Topic Modeling Time & Space

Shawn Graham, Carleton U
@electricarchaeo
Portable Antiquities Scheme

That’s a lot of data.
Full dump looks like this.
• Can a topic model fitted to this data provide insight?
  – Is it an appropriate technique for this kind of data?
  – Under what circumstances?
  – How do we visualize the results?
Can we do it?

Maybe.

Should we do it?
Workflow

1. Clean data
2. Extract Roman period records
3. Compile all records from a particular district into a single ‘document’.
4. Exploratory Topic Modeling
   (range of topic n, optimize interval, MALLET diagnostics, iterate, iterate, iterate!)
5. Exploratory Visualization – network graphs?
   (Gephi, Openheatmap.org)
6. PCA
Sure is some big data.

- 275 districts.
- Approximately 100,000 unique records.
- 8100 unique words, after tokenization.
- 50 topics
Netgraph: Districts – Districts via shared topics
Statistic: Gephi’s Modularity. 4 modules found
Roman roads (kmz file courtesy of Phil Mills)
Strongest topics

Each image has 8100 words!

Size according to MALLET word weights.
### Document Composition as Heatmap

#### Rows: Districts
#### Columns: Topics
#### Cells: % composition

- **Column is a very weak topic**
- **Column is a very strong topic**

The heatmap visually represents the composition of topics across different districts, with colors indicating the percentage of each topic's presence.
Principal Component Analysis

Ten strongest topics used; 2 components calculated. Covers about 60% of variation.
Principal Component Analysis

Ten strongest topics used; 2 components calculated. Covers about 60% of variation. Districts where topic is the largest % of the composition are shown.
Topic Models: A way through the maze?

- Network visualization – not all that useful
- Modularity metric – maybe useful
  For PAS: Seems to be some resonance with civitas/roads.
- PCA – useful
  w/ 2 components: x seems to correspond to date,
  y seems to correspond with material.
  And possibly a geospatial element too.
- Network Reading – useful
  Also helps spot connections between regions that might not be evident, a kind of ‘distant reading’ of the database.
Maybe this big data is just too big.

so let’s try something smaller.

Photo: http://bestandworstever.blogspot.ca/2012/05/best-way-to-store-extremely-small.html
Prescot Street

- Entire site excavation online
- 1813 ‘contexts’ – events which leave traces in the soil (cuts, fills, walls, burials, etc).
Topic model it blind: can a list of topic keywords tell us what the site *is*?

Shawn Graham @electricarchaeo
And this is what I got. major element to it was a Roman cemetery, lots of cremation burials, many skeletons, lots of later crap interfering.

Guy Hunt @GuyCGHunt
@electricarchaeo Your analysis is pretty much bang on the money.
Topic Model of Context Descriptions

2 mode – but probably appropriate; modularity find contexts with similar patterns of discourse
Topic Modeling a single excavation

- Particular excavators are associated with particular word choice, patterning of word usages

- Particular kinds of materials clump together quite nicely.

- Do individual excavators ‘see’ kinds of info that others don’t? Do they ‘specialize’ in certain kinds of info?

  - Topic model of contexts unpeels the ‘objectivity’ of the site description.

  - Topic model of subgroups: voice of excavator gone again.
Can this be mapped?

- This is a map of the 2d/3d spatial/chronological relationships between contexts.
- Size of nodes = high clustering coefficients; colour = modularity.
- Some correspondence with investigator-determined subgroups.
- Not sure how to blend the topic model relationships into this.

RAW TEXT: Can this be mapped?

- This is a map of the 2d/3d spatial/chronological relationships between contexts.
- Size of nodes = high clustering coefficients; colour = modularity.
- Some correspondence with investigator-determined subgroups.
- Not sure how to blend the topic model relationships into this.
So: data mining, eh?

- Topic models: deformations!
- Network viz: meh
- Network metrics: possibly useful for algorithmically interpreting archaeology.

See my series of posts on electricarchaeology.ca for details, headaches, disasters & triumphs.
Acknowledgements

Daniel Pett, Portable Antiquities Scheme
Phil Mills
Elijah Meeks
Jonathan Goodwin
Michael Widner
Scott Weingart
Ethan Watrall
Ian Milligan
Pete Warden & Openheatmap
Stu Eve
Guy Hunt
LP Archaeology
DH2013

....and everyone else who’s been subjected to my tweeting throughout this process.

Thank you also to Noun Project designers:
‘Bubble Graph’ designed by Sue Cardwell,
‘Archaeologist’ designed by Jorge Hernan Correa from The Noun Project
‘Question’ designed by Rémy Médard
‘Man’ designed by Clayton Meador
‘Maze’ designed by Leonardo Dri