

Paul Rosen

paul.rosen@utah.edu
@paulrosenphd
<https://cspaul.com>

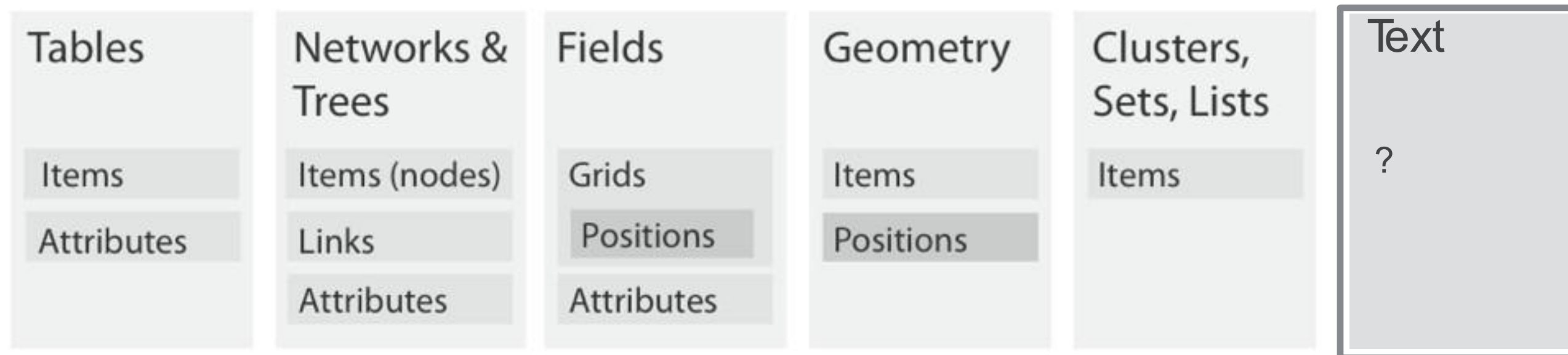


Visualization for Data Science

DS-4630 / CS-5630 / CS-6630

VISUALIZING TEXT

what does it mean to be an “item”?



text data type

- no numbers (implicitly)
- characters (ASCII)
- strings

USASCII code chart

					0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1			
					0	1	2	3	4	5	6	7			
b ₇	b ₆	b ₅	b ₄	b ₃	b ₂	b ₁	b ₀	Column							
								Row							
0	0	0	0	0	0	0	0	NUL	DLE	SP	0	@	P	\	p
0	0	0	0	1	0	0	0	SOH	DC1	!	1	A	Q	a	q
0	0	0	0	1	0	1	0	STX	DC2	"	2	B	R	b	r
0	0	0	0	1	0	1	1	ETX	DC3	#	3	C	S	c	s
0	0	0	1	0	0	0	0	EOT	DC4	\$	4	D	T	d	t
0	0	0	1	0	0	1	0	ENQ	NAK	%	5	E	U	e	u
0	0	0	1	0	0	1	1	ACK	SYN	&	6	F	V	f	v
0	0	0	1	1	0	0	0	BEL	ETB	'	7	G	W	g	w
0	0	0	1	1	0	0	1	BS	CAN	(	8	H	X	h	x
0	0	0	1	1	0	1	0	HT	EM	)	9	I	Y	i	y
0	0	0	1	1	0	1	1	LF	SUB	*	:	J	Z	j	z
0	0	1	0	0	0	0	0	VT	ESC	+	;	K	[	k	{
0	0	1	0	0	0	0	1	FF	FS	,	<	L	\	l	
0	0	1	0	0	0	1	1	CR	GS	-	=	M	]	m	}
0	0	1	0	1	0	0	0	SO	RS	.	>	N	^	n	~
0	0	1	0	1	0	1	1	SI	US	/	?	O	_	o	DEL

text data type

- no numbers (implicitly)
- characters (ASCII)
- strings

USASCII code chart

					0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1			
					0	1	2	3	4	5	6	7			
b ₇	b ₆	b ₅	b ₄	b ₃	b ₂	b ₁	b ₀	Column							
								Row							
0	0	0	0	0	0	0	0	NUL	DLE	SP	0	@	P	\	p
0	0	0	0	1	0	0	0	SOH	DC1	!	1	A	Q	a	q
0	0	0	1	0	0	0	0	STX	DC2	"	2	B	R	b	r
0	0	1	1	0	0	0	0	ETX	DC3	#	3	C	S	c	s
0	1	0	0	0	0	0	0	EOT	DC4	\$	4	D	T	d	t
0	1	0	0	1	0	0	0	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	0	0	0	0	ACK	SYN	&	6	F	V	f	v
0	1	1	0	1	0	0	0	BEL	ETB	'	7	G	W	g	w
1	0	0	0	0	0	0	0	BS	CAN	(	8	H	X	h	x
1	0	0	0	1	0	0	0	HT	EM	)	9	I	Y	i	y
1	0	1	0	0	0	0	0	LF	SUB	*	:	J	Z	j	z
1	0	1	0	1	0	0	0	VT	ESC	+	;	K	[	k	{
1	1	0	0	0	0	0	0	FF	FS	,	<	L	\	l	
1	1	0	0	1	0	0	0	CR	GS	-	=	M	]	m	}
1	1	1	0	0	0	0	0	SO	RS	.	>	N	^	n	~
1	1	1	0	1	0	0	0								DEL

[6:10]

0	1	2	3	4	5	6	7	8	9	10	11
M	o	n	t	y		P	y	t	h	o	n

-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
-----	-----	-----	----	----	----	----	----	----	----	----	----

[-12:-7]

text data semantics

- words
- lines
- sentences
- paragraphs
- chapters

love
I
.

