NetMapper

Dr. Kathleen M. Carley katheen.carley@netanomics.com (CEO, CSO)
Eric Malloy (Lead Software Engineer)

Netanomics

NetMapper

- Text mining tool for extracting networks and node attributes from texts
  - Based on a combination of language technology algorithms, machine learning, and thesauri
- NetMapper supports
  - Semantic network extraction
  - Meta-network extraction
  - CUES extraction
  - Sentiment extraction
- Produced by Netanomics
  - http://netanomics.com/

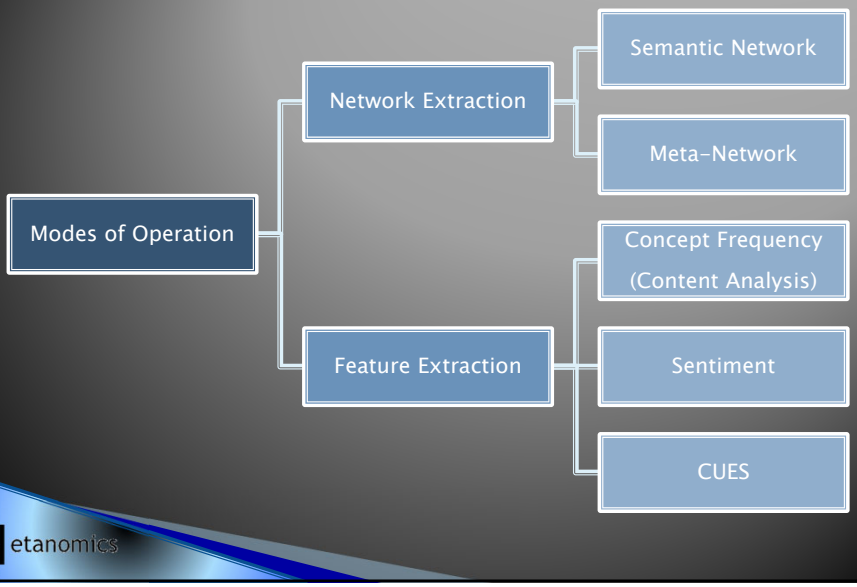
Enabling Users

- Users can apply defaults
- Users can also
  - Add a domain delete list
  - Add a domain thesauri/translation list
  - Add a list of keywords for sentiment identification
- Users can select special languages, level of classification, and can choose from specific delete lists and thesauri lists

IO & Interoperability

- Input format
  - Raw .txt files (automatically removes images)
  - PDF files (automatically removes images)
  - Json files
  - Csv or tsv files
- Output
  - Dynetml (dynamic version of graphml)
  - tsv

Two Modes of Operation



Netanomics

Thesauri – Illustrative

- Agent Specific
- Organization Specific
- Event Specific
- Emoji
- Emoticon
- Cities
- US State acronyms
- News agencies
- Ologies
- Abusive language
- pronouns

- Agent generic
- Organization generic
- Event generic
- Task generic
- Knowledge generic
- Resource generic
- Belief generic
- Location generic

Ontological Categories

Specific	Generic
Agents	Agents
Organizations	Organizations
Locations	Locations
Events	Events
	Tasks
	Resources
	Knowledge
	Beliefs

Delete

- Stop Words
- Connectors
- Measurement
- Modulators
- Negation
- Numbers
- Time
- In addition, strings of numerals and special characters are deleted, as are excess spaces

Other features

- Corrects for common English typos
- Converts from British English to American EnglishIIdentifies and translates emoji and emoticons

Netanomics

Translation – Illustrative

Light translation – over 3000 words in over 40 languages

- Arabic
- Armenian
- Belarusian
- Brazilian
- Chinese Mandarin
- Czech
- Danish
- Dutch
- Farsi
- Finnish
- French
- German
- Greek
- Haitian
- Hungarian
- Indonesian
- Italian
- Japanese
- Kazakh
- Kurdish
- Korean
- Latvian
- Lithuanian
- Malayalam
- Norwegian
- Pashto
- Polish
- Portugese
- Romanian
- Russian
- Serbo-Croatia
- Slovak
- Spanish
- Sudanese
- Swedish
- Tagalog
- Teluga
- Thai
- Turkish
- Ukrainian
- Urdu
- Uzbek
- Vietnamese
- English

Network Extraction

- Proximity based – default window of 2 sentences is used
  - Connections are made from first concept in window only to later to avoid double counting
  - All connections are treated as bi-directional
- User chooses whether to apply delete lists
- User chooses whether to apply thesauri/translation lists
- The default is universal lists are applied
- Language is auto-detected – but user can over-ride

Semantic Network

- First delete and thesauri/translation are applied if being used
- All identified concepts are put into the ontology – knowledge
- Links are number of times those two concepts appeared in the same window summed across all texts
- Output format is dynetml
- Output can be read by ORA

Meta-Network

- First delete and thesauri/translation are applied if being used
- All identified concepts are put into pre-defined ontological categories, unknown concepts are put in to the category “unknown”
- Links are number of times those two concepts appeared in the same window summed across all texts
- Output format is dynetml
- Output can be read by ORA

Feature Extraction

- Features are properties of the text
- Features fall in to many categories
- Counts
  - E.g., number of concepts
- Properties
  - E.g., reading level
- Presence of special feature
  - E.g., abusive terms
- When NetMapper extracts features it puts those in to TSV files
  - each row is a different document
  - each column is a different feature
  - These can be read in to ORA and used as attributes on the nodes

Concept Frequency

- Number of times that concept has occurred per document
- This is used for content analysis
- Thesauri and delete lists are applied prior to doing this count

Sentiment

- User supplies a list of key words
- Sentiment "in context" is calculated per document per key word
- A set of affect bearing words have a default sentiment in the thesauri
  - For most words the default is zero
  - These values are between 0 and 1
- The sentiment in context is calculated using a weight and adjustment scheme for proximate words
  - Only concepts near each other impact each other's sentiment
  - Negation words flip sentiment
  - Modulators can weaken or strengthen sentiment
  - Concepts near each other with similarly valenced sentiment can strengthen sentiment
  - Concepts near each other with oppositely valenced sentiment can weaken sentiment

CUES

- Subconscious tells in documents that signal the emotional state of the author and which can evoke a particular emotion in the reader
- Examples include
  - Use of first person pronouns
  - Use of abusive language
  - Use of exclusive or inclusive language

Support

- NetMapper User Guide
- Post problems to ORA google groups

http://www.casos.cs.cmu.edu/projects/ora/ORAGoogleGroup.php