

Text & Document

COMP8503

Visualization & Visual Analytics

Textual Information

- **Text** is the basic medium of communication
- Easy to get access to the ever-growing volume of documents with textual information
 - **Digital libraries**: E-book libraries (e.g., Google Books), Journals & magazines (e.g., IEEE, ACM, Elsevier)
 - **WWW**: blogs, Facebook, Twitter, webpages, Wikipedia
 - Many others (e.g., subtitles, lyrics)

Text Analytics

- How to find information from piles of text?
- Search results:

[Text Mining, SAS Text Miner | SAS](#)

www.sas.com/textminer ▾ 翻譯這個網頁

Text mining from SAS automatically finds information buried in unstructured text to save time and money. Get an interactive tour, white papers and more.

[National Centre for Text Mining — Text Mining Tools and Text ...](#)

www.nactem.ac.uk/ ▾ 翻譯這個網頁

Welcome to NaCTeM. The National Centre for Text Mining (NaCTeM) is the first publicly-funded text mining centre in the world. We provide text mining services ...

[Software: Text Analysis, Text Mining and Information Retrieval](#)

www.kdnuggets.com ▸ Software ▾ 翻譯這個網頁

超過 70 筆 - Software for text analytics, text mining and information retrieval.

keyword analysis and webmaster tools

part of KnowledgeStudio, allows users to merge the output of unstructured, text ...

[Content Analysis and Text Mining Software - Provalis Research](#)

provalisresearch.com/products/content-analysis-software/ ▾ 翻譯這個網頁

Content Analysis and Text Mining Software Tools for the Analysis of Unstructured Data.

[Text Mining with STATISTICA Video Series - YouTube](#)

www.youtube.com/playlist?list=PLA1B7C970F1803850 ▾ 翻譯這個網頁

The intent of this series is to familiarize you with text mining, so that you are empowered to st...

[Lexalytics: Text Analysis and Text Mining Software](#)

www.lexalytics.com/ ▾ 翻譯這個網頁

Analyze social media and social sentiment with text analysis and content analysis software from Lexalytics, including opinion mining & sentiment analysis tools.

Text Analytics

- **Text Summarization:**
 - From over 480K confirmed records of H1N1 (swine flu) influenza
 - Ask: What are the major sources of infection?
 - From over 5K customer reviews of your online store
 - Ask: What did the customers say about your services?

Text Analytics

- **Opinion Mining** (aka **Sentiment Analysis**)
 - From thousands of customer reviews
 - Ask:
 - How have your customers' opinions changed toward your services?
 - How do your customer feel about your new product?
 - What are the main features of your products that your customers like most?

Text Analytics

- Steps further:
 - How do the reviews on your products **compare** with your competitors'?
 - Any **correlations** between the blogger sentiments and movie box office?

The Challenges

- **Huge volume** of complex information
 - Textual data are unstructured
 - Difficult to understand the meaning
 - Hard to perform analysis
- Different **customs** on use of languages
- Different people want different things
- People may not know what they want
 - Show me first and I'll tell you
- Machines alone is not good enough for full analysis

Levels of Text Representations

1. **Lexical level:** To group a string of characters into **tokens**, which is the basic unit of text to be analyzed and is application dependent
 - Example of tokens:
 - Words
 - A word may be defined as a sequence of letters separated by space or punctuation
 - Are **apple** and **apples** the same word?
 - Is **@** a word? How about *** ?) . , ?**
 - Are **Mister** and **Mr** the same?
 - How many words are there in **aren't, dunno** ?
 - How about Chinese text? How to identify words?

李娜在中国乃至亚洲网球选手中属于攻击力比较强的类型。由于其上肢力量突出，相对于其他亚洲选手而言，可以打出更具威胁的底线进攻球。

Levels of Text Representations

- **Lexical level:** More example of tokens
 - **Word stems**
 - The inflected variants of a word
 - e.g., have, has, had, having
 - **Phrases**
 - A group of words used as a unit to express a concept
 - e.g., cool down, keen on
 - **Word n-grams**
 - A sequence of n words
 - e.g., 2-grams: I like, I go, I eat, I think
 - **Character n-grams**
 - A sequence of n characters
 - e.g., the, con, doc, thi

Levels of Text Representations

2. Syntactic level

- To parse the **purpose of tokens**
- Examples:
 - Grammatical category (nouns, adjectives, conjunctions, etc.)
 - Tense
 - Plurality
 - Richer purpose such as date, money, place, person, etc.

3. Semantic level

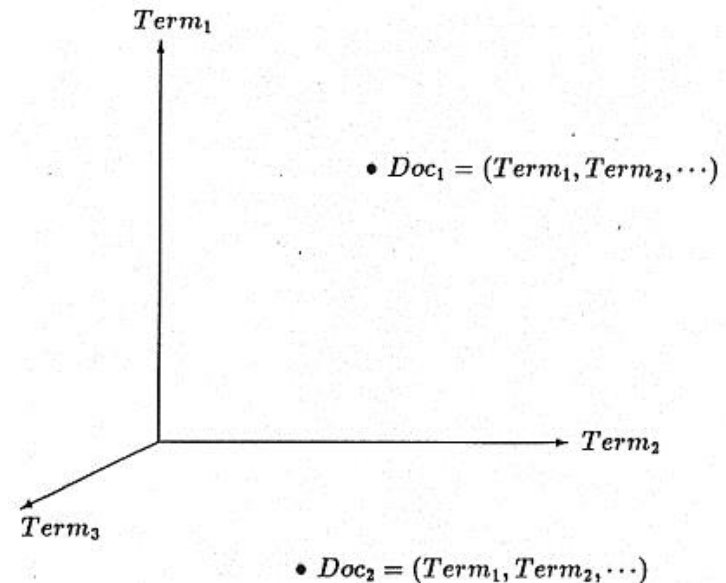
- To extract **meaning** of the syntactic structures of the full text

Document Statistics

- Number of words, paragraphs
- Word distribution or frequencies
- With document statistics, we may ask if there is any relationship between paragraphs or documents within a **corpus** (i.e., a collection of documents)?

Vector Space Model

- Each term in a corpus corresponds to a dimension in an n -dimensional space
- Each document is a vector with the i -th coordinates being the **weight** (or **importance**) of the i -th term
- Filtering — remove **stop words** (e.g., “the” or “a”)
- Stemming — group inflected forms of a word (e.g., “robot”, “robots”, “robotics”, “robotlike”)

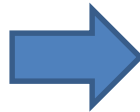


[Salton and McGill, *Introduction to Modern Information Retrieval*, 1983.]

Vector Space Model

- Using frequency of word occurrences as importance

The special kind of boredom from which modern urban populations suffer is intimately bound up with their separation from the life of Earth. It makes life hot and dusty and thirsty, like a pilgrimage in the desert. Among those who are rich enough to choose their way of life, the particular brand of unendurable boredom from which they suffer is due, paradoxical as this may seem, to their fear of boredom. In flying from the fructifying kind of boredom, they fall a prey to the other far worse kind. A happy life must be to a great extent a quiet life, for it is only in an atmosphere of quiet that true joy can live.



Word	Freq.	Word	Freq.
life	5	atmosphere	1
boredom	4	live	1
kind	3	urban	1
quiet	2	extent	1
suffer	2	dusty	1
population	1	due	1
way	1	thirsty	1
unendurable	1	enough	1
pilgrimage	1	paradoxical	1
far	1	bound	1
choose	1	seem	1
special	1	worse	1
Earth	1		

<http://www.wordcounter.com>

Ranking Documents

- Given a query (a string of words) q , how to retrieve the most relevant document?
- A solution:
 - Consider a document as a bag of words without ordering (using the vector space model)
 - Give a score for each document d w.r.t. the vocabularies t that appear in the query
$$S_{q,d} = \sum_{t \in q} W_{t,d}$$
 - Use the scores to rank the documents
- How to determine $W_{t,d}$, i.e., the weight (or importance) of a term?

