CDS 6324
DATA VISUALIZATION

Lecture 11: Text Visualization

Outline

- Text Visualization
- Single Document Visualization
- Document Collection Visualization
- Extended Document Visualization

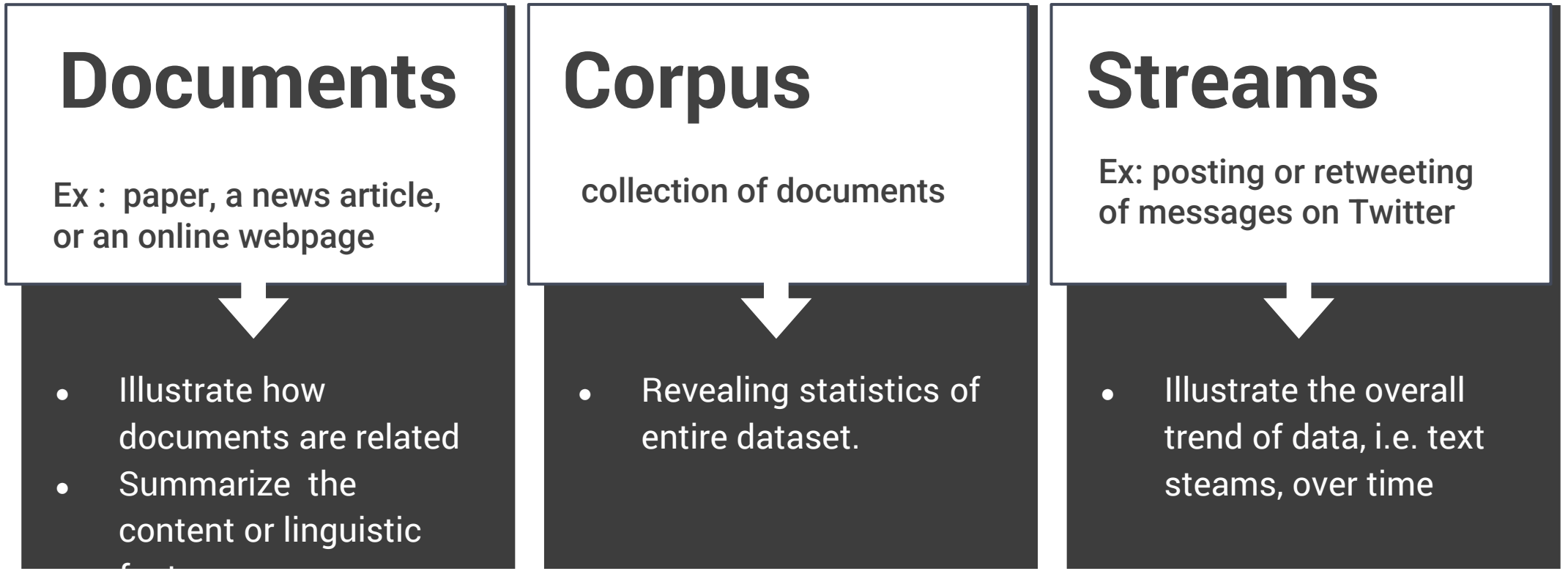
Text Visualization

Outline

- **Large collections of text documents** → daunting task of needing to understand the content of unfamiliar documents.
- **Text visualization**: Visualizing content-wised information in an intuitive manner and enable the discovery of actionable insights.

Text Documents

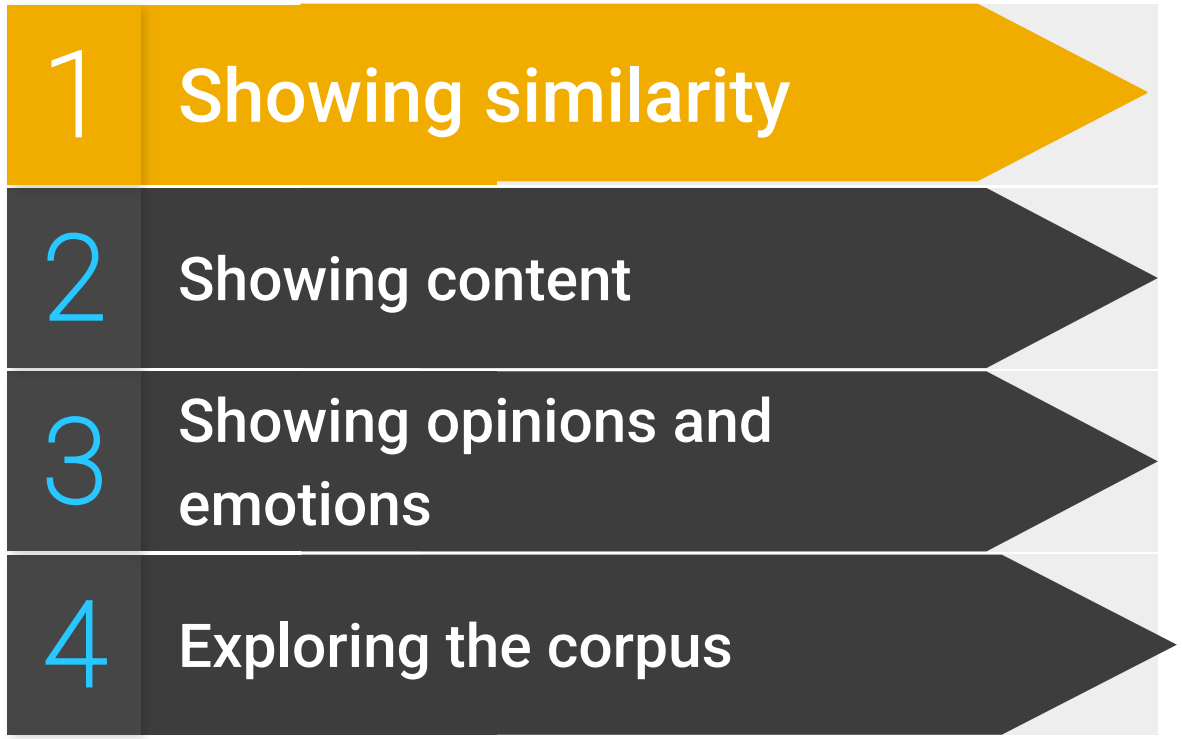
- Text visualization techniques largely designed to deal with the following three major forms of text data:



Text Visualization Techniques

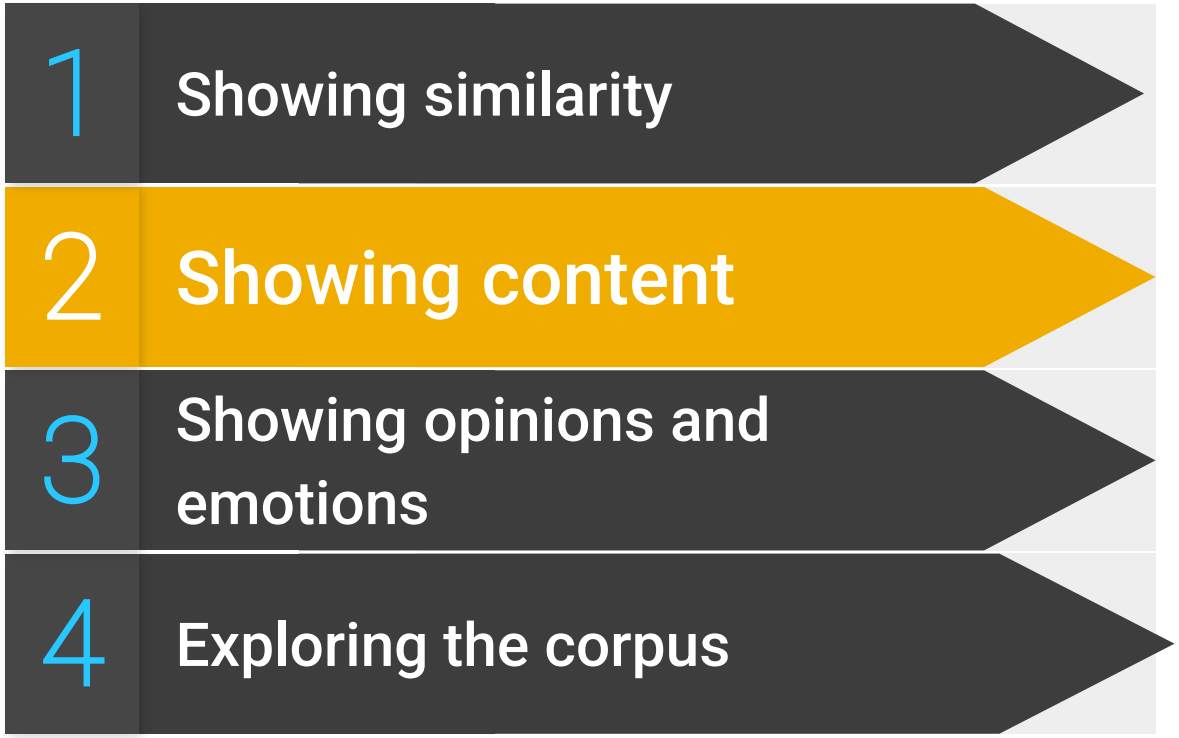
1. Showing similarity
2. Showing content
3. Showing opinions and emotions
4. Exploring the corpus