I love visualization.

text data semantics

- documents: books, papers, webpages, e-mails, twitter post
- corpus: collection of documents

Semantics-Based Analysis and Navigation of Heterogeneous Text Corpora: The Porpoise News and Blogs Engine

Bertjan Bovenot and Daniel Trappier

Department of Computer Science, R.U. Leuven, B-3000 Leuven, Belgium
Institute of Information Science, Ghent University, Ghent, Belgium
E-mail: bertjan.bovenot@kuleuven-leuven.be, daniel.trappier@kuleuven-leuven.be

Abstract. While information retrieval systems are more powerful and search engines offer up-to-date keyword search to help people grasp and identify relevant sources of information. However, often search engines are limited to local and mostly semantic queries (like date or source identity) or semantic criteria that are opaque ("semantic documents"). Thus, these interfaces do not support the economic grouping and navigation of different information sources that is needed to inform one's decision.

In this paper, we describe Porpoise, a system that provides users with a flexible way to group and navigate heterogeneous text corpora. Porpoise is able to analyze, cluster, and visualize heterogeneous text corpora and help to navigate them.

Keywords: text mining, information retrieval, semantic search, navigation

1. Introduction

Many information retrieval systems are based on keyword search. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1]. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1]. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1].

2. Related Work

Many information retrieval systems are based on keyword search. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1]. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1].

3. Conclusion

Many information retrieval systems are based on keyword search. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1]. These systems are often called "keyword search engines" as in Google, Bing, and Yahoo! [1].

References

[1] Google, Bing, and Yahoo! [1].

Chapter 3 Automatic Mapping of Social Networks of Actors from Text Corpora: Time Series Analysis

James A. Baskerville and Scott E. Fagan

Abstract. In this paper, we present a method for automatically mapping social networks of actors from text corpora. The method is based on the analysis of the text corpora and the mapping of the actors to the nodes of the network. The method is based on the analysis of the text corpora and the mapping of the actors to the nodes of the network.

Automatic Analysis of Large Text Corpora - A Contribution to Structuring Web Content

Gordon Hogg, Steve Goodwin, and Christian Wolff

University of York, Department of Computing Science, York, UK
g.hogg@york.ac.uk, s.goodwin@york.ac.uk, c.wolff@york.ac.uk

Abstract. This paper describes a corpus linguistic analysis of large text corpora based on a collection of the text corpora. The analysis is based on the collection of the text corpora and the mapping of the actors to the nodes of the network.

Examining the Use of Text Corpora and Online Dictionaries as Learning Tools: Pre-Service Teachers' Perspectives

Nikolaos Katsaros, George Katsaros, and George Katsaros

Abstract. This paper examines the use of text corpora and online dictionaries as learning tools by pre-service teachers. The study is based on the analysis of the text corpora and the mapping of the actors to the nodes of the network.

Constructing a Large Scale Text Corpus Based on the Grid and Trustworthiness

Yong Li, Qianqian Zou, Peng Qian, and Geoffrey C. Fox

School of Computer Science & Technology, Tsinghua University, Beijing, China
liyong@sem.tsinghua.edu.cn, zouqianqian@sem.tsinghua.edu.cn, qianpeng@sem.tsinghua.edu.cn, foxgc@sem.tsinghua.edu.cn

Abstract. The construction of a large scale corpus is a hard task. A novel approach is proposed to automatically build a large scale text corpus with low cost and shortening period based on the trustworthiness. A novel approach is proposed to automatically build a large scale text corpus with low cost and shortening period based on the trustworthiness.

Search for Text Corpora with GSP

Thomas H. Davenport

University of California, Berkeley, CA, USA
davenport@eecs.berkeley.edu

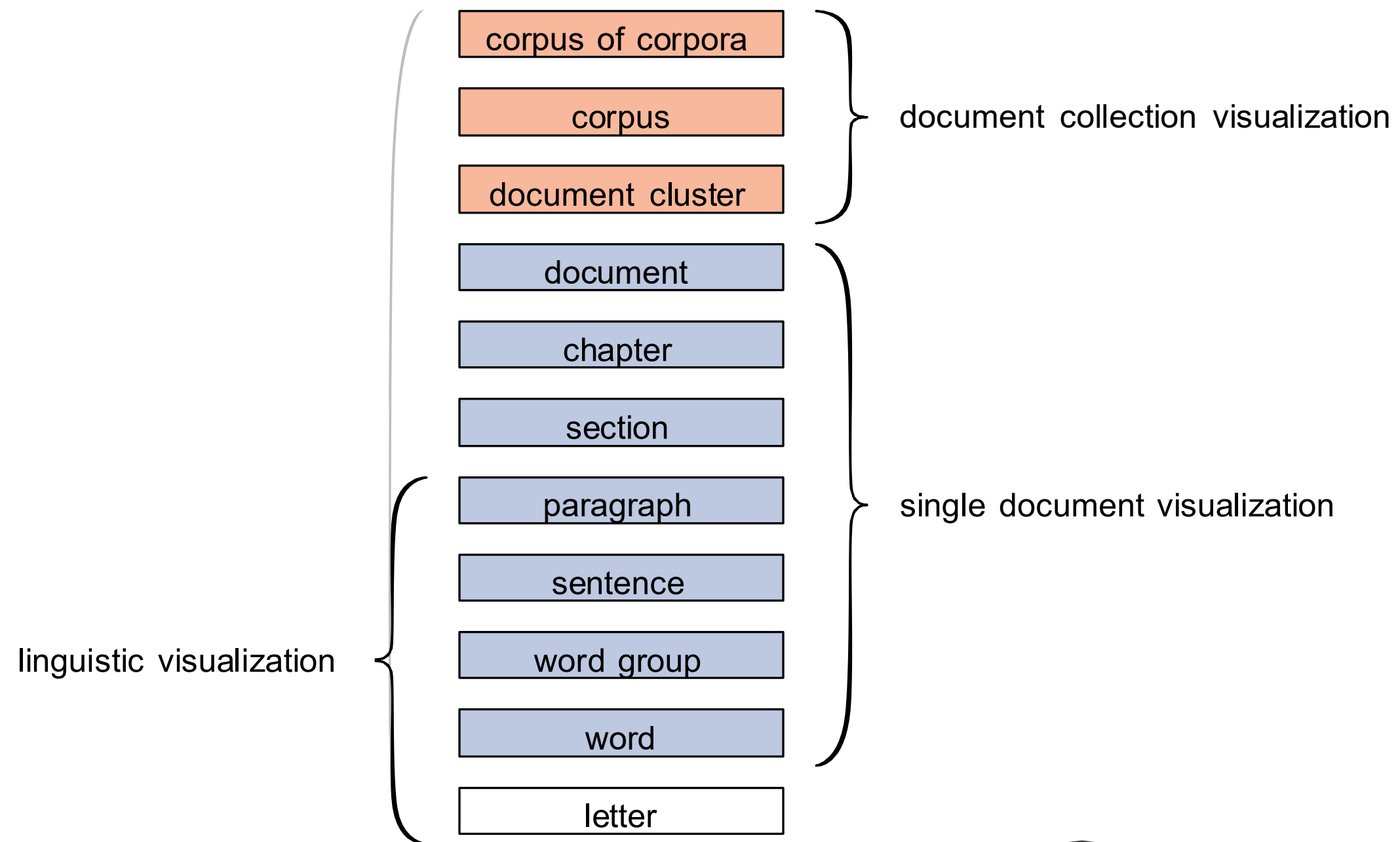
Abstract. This paper describes a method for searching for text corpora. The method is based on the analysis of the text corpora and the mapping of the actors to the nodes of the network.

