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

[Hearst 2009, Fig. 11.10]

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

[Wattenberg, Viegas 2008] http://hint.fm/projects/wordtree/ [D3 collapsible trees]

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

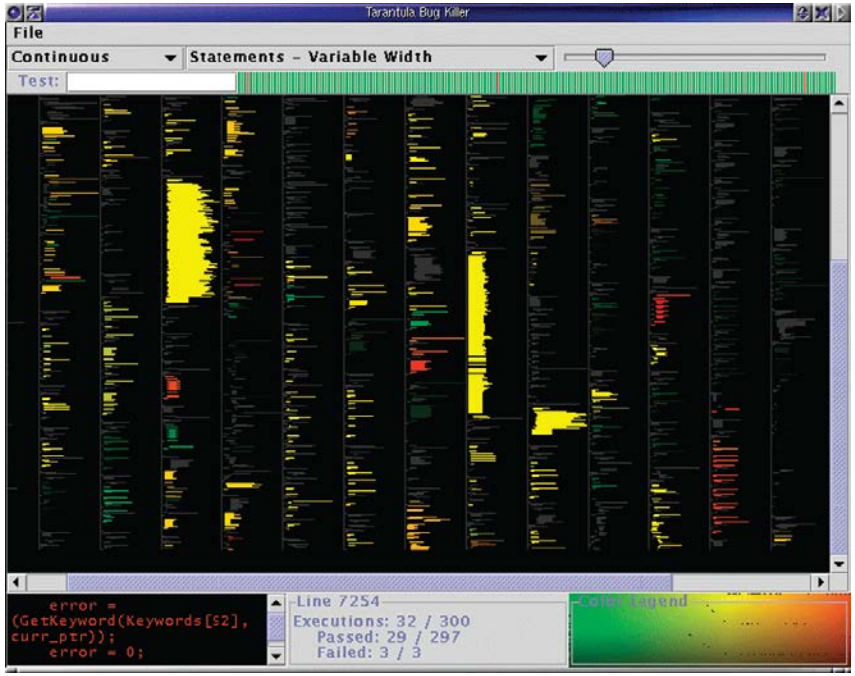
Words as nominal data

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

[*Tarantula*, Jones et al. 02, Figure 4].

Text Processing Pipeline

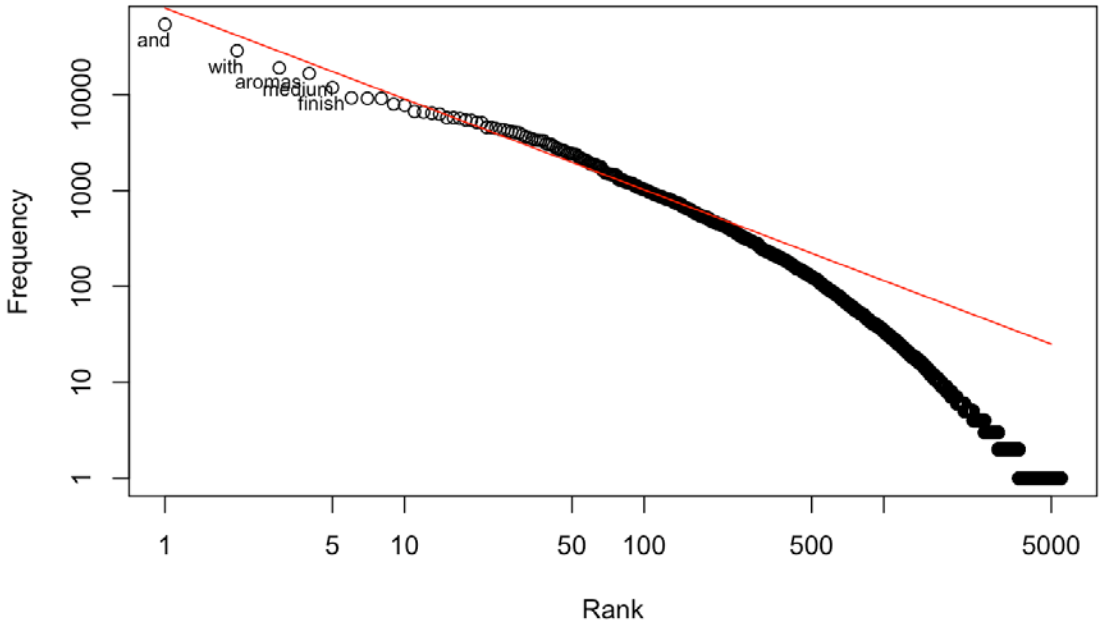
- *Tokenization*
  - Remove stop words
  - Numbers and symbols (like hashtags)
  - Entities (like British Columbia)
- *Stemming* to group different forms of a word
  - Porter Stemmer
    - visualizations, visualize, visually -> visual
  - Lemmatization
    - goes, went, gone -> go
- Ordered list of terms