Term Frequency

- **Term Frequency** ($tf_{t,d}$) of term t in document d is defined as the number of times that t occurs in d .
- A term has higher weight if it appears very often in a document
- Score of a document is given by

$$S_{q,d} = \sum_{t \in q} tf_{t,d}$$

Term Frequency

Query: who took computer science


$$S_{q,d} = \sum_{t \in q} \text{tf}_{t,d}$$

	Documents	$\text{tf}_{t,d}$				$S_{q,d}$
		who	took	computer	science	
D1	alan turing is considered the father of computer science			1	1	2
D2	who published in science	1			1	2
D3	computer science is not just science			1	2	3
D4	it was alan who took me to the science class	1	1		1	3

Term Frequency

- However, relevance is not directly proportional to term frequency. Hence, use $\log(1 + \text{tf}_{t,d})$ instead

	Documents	$\log(1 + \text{tf}_{t,d})$				$S_{q,d}$
		who	took	computer	science	
D1	alan turing is considered the father of computer science			0.301	0.301	0.602
D2	who published in science	0.301			0.301	0.602
D3	computer science is not just science			0.301	0.477	0.778
D4	it was alan who took me to the science class	0.301	0.301		0.301	0.903

$$S_{q,d} = \sum_{t \in q} \log(1 + \text{tf}_{t,d})$$


Term Frequency

- What's the problem in using only term frequency?
 - All terms have the same importance
 - Bigger documents have more terms
- Recall the stop words? Frequently appear in documents but less informative
- Doctrine: *A term that appears in every document is less important, because it has no discriminatory power*

Inverse Document Frequency

- We would like **rare terms** to have **higher weighting**
- df_t — **document frequency** of term t is defined as the number of documents that contain t
 - $df_t \leq N$, the total number of documents
- **Inverse document frequency**
$$idf_t = \log (N / df_t)$$

Inverse Document Frequency

Query: who took computer science

	Documents	Terms			
		who	took	computer	science
D1	alan turing is considered the father of computer science			1	1
D2	who published in science	1			1
D3	computer science is not just science			1	2
D4	it was alan who took me to the science class	1	1		1
	df	2	1	2	4
	idf	0.301	0.602	0.301	0

tf-idf weighting

- **Term frequency inverse document frequency**
- The **tf-idf** weight of a term t is the product of its **tf** and **idf** weights

$$\text{tf-idf}_{t,d} = \log(1 + \text{tf}_{t,d}) * \log(N / \text{df}_t)$$

- Higher weight for terms having high frequencies in a document
- Lower weight for terms appearing in more documents

Document Retrieval

- Matching score of document d to a query q :

$$S_{q,d} = \sum_{t \in q} \text{tf-idf}_{t,d}$$

- Increases with the number of occurrences **within** a document
- Increases with the rarity of the term **across** the whole corpus

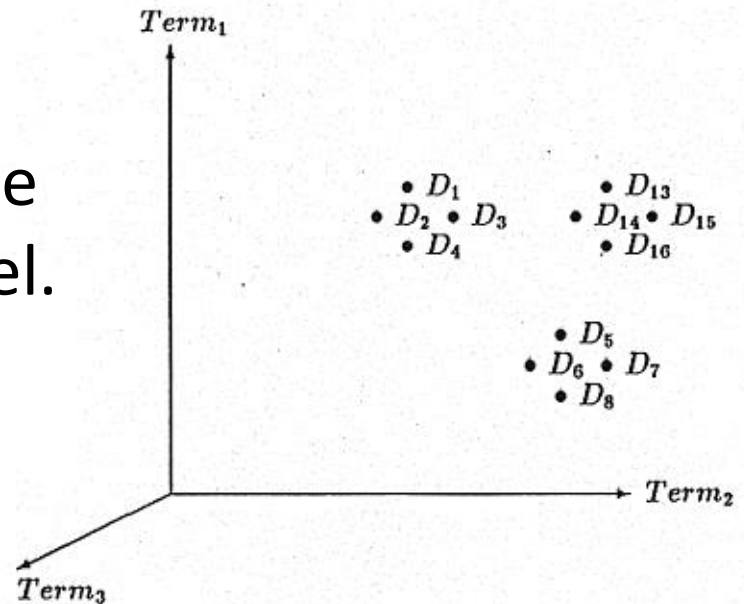
Document Retrieval

Query: who took computer science

	Documents	$\text{tf-idf}_{t,d}$				$S_{q,d}$
		who	took	computer	science	
D1	alan turing is considered the father of computer science			0.09		0.09
D2	who published in science	0.09				0.09
D3	computer science is not just science			0.09		0.09
D4	it was alan who took me to the science class	0.09	0.18			0.27

Document Clustering

- To group related or similar documents
- Need a measurement of how “close” two documents are
- A document is a **term vector** in the high-dimensional space under the vector space model.



Term Vector Similarity

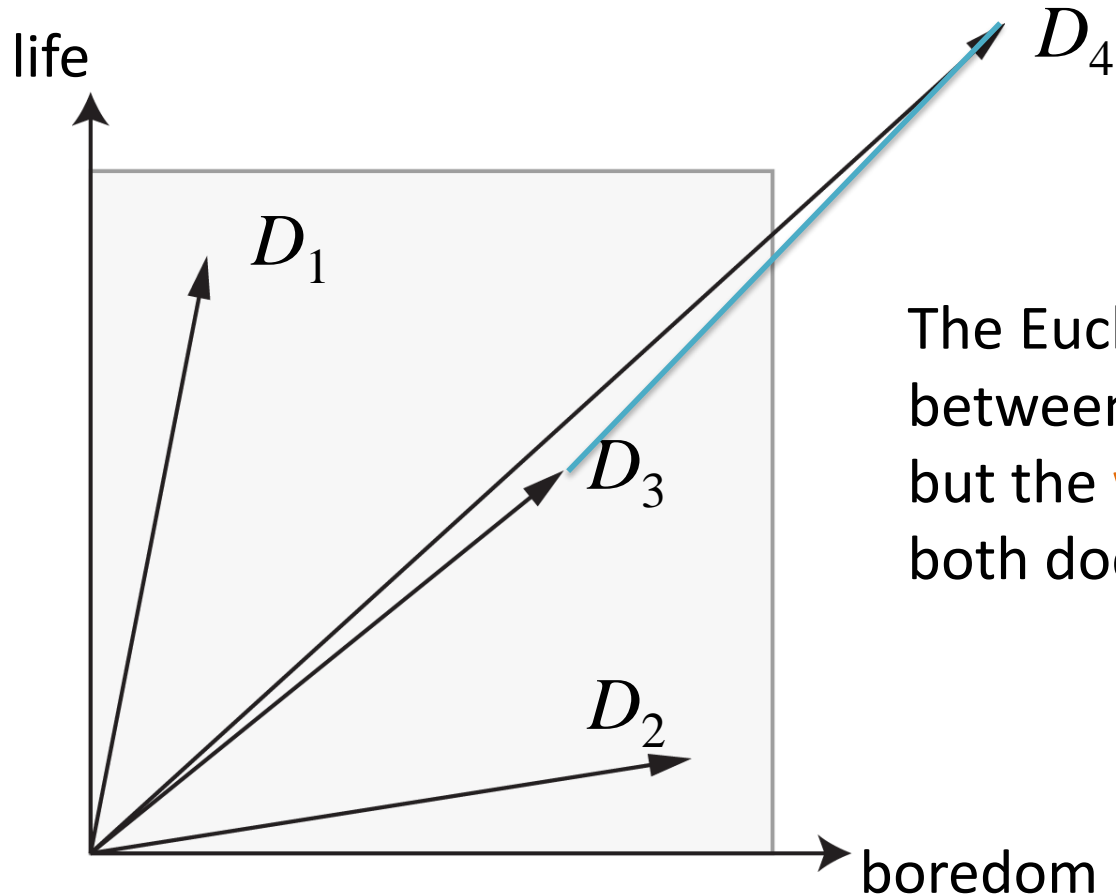
- Distance between two vectors
- Euclidean distance — vectors as points in n -dimensional space

Let $D_1 = (v_{1,i})$ and $D_2 = (v_{2,i})$, then

$$\text{dist}(D_1, D_2) = \sqrt{\sum_i (v_{1,i} - v_{2,i})^2}$$

- Why not?

Euclidean Similarity



The Euclidean distance between D_3 and D_4 is large, but the **word distribution** in both documents are similar.