Text Visualization Techniques



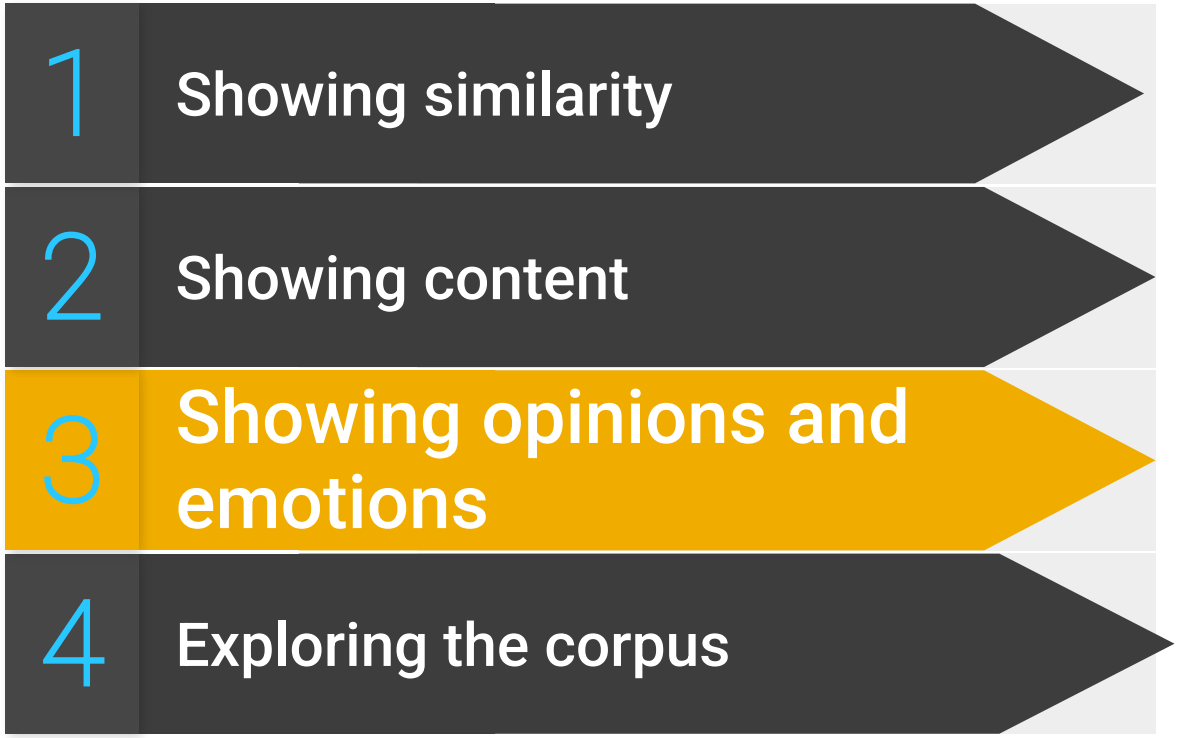
- Techniques in this category are developed to illustrate content-wised similarities of different documents.
- Various similarity measurements have been proposed based on two major types of techniques, which are projection-based and semantic-oriented.

Text Visualization Techniques



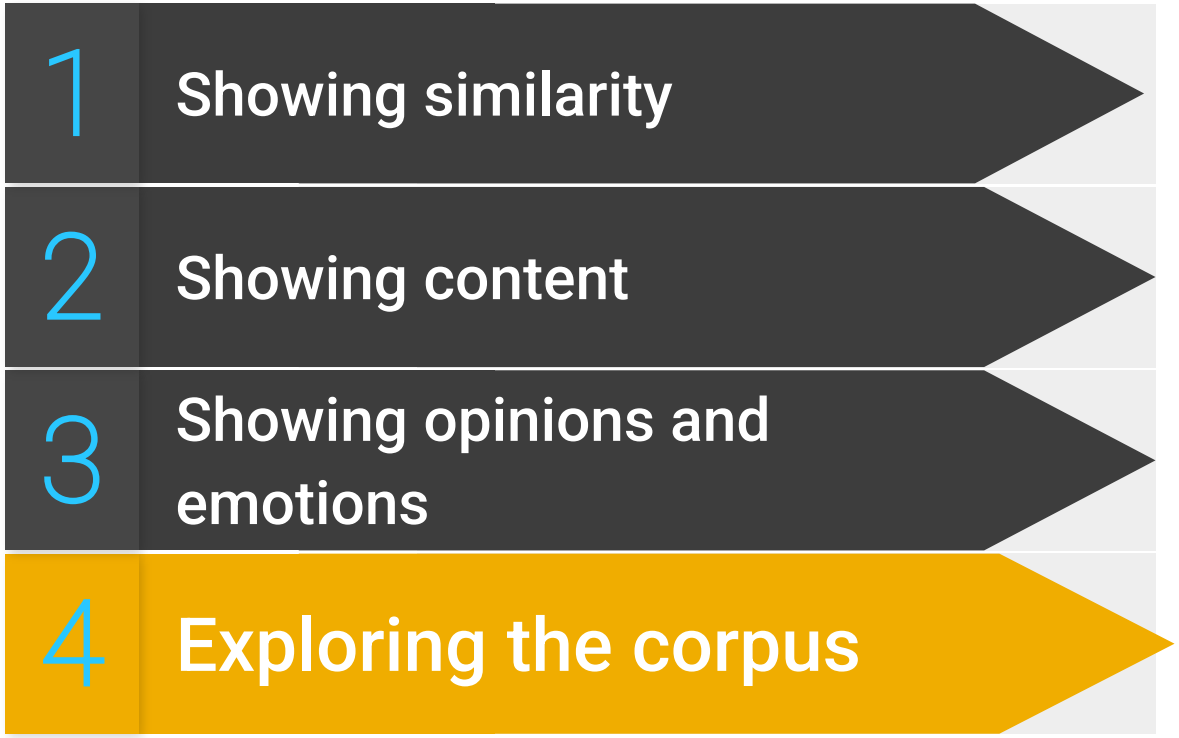
Most text visualization techniques have been proposed to illustrate different aspects of the content of text data, such as summarizing the content of a single document and showing the topics of a corpus.

Text Visualization Techniques



This category includes techniques that summarize the sentiment or emotional profiles of persons based on the text data they produced.

Text Visualization Techniques



Many text data exploration systems have been developed to help analysts or end users to efficiently explore text data.

Text Visualization Categories

Single document visualization	• Individual words and content
Document collection visualization	• Large collection - theme, concept and relations
Extended document visualization	• Comprehensive tasks, attributes beyond content

Single Document Visualization

Single Document Visualization

Vocabulary - basic unit of a document

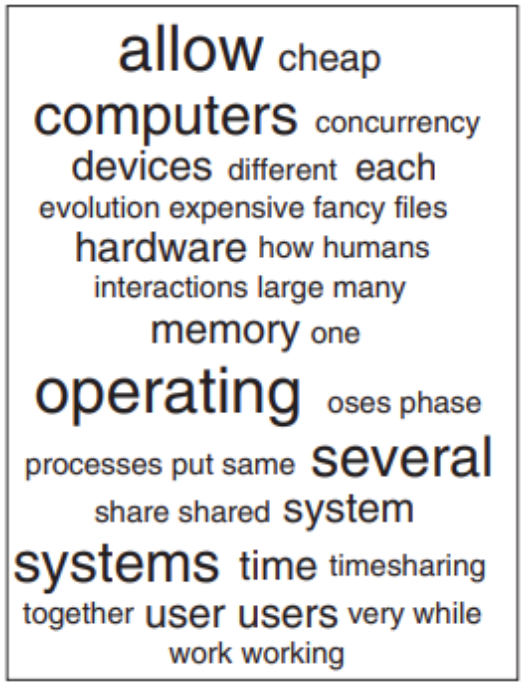
Single document visualization can be largely categorized into:

- Vocabulary-based visualization: visual representation of the document vocabulary features i.e.
  - word frequency
  - word distribution
  - lexical structure → general idea of contents and features in a document
- Document-based visualization