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(TF+1) or normalize
- TF-IDF = log TF x log(#docs / #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

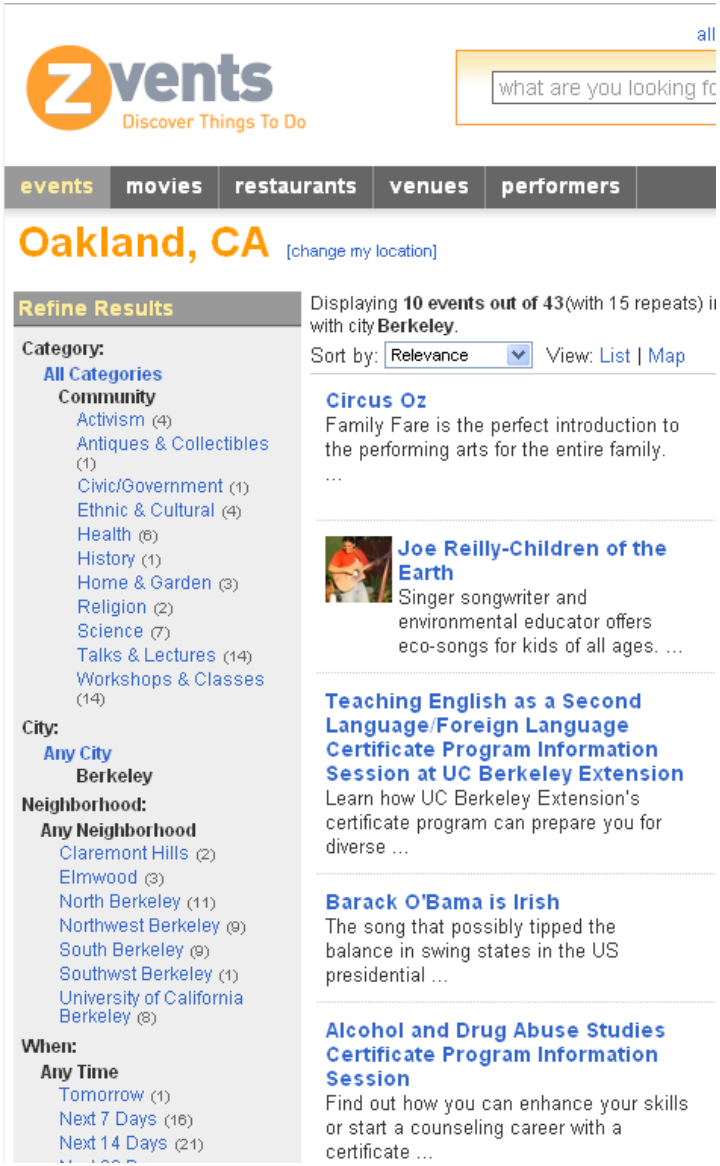
- Typically focused on unigrams
- Often favors frequent or rare terms only
  - May not provide best description
- BoW ignores information
  - Grammar / Part-of-speech
  - Position within document
  - Named Entity Recognition