Cosine Similarity

- **Cosine similarity**: Similarity between documents = cosine of the angle between their vectors

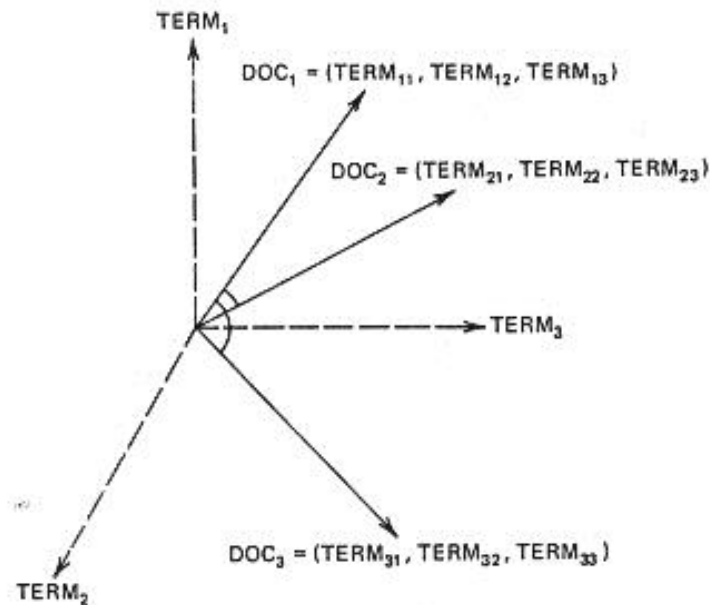


Figure 4-2 Vector representation of document space.

Let $D_1 = (v_{1,i})$ and $D_2 = (v_{2,i})$, then

Similarity (D_1, D_2)

$$= \cos(\theta)$$

$$= \frac{D_1 \cdot D_2}{\|D_1\| \|D_2\|}$$

$$= \frac{\sum_i v_{1,i} \cdot v_{2,i}}{\sqrt{(\sum_i v_{1,i}^2)(\sum_i v_{2,i}^2)}}$$

Visualization

- With the vector space model, we may derive patterns and structures from a volume of text which need to be visualized
- Visualization by
 - **Words or source text** (words / images / phrases / paragraphs)
 - tag clouds, word tree, phrase net, themail
 - **Computed metrics or features**
 - arc diagram
 - **Models and abstractions**, i.e., concepts or **themes** over documents
 - ThemeRiver, TIARA

Tag Clouds

- A visualization of **word frequencies**
- Size / color determined by frequencies
- Simple and easy to read, good data density
- Potential problems:
 - long words are emphasized over short words
 - **Ascenders** (e.g., b, d, f, h, k, l, t) and **descenders** (e.g., g, j, p, q, y) receive unconscious attention
 - Alternatives: bar charts, tables

Tag Clouds

'come ago although atmosphere attraction **boredom** bound boy **bring** called
cease choose civilised consider contact difference displayed
distinction due **dusty** **earth** element emotion exciting existed experience fall far
form found gambling grass **happy** hark instant intensity intercourse intimately
joy lark leave **life** live **love** lyrics man mere moment **nothing**
occupations pilgrimage plants **pleasures** populations possible prey primitive
quiet rain remains renewed sense separation **sex** simplest something **suffer**
supreme true **two-year-old** unendurable union utterance **whole** yellow

A paragraph from Chapter 4, “The Conquest of Happiness” by Bertrand Russell.

Tag cloud generated by <http://tagcrowd.com/>

Tag Clouds

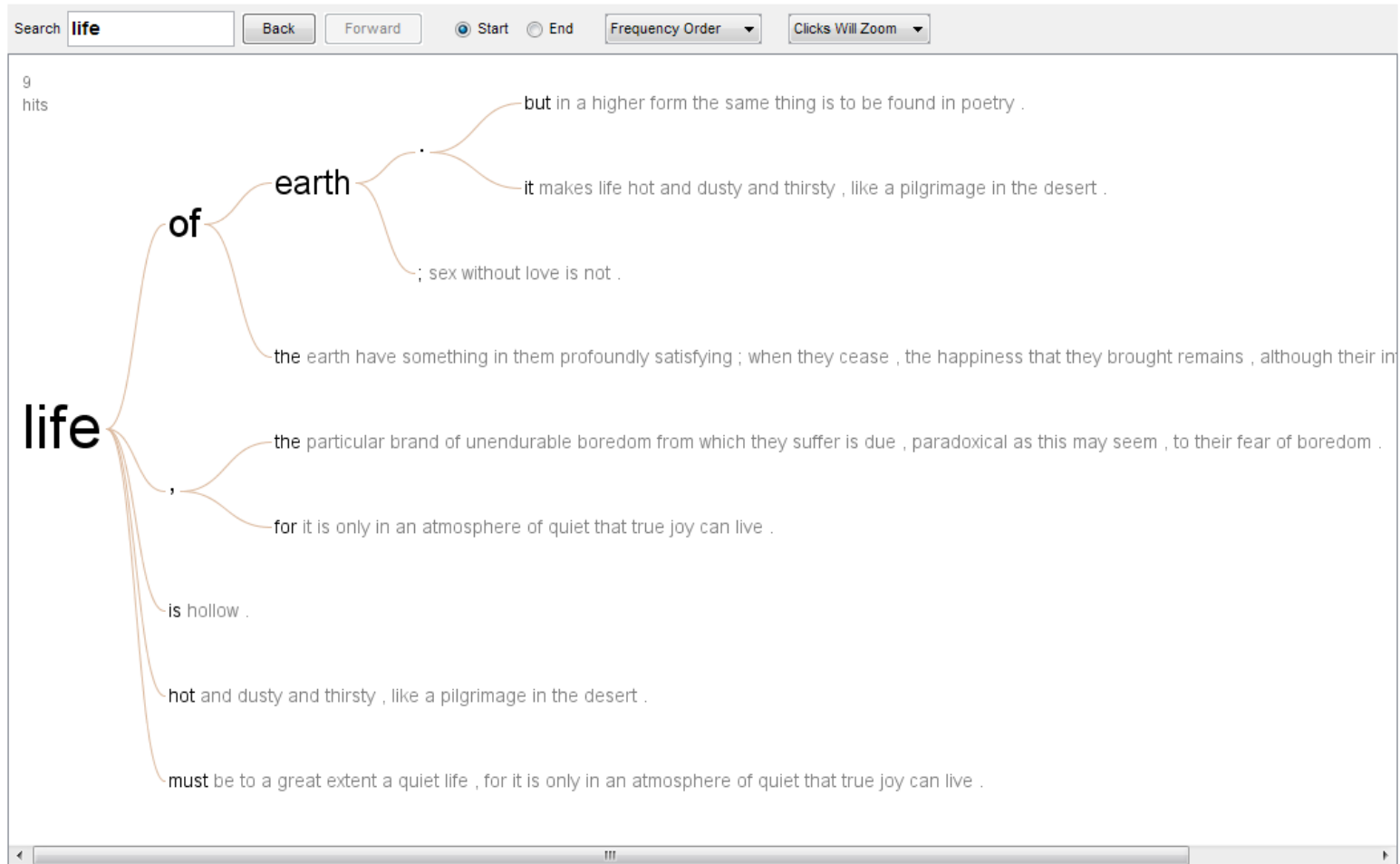


Generated from <http://www.wordle.net>

Word Tree

- Visualization of **term frequencies**, with context
- Document flow visualization
- **Root** of a tree = user-specified words (***W***)
- **Branches** = phrases that appear after ***W***
- Similar to the suffix tree

