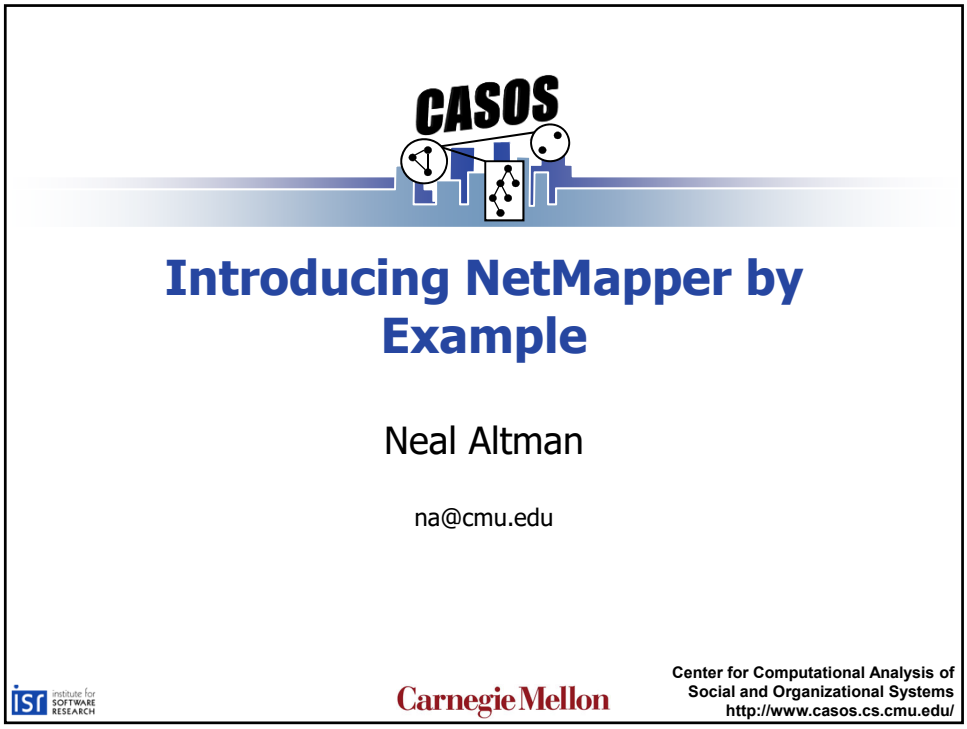
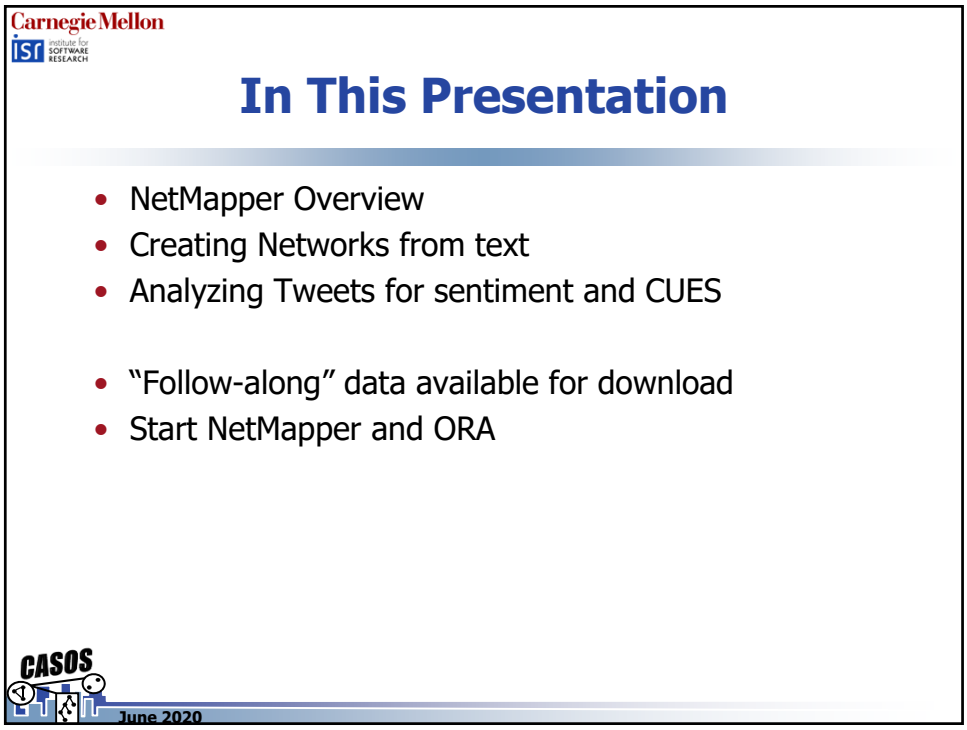