AIML Knowledge Base Construction from Text Corpora

Giuseppe De Gregorio, Roberto Di Stefano, and Vito Portale

Abstract. This paper describes a method for constructing a knowledge base from text corpora. The method is based on the analysis of the text corpora and the mapping of the actors to the nodes of the network.

Text Units Hierarchy



single document

Tag Cloud / Word cloud

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

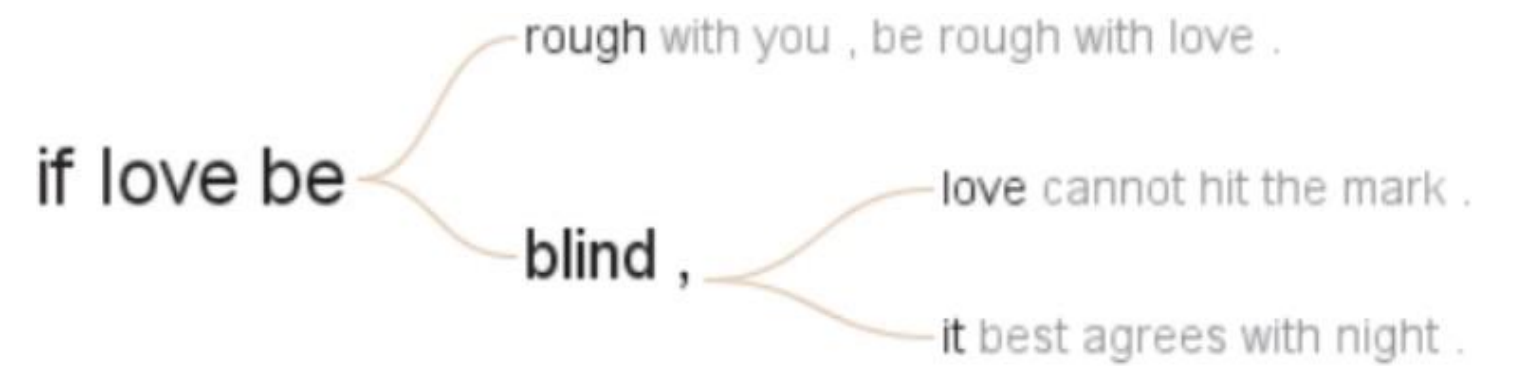


Word Tree

Text

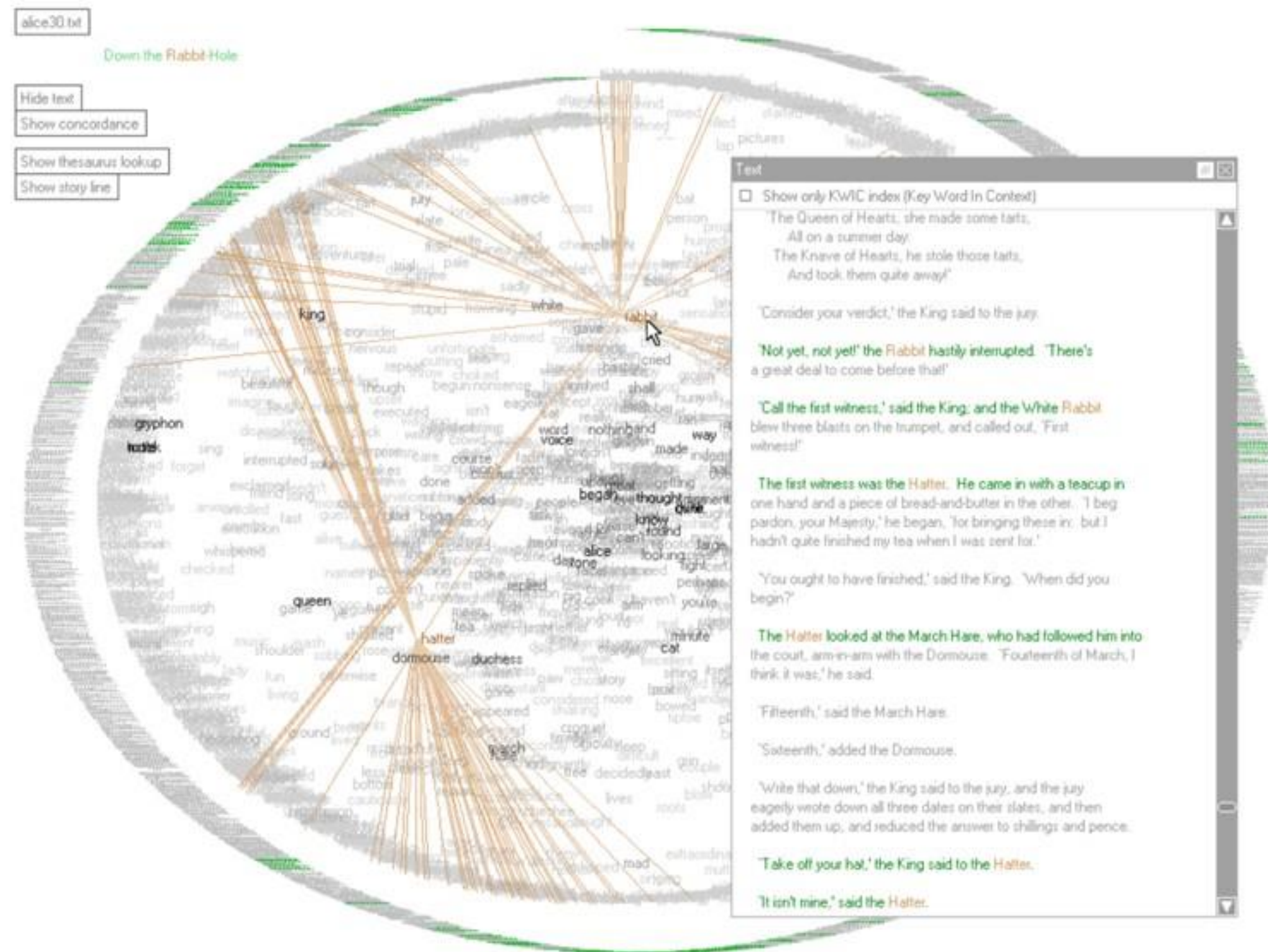
if love be rough with you , be rough with love .
if love be blind , love cannot hit the mark .
if love be blind , it best agrees with night .