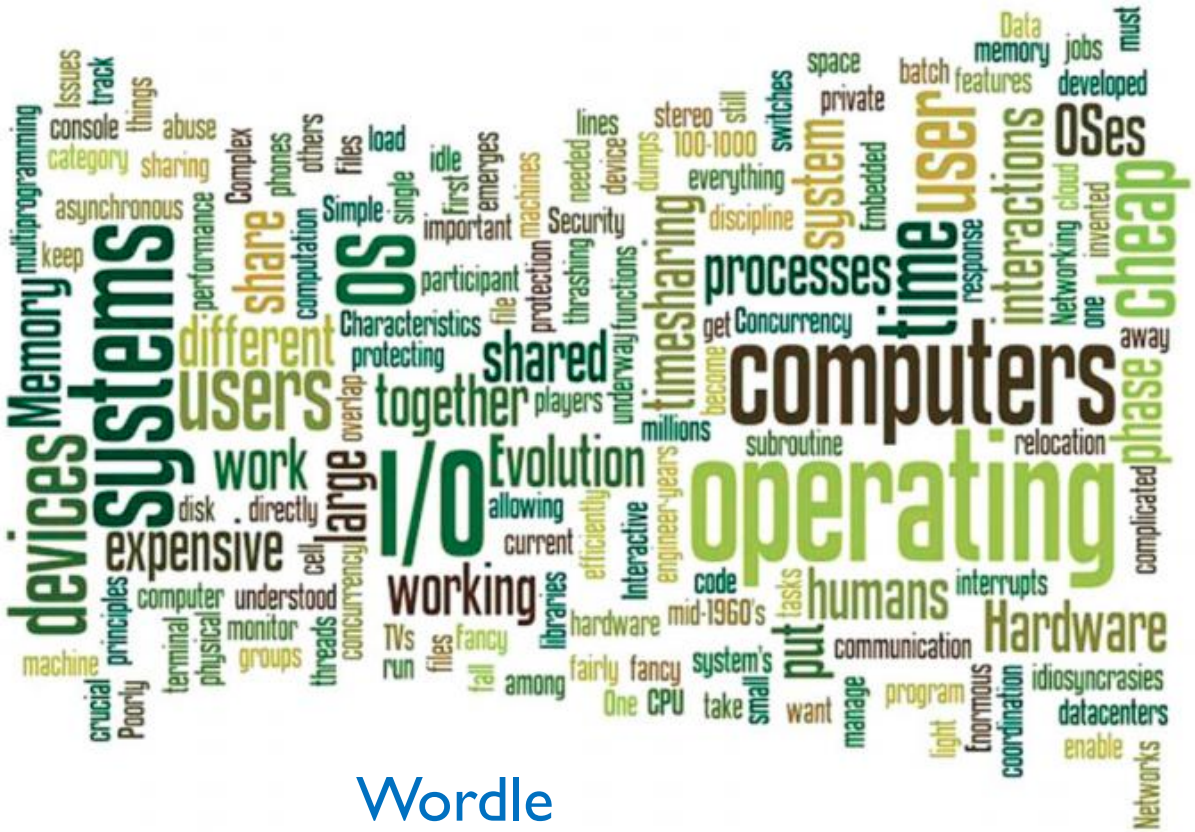
Visualization Based on Frequency

Tag Clouds & Wordle: Representative methods mainly visualizing **word frequency.**

Other methods that extend the tag/word cloud, e.g. parallel tag clouds (PTC), ManiWordle, context preserving dynamic word cloud



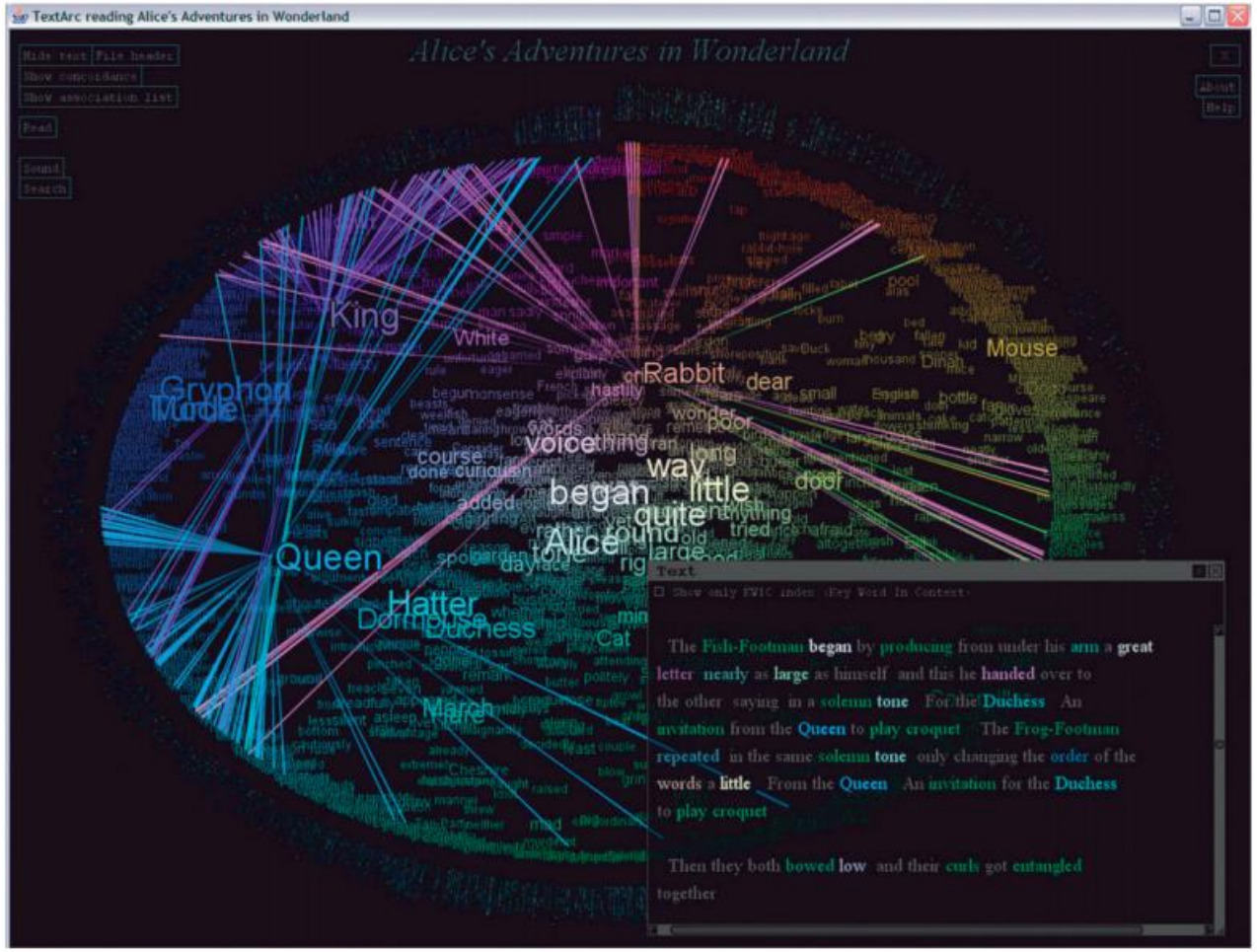
Tag Cloud



Wordle

Visualization Based on Frequency & Distribution

TextArc: Presents a view of a text that concisely **reveals the frequency and distribution of words** of an entire book



Visualization Based on Semantic Structure

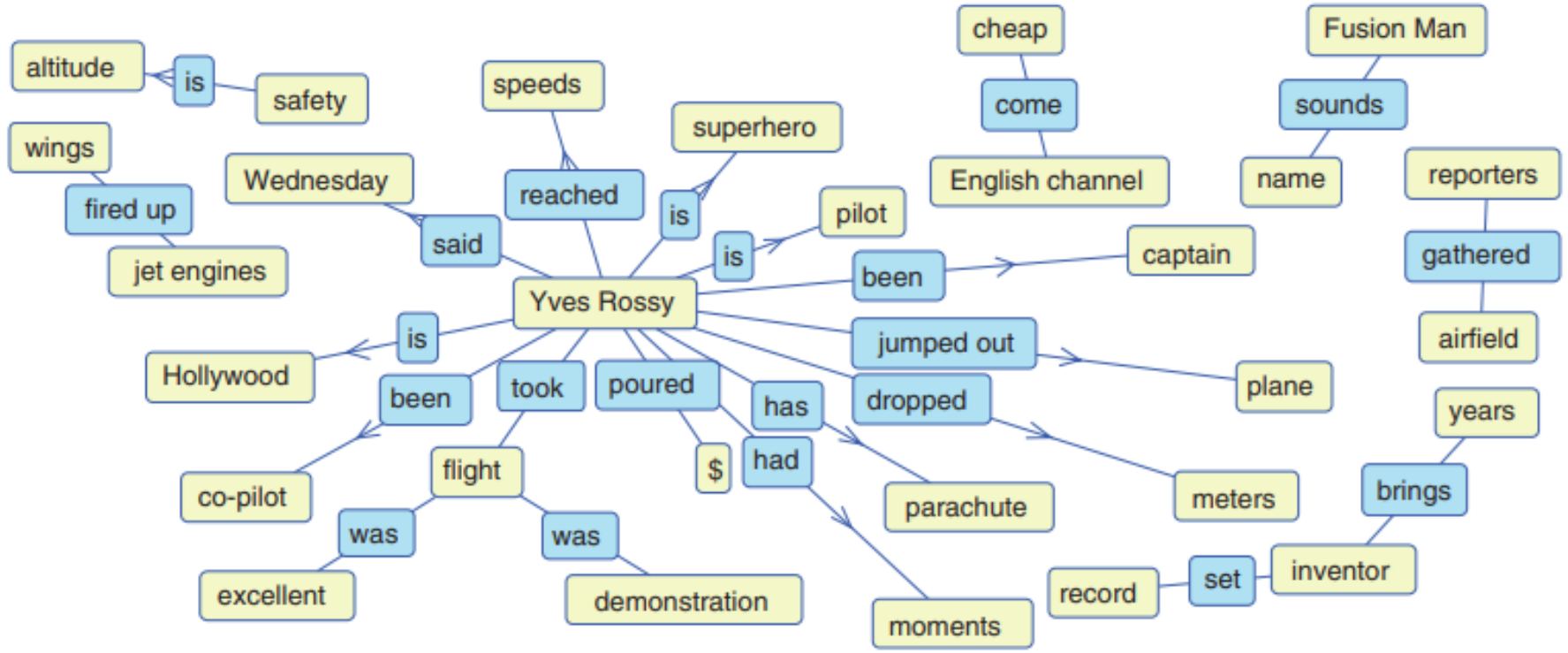
Semantic structure:

- the semantic representation associated with language
- overview of the key content of the document without entirely reading it
- use entities and their relationships to reveal the semantic content

Visualization Based on Semantic Structure

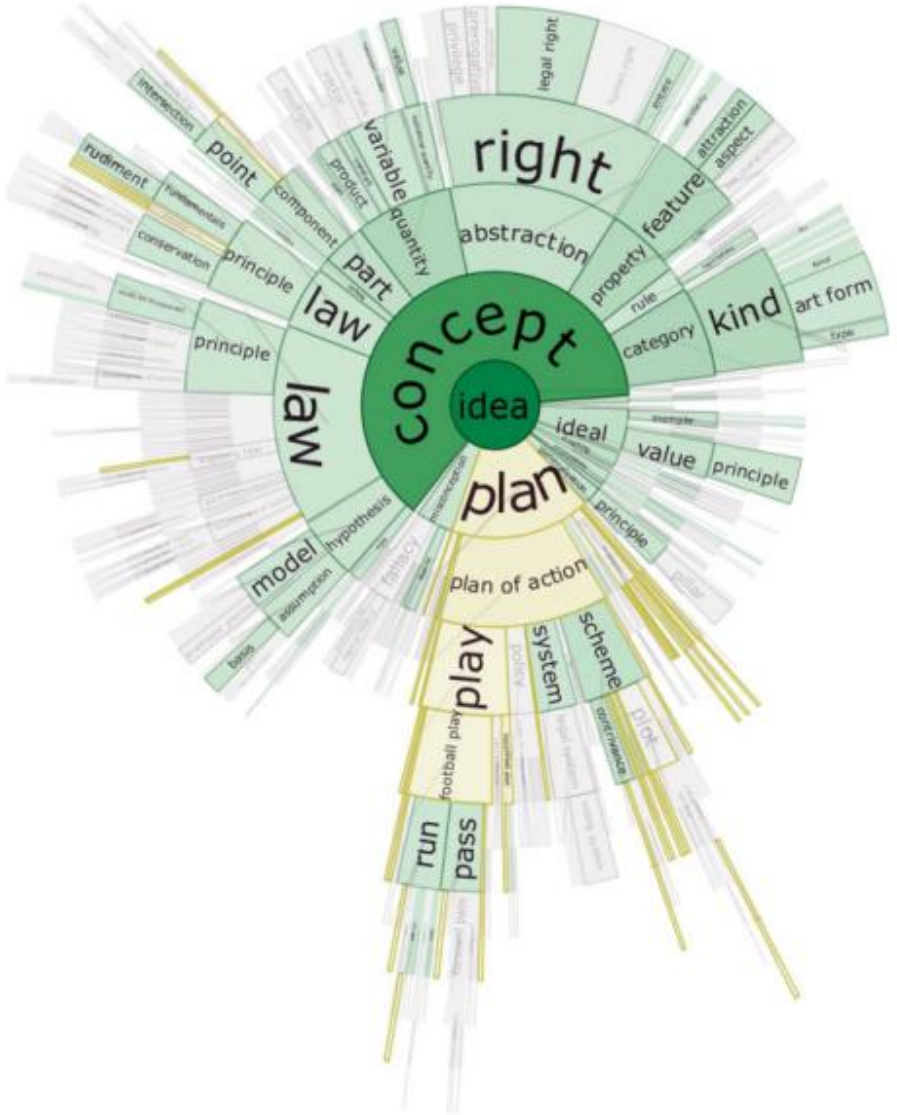
Semantic graph: visualization based on the semantic representation

1) Extracts **subject–verb–object (triplets)** for each sentence
2) Links the triplets to their corresponding entity



Visualization Based on Frequency & Semantic

DocuBurst: combines word frequency with the human-created structure in lexical databases to create a visualization that reflects both document content and semantic content.



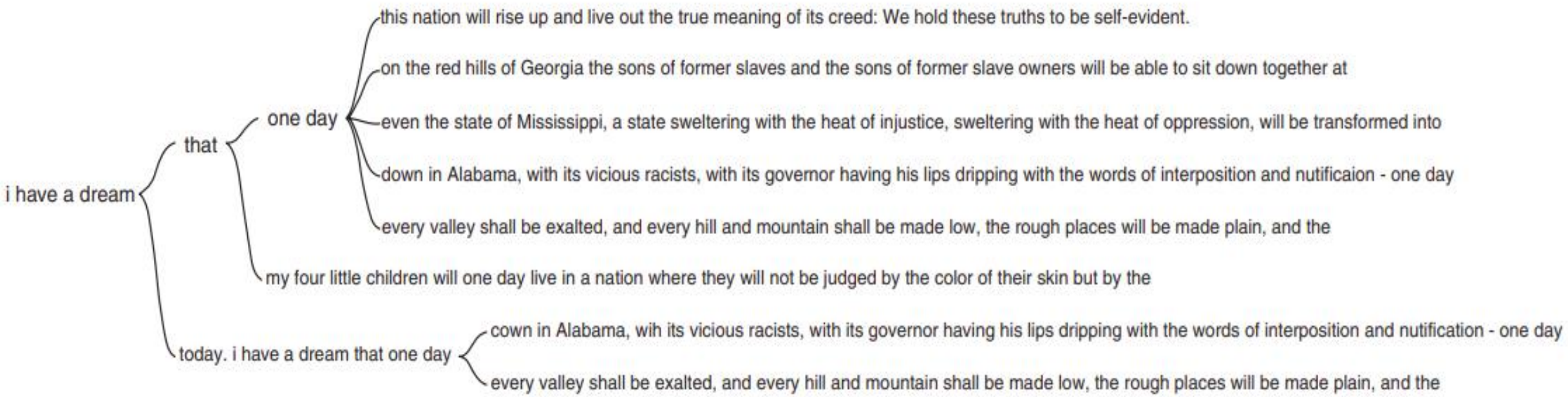
Visualization Based on Document Content

Goals:

- Search for specific words
- Obtain characteristics and relationships of the contents