CASOS
Introducing NetMapper by Example
Neal Altman
na@cmu.edu
institute for SOFTWARE RESEARCH
Carnegie Mellon
Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/


Carnegie Mellon
institute for SOFTWARE RESEARCH
In This Presentation
NetMapper Overview
Creating Networks from text
Analyzing Tweets for sentiment and CUES
"Follow-along" data available for download
Start NetMapper and ORA
CASOS
June 2020

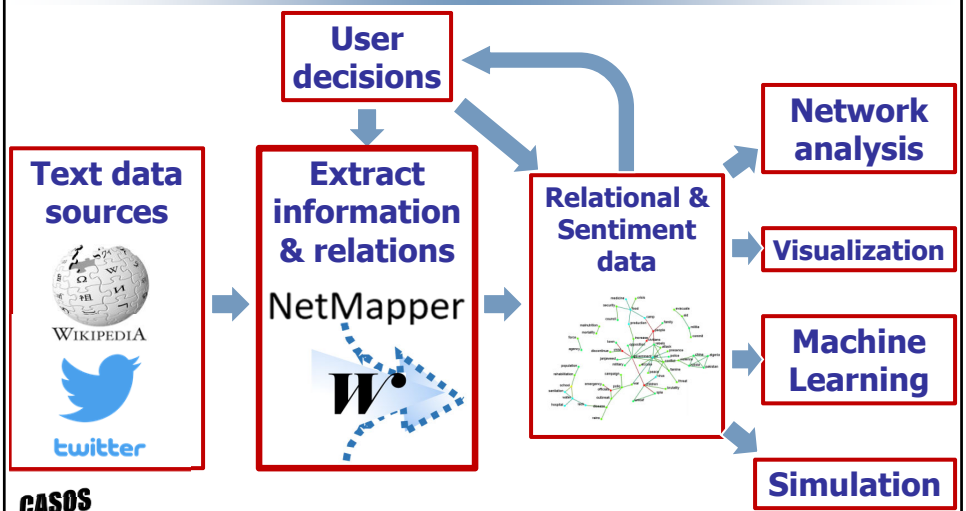
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

NetMapper

- NetMapper is a tool that supports extracting concepts from texts and assigning sentiment at the concept level.
- NetMapper's principal input types:
  - Plain text documents.
  - Twitter tweets.
- NetMapper processes text to identify concepts and their relationships.
- NetMapper's principal outputs
  - Networks – concepts and the links between them
  - Statistics about concepts
    - Sentiment Analysis
    - CUES
- NetMapper is interoperable with ORA

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

The NetMapper Workflow



Carnegie Mellon
institute for SOFTWARE RESEARCH

"User Decisions"

- Principal user tasks:
  - Collecting and preparing text
  - Augment built-in universal translators with domain specific mappings
    - Define domain specific concepts and common concepts
    - List unwanted text/concepts.
  - Selecting options for text processing and output
  - Evaluating outputs (using ORA)
- Illustrate with two operational examples
  - Creating networks from plaintext
  - Extracting sentiment from social media

Carnegie Mellon
institute for SOFTWARE RESEARCH

NetMapper Example 1: Creating Networks from Text

- Netmapper:
  - Load text
  - Set parameters
  - Create networks
- ORA
  - Load networks
  - Visualize results

Carnegie Mellon
institute for SOFTWARE RESEARCH

Text Data for NetMapper

- Text data is a series of files, containing content:
  - News stories
  - Journal articles
  - Blog posts…
- Text should be "plain":
  - Content only (no HTML tags, images, etc.)
- Supported text encodings:
  - ANSI (US-ASCII)
  - UTF-8
  - UTF-16
  - UTF-32

Carnegie Mellon
institute for SOFTWARE RESEARCH

Why Use Text as 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. Syria, 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).

Carnegie Mellon
institute for SOFTWARE RESEARCH

Network Creation

- NetMapper identifies concepts (a word or works identifying an idea)
- NetMapper prunes and modifies text:
  - Removes "noise" (e.g. stop words, punctuation, numbers).
  - Deletes specified words.
  - Translates synonyms and n-grams
- NetMapper treats the remaining concepts as **nodes** in a meta-network.
- NetMapper creates **links** between concepts which are sufficiently close:
  - Within a specified window of words/sentences of size N.
  - Within entire document.

