Introducing AutoMap and the Data-to-Model (D2M) Wizard

Neal Altman
CASOS Summer Institute, June 2018

(Based on slides developed by Peter Landwehr and Dr. Ju-Sung Lee)

In This Section

- AutoMap Overview
  - Discuss text to network concepts
  - Introduce AutoMap tools and procedures
- Data to Model (D2M) Wizard
  - Overview
  - Run the Data To Model Wizard
  - Review the Data To Model outputs
  - Create and process a simple text corpus
INTRODUCTION

Why Use Text Data for Network Analysis?

- Information about socio-technical networks often resides in unstructured or semi-structured natural-language text data. What are our options?
  - Ignore it or store it (e.g., in a database and let it sit there)
  - Sampling
  - Qualitative, in-depth studies of subsets
  - Analyze separately or jointly
- Networks that don’t exist anymore, e.g. former regimes, bankrupt companies.
- Large-scale networks in which survey within network boundaries is prohibitively dangerous (e.g. Sudan, Gaza, Iraq) and/or expensive/time-consuming (e.g., Twitter feeds, news articles, courtroom proceedings, etc.).
- Covert networks (e.g. white-collar crime syndicates, adversarial organization).
- Networks that lack underlying real-world network or are the same as the data traces produced by or within them. WYSIWII (What-You-See-Is-What-It-Is) (Diesner & Carley, 2009).
There are TWO kinds of networks that you can get from text using AutoMap

Semantic Networks

- Networks of words linked to each other based on co-occurrence.
  - Each link is concept-to-concept, e.g., in Shakespeare’s Romeo and Juliet
    - Romeo Montague ↔ Juliet Capulet
  
- Networks of words linked to the documents in which they appear.
  - Each link is concept-to-document, e.g.,
    - Romeo ↔ Shakespeare’s Romeo and Juliet
    - Juliet ↔ Shakespeare’s Romeo and Juliet
Conventional Meta-Networks

- Collections of multiple networks linking together agent (actors), events, organizations, and other node classes.
  - One-mode: agent x agent (links of agents to agents)
  - Two-mode: agent x organization (links of agents to orgs)

Why The Distinction?

- Sometimes a text is just a text, not a detailed map of specific relationships.
- But, more often than not, texts contain entities that qualify into one of our node classes.
- Some of ORA’s metrics are contingent on the existence of particular types of nodes and networks.
  - For example, **Knowledge Negotiation** measures the extent to which individuals (Agent nodes) need to negotiate with each other for information (Agent x Knowledge) to complete assignments (Agent x Task; Knowledge x Task).
The AutoMap Workflow

AutoMap Network Creation and Refinement Cycle

1. Create a body of plain text
   - Harvest text
   - Divide text into “sensible” document units
   - Remove non-content elements
2. Isolate meaningful text elements
   - Eliminate common terms
   - Merge synonymous terms
   - Remove non-relevant elements
3. Create networks
   - Define network creation parameters
   - Process text to networks
4. Evaluate outcome networks and iterate
Interacting with AutoMap to Build Networks

- AutoMap supports several modes for network creation:
  - Graphical User Interface (GUI)
  - Script Commands
  - Data to Model (D2M) Wizard

- Modes work “cooperatively”:
  - GUI interaction can be recorded as scripts.
  - Scripts can be created interactively.
  - Text and text mappings are held in common.

RUNNING THE DATA-TO-MODEL WIZARD
Data to Model – 10,000 Feet

- Deletes unnecessary elements from text
  - Symbols (punctuation, capitalization, etc.)
  - Noise words (“the”, “and”, etc.)
- Recodes words
  - Tries to create N-Grams (“President Lincoln” = “President_Lincoln”)
  - Tries to relabel common references to terms
- Tries to create appropriate meta-network labels
- Creates networks by linking text within common windows.

Consider this sample text...

Apollo 11 was the spaceflight which landed the first humans, Neil Armstrong and Edwin "Buzz" Aldrin, Jr, on Earth's Moon on July 20, 1969, at 20:17:39 UTC. The United States mission is considered the major accomplishment in the history of space exploration.

Launched from the Kennedy Space Center Launch Complex 39 in Merritt Island, Florida on July 16, Apollo 11 was the fifth manned mission, and the third lunar mission, of NASA's Apollo program. The crew consisted of Armstrong as Commander and Aldrin as Lunar Module Pilot, with Command Module Pilot Michael Collins. Armstrong and Aldrin landed in the Sea of Tranquillity and became the first humans to walk on the Moon on July 21. Their Lunar Module, Eagle, spent 21 hours 31 minutes on the lunar surface, while Collins remained in orbit in the Command/Service Module, Columbia.[2]

The three astronauts returned to Earth on July 24, landing in the Pacific Ocean. They brought back 47.5 pounds (21.5 kg) of lunar rocks.

What did it take?
1. 1 min of ORA-based cleaning
2. No 2nd iteration (did not use suggested n-grams or acronyms)
In Class Exercise
D2M in AutoMap

In AutoMap, the data to model (D2M) workflow can automate initial network creation

1. Unpack the sample data set “AutoMap_Data.zip”
   (an unpacked copy is available in “2 – AutoMap_Data”)
2. Start AutoMap
3. Run Data to Model Wizard
4. Explore D2M outputs

A quick note: If you have 10,000 texts, use batch mode!