Word Tree



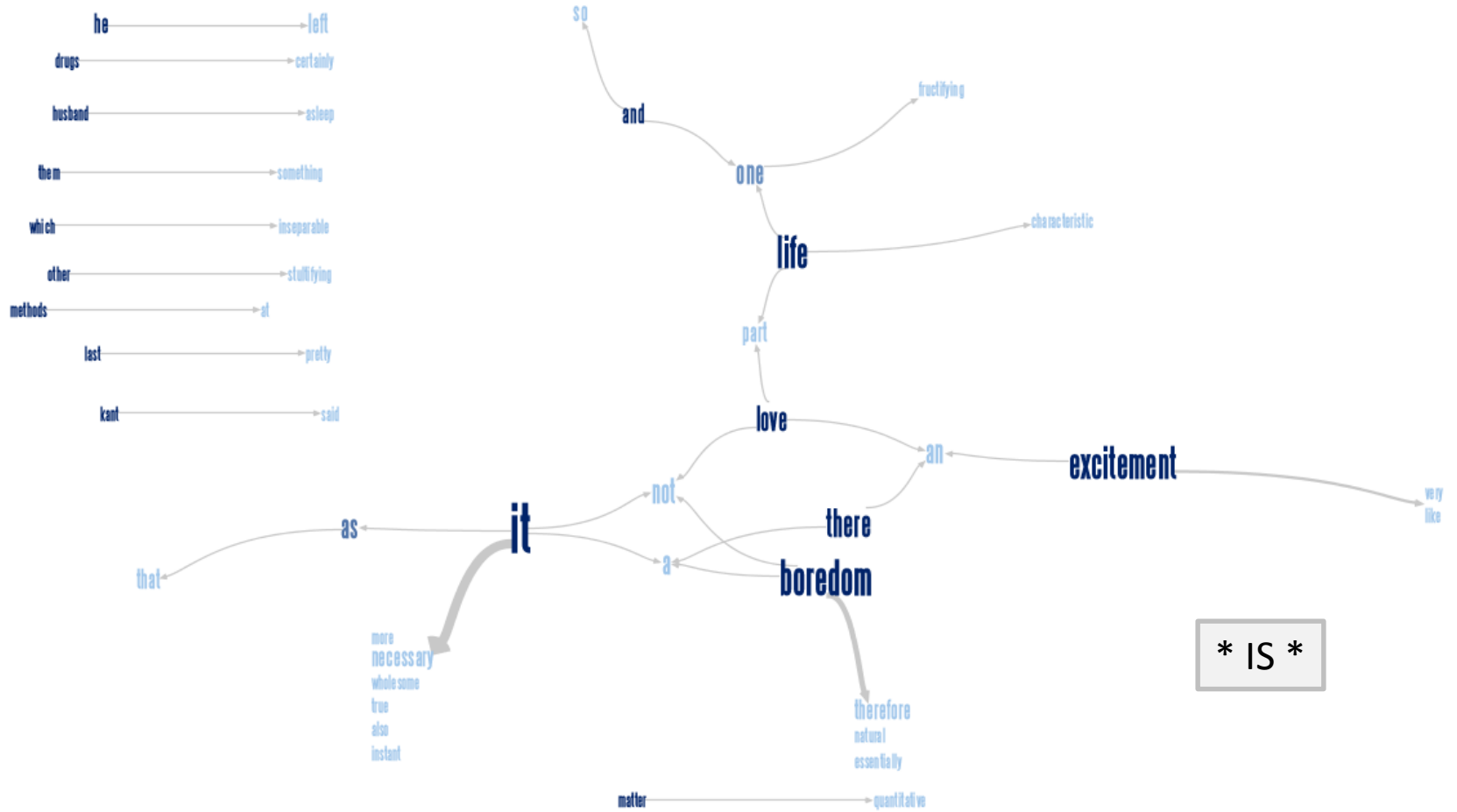
Generated by Many Eyes

[<http://www-958.ibm.com/software/analytics/manyeyes/visualizations/bertrand-russell-the-conquest-of-h>]

Phrase Net

- Pairs of words that fit a particular **pattern**
 - e.g., pride AND prejudice, spring AND summer
- Visualize in the form of a network diagram
- Two words are connected if they occur in the same phrase; line thickness depicts frequency of match
- Size of word proportional to the number of times it occurs in a match
- Color indicates whether a word is more likely to be found in the first or second slot of a pattern

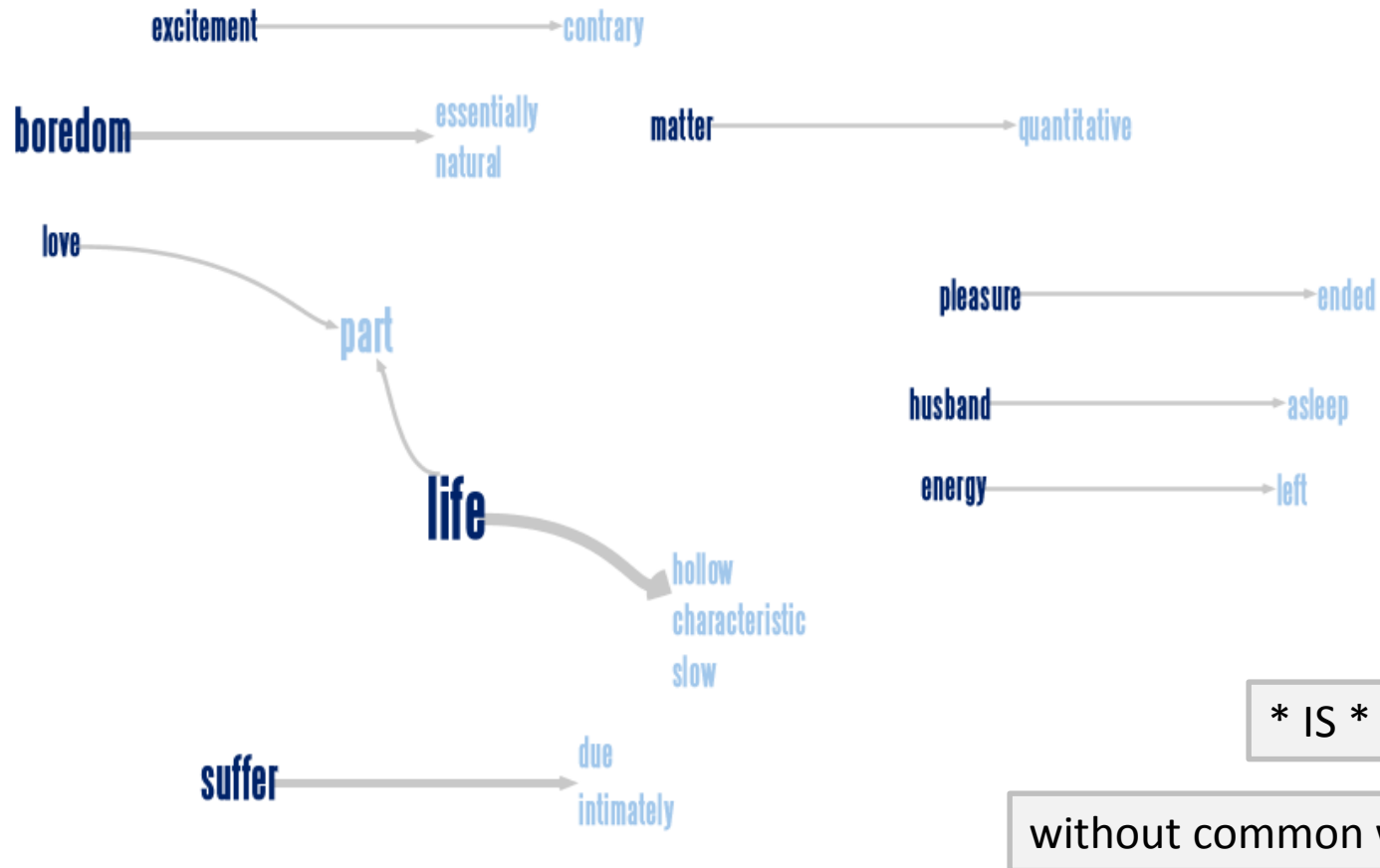
Phrase Net



Generated by Many Eyes

[\[http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/phrase-net-visualization-of-russel\]](http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/phrase-net-visualization-of-russel)