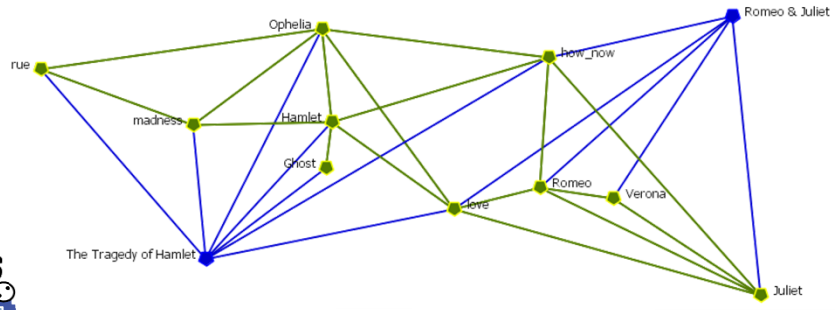
Carnegie Mellon
institute for SOFTWARE RESEARCH

There are TWO kinds of networks that you can extract from text using NetMapper

Carnegie Mellon
institute for SOFTWARE RESEARCH

Semantic Networks

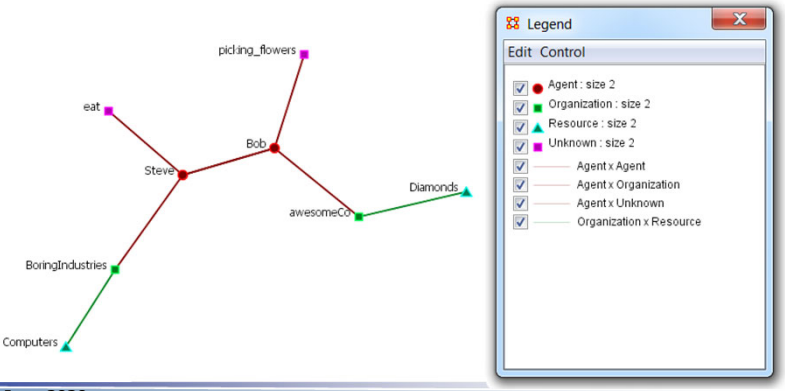
- Networks of words linked to each other based on co-occurrence.
  - Each link is concept-to-concept, e.g., in Shakespeare's Romeo and Juliet
    - Romeo Montague ↔ Juliet Capulet
- Networks of words linked to the documents in which they appear.
  - Each link is concept-to-document, e.g.,
    - Romeo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



Carnegie Mellon
institute for SOFTWARE RESEARCH

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)



Carnegie Mellon
institute for SOFTWARE RESEARCH

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).

Carnegie Mellon
institute for SOFTWARE RESEARCH

Step-by-Step Example

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Delete List

- Delete list – defines a set of concepts that should not be included in a network
- Format is a one column list of concepts.
- Two types of delete lists in NetMapper:
  - Universal Delete Lists – built in to NetMapper, applied to text by default (use can choose not to use them)
  - Domain Delete List – user provided list tailored to the input text
- Two ways to treat deleted concepts during link creation:
  - Ignore deleted concepts for distance determination.
  - Count deleted concepts when determining distance.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Thesaurus

- A thesaurus provides a translation of word(s) in the text to specified concept.
- Two main uses:
  - Merge synonyms and alternates to a common concept, reducing complexity:
    - "Rob", "Robert", "D. Robert Smith" → "Robert_Smith"
    - "Amazon", "Newegg", "eBay" → "online_vendor"
  - Group a series of adjacent words (*n*-grams) as one concept:
    - "Abraham Lincoln" → "Abraham_Lincoln"
    - "Torpedo boat destroyer" → "torpedo_boat_destroyer"
- Two types of thesauri in NetMapper:
  - Universal
  - Domain

