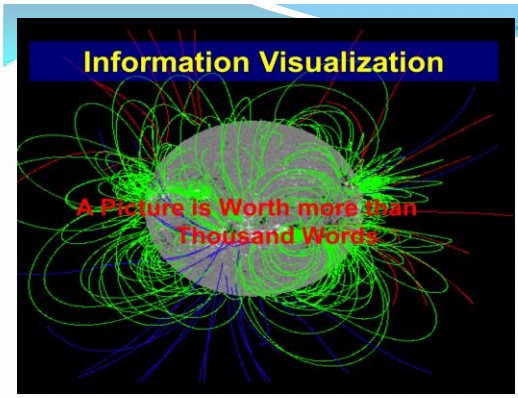
Text Visualization and Knowledge Discovery for Text

Do Phuc, Ph. D, Assoc. Prof.
Faculty of Information Systems,
University of Information Technology, VNU-HCM



Text Visualization and Knowledge Discovery for Text

- Text is one of the greatest data and popular around us. Text contains our history and is a major mean to recording information and knowledge. With the Internet , text data has been generated with significantly speed. Currently, millions of websites are generating extraordinary amount of online text data everyday. The users of the social network like FaceBook, Twitter,.. produce billions of posting messages everyday.
- The explosion of the data makes a lot of challenges to discover and understand the knowledge in text.
- Text visualization techniques can be helpful for solving these problems. Various visualizations have been designed for showing the similarity of text documents, revealing and summarizing text content, and helping discover the big text data exploration.
- In this talk, we will show the techniques of text visualization and the application of text visualization in knowledge discovery for text. We also address the challenges for text visualization with big text data in big data era.

Outline

- Data Visualization
- Examples
- Tree Visualization
- Graph Visualization
- Text Visualization

Data Visualization

Visualization & Information Visualization:

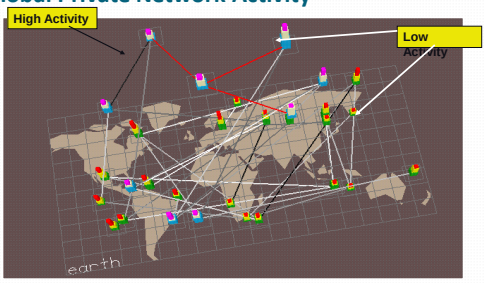
- Visualization:
  - "The use of computer-supported, interactive, visual representations of data to amplify cognition."
  - Goal: discovery, decision making, explanation
- Information Visualization:
  - "The use of **computer-supported?**, interactive, visual representations of abstract data to amplify cognition."

Visualization Amplifies Cognition

1. Increase memory and processing resources
2. Reduce information searching
3. Enhance detection of patterns
4. Enable perceptual inference operations
5. Use perception attention for monitoring
6. Encode information in a manipulative medium

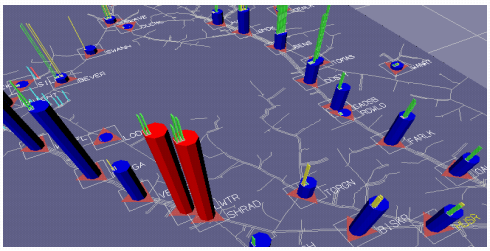
Data Visualization Examples

Global Private Network Activity



Phuc Do

Natural Gas Pipeline Analysis



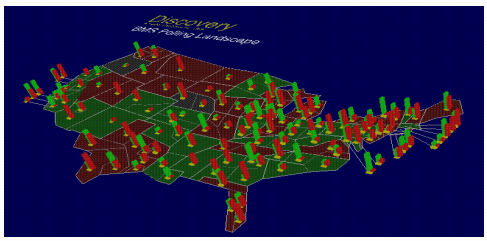
Height shows total flow through compressor stations.

Phuc Do

Risk Analysis Report

Phuc Do

Telephone Polling Results

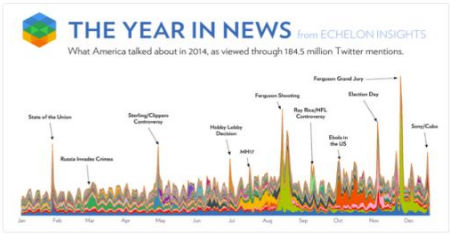


Note: On the "live" map, clicking on an area allows the user to drill down and see results for smaller areas.

Phuc Do

12) The Year in News

The best data visualizations communicate information and do so in an intuitive and beaut way. Echelon Insights nailed it with this piece, which visualizes the most talked-about nev stories of 2014 on Twitter. What do 184.5 million tweets look like? Rad spin art.



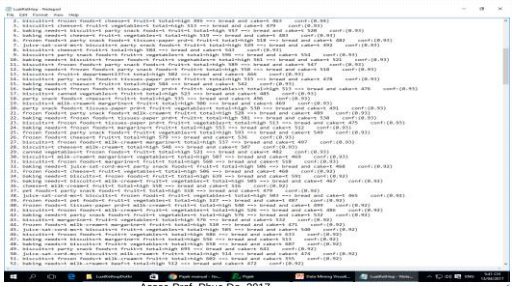
Phuc Do

Akami Real-time Web Monitor shows a map of the world with real-time information about internet traffic, latency and current network attacks. The maps are color-coded and easy to read, but only give very generalized information.