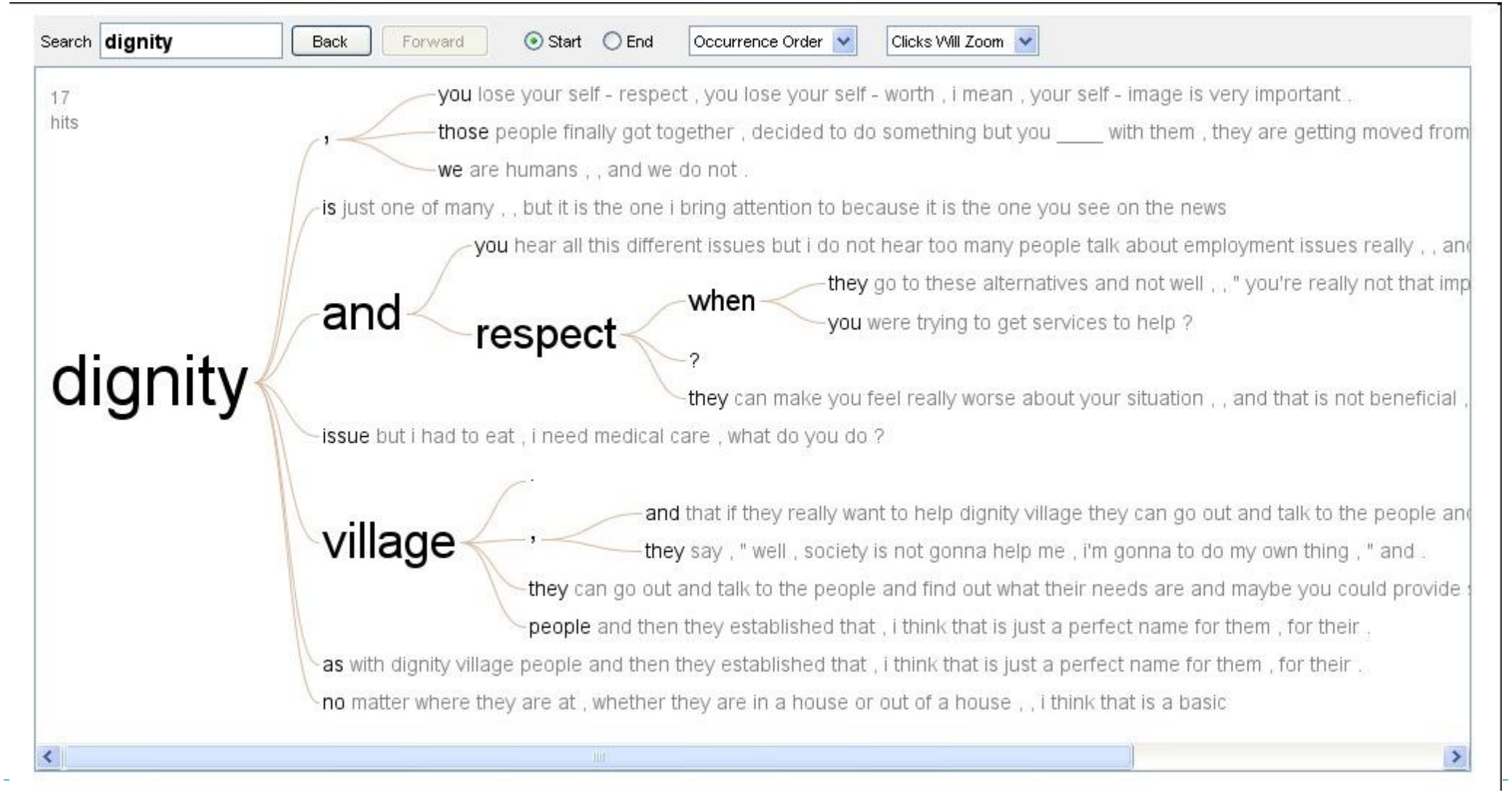
Visualization Based on Document Content

In addition to searching for specific words, it obtain the characteristics and relations of the contents in the document.



WordTree shows occurrences of 'I have a dream' in Martin Luther King's historical speech

Visualization Based on Document Content



From https://www.betterevaluation.org/en/evaluation-options/wordtree

Visualization Based on Document Content

Document Collection Visualization

Document Collection Visualization

Document Collection Visualizations are designed to represent the corpus usually focus on revealing statistics, such as topics/ themes, of the entire dataset.

It can be largely categorized into:

- Visualization by document themes
- Visualization of core content
- Visualization for changes over different version

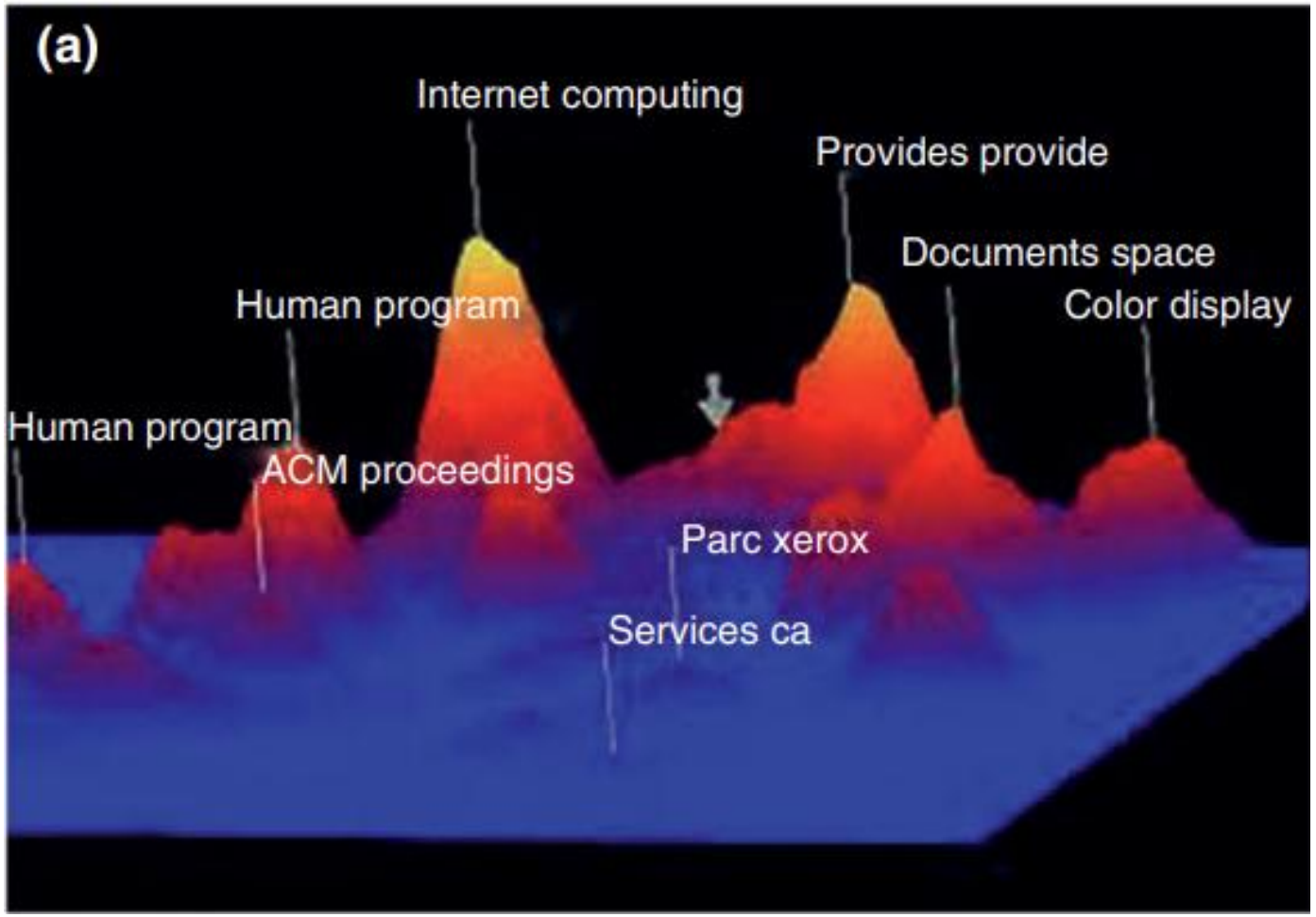
Visualization of Document Themes

Common pattern for large scale documents.

Goals:

- To discover specific topics
- To reflect the relationships among various topics

Applications: Find hot disciplines, evolutions, and trends.