WordTree

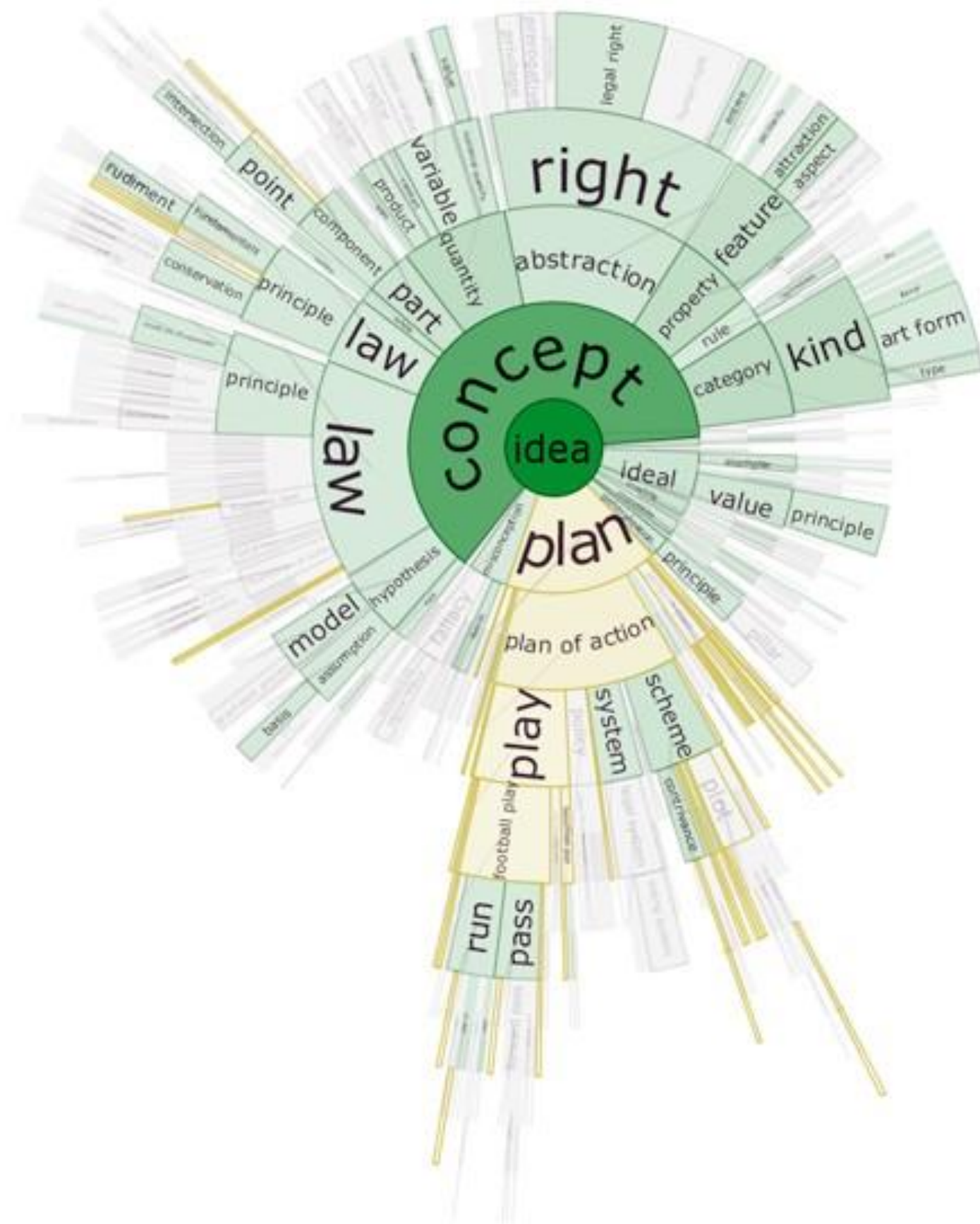


[Wattenberg 2008]