Phuc Do

An example of Data Mining Visualization: Association rules

List of Associations



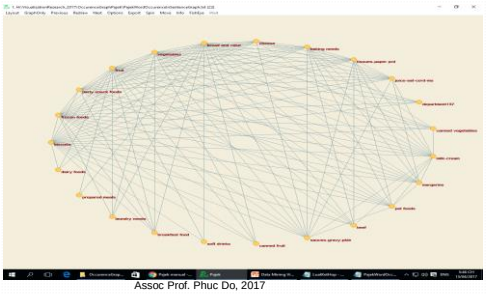
Drawbacks

- Long list of rules
- Many rules are redundant
- Very hard to capture the important rules
- How to access association rules effectively

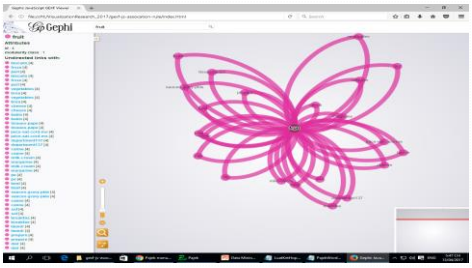
Graph of Association rule

- Association rules X -> Y
- Edge
  - Source: X
  - Target: Y
- Set of Association rules
  - Graph (V,E)
  - V: set of frequent itemsets
  - E:set of edges created from association rules

Graph visualization by Pajek software



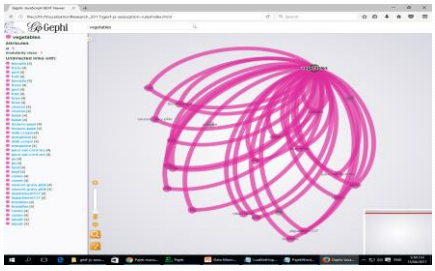
Graph Visualization by Gephi on the Web



Interactive mode in Graph visualization

- Very easy to capture the important association rules
- Click on node to discover the co-occurrence items
- Move around to discover the association rules

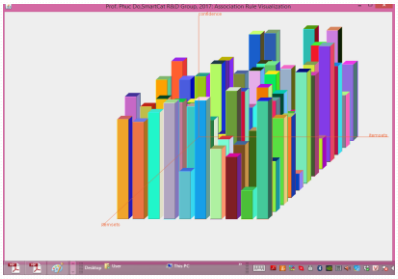
Discover the items co occurrence with vegetables



Using cube to visualize AR

- Each rule is a cube
- Click on Cube to visualize the association rule
- The height of Cube: the confidence of AR.
- Easy to discover the important rules based on the height of cube.

AR visualization, each AR is a cube



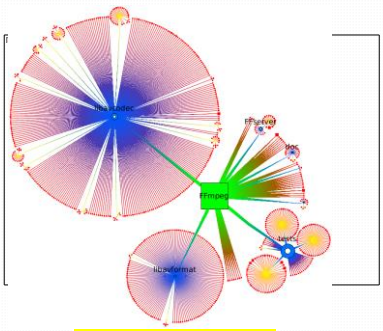
Visualization of Relations

- Tree Visualization
- Graph Visualization

Tree Visualization

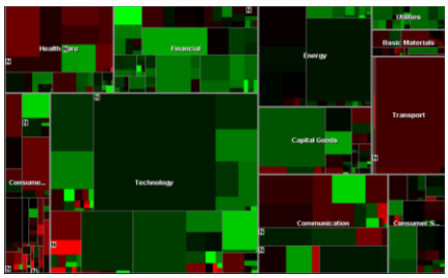
- Treemap method: visualize the tree structure that use virtually every pixel of the display space to convey information
- Every subtree is represented by a rectangle, that is partitioned into smaller rectangles with correspond to its children.
- The position of the slicing lines determines the relative sizes of the child rectangles.
- For every child, repeat the slicing recursively, swapping the slicing direction from vertical to horizontal or conversely

Tree Visualization



Bubble-Tree Layout

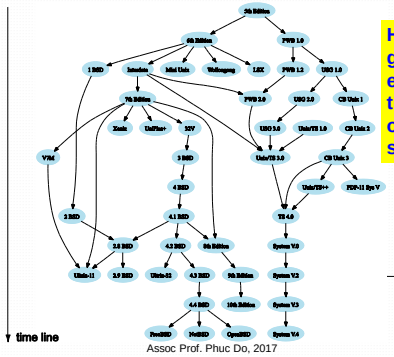
Tree Visualization



Graph Visualization

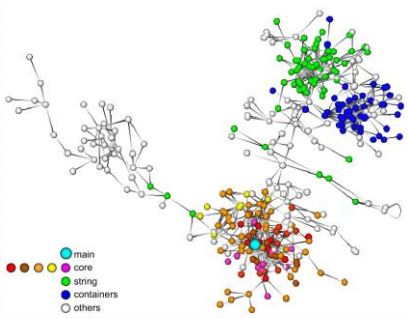
- Graphs are the most general type of relational data
- Similar to a tree, it consists of nodes and edges
- Different from a a tree, a child node may have multiple parent nodes
- A graph contains loops, or multiple paths between two nodes in the graph.

