Text Visualization

Naveed Kamran, Jan Kalkan
AIM3 - Scalable Data Science
DIMA / TU Berlin
16.6.2017

Agenda

- Introduction
- Taxonomy of Text Visualization Techniques
- Text Visualization Browser
- Presentation of Selected Text Visualization Techniques
- Conclusion

Introduction

- Motivation: Increasing availability of text data in various forms:
  - e.g. web pages, blogs, publications etc.
- Challenge:
  - How to discover and retrieve useful information without going through all the sources?
  - Semi-structured nature of text
  - Understanding often requires world knowledge and interpretation
- Solution: Text Visualization
  - Combine the efficiency of machines to analyze and process large data sets with human creativity and intuition

Taxonomy of text classification techniques

1. Visualizing Document Similarity
    a. Projection Oriented Techniques
    b. Semantic Oriented Techniques
2. Revealing Text Content
    a. Summarizing a Single Document
    b. Showing Content at the Word Level
    c. Visualizing Topics
    d. Showing Events and Storyline
3. Visualizing Sentiments and Emotions
4. Document Exploration Techniques
    a. Distortion Based Approaches
    b. Exploration Based on Document Similarity
    c. Hierarchical Document Exploration
    d. Search and Query Based Approaches

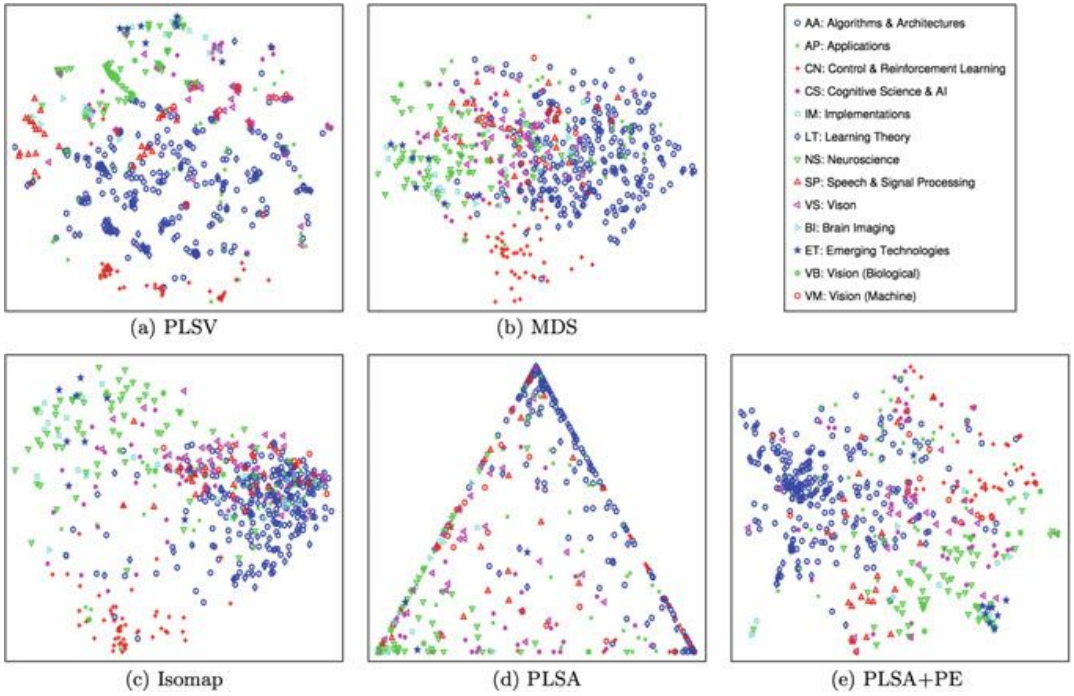
1. Visualizing Document Similarity

- Traditional technique for summarizing document collections.
- Documents visualized as points on a low dimensional (2D or 3D) visualization pane.
- The distance between each pair of points represents the similarities between the corresponding two documents.
- Two techniques:
  a. Projection Oriented Techniques
     - Visualize through a **dimension reduction procedure**
     - Document is represented as a bag of words and formally described by an **N-dimensional feature vector**.
  b. Semantic Oriented Techniques
     - Represent the document similarity via latent topics extracted from text data

1. Visualizing Document Similarity Example

Visualizations of the same data:

- Visualizing document similarity based on semantic oriented techniques (a, d, e) and nonlinear projection (b, c).
- Each point in the diagram is a document colored by their primary topics that are extracted based on topic modeling.
- The distances between documents encode their similarities, following the rule of “the more similar, the closer”.



