Text Visualization

Text & Document Visualization

- Text not pre-attentive
- Text = Abstract Concepts = Very High Dimensionality
  - Multiple & ambiguous meanings
  - Combinations of abstract concepts more difficult to visualize
  - Different combinations imply different meanings
  - Language only hints at meaning
    ➔ based on common understanding "How much is that doggy in the window?"
- Facilitate Information Retrieval
  - Collection Overview
  - Visualize which parts of query satisfied by document / collection
  - Understand why documents retrieved
- Cluster Documents Based on Words in Common
  - Finds overall similarities among groups of documents
  - Picks out some themes, ignores others
- Map Clusters onto 2D or 3D Representation
  - Minimize time/effort to decide which documents to examine

What is text data?

Documents

- Articles, books and novels
- Computer programs
- E-mails, web pages, blogs
- Tags, comments

Collection of documents

- Messages (e-mail, blogs, tags, comments)
- Social networks (personal profiles)
- Academic collaborations (publications)

Text as Data

Words are (not) nominal?

- High dimensional (10,000+) More than equality tests
- Words have meanings and relations
  - Correlations: Hong Kong, San Francisco, Bay Area
  - Order: April, February, January, June, March, May
  - Membership: Tennis, Running, Swimming, Hiking, Piano
  - Hierarchy, antonyms & synonyms, entities, …

Text Processing Pipeline

- Tokenization: segment text into *terms*
  - Special cases? e.g., “San Francisco”, “L’ensemble”, “U.S.A.”
  - Remove stop words? e.g., “a”, “an”, “the”, “to”, “be”?
- Stemming: one means of normalizing terms
  - Reduce terms to their “root”; Porter’s algorithm for English
  - e.g., *automate(s), automatic, automation* all map to *automat*
  - For visualization, want to reverse stemming for labels
  - Simple solution: map from stem to the most frequent word
- Result: ordered stream of terms

The Bag of Words Model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
  - Each dimension corresponds to a term (10,000+)
  - Each value represents the relevance
  - For example, simple term counts
- Aggregate into a document x term matrix
  - Document vector space model

Document *x* Term matrix

- Each document is a vector of term weights
- Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

WordCount (Harris 2004)

WordCount™ is an interactive presentation of the 86,800 most frequently used English words.