Graph Visualization



Hierarchical graph of the evolution of the UNIX operating system

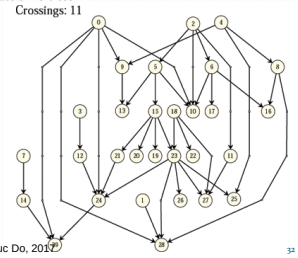
Graph Visualization



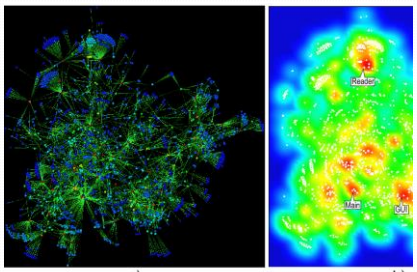
Call Graph using a Force-Directed Layout

Layout of Directed Graphs

- Layering (http://www.csus.yk.ue/staff/NikolaNikolov/#phd)
- Algorithm to calculate the co-ordinate of vertices



Graph Visualization



Force-Directed Layout

Splatting; Dense Representation

Text Visualization and Knowledge Discovery for Text

Why visualize text?

- **Understanding** – read a document
- **Summaries** – get the "key" topic of a document
- **Clustering** – group together similar contents
- **Correlate** – compare patterns in text to those in other data, e.g., correlate with social network

Summarize with Words

Words are (not) nominal?

High dimensional

- High dimensional (10,000+)
- Words have meanings and relations
  - Correlations: *Hong Kong, San Francisco, Bay Area*
  - Order: *April, February, January, June, March, May*
  - Hierarchy, antonyms & synonyms, entities, ...(ontology)

Tips: Tokenization and Stemming

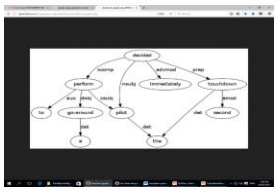
- Tokennizer, Vntokennizer
  - "Thuyền ai thấp_thoáng bên sông",
  - "Đưa câu mái đẩy chạnh_lòng nước non"
    (Hue Fork Song)
- Stemming
- Name Entity Recognition

Bag of Words and Vector 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-term matrix
  - Document vector space model
  - Frequency of word in document

Graph based text representation

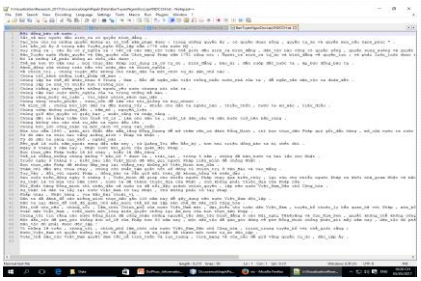
- Frequent words: Vertices
- Link: the co occurrence words in sentence.
- Co occurrence graph
- Preserve the order of words in sentence



Analyzing document: word cloud

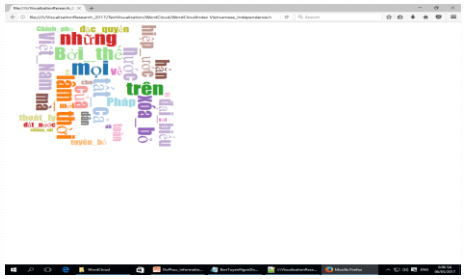
- **Strengths**
  - Can help with initial idea about text
- **Weaknesses**
  - Term frequency may not be meaningful
  - Does not show the structure of the text

Proclamation of Independence of the Democratic Republic of Vietnam



Word Cloud

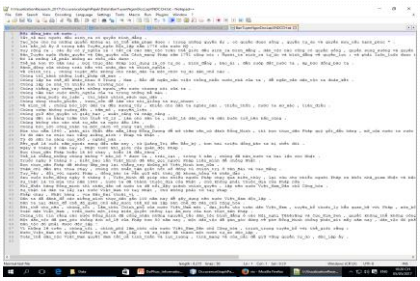
Word Cloud



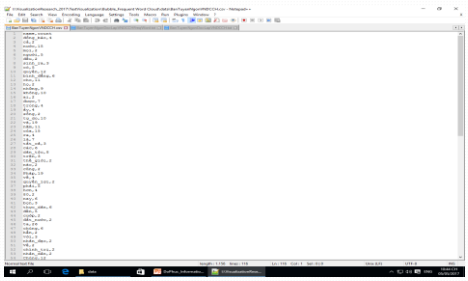
Word Cloud

Analyzing document
Bubble Cloud
Visualization of frequent words

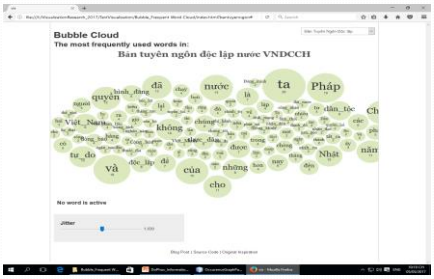
Proclamation of Independence of the Democratic Republic of Vietnam