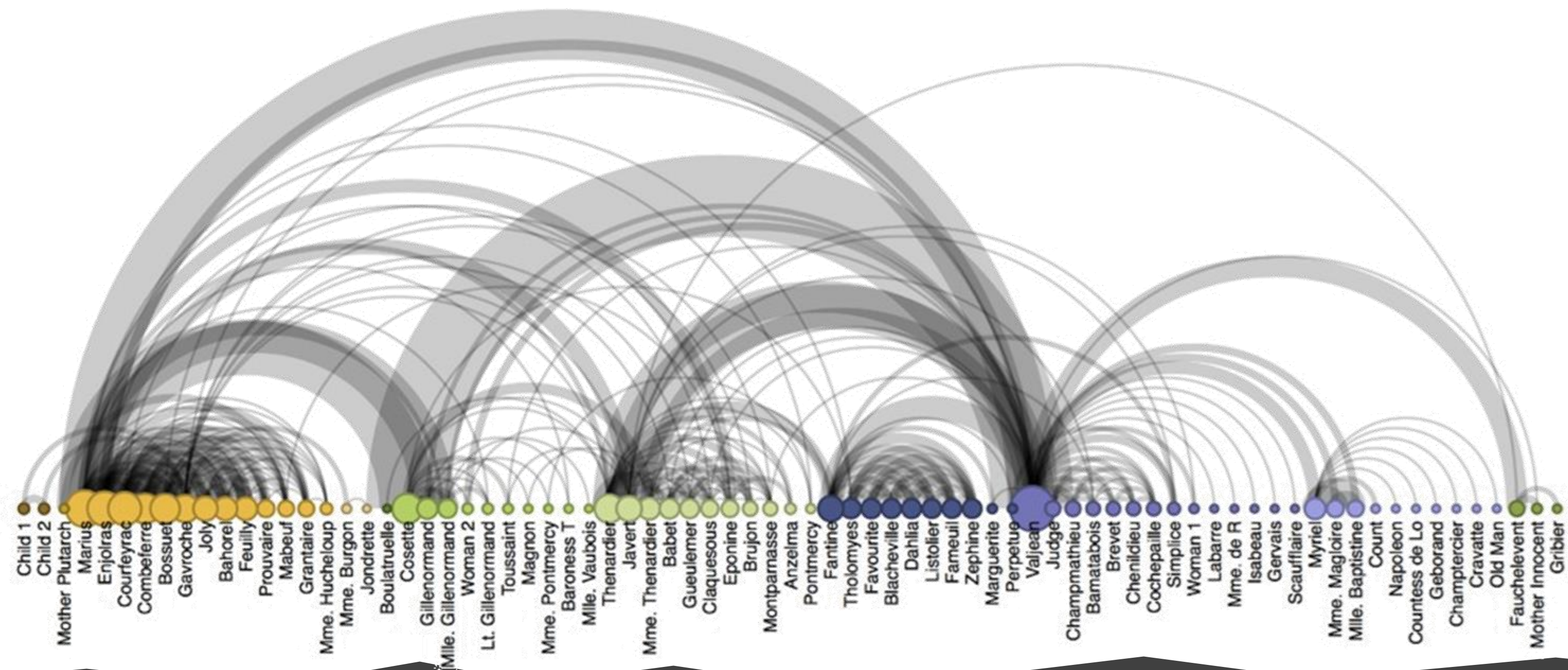
Text Arc



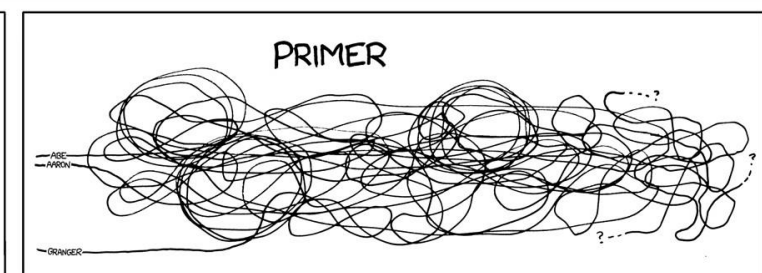
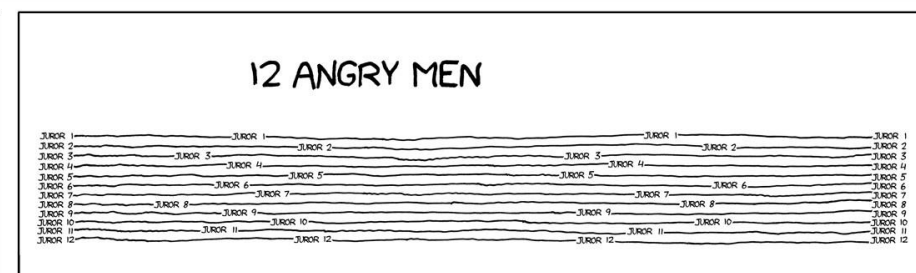
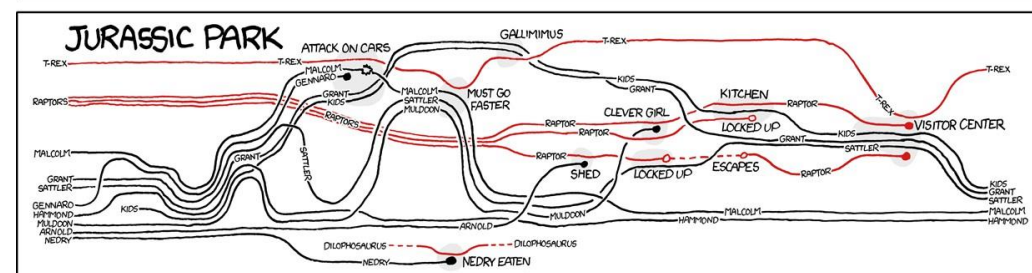
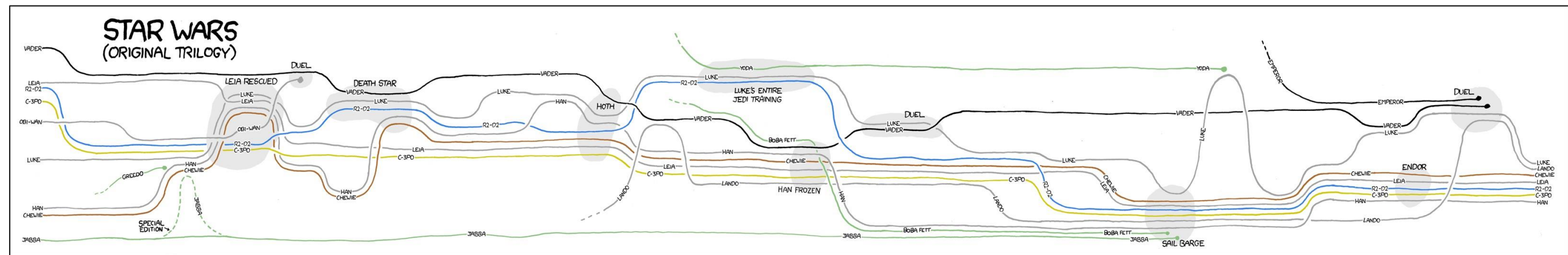
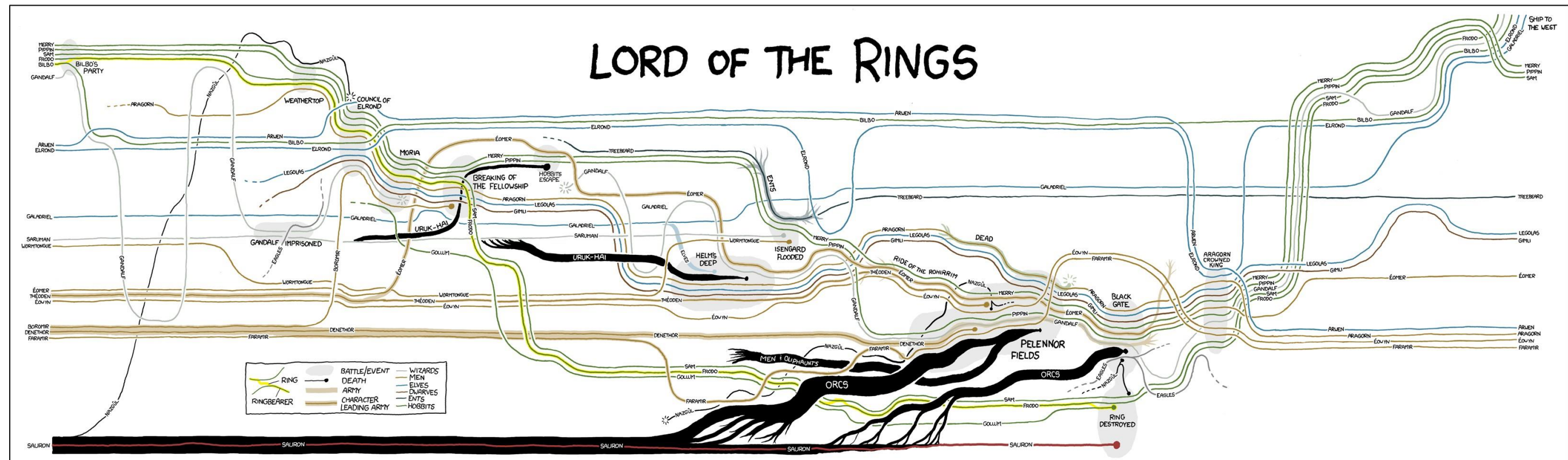