Carnegie Mellon
institute for SOFTWARE RESEARCH

The Four Required Fields

- A NetMapper thesaurus is a tab-separated value (TSV) file containing a set of predefined columns:

conceptFrom	conceptTo	metaOntology	nodetype
Ken Macdonald, director of public prosecutions	Ken_Macdonald	agent	specific
2nd Battalion Royal Anglian Regiment	2nd_Battalion_Royal_Anglian_Regiment	organization	specific
Iraqi Finance Minister Rafi al-Isawi	Rafi_al-Isawi	agent	specific
Islamic Human Rights Commission	Islamic_Human_Rights_Commission	organization	specific
Lord Goldsmith, attorneygeneral	Peter_Goldsmith	agent	specific
Bow Street Magistrates' Court	Bow_Street_Magistrates'_Court	organization	specific
Liverpool John Lennon Airport	Liverpool_John_Lennon_Airport_UK	location	specific
Chief Editor Tariq al-Humayd	Tariq_al-Humayd	agent	specific
Iraqi Deputy Sabah al-Sa'idi	Sabah_al-Sa'idi	agent	specific
Crown Prosecution Service's	Crown_Prosecution_Service	organization	specific

- The file layout is:
  - Header line with fixed header fields separated by tabs.
  - Encoding is UTF-8 (without BOM)
  - One line per concept mapping.
  - Sorted by **conceptFrom** field length.

Carnegie Mellon
institute for SOFTWARE RESEARCH

The Four Required Fields

- **conceptFrom** – the match text in the input files
- **conceptTo** – the replacement concept (spaces replaced by underscores)
- **metaOntology** – one of the standard ORA node classes (more later)
- **nodetype** – note if the concept is general or explicit (allowed only for metaOntology types agent, organization, location and event):
  - **generic** - the concept applies to a class or group of things (e.g. "pilot", "government", "river", "depression").
  - **specific** – the concept applies to a particular instance (e.g. "Blériot", "Thailand", "Mississippi", "The_Great_Depression")
  - *<blank>* - other metaOntology types or unknown.

Carnegie Mellon
institute for SOFTWARE RESEARCH

Required and Optional Fields

Required Columns

1. conceptFrom
2. conceptTo
3. metaOntology
4. nodetype

Optional Columns

5. Category 1
6. Category 2
7. Category 3
8. Country
9. First Name
10. Last Name
11. Gender
12. Suffix
13. Language
14. Acronym
15. Valence
16. Evaluation
17. Potency
18. Activity
19. Affect Mean
20. Military Role
21. Political Role
22. Religious Role
23. Abusive
24. Exclusive
25. PowerAnger
26. PowerEncourage
27. PowerFear
28. PowerForbidden
29. PowerGreed
30. PowerLust
31. PowerSafety
32. Absolutist
33. Equivocal
34. Connective
35. NamedEntity
36. Pronoun_Level
37. Adverb
38. OtherUsage
39. Inclusive

Carnegie Mellon
institute for SOFTWARE RESEARCH

MetaOntology (Node Classes)

- NetMapper supports the standard ORA node classes:
  - **Agent** refers to single actors.
  - **Organization** refers to actors that consist a group of agents.
  - **Knowledge** describes cognitive capabilities and skills.
  - **Resource** refers to things that can be owned or acquired.
  - **Belief** identifies attitudes, positions or beliefs.
  - **Event** identifies occurrences or phenomena.
  - **Task** refers to actions than an actor can, or cannot take.
  - **Location** refers to places, real or conceptual.
  - **Role** is a *deprecated* identifier for position, function, or purpose.
  - **Action** is a *deprecated* synonym for Task.
  - **Unknown** is used when a nodeset is not otherwise classified.

Carnegie Mellon
institute for SOFTWARE RESEARCH

Plain Text Output Files

File Type	Content
X.emoticon.tsv	Emoticons and emoji found in text.
X.hashtag.tsv	Hashtags found in text.
X.meta.xml	Conventional meta-network built from text.
X.phone_number.tsv	Phone numbers found in text.
X.rnmf.tsv	List of all concepts in each text with statistics for each.
X.semantic.xml	Semantic meta-network built from text.
X.tsv	List of all concepts and assigned ontological category.
X.twitter_handle.tsv	Twitter handle to author map.
X.url.tsv	URLs found in text.
X.usage_measures.tsv	List of all concepts in each text with statistics for each.
X.zip_code.tsv	ZIP codes found in text.