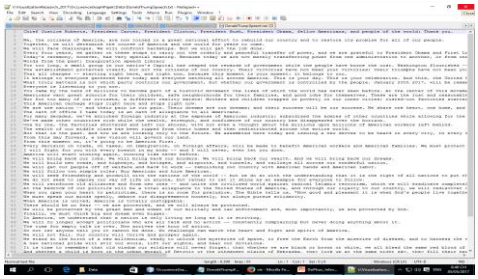
List of frequent words



Bubble Cloud



President Donald Trump inauguration speech



https://qz.com/889611/the-most-frequently-used-words-in-donald-trumps-inauguration-speech-and-every-previous-one/

- **The most frequently used words in Donald Trump's inauguration speech—and every previous one**
- After a bitter contest and controversial win, Donald J. Trump is now officially president of the United States. In his inaugural address, Trump continued the message of populism he maintained throughout his campaign. "We are transferring power from Washington, D.C.," he said, "and giving it back to you, the people." Trump also said he would invest in improving the country's infrastructure, lamented the shuttering of factories, and referred to crime as "American carnage."
- But the word Trump used most in his speech wasn't "people," "power," or even "great." It was "dreams." At one point, Trump said he would "bring back our dreams." Later, when talking about the victims of the aforementioned carnage, he said "their dreams are our dreams."
- We counted the words used in Trump's speech, as well as the words in every inaugural address ever delivered, to find which were used the most. (This list filters out common articles like "the," "and," and "or," as well as some terms that are inherently ubiquitous in inaugural speeches, like "government" and "America.") The result gives us a sense of what each president, going back to 1789, focused on at the start of his term.

List of frequent words

Bubble Cloud

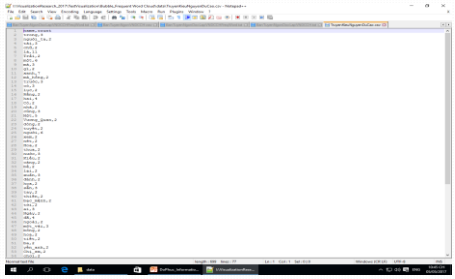
Word Cloud

Word Cloud

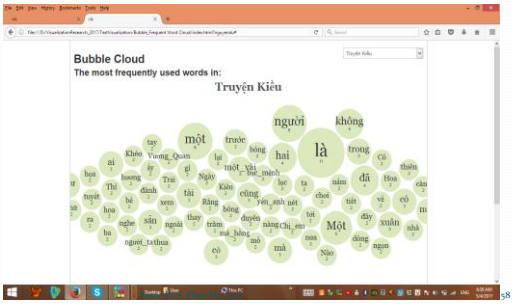
Kieu Legend, Nguyen Du

- Trăm năm trong cõi người ta ,
- Chữ tài chữ mệnh khéo là ghét nhau .
- Trải qua một cuộc bể_dâu ,
- Những điều trông thấy mà đau_đớn lòng .
- Lạ gì bỉ_sắc_tư_phong ,
- Trời xanh quen_thói má_hồng đánh_ghen .
- Cảo thơm lần giở trước đèn ,
- Phong_tình cổ lục còn truyền sử_xanh .
- Rằng năm Gia_Tĩnh triều Minh ,
- Bốn_phương phẳng_lặng , hai kinh vững_vàng .
- Có nhà viên_ngoại họ Vương ,
- Gia_tư nghĩ cũng thường_thường bực trung .
- Một trai con_thứ rốt_lòng ,
- Vương_Quan là chữ , nối dòng nho_gia .
- Đầu_lòng hai ả tố_nga ,
- Thúy_Kiều là chị , em là Thúy_Vân .
- Mai cốt_cách , tuyết tinh_thần ,
- Một người một vẻ , mười_phân_vẹn_mười .
- Vân xem trang_trọng khác vời ,
- Khuôn trăng đầy_đặn , nét ngài nở_nang .

List of Frequent words



Bubble Cloud



Latent Dirichlet Allocation

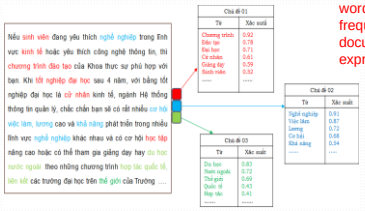
Discover the implicit topics of a corpus

Latent Dirichlet Allocation (LDA)

- Invented by Prof. David Blei, 2003
- Probabilistic Graphical Model (PGM)
- Generative probabilistic modeling
  - Treats data as observations
  - Contains hidden variables
  - Hidden variables reflect thematic structure of the collection.
- Infer hidden structure using posterior inference
  - Discovering topics in the collection.

Using LDA to discover the latent topics of messages

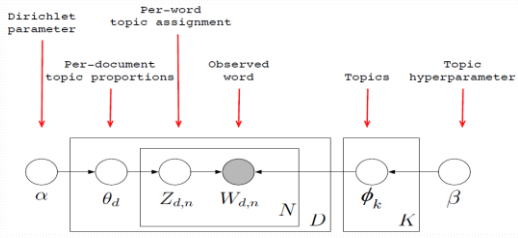
Topic is a group of co-occurrence words occur frequently in documents that express a concept.



LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words.