DocuBurst



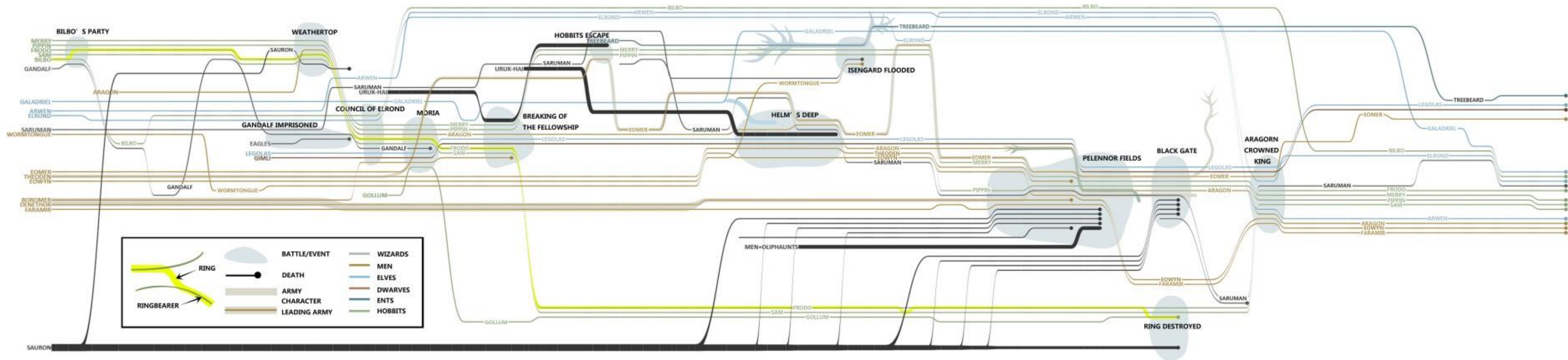
Arc Diagram



THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS.
THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE
LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.