Phrase Net

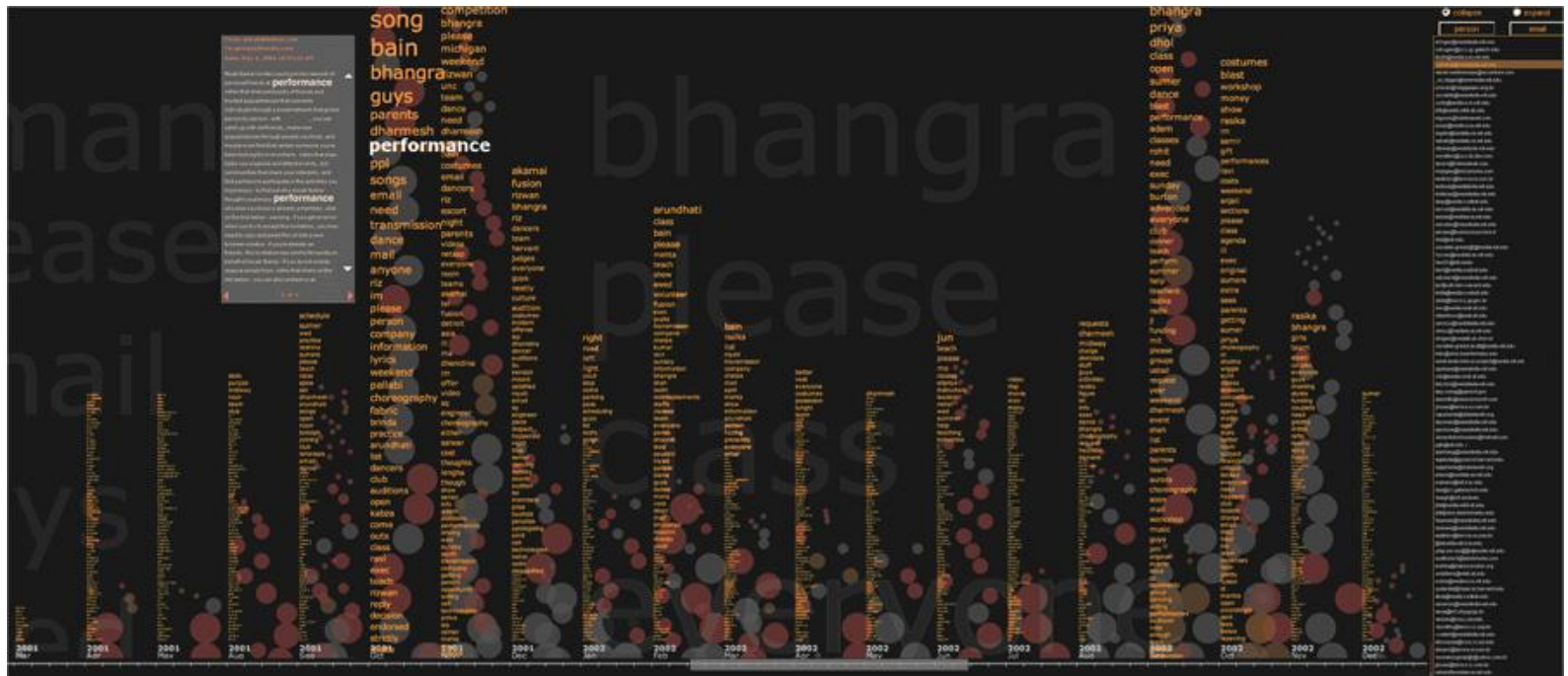


Generated by Many Eyes

<http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/phrase-net-visualization-of-russel>

Themail

- Shows the words that characterize one's correspondence with another and how they **change over time**



Word importance ↔ size of words

Circles ↔ email messages

Circle size ↔ length of email

Circle color ↔ direction of emails

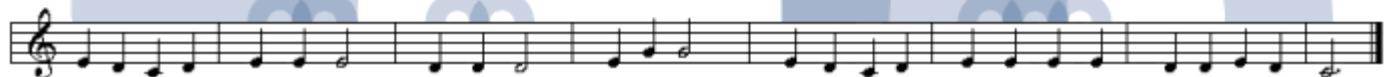
<http://fernandaviegas.com/themail/>

Arc Diagrams

- Show **repetitive structures** by connecting repeated subsequence with circular arcs
- Good for visualizing musical pieces
- Thickness of arc represents length of the repeated subsequence and height of arc represents distance between the subsequences

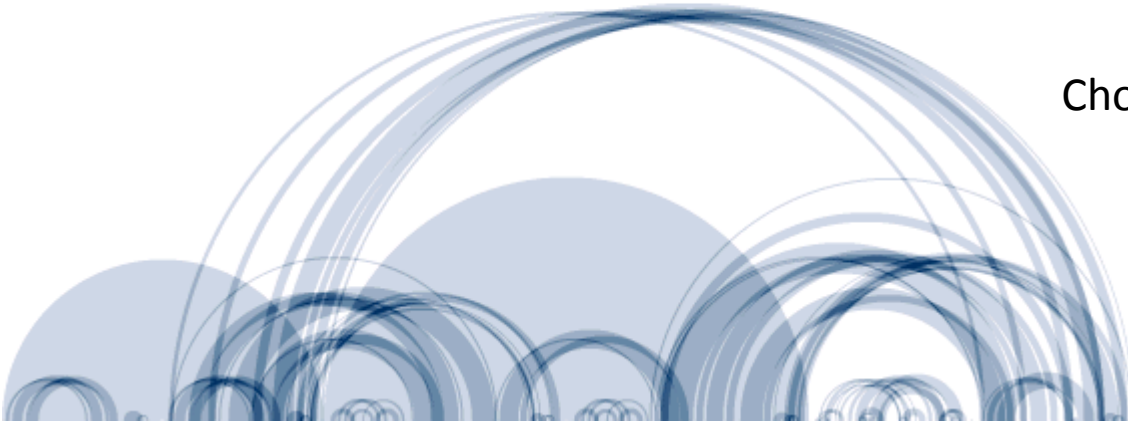
The Shape of Songs

[\[http://www.turbulence.org/Works/song/index.html\]](http://www.turbulence.org/Works/song/index.html)