61

LDA as Graphical Model



$$P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}|\theta_j)P(W_{j,t}|\varphi_{Z_{j,t}}),$$

62

Generating Process of LDA

- Choose $N \sim Poisson(\xi)$
- Choose $\theta \sim Dir(\alpha)$
- For each of the N words w_n
  - Choose a topic $z_n \sim Multinomial(\theta)$
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n, β is a k by n matrix parameterized with the word probabilities.

$$[\beta]_{k\times V} \quad \beta_{ij} = p(w^j = 1 | z^i = 1)$$

Inference and Parameter Estimation

- Parameters:
  - Corpus level: α, β: sampled once in the corpus creating generative model.
  - Document level: θ, z: sampled once to generate a document
- Inference: estimation of document-level parameters.

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

- However, it is intractable to compute, which needs approximate inference.

Gibbs Sampling for LDA

- Gibbs Sampling can be used to sample ϕ and θ They are calculated by.

$$\hat{\phi}_k^{w_d} = \frac{N_k^{w_d} + \beta}{N_k^{(\cdot)} + W\beta} \qquad \hat{\theta}_k^d = \frac{N_k^d + \alpha}{N_{\cdot}^d + K\alpha}$$

ϕ is a probability dist. of topic x word
θ is a probability dist. of doc x topic

65

Latent topics without label

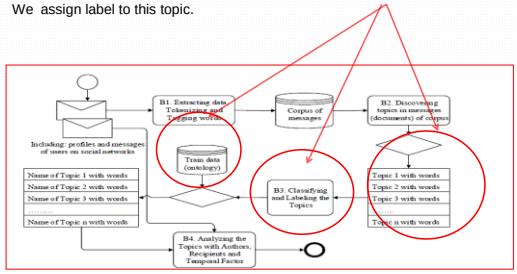
Topic 4		Topic 12		Topic 17	
đoàn	0.03189	phòng	0.03925	công_ty	0.03245
hội	0.01957	quản_lý	0.02322	việc_làm	0.01767
Bác	0.01769	tín_chỉ	0.01933	kiến_tập	0.01485
hoạt_động	0.01689	học_phí	0.01402	nhân_viên	0.01093
đoàn_viên	0.01433	giảng_dạy	0.01021	tham_quan	0.00952
nội_dung	0.01277	thi	0.00967	lương	0.00711
công_tác	0.00677	học_lại	0.00833	người	0.00429
năm_học	0.00597	thời_khoá_biểu	0.00790	môi_trường	0.00641
qui_định	0.00535	phòng_học	0.00718	hoạt_động	0.00500
phong_trào	0.00499	máy_chiếu	0.00554	học_hỏi	0.00359
Topic 2		**Topic 7**		**Topic 14**	
chất_lượng	0.04213	cậu	0.06212	ngoại_ngữ	0.04539
sinh_viên	0.02543	tớ	0.05423	hợp_tác	0.03478
hội_thảo	0.01456	yêu_quí	0.04562	du_học	0.03941
toạ_đàm	0.01213	tình_cảm	0.03412	nước_ngoài	0.03415
kinh_nghiệm	0.01122	tình_yêu	0.002431	học_phí	0.02727
hội_nghị	0.00992	bạn_bè	0.001321	qui_định	0.01325
khoa_học	0.00782	cảm_giác	0.008722	tiến_sĩ	0.00932
nghiên_cứu	0.00612	thầy	0.007890	học_bổng	0.00823
phong_trào	0.00467	cô	0.006532	nhà_nước	0.00763
số_liệu	0.00421	quà_tặng	0.004532	đối_ngoại	0.006349

66

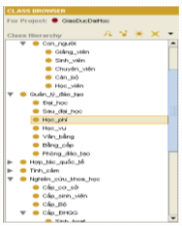
Automatic labeling of topics

Automatic labeling by classification

Latent topic: just a group of words, no topic label
We assign label to this topic.



A topic taxonomy of university



We build a topic taxonomy of university
Then we collect 200-300 messages for each topic
We use LDA to discover the words representing for this topic
SVM is used for classification two sets of words.

Table 1. List of labeled topics and words distribution with probability

Topic 4 "social activity"		Topic 12 "management training"		Topic 17 "recruitment jobs"	
union	0.03189	Office	0.03925	company	0.03245
associate	0.01957	management	0.02322	job	0.01767
Uncle Ho	0.01769	credits	0.01933	fresher	0.01485
activity	0.01689	school fee	0.01402	staff	0.01093
union member	0.01433	teaching	0.01021	visit	0.00952
content	0.01277	test	0.00967	Salary	0.00711
mission	0.00677	leaning	0.00833	people	0.00429
school year	0.00597	schedule	0.00790	environment	0.00641
regulation	0.00535	classroom	0.00718	activity	0.00500
movement	0.00499	projector	0.00554	learning	0.00359

Visualize Themes in a Document Collection

Topical Analysis

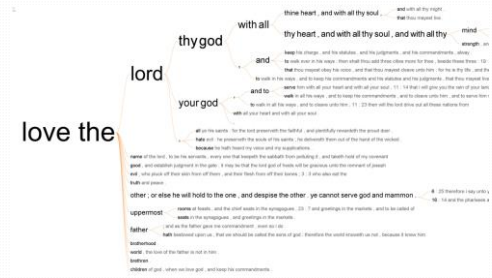
- Large document collections
  - Too large to manually read the source documents
  - Deeper analysis than the most common theme
- Statistical topic modeling
  - Analysis of word relationships
  - Extract *latent topics* belonging to the documents
- Latent Dirichlet Allocation (LDA)