StoryFlow: Tracking the Evolution of Stories

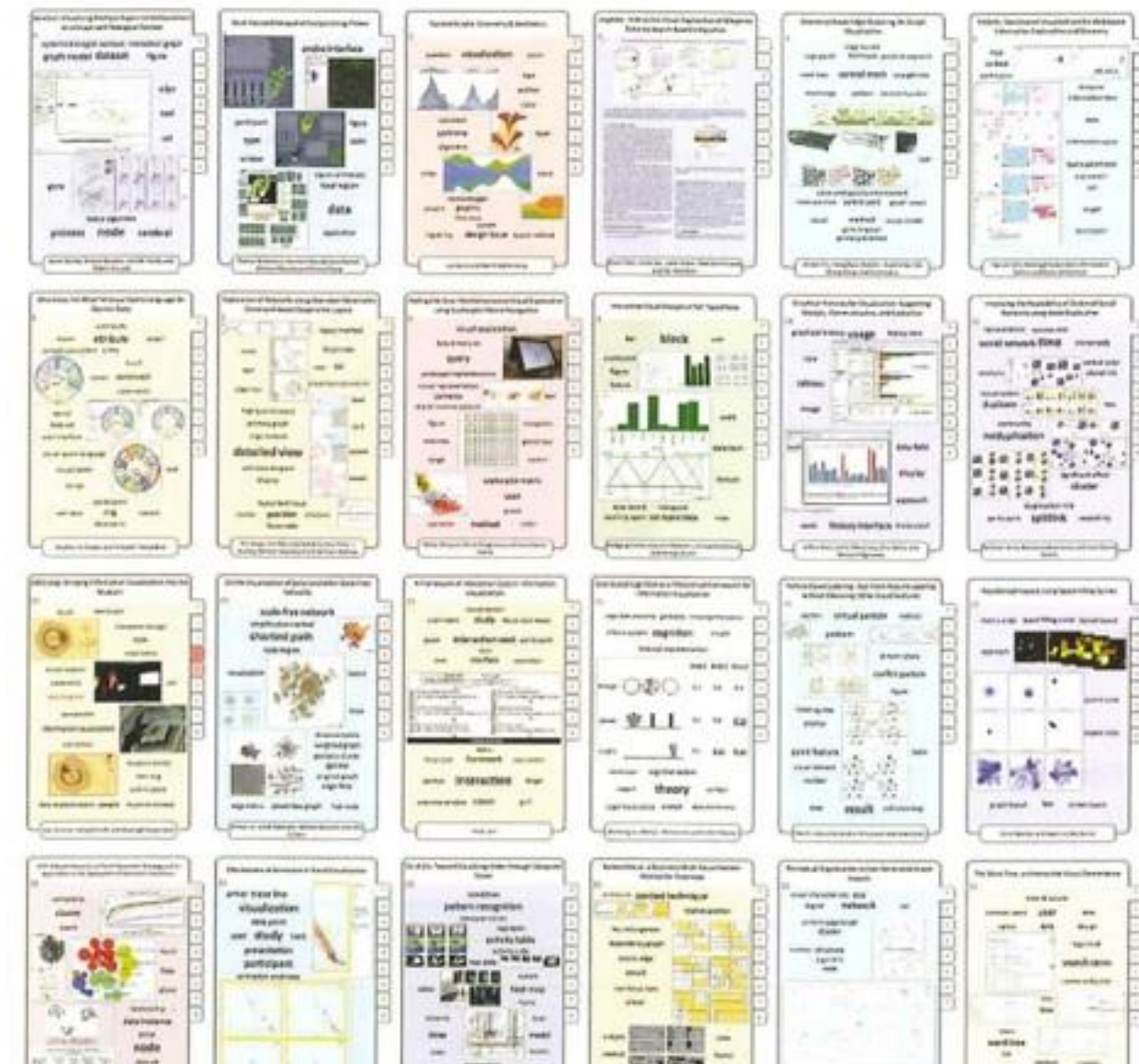


collection of documents

Text Corpora

- Varied Goals:
 - Discover interesting documents Summarize Documents Classify Documents
 - Extract Facts (Intelligence Analysis)
- Rich Information:
 - Document Metadata Authors, date, type,
 - Paragraphs, figures...
 - Revisions, annotations, comments,

Document Cards (small multiples)



MDS Approaches

- use bag-of-words to project documents w.r.t. text similarity into a landscape
- (only) one example