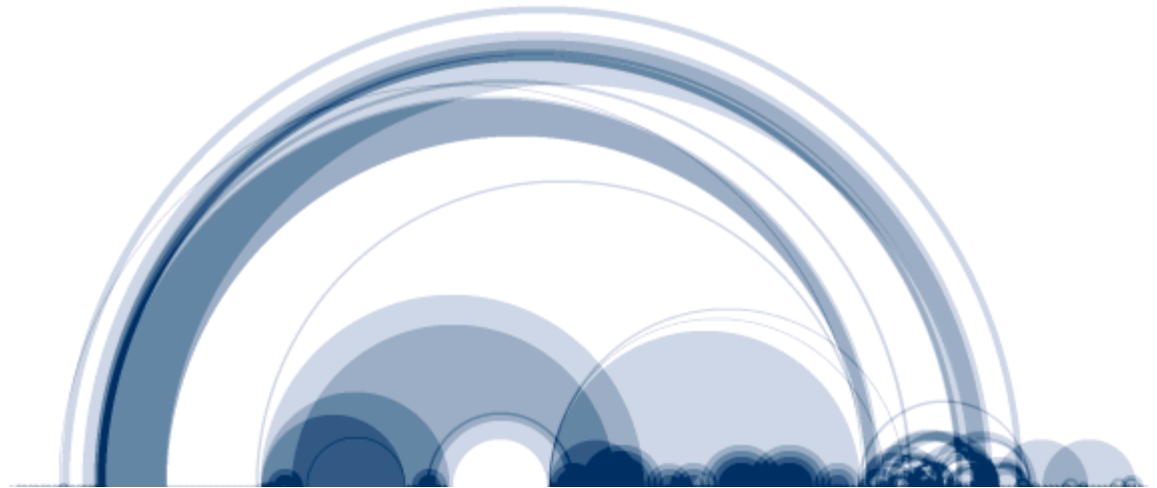


Arc Diagrams

Chopin, Mazurka in F# Minor

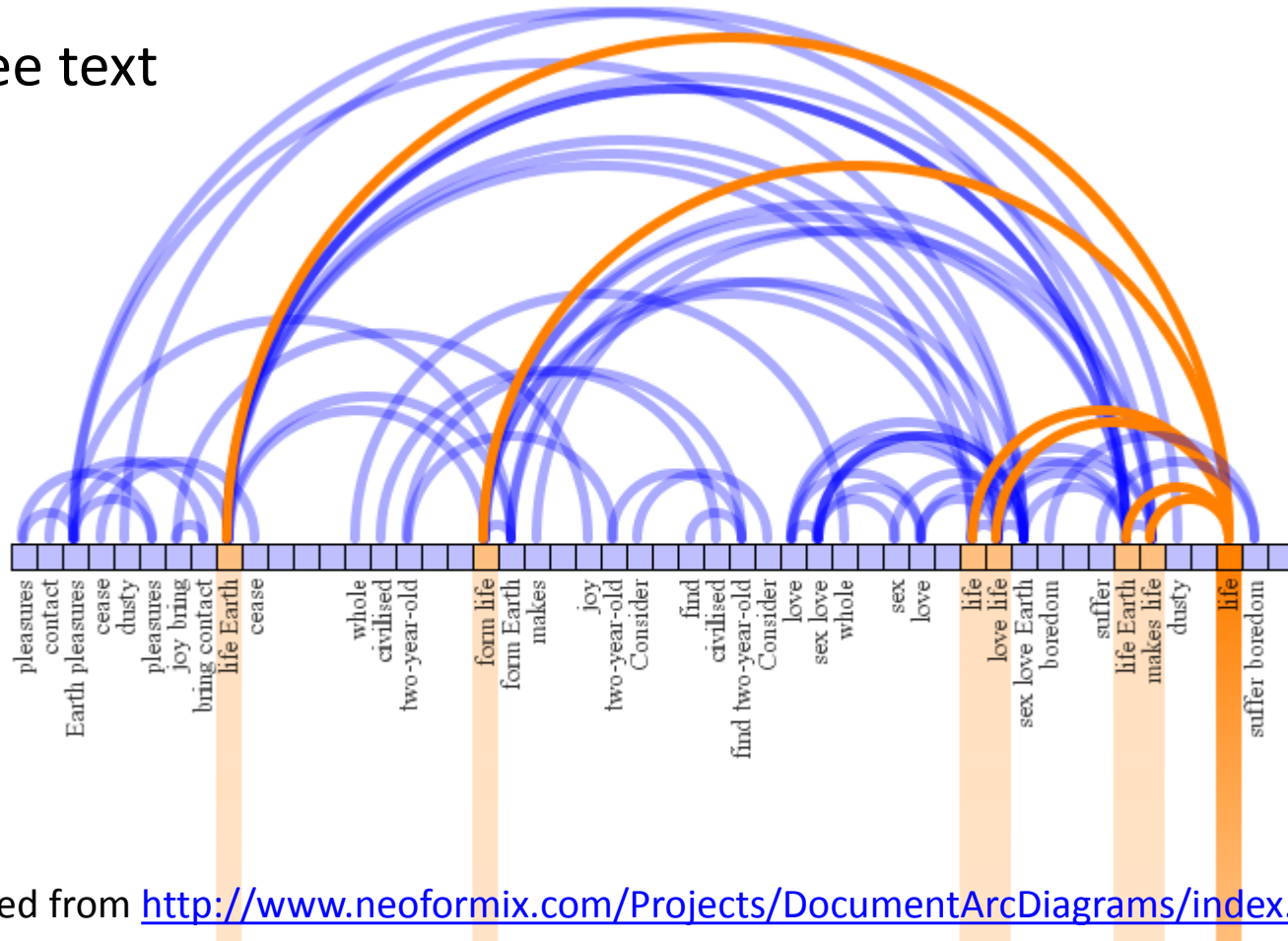


Madonna, Like A Prayer



Arc Diagrams

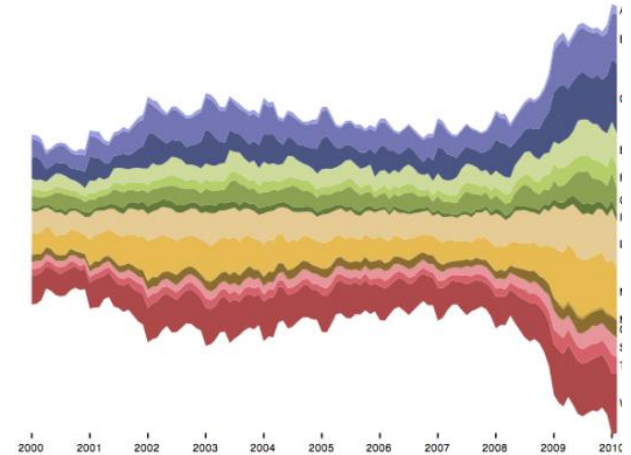
For free text



Generated from <http://www.neoformix.com/Projects/DocumentArcDiagrams/index.html>

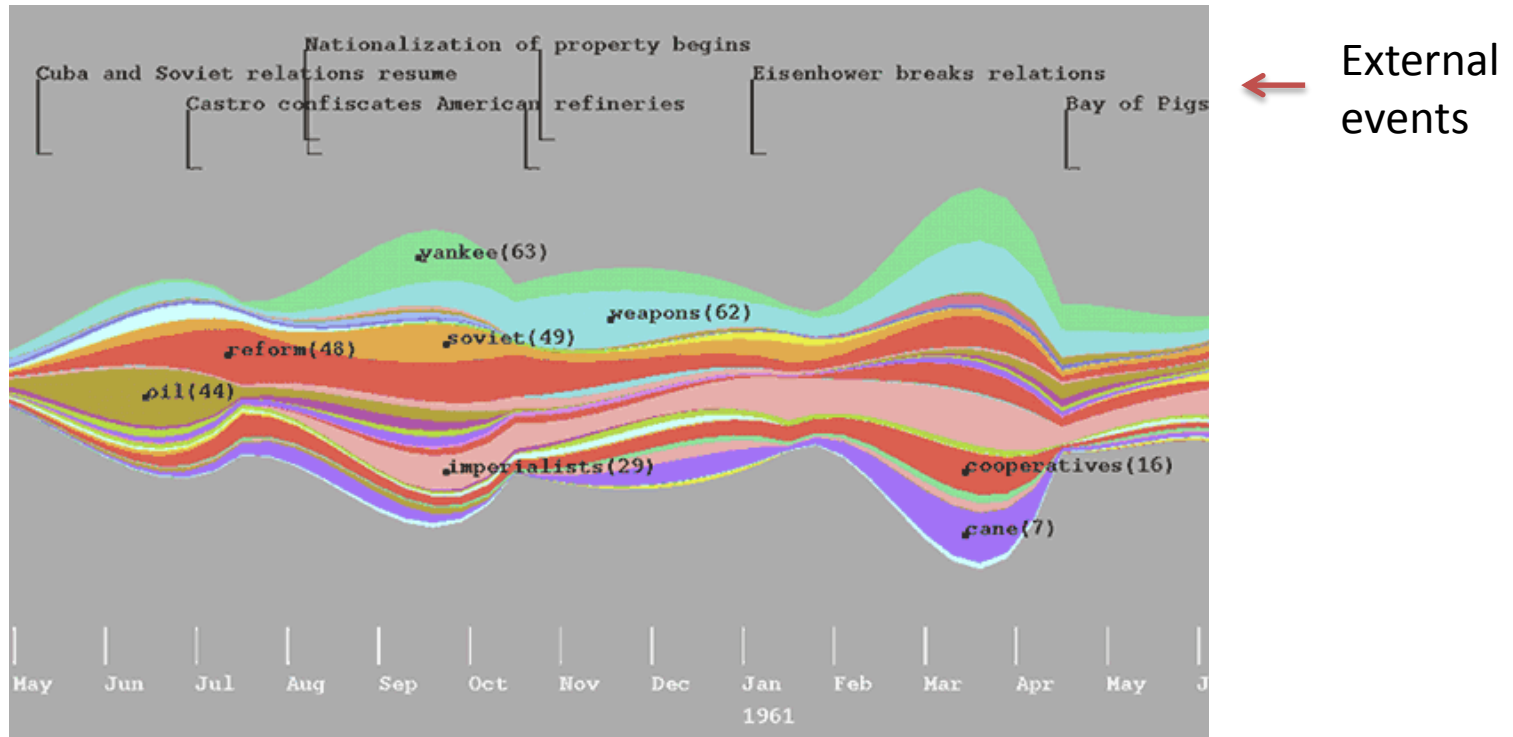
ThemeRiver

- Show thematic variations over time within a corpus.
- A river metaphor:
 - Direction of flow: time line
 - Width: strength of a theme
- Idea based on the stacked graph