Open AutoMap
(Start/Windows Icon→All Programs→AutoMap→AutoMap)
- A brief aside about help -

Data to Model (D2M) Wizard

1. Procedures menu
2. Data To Model Wizard menu item
3. Pick an output directory
4. Check “Location of text…”
5. Select the input file directory
Apply Optional Project Thesaurus and Delete Lists

6. Check “Select a project thesaurus…” and select the project thesaurus “AMnews_thes.csv”

7. Check “Select a project delete list…” and select the project delete list “AMnews_delete.csv”

8. Start the run

D2M Wizard running

1. Close “Complete” notice
2. Close D2M Wizard Window
D2M – What did it do?

- "Cleans Text" (removes extra whitespace, corrects British to American spelling, fixes common typos, expands contractions)
- Generates named thesauri
- Generates deplural thesauri
- Applies deplural thesauri
- Data preparation (pronoun resolution, delete singleton letters, n-gram conversion, removes noise words)
- Generates "Suggested" thesauri for uncategorized items using Part of Speech (POS)
- Applies general and project thesauri and delete lists
- Generates...
  - Networks: Concept, Semantic, Meta (unioned across all input files)
  - Concept List (unioned across all input files)
  - Suggested Acronym List
  - Suggested N-Gram List

D2M Wizard Outputs - Networks

1. Use Explorer (⊞️ + e) to open AMnews_for_SI_D2M_sample_output_w_project_thes
2. Open README_D2M.txt and read/scan its contents
**D2M Wizard Outputs – Meta-Network**

1. Open ORA
2. Drag and drop `union_meta_network.xml`
3. To ORA’s Meta-Network Manager
4. Click “+” mark

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td>26</td>
</tr>
<tr>
<td>Event</td>
<td>5</td>
</tr>
<tr>
<td>Knowledge</td>
<td>1</td>
</tr>
<tr>
<td>Location</td>
<td>23</td>
</tr>
<tr>
<td>Organization</td>
<td>22</td>
</tr>
<tr>
<td>Resource</td>
<td>2</td>
</tr>
<tr>
<td>Unknown</td>
<td>489</td>
</tr>
</tbody>
</table>

**“Reduced-Form Meta-Networks”**

- These node sets
- Undirected networks
- One network for any pair of entities.
- Why?
  - Can save space
  - Certain metrics depend on orientation.

- Agent
- Event
- Knowledge
- Location
- Organization
- Resource
- Unknown
Visualize the Meta-Network

1. Select the “Agent x Agent” network
2. Click “Visualize Only this Network”.

D2M Wizard Outputs – Meta-Network

3. Press “Play” to re-arrange the network
4. Press “Magnifying Glass” to re-center the network
Now let’s look at the semantic network:

1. Click “-“ to contract the `union_meta_network`
2. Drag and drop `union_semantic_network.xml` to:
3. ORA’s Meta-Network Manager
4. (If you see an “Import Menu” pop-up, select “Append as Additional Meta-Network”.)
5. Click Visualize then...

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>568</td>
</tr>
<tr>
<td>Text</td>
<td>7</td>
</tr>
</tbody>
</table>

Select Network Display Options

7. Press “Continue”. 
D2M Wizard Outputs – Semantic-Network

1. Color the ‘Text’ nodes blue: left click ‘Text’ and select from color menu
2. Press “Play” to re-arrange the model
3. Press “Magnifying Glass” to re-center the network

D2M – Two (2) Thesauri

1. Generated thesaurus (A & B)
2. Final thesaurus = combination of all thesauri used to generate networks (not shown in slide, may be empty)
D2M – ReviewList.csv

1. Review List
2. AutoMap’s low-confidence suggestions and ‘I give up’ declaration

D2M – Suggestions

1. Open Suggestions directory
2. Possible acronyms
3. Suggested n-grams
D2M – fullunion.csv

- A “concept list”, providing frequencies for all concepts present in the text after processing by the D2M.

Other useful locations

- C:\temp\am3temp
  - Default temporary workspace, cleared when you exit AutoMap
  - Intermediate files created by AutoMap are here.
    - Networks for individual documents...

- C:\Users\<you>\automap3\logFiles OR C:\Documents and Settings\<you>\automap3\logFiles
  - Log files reporting the steps in the cleaning and network generation

- ...(as above)\timingFiles
  - Timing files, times individual steps
D2M – What Next?

- Begin iterations of clean→assess→clean→assess
- Stop iterations when:
  - Key Entity Reports in ORA have no entries that make no sense in overall network and per-nodeset
  - SME or other validation of the model is in the realm of reasonable for your research topic
  - Out-of-resources (e.g., time, money, person-hours)
  - Research question answerable
  - ???

Text data Sources

D2M Wizard → ORA → Edit

D2M Wizard
Apply results
Clean outputs

D2M Wizard
Apply results

Custom AutoMap Run
Apply changes
Clean

ScriptRunner Run
Apply changes
Clean

Relational data
EXERCISE:
FROM NEWS TO NETWORK

Exercise Overview

• Create a network from one or more articles downloaded from the web.

• Steps:
  1. Create an empty directory for your articles.
  2. “Scrape” one or more articles from the web, saving them as text.
  3. Create an output directory.
  4. Run the D2M process.
  5. Create and apply a project thesaurus based on D2M suggests (or, work from your original text).

• Group work is allowed.
Notes

- Text file format:
  - Use UTF-8 (without BOM) or ANSI encoding
  - File extension ".txt"

- Scraping text from a web browser
  - “Save as text...” generally carries excess text along.
  - Simplest strategy is to highlight and copy text