Further text vis tasks

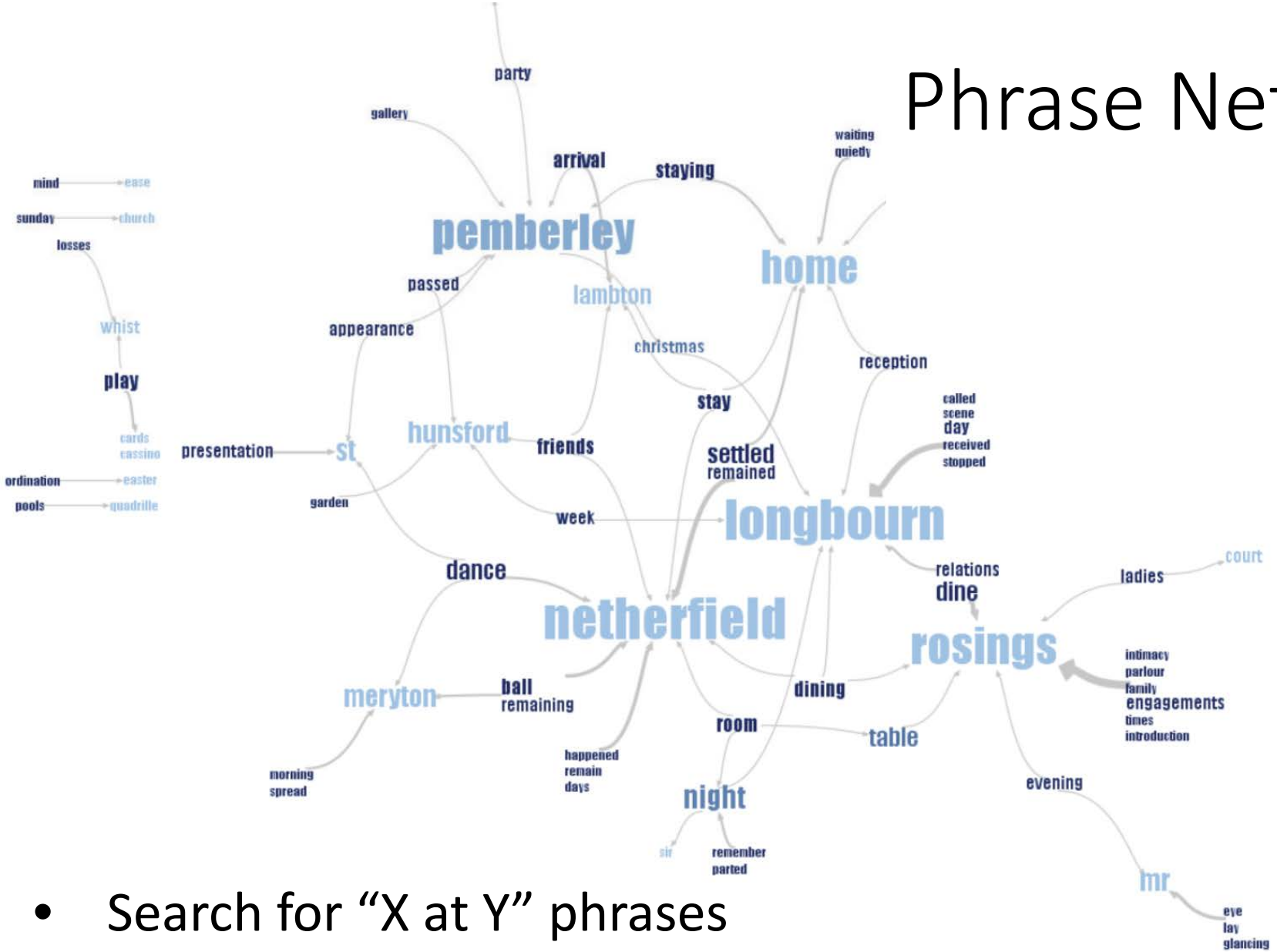
- Categorization
- Phrase Nets
  - Node Grouping
- Comparison, Trends
  - Parallel Tag Clouds: linguistic differences faceted over time
  - Theme River - Stacked area charts of word count sequence
- Similarity & Clustering
  - Vector distances among docs, use to cluster
  - Topic Modeling using LSA and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