LDA: visualize topics

LDA: visualize topics

Word Map

WordTree (Wattenberg et al)



Biển nhớ, Trịnh Công Sơn

- "Ngày_mai em đi , biển nhớ tên em gọi về",
- "Ngày_mai em đi , đồi_núi nghiêng_nghiêng đợi_chờ","

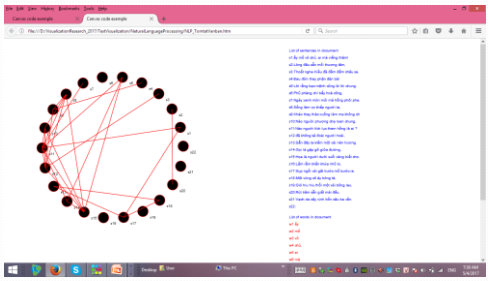
Word Tree

Graph based document summarization

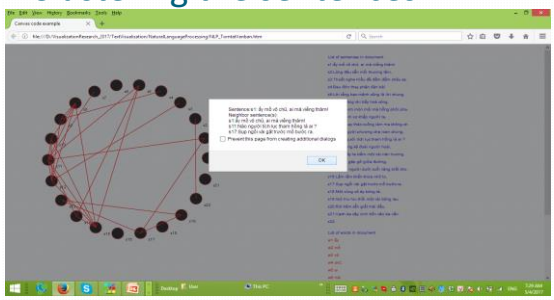
Text Visualization Summary

- High Dimensionality
  - Where possible use **text to represent text**...
  - ... which terms are the most descriptive?
- Context & Semantics
  - Provide **relevant context** to aid understanding.
  - Show (or provide access to) the **source text**.
  - Modeling Abstraction
- Determine your **analysis task**.
  - Understand abstraction of your **language models**.
  - Match analysis task with appropriate tools and models.

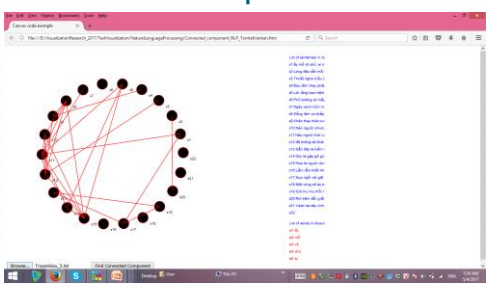
Kieu legend of Nguyễn Du



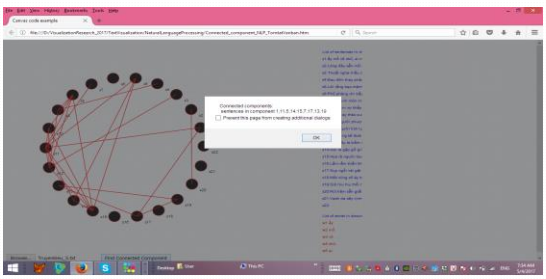
Clustering the sentences



Connected Components



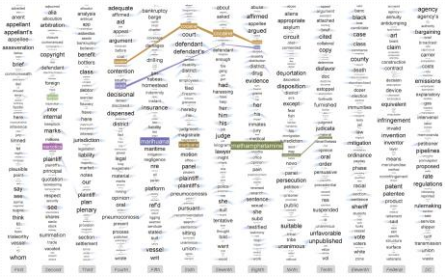
Discover the connected component



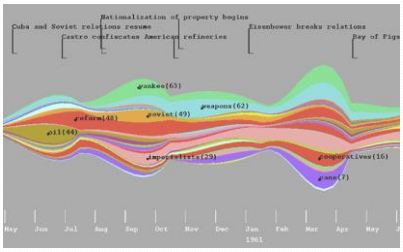
Doc. Similarity & Clustering

- In vector model, compute distance among docs
  - For TF.IDF, typically cosine distance
  - Similarity measure can be used to cluster
- Topic modeling approaches
  - Assume documents are a mixture of topics
  - Topics are (roughly) a set of co-occurring terms
  - Latent Semantic Analysis (LSA): reduce term matrix
  - Latent Dirichlet Allocation (LDA): statistical model

Parallel Tag Clouds [Collins et al 09]



ThemeRiver [Havre et al 99]



What topics they used in their email

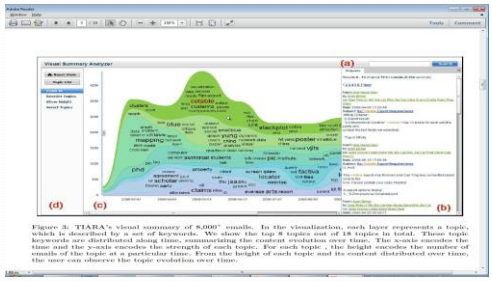


Figure 3: TIARA's visual summary of 8,000 emails. In the visualization, each layer represents a topic, which is described by a set of keywords. We show the top 8 topics out of 18 topics in total. These topic keywords are distributed along time, summarizing the content evolution over time. The x-axis encodes the time and the y-axis encodes the strength of each topic. For each topic , the height encodes the number of emails of the topic at a particular time. From the height of each topic and its content distributed over time, the user can observe the topic evolution over time.