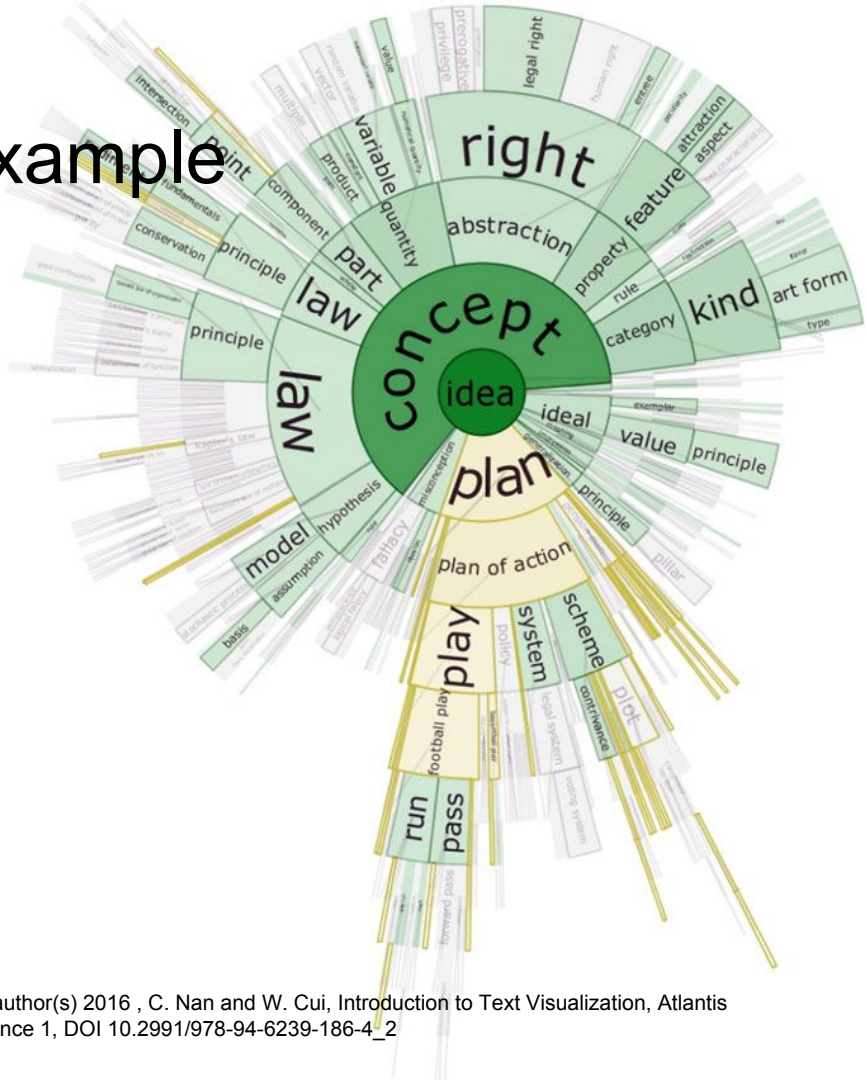
2. Revealing Text Content

Showing the content of the documents from different aspects and at different level of details.

- **Summarizing a Single Document** - e.g DocBurst decompose into a tree via its innate structures such as sections, paragraphs, and sentences
- **Showing Content at the Word Level** - e.g TagCloud summarizes input text data in a cloud form where words with font size indicate their importance are packed together without any overlap.
- **Visualizing Topics** - (1) summarize and explore static topic information, (2) illustrate the topic dynamics over time, (3) help with topic comparison, and (4) illustrate events and storylines
- **Showing Events and Storyline** -e.g, LifeFlow and Outflow aggregate multiple event sequences into tree or graph visualizations and provide users with highly scalable overviews

2. Revealing Text Content Example

- DocuBurst visualization of a science textbook rooted at idea.
- A search query for words starting with pl.
- Nodes matching the query are highlighted in gold