http://wordcount.org

Visualizing Document Content

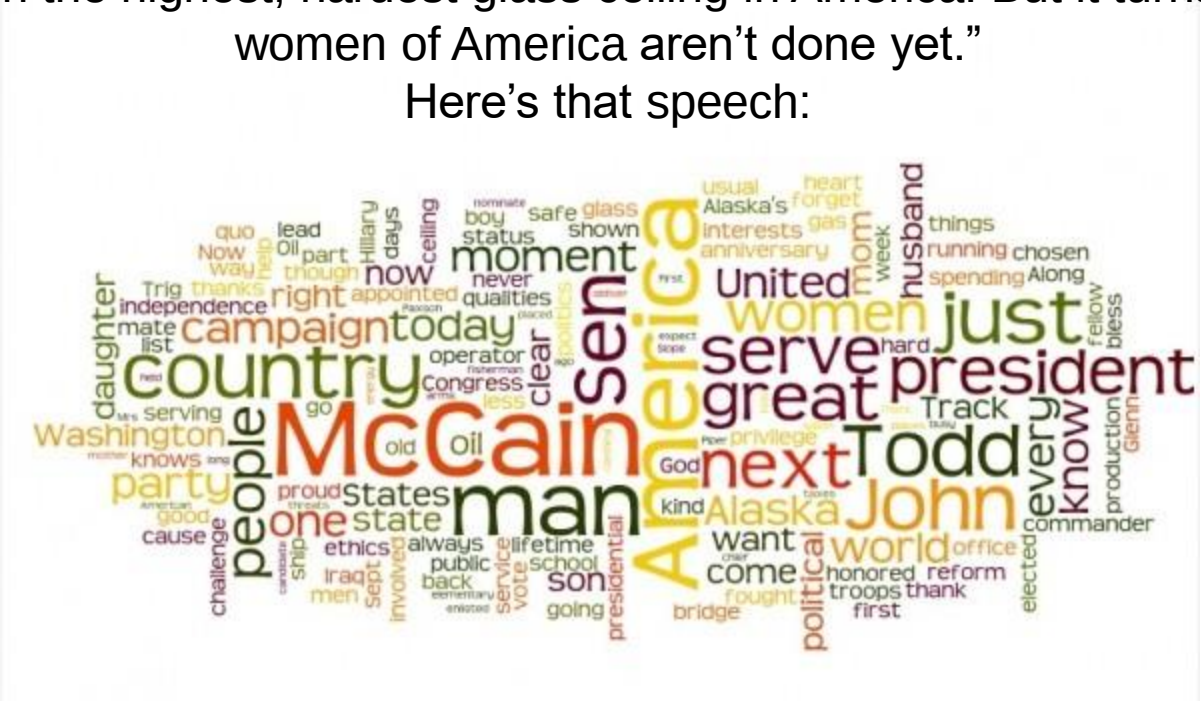
Tag Cloud: Word Counts

Wordle

http://www.wordle.net/create

During the campaign, Palin gave an energetic speech in Dayton, Ohio. She appealed to women voters by evoking, of all things, the presidential campaign of Democrat Hillary Clinton, saying, “Hillary left 18 million cracks in the highest, hardest glass ceiling in America. But it turns out the women of America aren’t done yet.”

Here’s that speech:

