – The *presentation* of the DHH15 key concepts of socialism group
  – The *presentation* of the DHH15 Finnair Blue Wings multimodality group
  – If you understand Finnish, the *election questionnaire analysis and visualisation*
• Explore this *topic model* of CEEC and read the *explanation* on topic modelling