

Text

Cmpt 767 - Visualization

Steven Bergner

sbergner@sfu.ca

Overview – Text Vis

- Language Model Vis
- Words as nominal data
- Text Processing Pipeline
- Bag of Words Model
- Keyword Weighting

Reading / Sources

- Jeffrey Heer's [CSE512](#) – Data Visualization – Wk 7
- Notes from Directed Reading with Mayank Vaccher
- [Chapter 10 + 11: Info Vis for Search Interfaces and Text Analysis, in *Search User Interfaces*.](#)
[Marti Hearst. 2009](#)

Why visualize text?

- Understanding
 - Get “gist” of document
- Grouping
 - Cluster for overview or classification
- Comparison
 - Compare document collections
 - Show evolution over time
- Correlation
 - Compare patterns in text to those in other data

Language Model Vis

Many Text Visualizations represent the language model learnt and not the text itself

- How well does visualization represent properties of the model?
- Does the model enable reasoning about the text?

Word Clouds

abstract accepted analogue applications applying attuned bar burgeoning challenging
chapters chart collections combine communicate conducted convert **data** date difficult
discussed earlier effectively end evaluation evocative familiar field focus focused form
general goal graph highly human hundreds ideas images improve
information innovative insight kinds line makes means
meta-analysis nature new numbers order ost perceive perceptual points positive
problems providing purpose range rapidly read reading reasons representations **results**
retrieval robust **search** shortciten{chen2000esi} shortcite{larkin1987dsw} shown space
studies successful system table task tasks **text** textual time translate underlying
usability vibrant **visual visualization** visually web wide widely

Tag Clouds based on Word Count

Strength

- Can help with gisting and initial query formation

Weakness

- Sub-optimal visual encoding
- Inaccurate size encoding
- May not facilitate comparison
- Term frequency may not be meaningful
- Does not show text structure

Word Tree based on Word Sequences

Challenges of text vis

- High Dimensionality
 - Use text to represent text, if possible
- Context and Semantics
 - Show context to aid understanding
 - Provide access to source text
- Modeling Abstraction
 - Determine analysis task
 - Understand language model abstraction
 - Match analysis task with appropriate models and tools

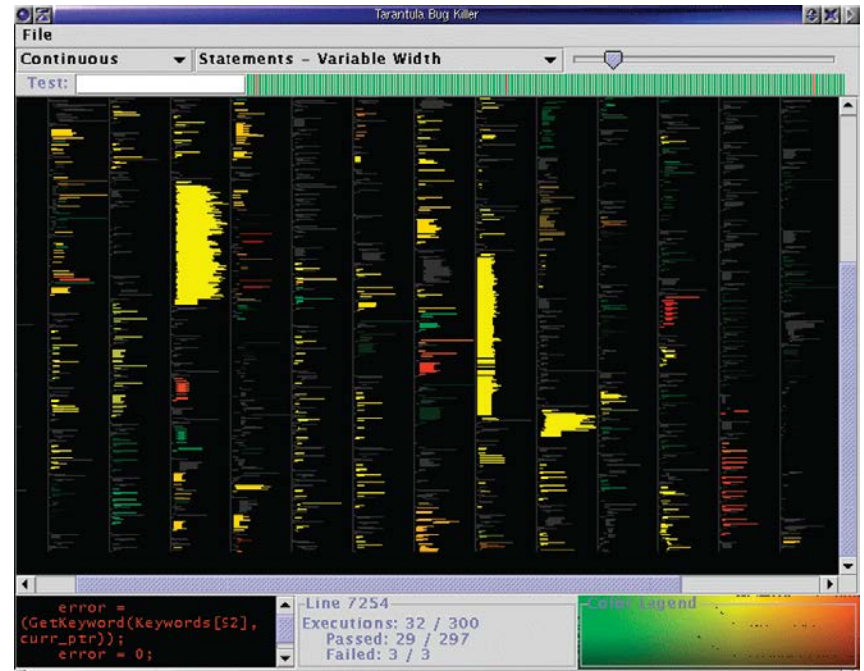
Words as nominal data

Have meanings and relations

- Correlations: Hong Kong, Bay Area
- Order: April, May, June, ...
- Membership: Tennis, Running, Swimming, Piano
- Hierarchy, antonyms and synonyms, entities, ...