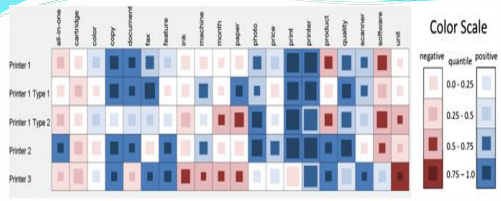
Visualizing Sentiments and Emotions

Sentiment analysis

- Sentiment analysis, also known as opinion mining, is one of the most important text mining tasks and has been widely used for analyzing, for example, reviews or social media data for various of applications, including marketing and customer service.
- The result of sentiment analysis is negative, neutral, positive

Sentiment indexing of Twitter data in a time-series diagram. This figure shows that the public sentiment regarding different events in real-life may change dramatically

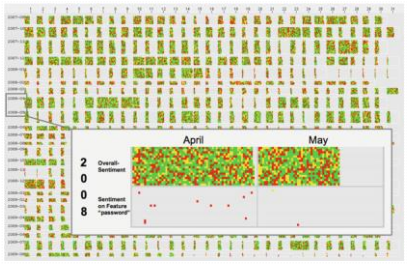
C. Nan and W. Cui, *Introduction to Text Visualization*, Atlantis Briefs in Artificial Intelligence 1, DOI 10.2991/978-94-6239-186-4_6



Summary report of printers: each *row* shows the attribute performances of a specific printer. *Blue color* represents comparatively positive user opinions and *red color* comparatively negative ones (see *color scale*). The size of an *inner rectangle* indicates the amount of customers that commented on an attribute. The larger the rectangle the more comments have been provided by the customers

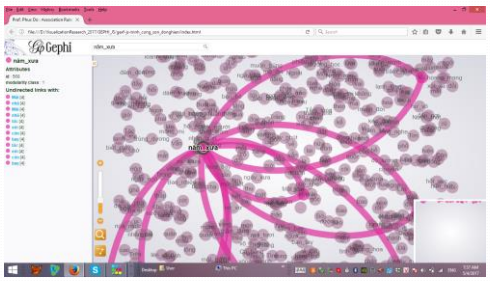
C. Nan and W. Cui, *Introduction to Text Visualization*, Atlantis Briefs in Artificial Intelligence 1, DOI 10.2991/978-94-6239-186-4_6

Pixel map calendar: each pixel indicates a document with the color encoding its overall sentiment, i.e., the average of all contained feature sentiments. Here, *red* indicates negative, *green* indicates positive, and *yellow* indicates neutral. In the background, the x-axis bins are days and y-axis bins are years with months

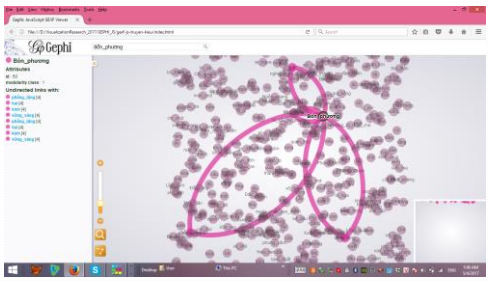


Gephi Viewer

Trịnh Công Sơn
Co-occurrence Word Graph

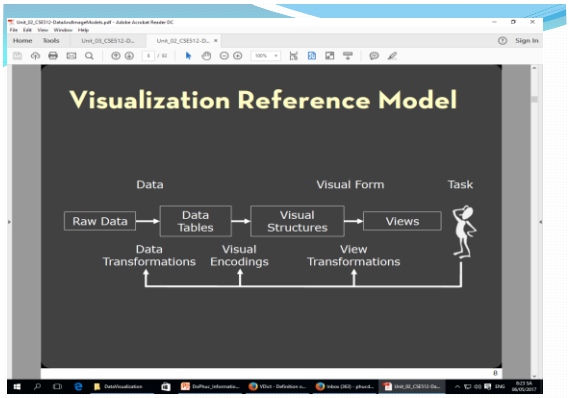


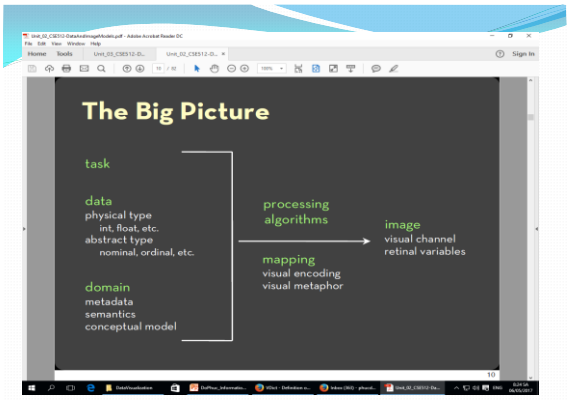
Kiều Legend, Nguyễn Du
Co-occurrence Word Graph



Visualization Design

Jeffrey Heer University of Washington, 2015





Formalizing Design (Mackinlay 1986)

Visualization components

- Color
- Size
- Texture
- Proximity
- Annotation
- Interactivity
  - Selection / Filtering
  - Zoom
  - Animation

Visual Encoding Variables

- Position
- Size
- Value
- Texture
- Color
- Orientation
- Shape
- Others?