The stacked graph [Heer 2010]

ThemeRiver



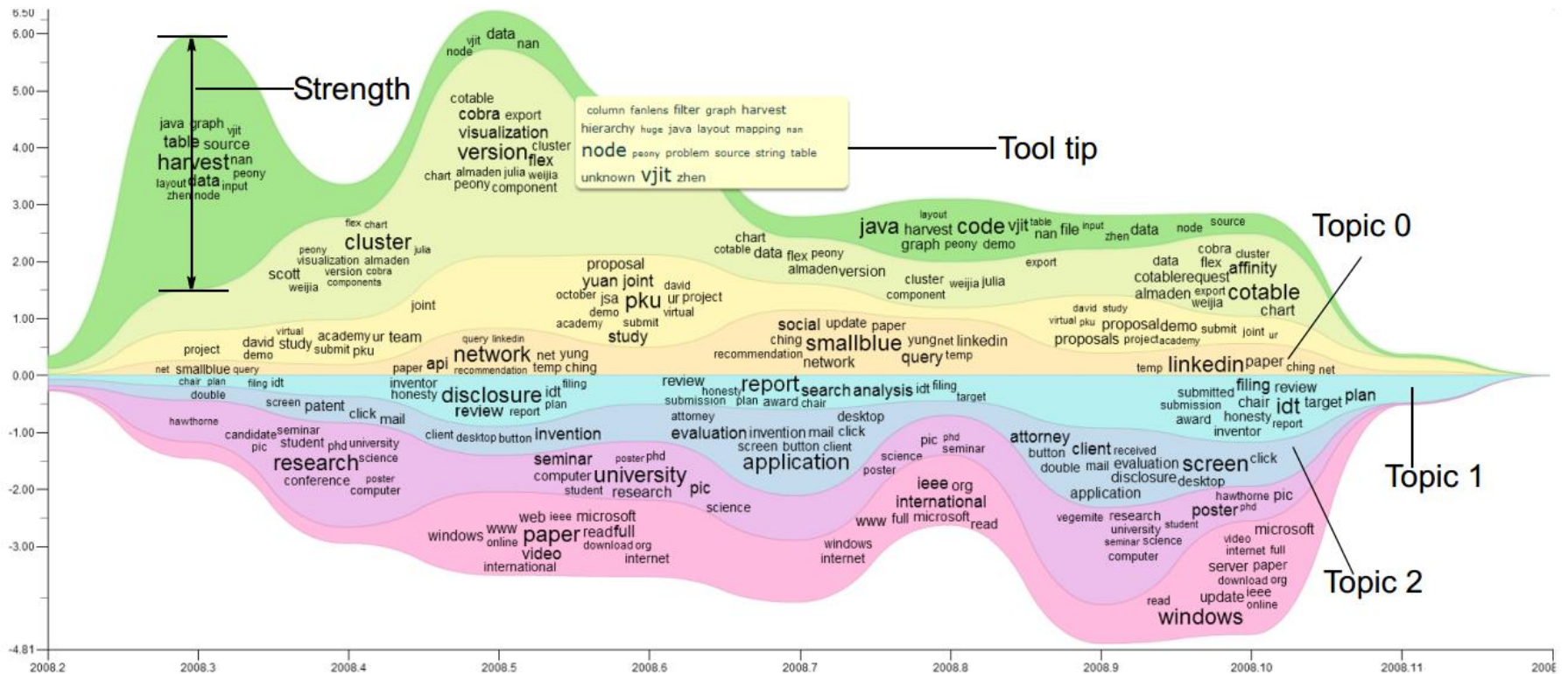
- The sudden change of thematic strength following an external event may indicate a causal relationship

[Havre et al. "ThemeRiver: visualizing thematic changes in large document collections," *TVCG* 2002]

Tiara

- Illustrates thematic variation over time like ThemeRiver
- Depicts also detailed thematic content in **keywords**
 - Each topic layer is filled with keyword clouds at different time points to informatively convey the content changes over time.

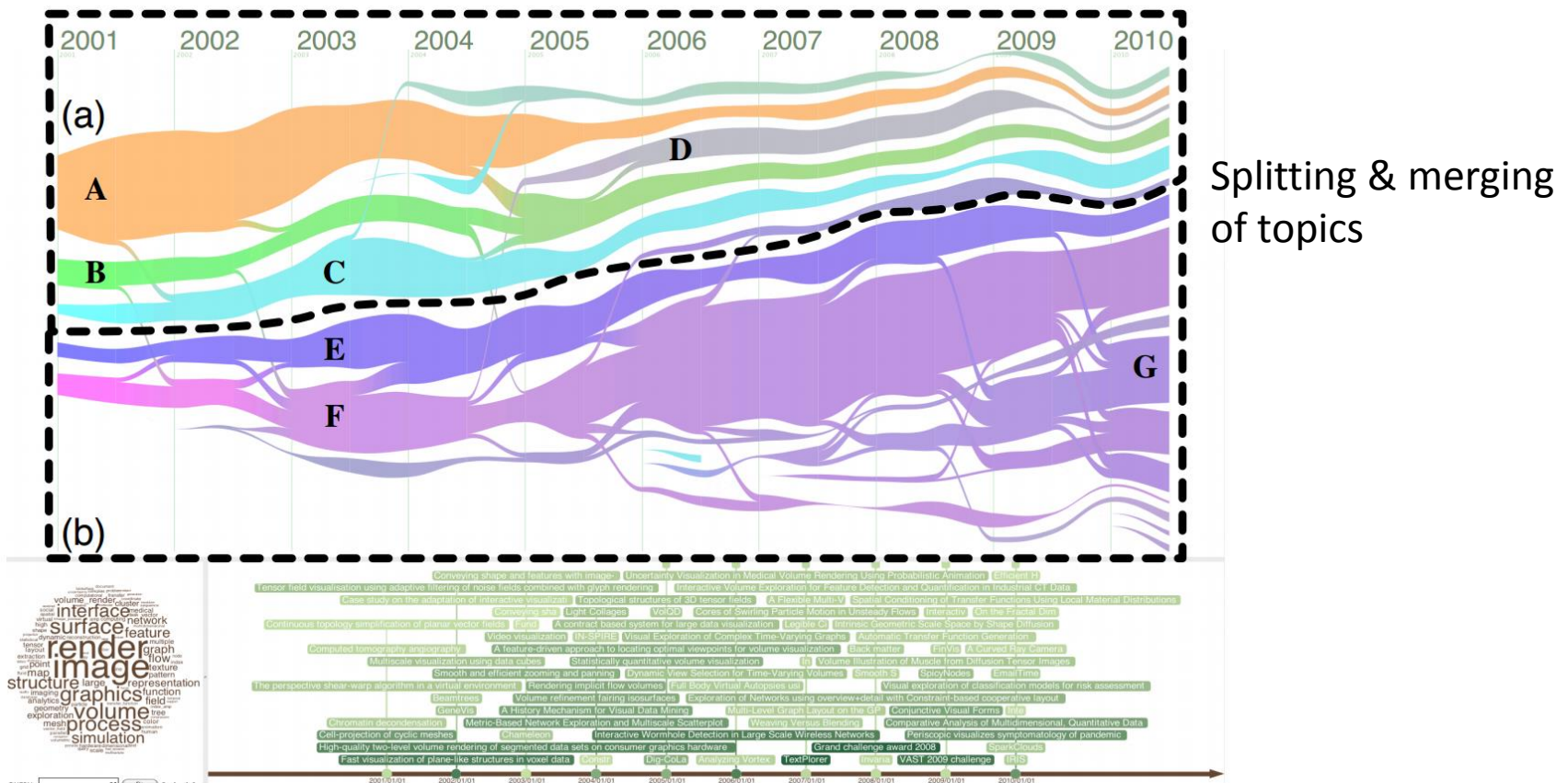
Tiara



[Liu et al. "TIARA: Interactive, Topic-based Visual Text Summarization and Analysis", Transactions on Intelligent Systems and Technology, 2012.]