Text data sets

- Dense layout of sequential, categorical data
- Grouped: character, word, sentence, ...
- Example: source code
 - Quantitative derived variables
 - Test coverage (brightness)
 - Test pass rate (hue)



Dense overview of source code with lines color coded by execution status of software test suite

Text Processing Pipeline

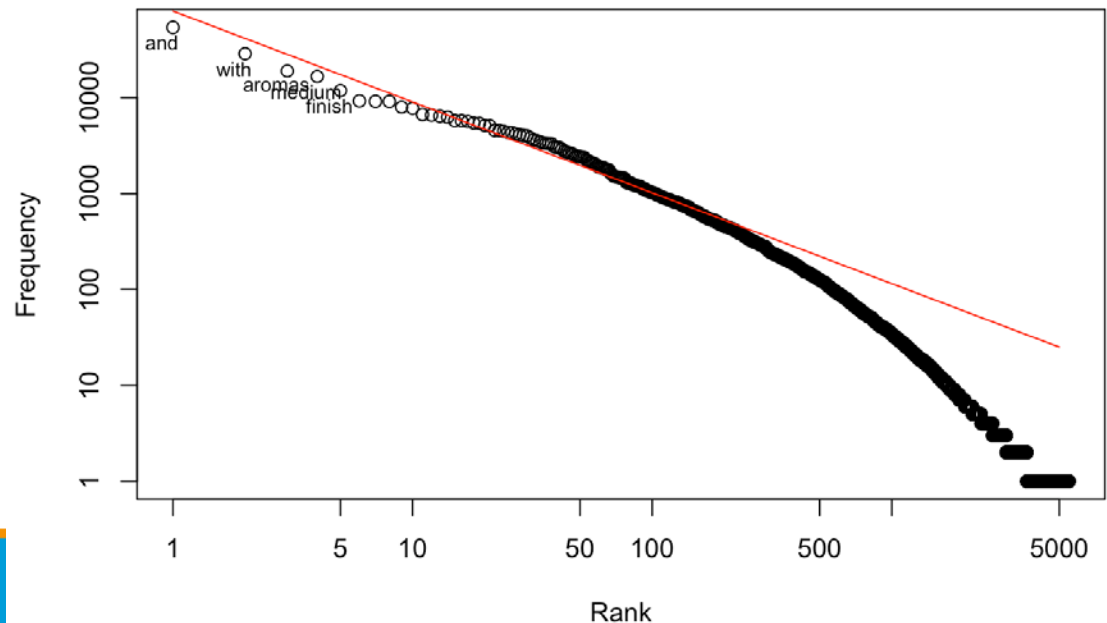
- *Tokenization*
 - Remove stop words
 - Numbers and symbols (like hashtags)
 - Entities (like British Columbia)
- *Stemming* to group different forms of a word
 - Porter Stemmer
 - visualizations, visualize, visually -> visual
 - Lemmatization
 - goes, went, gone -> go
- Ordered list of terms

Bag of Words Model

- Document -> vector of term weights
- Aggregate into a *document-term matrix*
 - In *xyz* document, *abc* term occurred *N* times
 - *N* can be any type of relevance weight

- Zipf-plot
rank proportional to 1/N

- [Ngram viewer](#)



Keyword Weighting

- Term Frequency (TF)
 - Count of term T in document
 - Can take $\log(\text{TF}+1)$ or normalize
- TF-IDF = $\log \text{TF} \times \log(\# \text{docs} / \# \text{docs containing T})$
- Term Commonness
 - normalized TF
relative to most frequent n-gram i.e. the word “the”
- G^2 probability of different word frequency

Limitations of Frequency Stats

- Typically focused on unigrams
- Often favors frequent or rare terms only
 - May not provide best description
- BoW ignores information
 - Grammar / Part-of-speech
 - Position within document
 - Named Entity Recognition

Further text vis tasks

- Categorization
- Phrase Nets
 - Node Grouping
- Comparison, Trends
 - Parallel Tag Clouds: linguistic differences faceted over time
 - Theme River - Stacked area charts of word count sequence
- Similarity & Clustering
 - Vector distances among docs, use to cluster
 - Topic Modeling using LSA and LDA
 - Text is mixture of topics

Categorization

- Flat
 - Small set, scannable
- Hierarchical, tree-structured
- Faceted
 - Many small trees
 - Item in collection can be labelled in several facet hierarchies

The screenshot shows the Zvents website interface. At the top, there is a search bar with the text "what are you looking for" and a "Zvents Discover Things To Do" logo. Below the search bar is a navigation menu with tabs for "events", "movies", "restaurants", "venues", and "performers". The current location is set to "Oakland, CA" with a "[change my location]" link. A "Refine Results" sidebar on the left lists various categories and neighborhoods with their respective counts. The main content area displays a list of events, including "Circus Oz", "Joe Reilly-Children of the Earth", "Teaching English as a Second Language/Foreign Language Certificate Program Information Session at UC Berkeley Extension", "Barack O'Bama is Irish", and "Alcohol and Drug Abuse Studies Certificate Program Information Session".