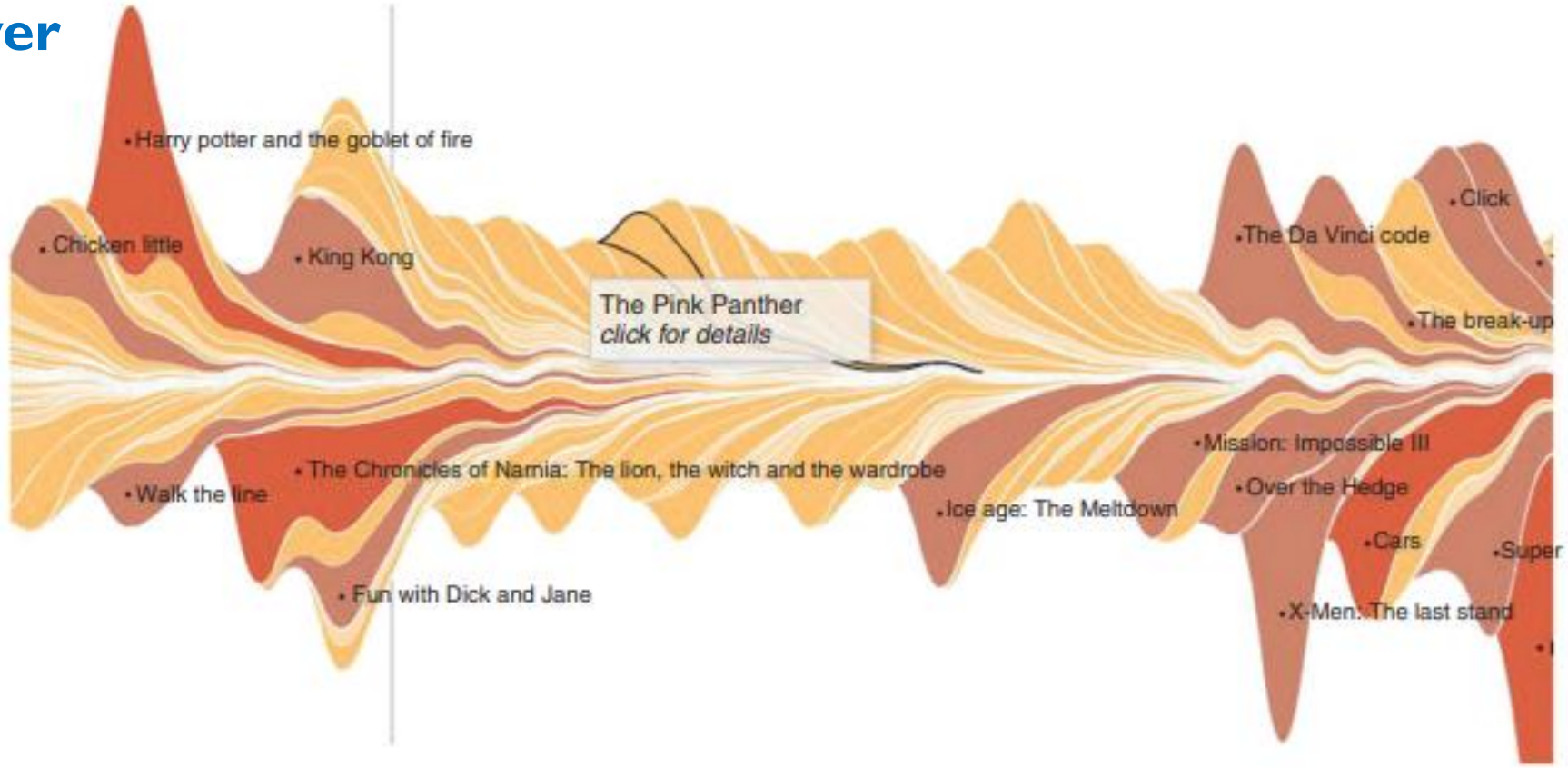


Example: **ThemeView**

Visualization of Document Themes

Example: **ThemeRiver**

generated by the box office receipts from 1986 to 2007.

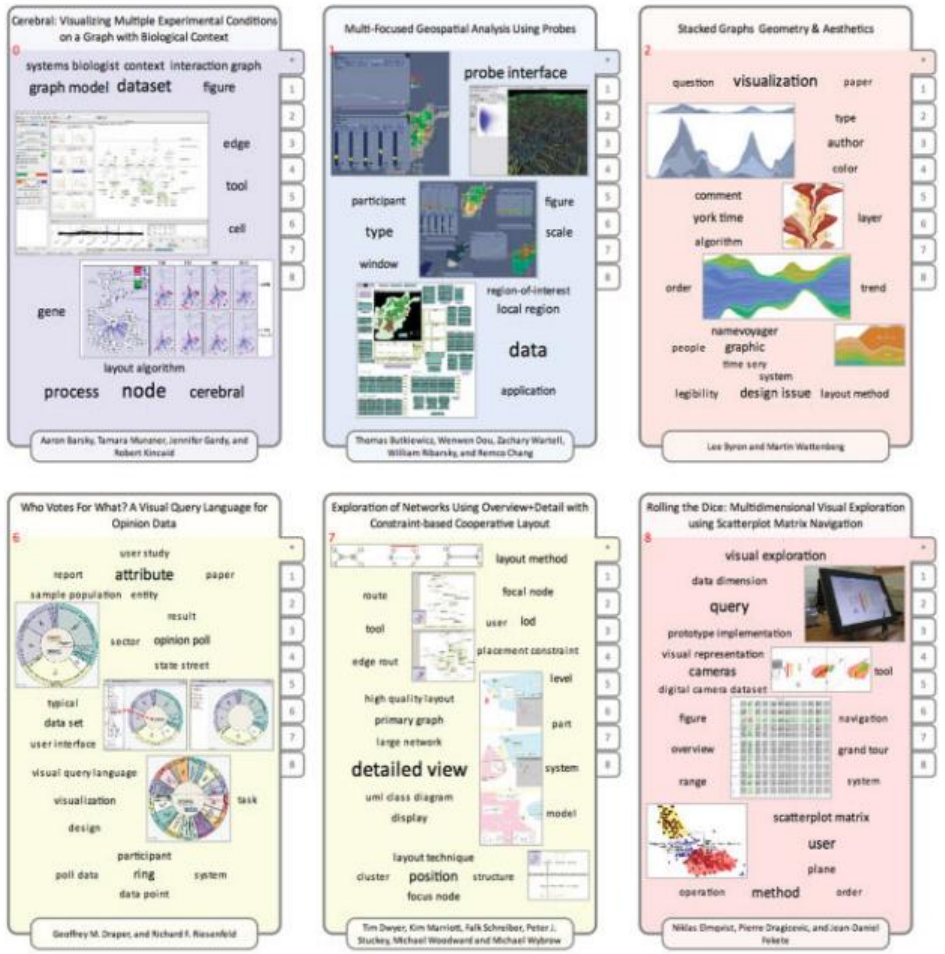


- Has greater emphasis on the time factor, focusing more on visualizing thematic variations over time within a collection of documents

Visualization of Document Core Content

Document Cards:

- Visualizes large document collections, such as paper collections and news reports, which contain both texts and images to describe facts, methods, or stories.
- Represents the document's key content as a mixture of images and important terms, similar to cards

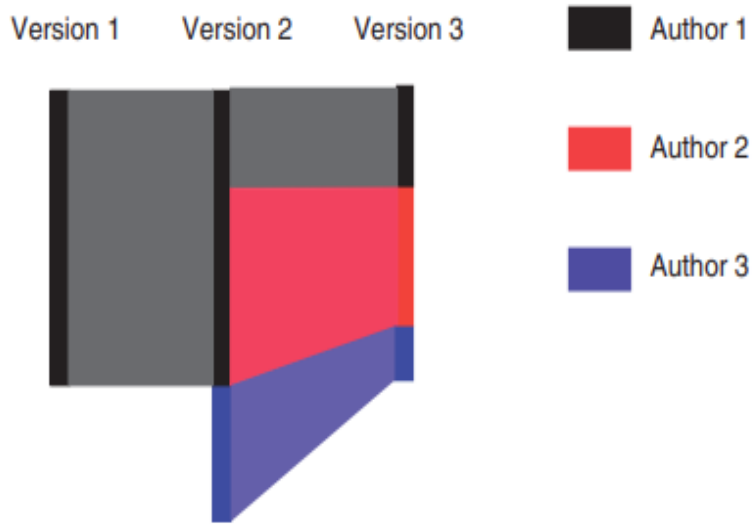


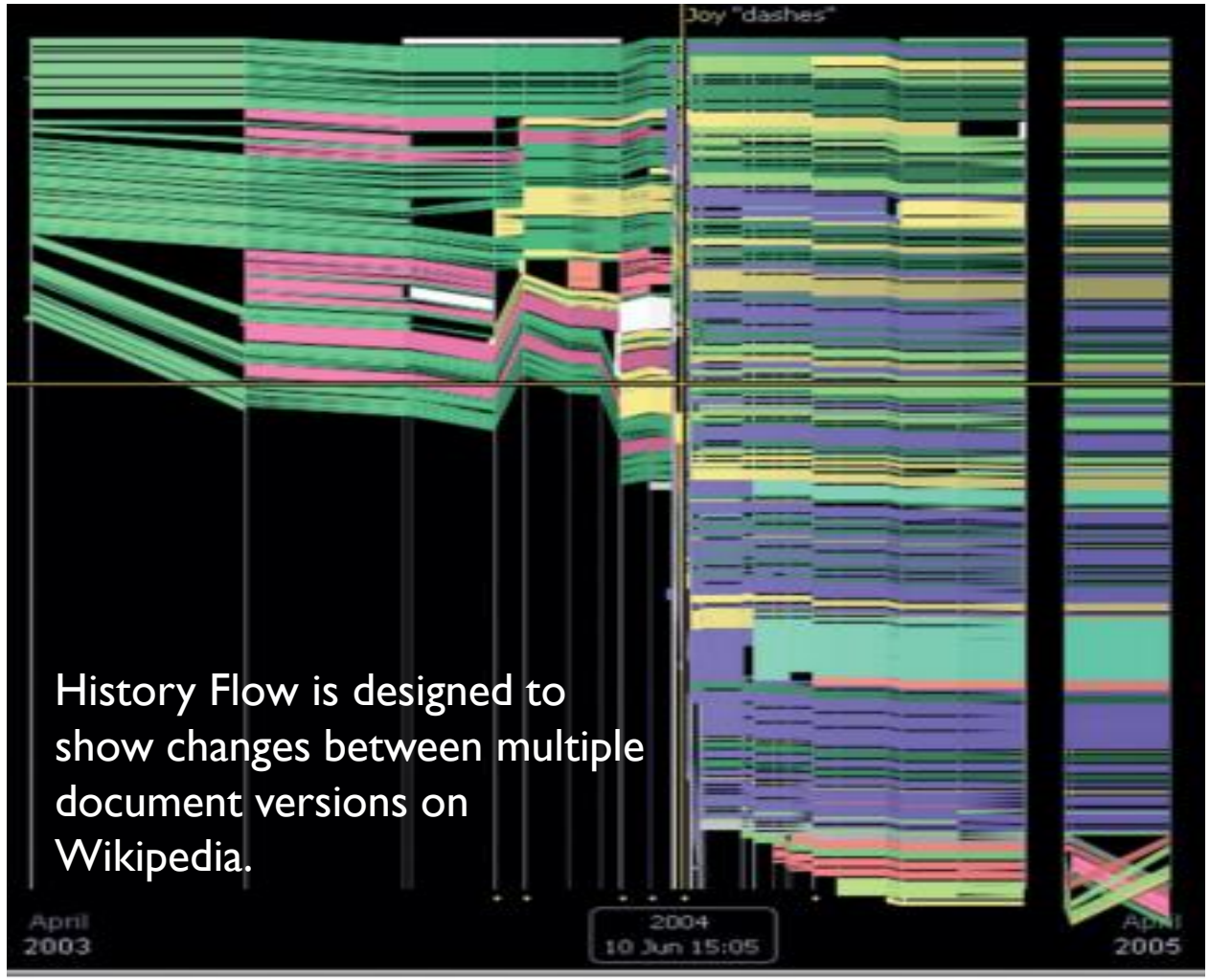
Example: The IEEE InfoVis 2008 proceeding corpus represented

Visualization of Changes over Different Versions

Goal: To visualize differences among multiple document versions over time.

