Where Does the Data Come From?

Prof. Kathleen M. Carley

kathleen.carley@cs.cmu.edu

Data Sources

- Pre-existing data sets – structured
- Questionnaires – semi-structured
  - Most tools don't have auto-features for networks
- Citation data – semi-structured
  - APIs or scrape
- Email – semi-structured
- Social Media & MMOG – semi-structured
  - APIs
  - Buy from provider
  - Constraints on data sharing and amount of data
  - Freeware – TweetTracker, BlogTracker
- Text
  - Qualitative or hand coding – e.g., invivo
  - Text mining – e.g., AutoMap, NetMapper
- Video
  - No tools
Many Types of Social Media

Many Situations Where Analysts Need to Examine Open Source Data
Motivation

- **Fact:** Collection and storage of large volumes of text data are cheap, easy, and efficient
  - Book, legal documents, news, emails, web sites, blogs, chats, annual reports, political debates, mission statements, interviews
- **Need:** Techniques, measures, and tools for automated knowledge discovery and reasoning about relational and sequential structures derived from linear data
- **Challenge:** Effective, efficient, and controlled extraction of relevant, user-defined instances of node and edge classes from unstructured, natural language text data

Data Formats

- **Unstructured**
  - Raw data often in text, audio, video, or mixed media form
  - E.g., news articles
- **Semi-structured**
  - Meta-data is in near network format
  - A partial structure that can be parsed
  - E.g., email, twitter, questionnaire
- **Structured**
  - Already in network format
  - E.g., network data in csv format
Critical Steps

- Data Pedigree and Information Assurance
  - Tracking source and modification steps
- Storage and Retrieval
  - SVN repositories and large databases
- Data Cleaning
  - Process of removing erroneous data, creating consistent coding formats, removing typos, etc.
- Data Fusion
  - Process of merging data from multiple sources
  - Often data cleaning is done before and after
  - Requires creation of common ontology
- Data Reduction
  - Deleting un-needed data
  - Merging data into larger granules

New companies are emerging as data providers that specialize in these steps

Text Mining

- Entity Extraction
  - Who, what, where
- Entity Disambiguation
  - When do two phrases or words refer to the same entity
  - Handling pronouns, mis-spellings, etc.
- Entity Classification
  - What ontological category does a concept fall in to
- Locating Links
  - When are two “concepts” linked
  - Semantics (meaning), syntax (order), proximity
- Text Similarity
  - Are these texts the same or about the same thing
- Theme Extraction
  - What ideas and authors/texts hang together
- Sentiment Mining
  - What is the prevailing "attitude" or "belief"
Text Analysis

- Content Analysis
  - Hot Topics
  - Themes
- Author identification
  - Pattern or “signature”
- Semantic Network Analysis
  - Mental Model
- Implied meta-network

- Activity Analysis
  - KEDS – focus on nouns
- Protocol Analysis
  - Logic reasoning
- Abstraction
  - Generate synopsis

Tools

- Lots of tools
  - Many focus on entity extraction or theme extraction
  - Many focus on only verbs or nouns
- Many tools only process part of the text
  - For news stories often the focus is only on headlines
  - For web pages often the focus is only on links or header
- Unresolved issues
  - Many
    - Time
    - Meta-data
    - Lists
    - Inferred meaning
    - Belief extraction
Basic Approach

**Focus: Meaning**
Do people use the same words
Do people use the same words in the same way

**Method: Textual Analysis**
Multiple sources
Verbal data

**Result: Rich Account**
Graphic or Quantified

Shared meaning - across people
Shifts in meaning - over time
Topics – relations among concepts

Levels of Analysis for Concepts

- **Node Level - Concept Based Techniques**
  - Traditional content analysis
  - Occurrence and frequency of concepts
  - Explicit and implicit concepts

- **Graph Level - Map Based Techniques, Network, Link**
  - Focus on meaning and relation between concepts
  - Occurrence and frequency of concepts and statements
  - Explicit and implicit concepts and statements
Tools and Workflows Exist and Are Improving for extracting, analyzing, forecasting, ...

Concept Circle - Example

Clustering Task #2: April 26, 1989

Directions: These words have been mentioned in class lectures over the past semester. Please draw a line between pairs of words which you believe should be connected. It is important that all connections that you intend to make be clear and easy to see. Please do not draw so many lines on any one worksheet that you cannot easily see how you've connected those words.

analysis, approaches, argument, aspects, attributes, authority, background, body, conclusion, community, contribution, converge, define the problem, detailing the problem, differences, directions, disclaimer, elaboration, explanation, fuzzy path, focus, framework, goal, progress, introduction, account, introduction, irrelevant, visual, visual, seeing an issue, one of argument, main path, main point, new, milestone(s)
Concept Circle - Cont.

Variations:

When Respondent Draws Lines

- Place strength on lines
- Place arrows on lines for causality
- Place marker on lines for type of link

Application Process

- Can be applied by interviewer during interview
- Can be done as reading text

Meta-Data

- **archivesource**: API source of tweet (twitter-search or twitter-stream)
- **text**: contents of tweet
- **to_user_id**: numerical ID of tweet recipient (for @replies)
- **from_user**: screen name of tweet sender
- **id**: numerical ID of tweet
- **from_user_id**: numerical ID of tweet sender
- **iso_language_code**: code (e.g. en, de, fr, ...) of sender’s default language
- **source**: name or URL of tool used for tweeting (e.g. Web, Tweetdeck, ...)
- **profile_image_url**: URL of tweet sender’s profile picture
- **geo_type**: form in which sender's geographical coordinates are provided
- **geo_coordinates**: first element of geographical coordinates
- **geo_coordinates_1**: second element of geographical coordinates
- **created_at**: tweet timestamp in human-readable format
- **time**: tweet timestamp as numerical Unix timestamp
Twitter Ties

- One mode directed
  - A follows b
    - Reflection of offline social relationships
      - Apx 22.1% follow each other
    - Subscriptions
      - Bulk
    - Makes it more like a news service
  - A retweets b
    - Retweets attached to sender creating social games
  - A mentions b
    - Retweets attached to sender creating social games
- Two mode
  - Hashtag usage
- Two mode undirected
  - Co-hashtag network

Text Mining to Extract Networks

- Network Text Analysis --- Encode links between words in texts and construct network of linked words
- Content Extraction (a.k.a. Content Analysis)
- Semantic Network Extraction (a.k.a. Mental Model Analysis)
- Meta-Network Extraction (a.k.a. Structural Analysis)
- Belief Extraction (a.k.a. Context based Sentiment Analysis)
- CUES