Zvents
Discover Things To Do

all
what are you looking for

events movies restaurants venues performers

Oakland, CA [change my location]

Refine Results

Category:

- All Categories
- Community
 - Activism (4)
 - Antiques & Collectibles (1)
 - Civic/Government (1)
 - Ethnic & Cultural (4)
 - Health (6)
 - History (1)
 - Home & Garden (3)
 - Religion (2)
 - Science (7)
 - Talks & Lectures (14)
 - Workshops & Classes (14)

City:

- Any City
- Berkeley

Neighborhood:

- Any Neighborhood
- Claremont Hills (2)
- Elmwood (3)
- North Berkeley (11)
- Northwest Berkeley (9)
- South Berkeley (9)
- Southwest Berkeley (1)
- University of California Berkeley (8)

When:

- Any Time
- Tomorrow (1)
- Next 7 Days (16)
- Next 14 Days (21)

Displaying **10 events out of 43**(with 15 repeats) in with city **Berkeley**.

Sort by: Relevance View: List | Map

Circus Oz
Family Fare is the perfect introduction to the performing arts for the entire family. ...

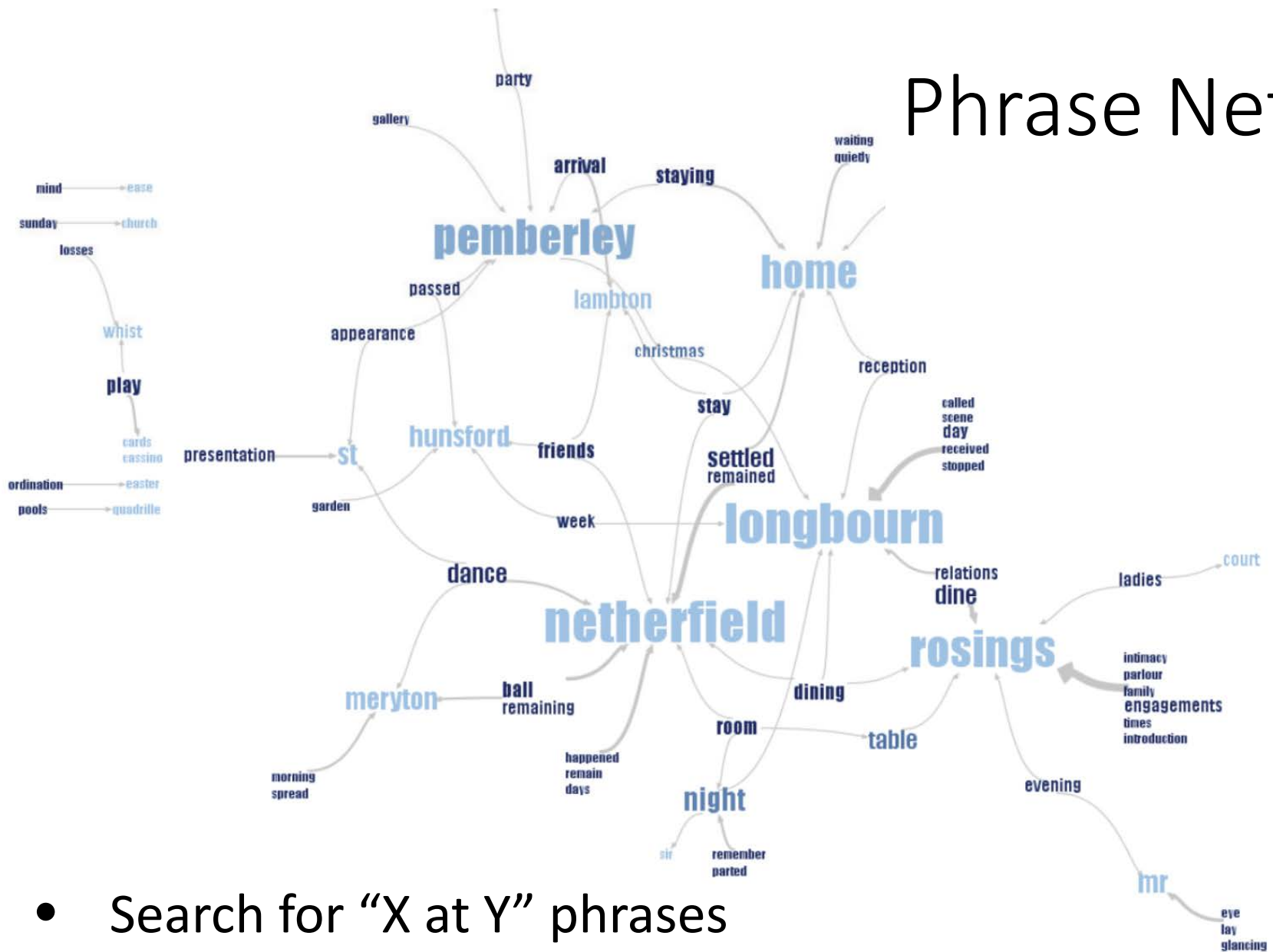
Joe Reilly-Children of the Earth
Singer songwriter and environmental educator offers eco-songs for kids of all ages. ...