Example: Visualization of changes over different versions is used to visualize differences among multiple document versions that are generated over time





History Flow is designed to show changes between multiple document versions on Wikipedia.

Extended Document Visualization

Extended Document Visualization

Deals with comprehensive tasks, involves other attributes beyond the content of documents

Goals:

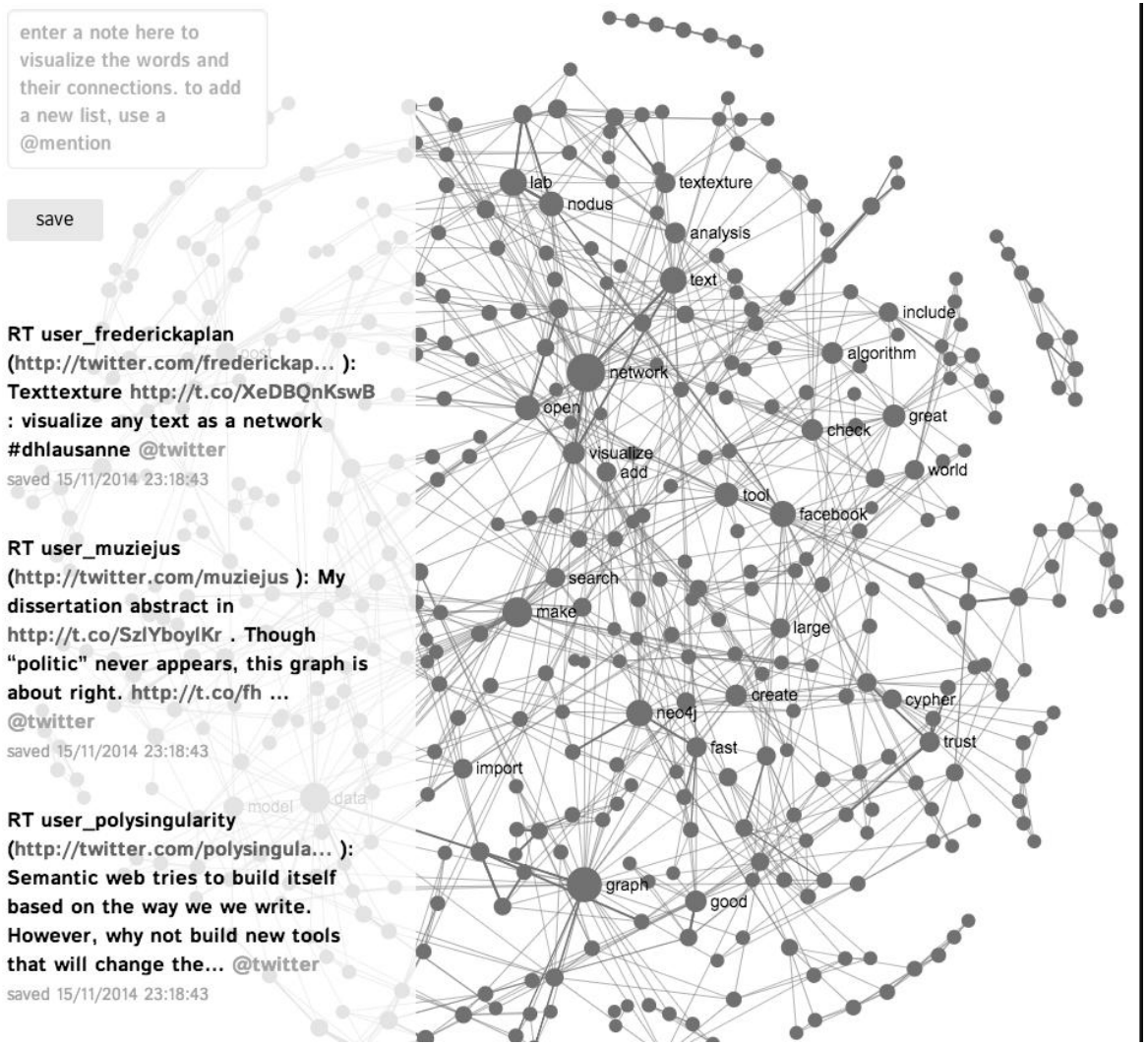
- Statistical analysis and presentation of topics or terms
- Emergence of topic events, and visualizing the text messages themselves

Applications: Always applied to specific field such as social media, such as t**ext stream** in Twitter

Extended Document Visualization

Text Network Analysis: useful tool to make sense of Twitter's ever-expanding newsfeed.

- Can be used to visualize a user's feed of tweets or visualize one's own newsfeed as a network to be able to see what the tweets are about and how they connect

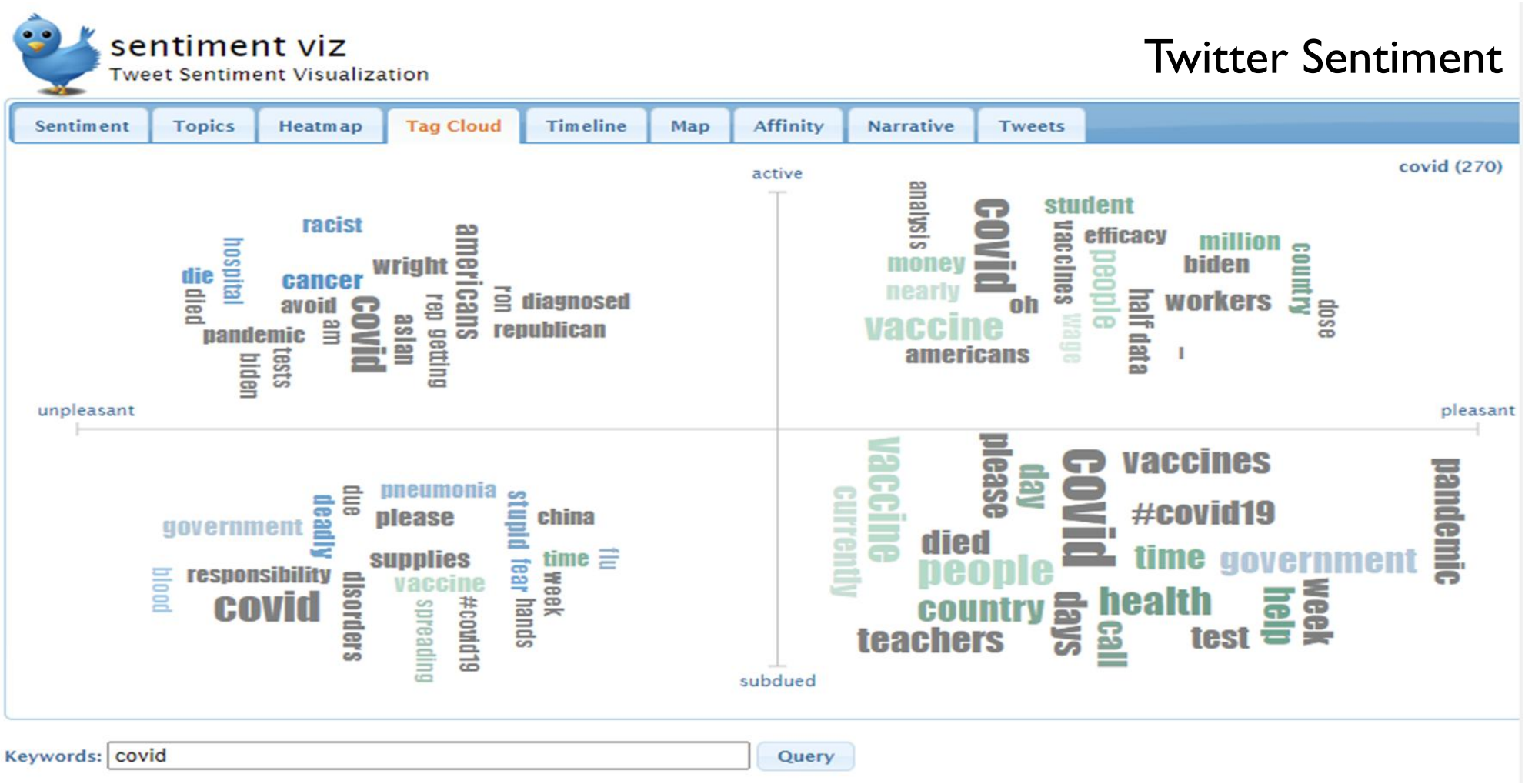


https://noduslabs.com/cases/twitter-xray-text-network-analysis/

Extended Document Visualization

https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Extended Document Visualization