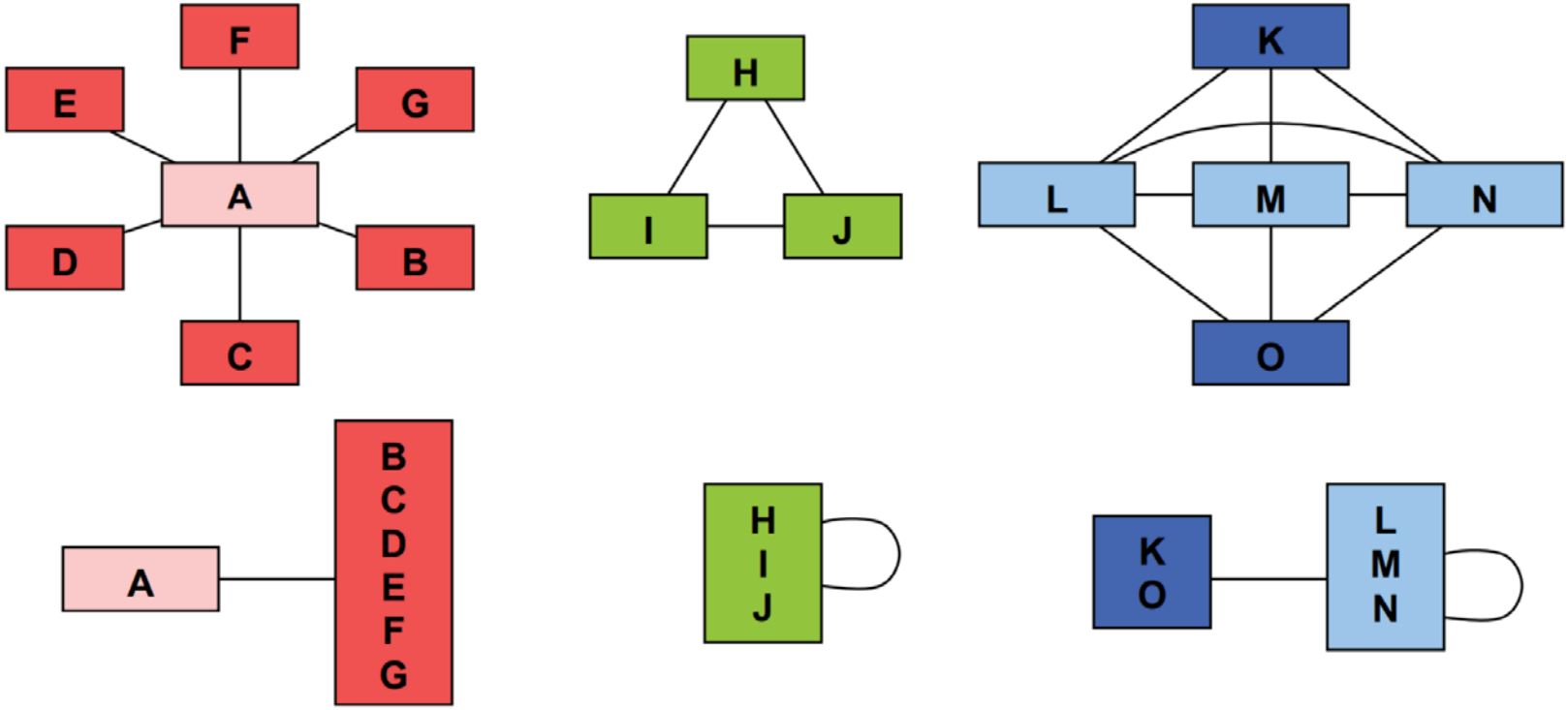
Phrase Nets



- Search for “X at Y” phrases

[Van Ham, Wattenberg, Viegas 2009] http://hint.fm/papers/phrase-net-rev5.pdf

Phrase Nets – Edge Compression



Collapsing networks based on identical network neighborhoods

Thanks for your attention!