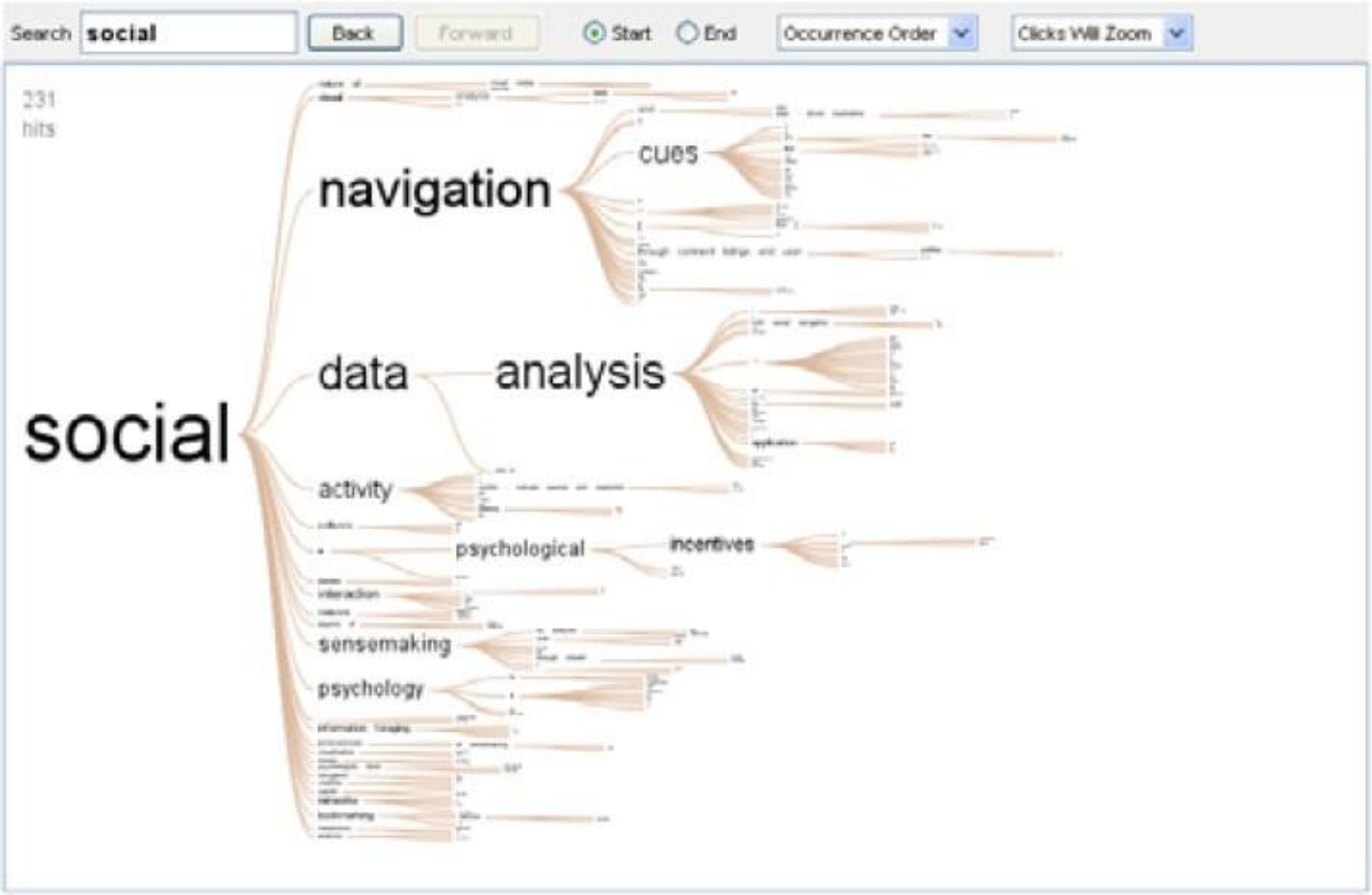


Text from: California Watch

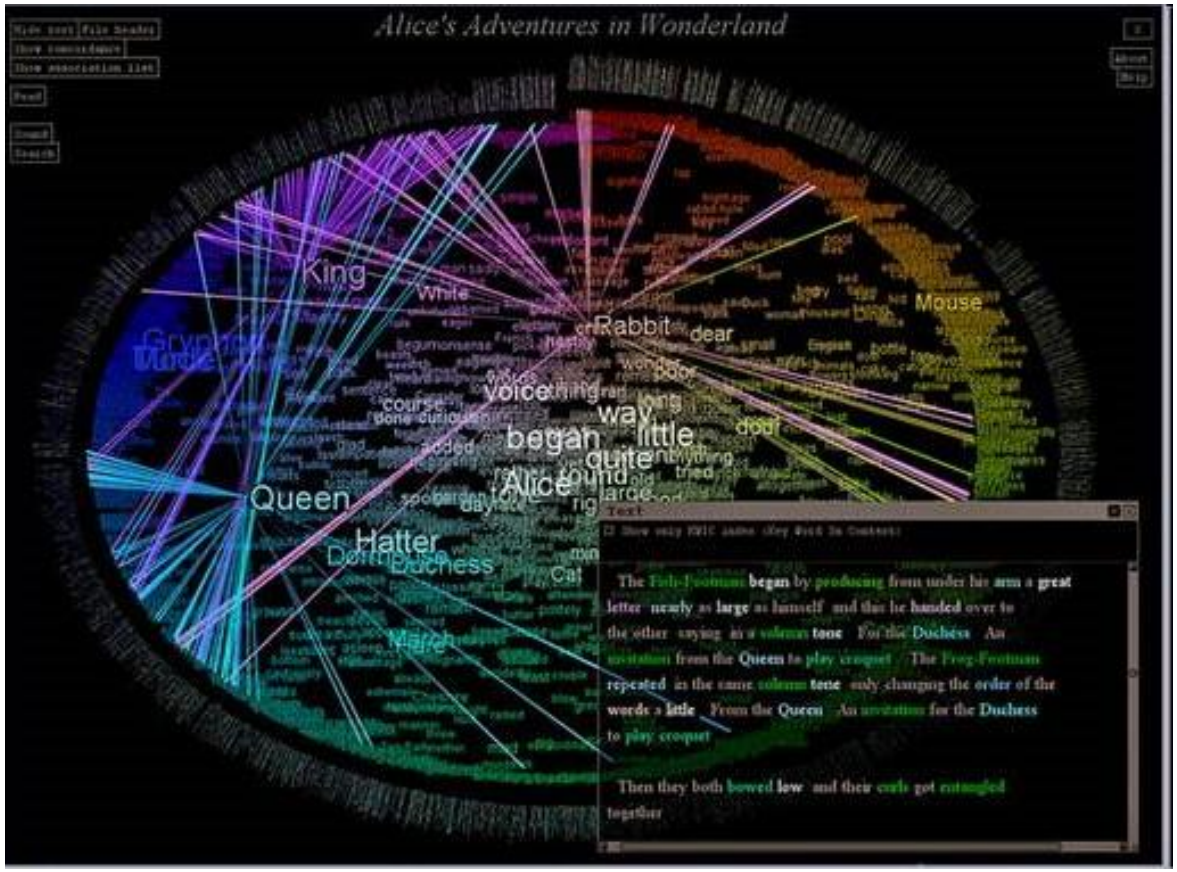
Weaknesses of Tag Clouds

- Sub-optimal visual encoding (size vs. position)
- Inaccurate size encoding (long words are bigger)
- May not facilitate comparison (unstable layout)
- Term frequency may not be meaningful
- Does not show the structure of the text