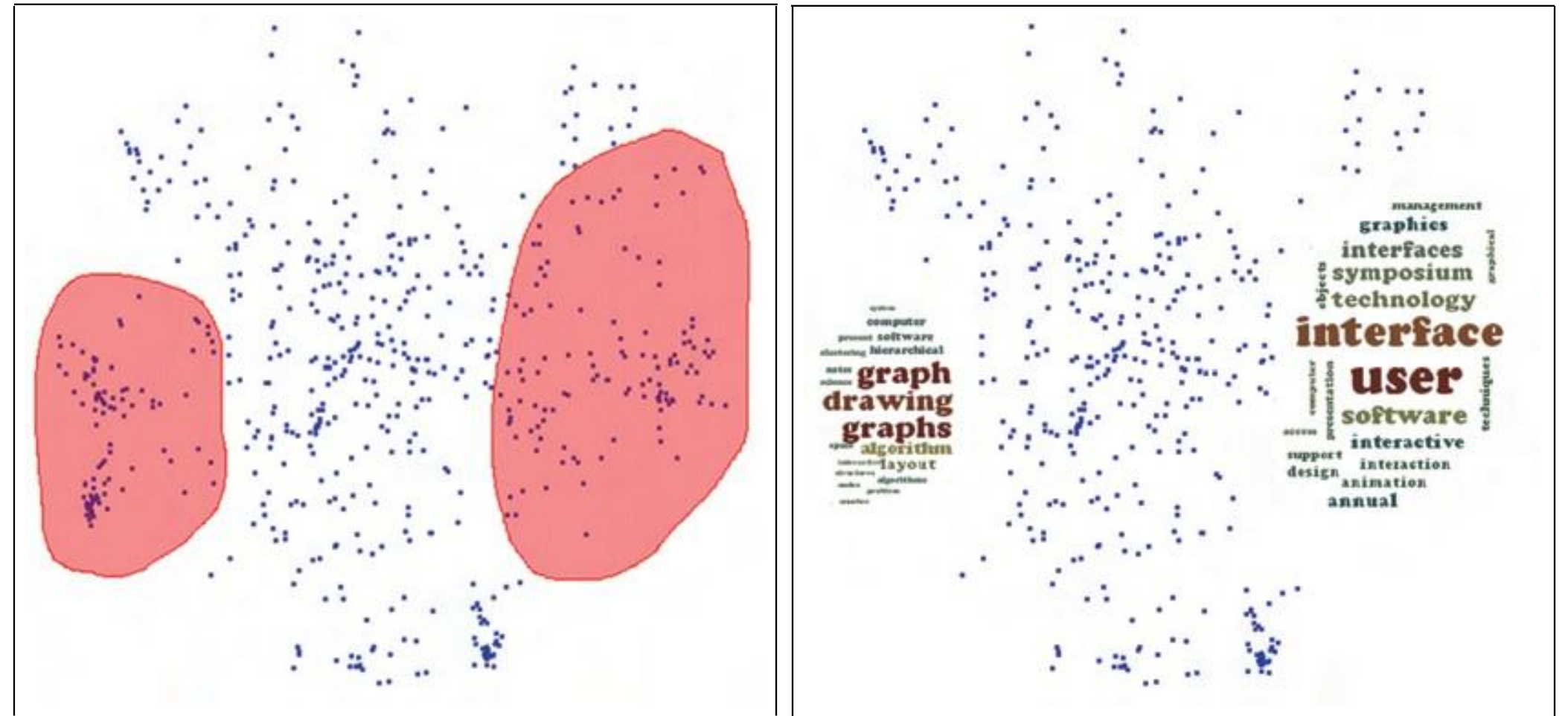
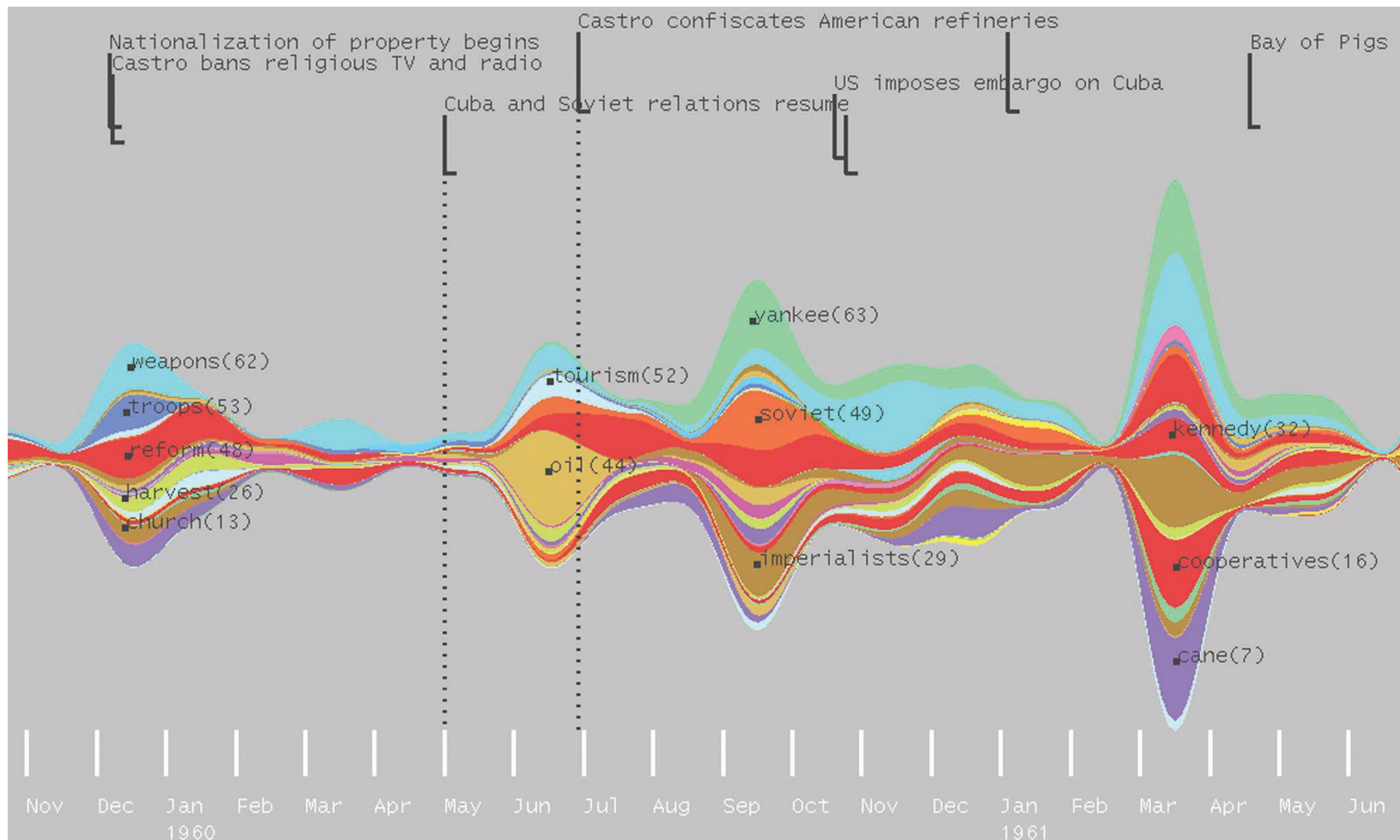


Figure 5: A user can interactively draw a region (polygon) containing a subset of documents of interest (top figure). Keywords are extracted from the selected document and their corresponding word cloud is built inside the user-defined region (bottom figure).



Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora

Christopher Collins
Fernanda B. Viégas
Martin Wattenberg

0:01 / 7:40

CC HD

Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora



Christopher Collins

Subscribe 26

2,768

+ Add to Share ... More

12 0

Uploaded on Aug 19, 2009

This video accompanies a full publication which appeared in the Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and

Jigsaw: Many Linked Views

Visual Analytics Support for Intelligence Analysis Case Study: The 9/11 Report

Carsten Görg
Youn-ah Kang
Zhicheng Liu
John Stasko



Information Interfaces Group
Georgia Institute of Technology

<http://textvis.lnu.se/>

Text Visualization Browser

A Visual Survey of Text Visualization Techniques

Provided by ISOVIS group

About Add entry

Techniques displayed:

141

Search:

Time filter:

1976

2014

Analytic Tasks



Visualization Tasks



Data

Source



Properties