Textflow

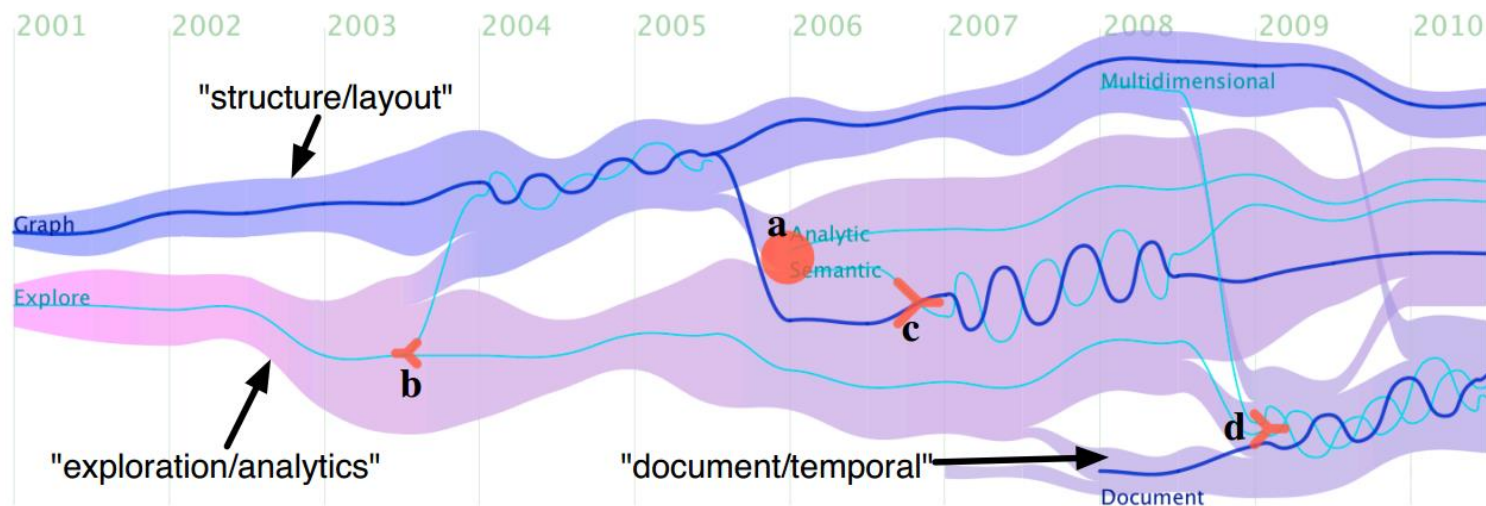
- Extract also the connections between evolving topics



[Cui et al. "TextFlow: Towards Better Understanding of Evolving Topics in Text", InfoVis 2011.]

Textflow

- Critical events (**topic birth, death, split, merge**) in relation to keyword changes



Visual Design



source



sink

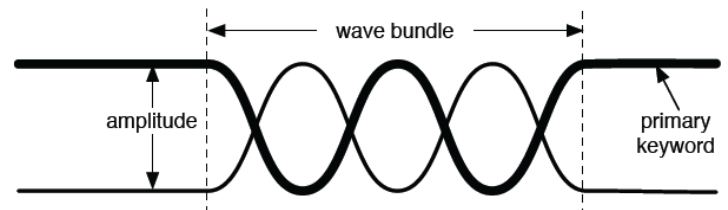


split



merge

Critical events

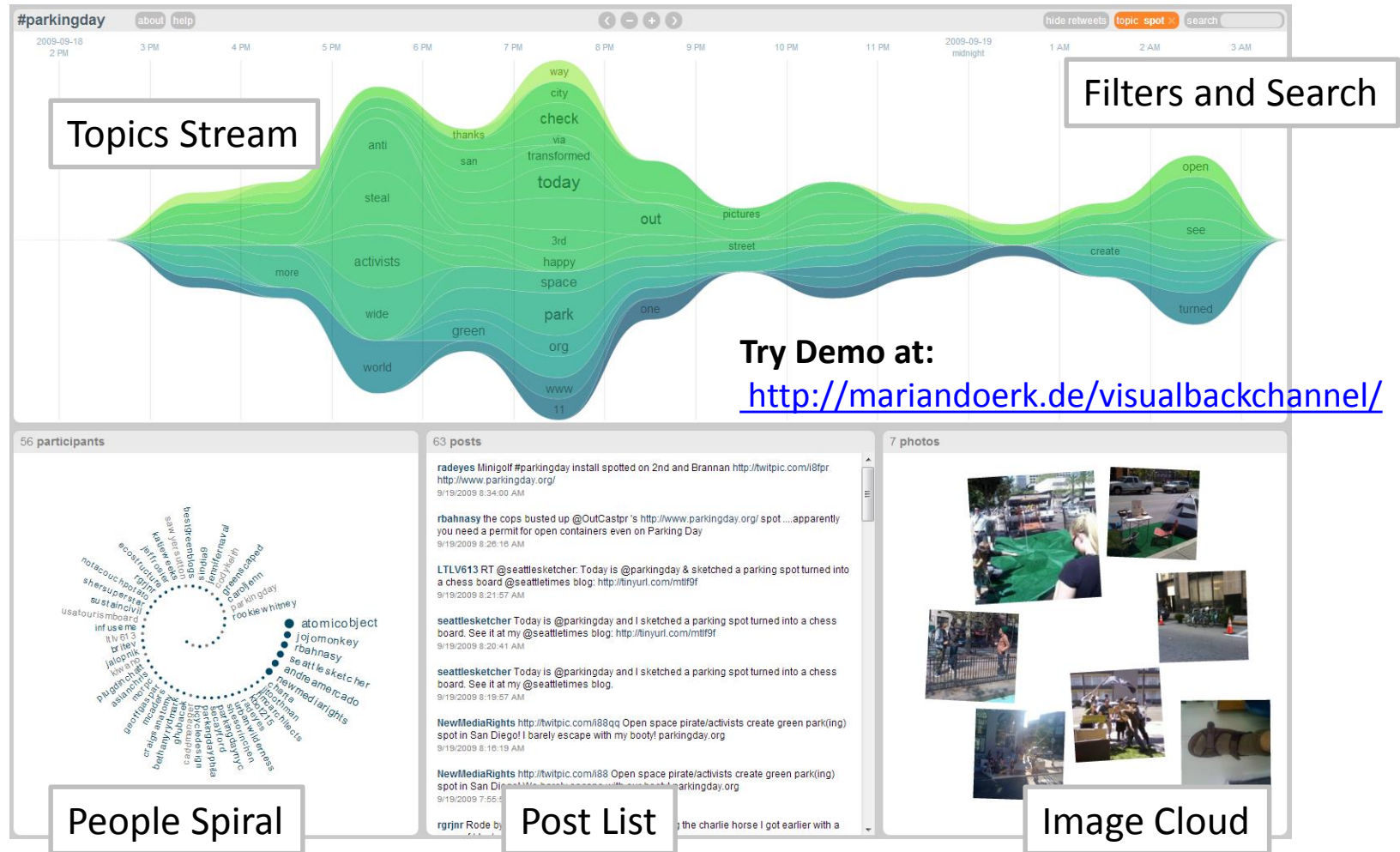


Keyword threads

Visual Backchannel

- Digital backchannel (e.g. Twitter): a communication channel to share brief and timely information (e.g., impressions / suggestions / comments) on a social event.
- Aims to represent conversation topics in the context of their temporal development together with participants' activity and pictorial impressions

Visual Backchannel



dot size – number of activities
saturation – originality of posts

[Dörk et al. "A Visual Backchannel for Large-Scale Events ", InfoVis 2010]

Text & Document

Reference

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "*Introduction to Information Retrieval*", Cambridge University Press, 2008 [Chapter 6]
<http://nlp.stanford.edu/IR-book/>
- Matthew Ward, Georges Grinstein and Daniel Keim, "*Interactive Data Visualization: Foundations, Techniques, and Applications*", 2010 [Chapter 9]