Carnegie Mellon
institute for SOFTWARE RESEARCH

Example 2: Sentiments and CUES from Micro-Text

- Netmapper:
  - Load tweets (text micro-blocks)
  - Set parameters
  - Calculate statistics
- ORA
  - Sentiment
    - Read sentiment values as meta-network
  - CUES
    - Load tweets
    - Append CUES values as attributes
    - Analysis
    - Sentiment

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Processing Tweets with NetMapper

- Twitter "is an American microblogging and social networking service" (Wikipedia).
- Tweets can be downloaded in structured formats.
- NetMapper supports several JSON formats for reading in Tweets
- We will process tweets for:
  - Sentiment – how positive or negative the post is about a concept
  - CUES – estimates various emotional states

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Twitter Output Files

File Type	Content
X.emoticon.tsv	Emoticons and emoji found in tweets.
X.hashtag.tsv	Hashtags found in tweets.
X.indexed_sentiment.tsv	Concepts and sentiment values grouped by tweet.
X.phone_number.tsv	Phone numbers found in tweets.
X.rnmf.tsv	Concepts with tweet ID and sentiment values.
X.tsv	List of all concepts and assigned ontological category.
X.twitter_handle.tsv	Twitter handle to author map.
X.url.tsv	URLs found in tweets.
X.usage_measures.tsv	Statistics per tweet.
X.zip_code.tsv	ZIP codes found in tweets.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Step-by-Step Example

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Processing for Sentiment Data

- Domain thesaurus to coalesce hashtags

conceptFrom	conceptTo	metaOntology	nodetype
#EllicottCityFlood2018	#EllicottCityFlood	Event	specific
#ellicottcityflooding	#EllicottCityFlood	Event	specific
#Ellicott_City	#EllicottCityMD	Location	specific
#EllicotCity	#EllicottCityMD	Location	specific
#Ellicott	#EllicottCityMD	Location	specific
#ECFlood	#EllicottCityFlood	Event	specific
#Ellicot	#EllicottCityMD	Location	specific

- Concept list to limit analysis to selected hashtags

#BCFloods
#climatechange
#EllicottCityFlood
#EllicottCityMD
#ExtremeWeather
#flood
#flooding
#FloodRisk
#FloodWatch
#Hurricane
#Wuppertal

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Twitter Output Files (Sentiment)

File Type	Content
X.emoticon.tsv	Emoticons and emoji found in tweets.
X.hashtag.tsv	Hashtags found in tweets.
X.indexed_sentiment.tsv	Concepts and sentiment values grouped by tweet.
X.phone_number.tsv	Phone numbers found in tweets.
X.rnmf.tsv	Concepts with tweet ID and sentiment values.
X.tsv	List of all concepts and assigned ontological category.
X.twitter_handle.tsv	Twitter handle to author map.
X.url.tsv	URLs found in tweets.
X.usage_measures.tsv	Statistics per tweet.
X.zip_code.tsv	ZIP codes found in tweets.

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Adding CUES to ORA

- CUES are metrics that estimate various emotional states in messages based on subconscious cues in the text.
- CUES can be appended to ORA meta-networks

Carnegie Mellon
institute for SOFTWARE RESEARCH

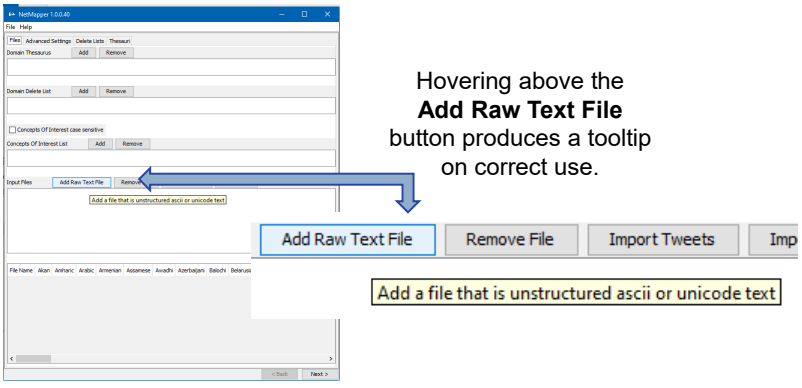
Twitter Output Files (CUES)

File Type	Content
X.emoticon.tsv	Emoticons and emoji found in tweets.
X.hashtag.tsv	Hashtags found in tweets.
X.indexed_sentiment.tsv	Concepts and sentiment values grouped by tweet.
X.phone_number.tsv	Phone numbers found in tweets.
X.rnmf.tsv	Concepts with tweet ID and sentiment values.
X.tsv	List of all concepts and assigned ontological category.
X.twitter_handle.tsv	Twitter handle to author map.
X.url.tsv	URLs found in tweets.
X.usage_measures.tsv	Statistics per tweet.
X.zip_code.tsv	ZIP codes found in tweets.

Carnegie Mellon
institute for SOFTWARE RESEARCH

Tooltips

- Moving the cursor above NetMapper control often provides a description of the control's function.



Hovering above the **Add Raw Text File** button produces a tooltip on correct use.