Teaching English as a Second Language/Foreign Language Certificate Program Information Session at UC Berkeley Extension
Learn how UC Berkeley Extension's certificate program can prepare you for diverse ...

Barack O'Bama is Irish
The song that possibly tipped the balance in swing states in the US presidential ...

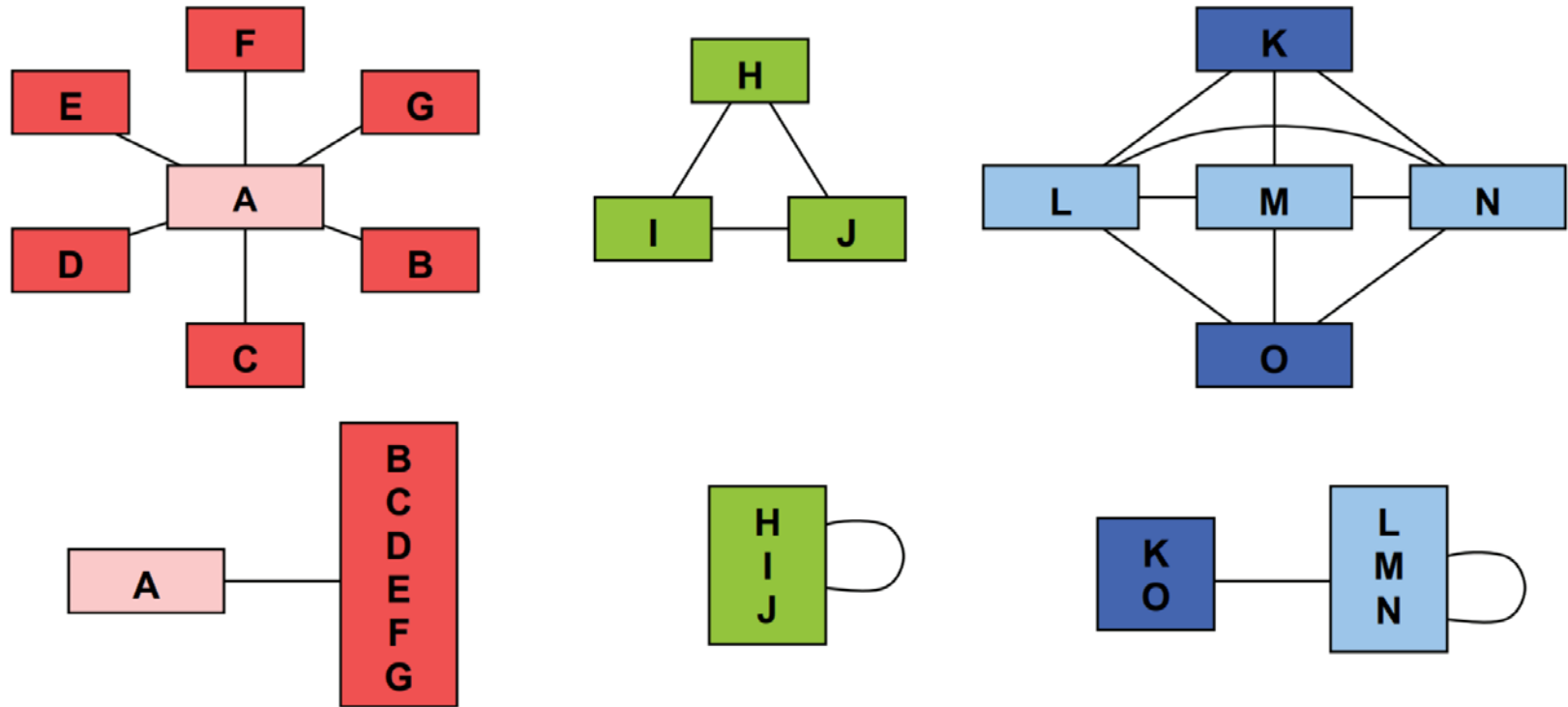
Alcohol and Drug Abuse Studies Certificate Program Information Session
Find out how you can enhance your skills or start a counseling career with a certificate ...

Phrase Nets



- Search for “X at Y” phrases

Phrase Nets – Edge Compression



Collapsing networks based on identical network neighborhoods

Thanks for your attention!