https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Summary

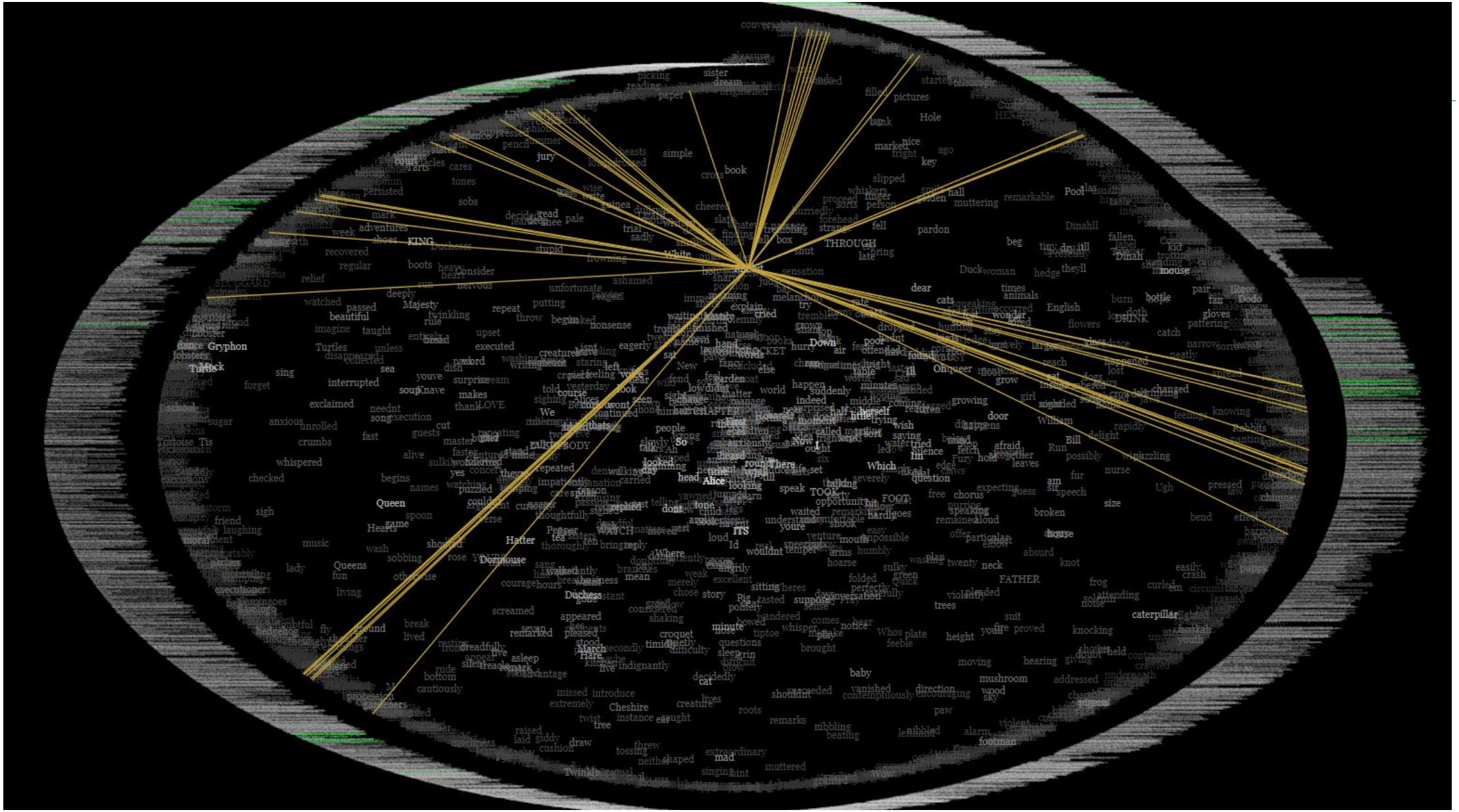
- Single Document Visualization
  - By word vocabulary: frequency, distribution, semantic
  - By document content
- Document Collection Visualization
  - By theme
  - By document core content
  - By changes over different versions
- Extended Document Visualization
  - Deal with comprehensive tasks → attributes beyond the content

Interesting Examples

30 Years of American Anxieties

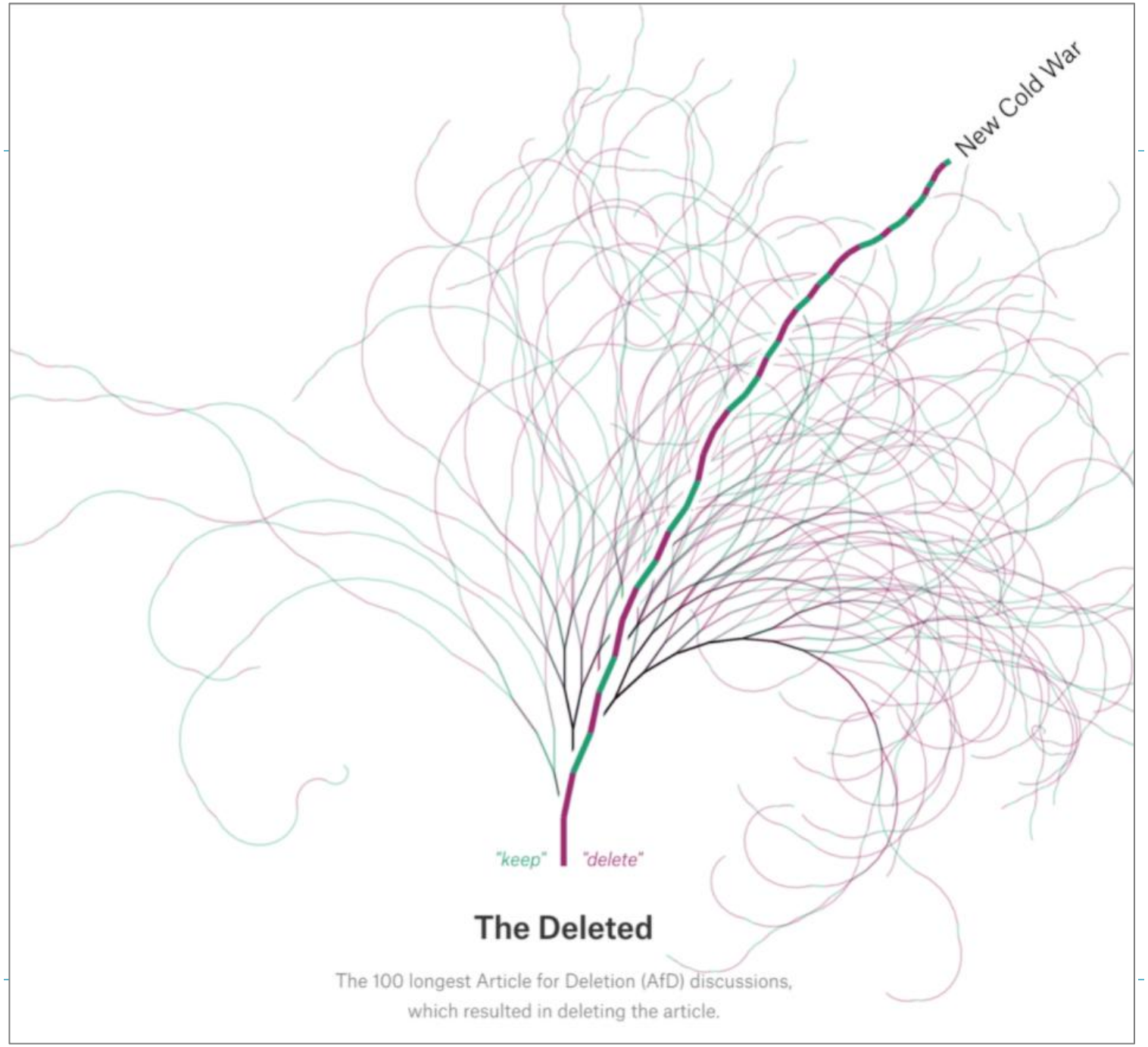
What 20,000 letters to an advice columnist tell us about what—and who—concerns us most.

- https://pudding.cool/2018/11/dearabby/



- http://vallandingham.me/textarc/

- https://fathom.info/traces/



- http://notabilia.net/

References

Gan, Q., Zhu, M., Li, M., Liang, T., Cao, Y., & Zhou, B. (2014). Document visualization: an overview of current research. Wiley Interdisciplinary Reviews: Computational Statistics, 6, 19-36.