Word Tree: Word Sequences

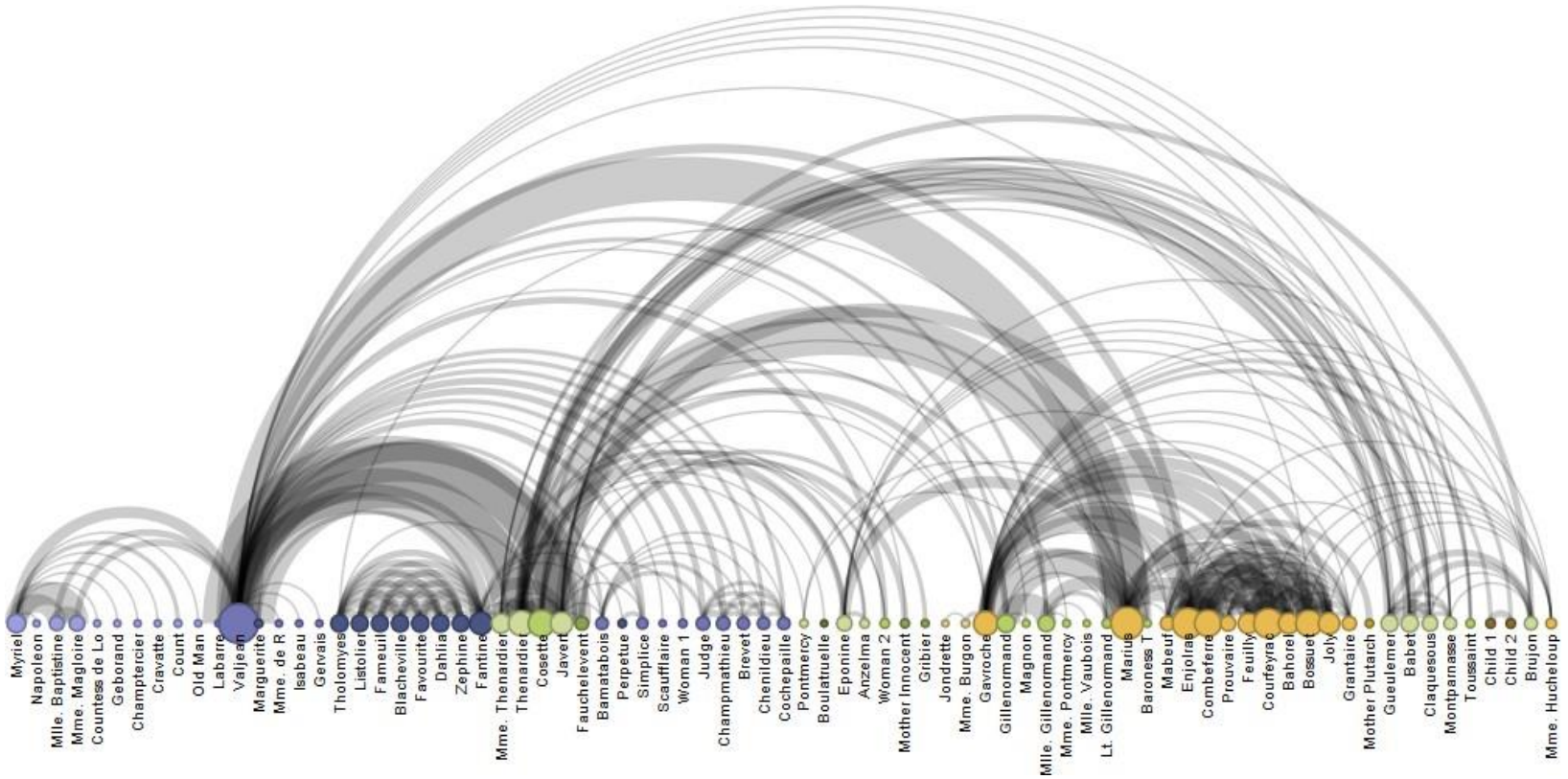


TextArc – Brad Paley



http://textarc.org/

Arc Diagrams – M. Wattenberg



Les Misérables character interaction. Each character is represented by a circle and the connecting arc represents co-occurrence in a chapter. The character's size indicates the number of appearances they have over the entire work.