Choosing Visual Encodings

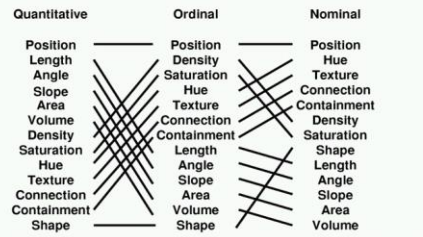
- **Challenge:**
- Assume 8 visual encodings and *n* data attributes.
- We would like to pick the "best" encoding among a
- combinatorial set of possibilities with size (*n*+1)8
- **Principle of Consistency:**
- The properties of the image (visual variables) should
- match the properties of the data.
- **Principle of Importance Ordering:**
- Encode the most important information in the most
- effective way.

Design Criteria (Mackinlay)

- **Expressiveness**
- A set of facts is expressible in a visual language if the
- sentences (i.e. the visualizations) in the language
- express all the facts in the set of data, and only the
- facts in the data.
- **Effectiveness**
- A visualization is more effective than another
- visualization if the information conveyed by one
- visualization is more readily perceived than the
- information in the other visualization.

Channel ranking varies by data type

spatial position best for all types



[Mackinlay, Automating the Design of Graphical Presentations of Relational Information, ACM TOG 5:2, 1986] 12

Visualization software resources

Visualization software

- Host language (C/C++/Java/Python) plus OpenGL
- Stat/math package with graphics
  - R
  - MATLAB
- Special-purpose info visualization software
  - Earth mapping, biological network visualization, etc.
- Browser-enabled graphics/info visualization packages
  - Google Charts
  - Processing / Processing.js
  - D3
  - Java + Flash

Hands-on

- HTML intro*
- Google charts
- D3

Resources

- Books
  - Visual Complexity, Mapping Patterns of Information , Manuel Lima
  - The Visual Display of Quantitative Information, Edward Tufte
  - Information Visualization: Beyond the Horizon, Chaomei Chen
  - JavaScript: The Definitive Guide, David Flanagan
  - Getting Started with D3, Mike Dewar
  - Visualizing Data, Ben Fry
  - Interactive Data Visualization for the Web, Scott Murray
- Websites
  - http://processingjs.org/
  - http://d3js.org/, https://github.com/mbostock/d3/wiki/API-Reference
  - http://code.google.com/apis/ajax/playground/
  - http://www.edwardtufte.com/tufte/
  - http://www.visualcomplexity.com/
  - http://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/

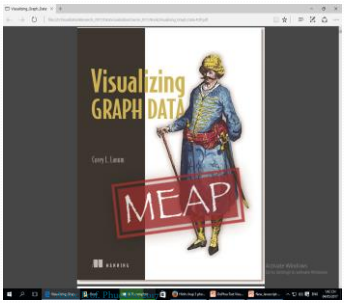
Conclusion

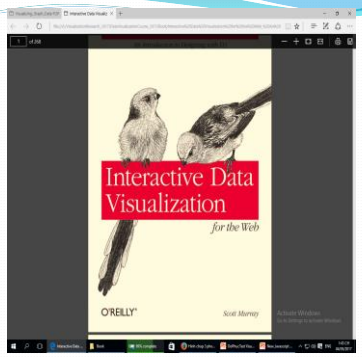
- Data Visualization , Text visualization are effective tools to visualize and discover knowledge from data and text.
- Big data is a motion to promote knowledge discovery and data visualization is a good tool to support big data discovery
- Research in big data analysis is concurrently process with data and text visualization

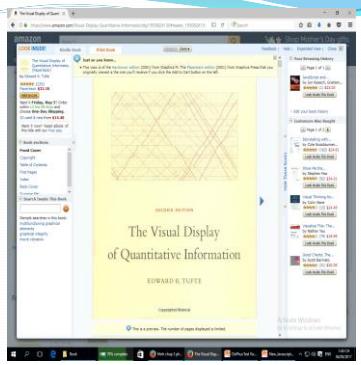
Some Challenges in data and text Visualization

- Interactive Visualization of Very Large Graphs
- Develop novel interactive visualizations that are tightly coupled with automatic natural language processing methods to enable human users to explore their data.
- Develop Real-Time Visualization of Streaming Text
- How interactive visualization can assist investigative analysis for large corpus
- Optimizing the facilities of hardware and software for big data visualization and discovery.

References







Short Bio of Phuc Do

- Phuc Do is an Associate Professor of Information Systems at the University of Information Technology, Vietnam National University of Ho Chi Minh City.
- He currently works mostly on knowledge discovery technologies and their application in social network analysis by using topic modeling .
- Phuc Do has written about 80 research papers and books. He has been the Principal Investigator of several projects in Knowledge Discovery From Data, Text Analysis and Summarization, Topic Based Social Network Analysis, Citation Network Analysis.

Thank you for your attention!