2. Revealing Text Content Example

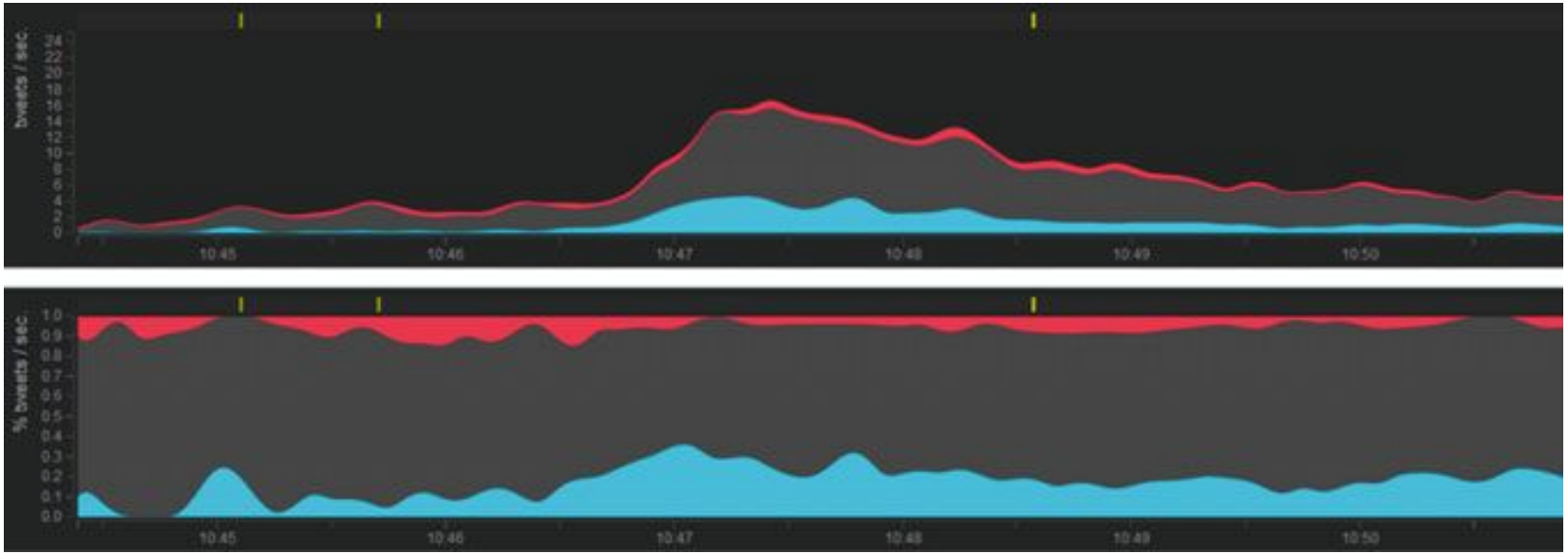
1. TagCloud is a techniques for visualizing words.
2. It illustrates a bag of words that summarize the content of the input text data in a cloud form, in which **words, with font size indicating their importance, are packed together without any overlap**

Wordle visualization of a bag of words extracted from text data.

3. Visualizing Sentiments and Emotions

- Illustrates the change of sentiments over time
  - regarding to a given streaming text corpus such as news corpus, review comments, and Twitter streams
- Techniques have been introduced to illustrate and interpret the **sentiment dynamics from different perspectives**

3. Visualizing Sentiments and Emotions Example



Sentiment stream graphs for the keyword search Flacco, the Super BowlMVP in a Twitter dataset using Agave. Negative is red, neutral is gray, and positive is blue. Top overall frequency of tweets, divided by sentiment type. Bottom sentiment as percent of overall volume

4. Document Exploration Techniques

Explore the data to find useful information or insightful data patterns

- **Distortion Based Approaches**
  - e.g Document Lens, focus+context design inspired by the magnifier lens.
  - The focused content of a document is shown in details in the view center, surrounded by other parts of the content that provides an overall impression of the text data
- **Exploration Based on Document Similarity**
  - e.g, InfoSky, I N-SPIRE, and ForceSPIRE, use an overview that summarizes the entire document collection
  - based on document similarities and employs a multiple coordinated view to reveal the document details from various aspects such as keywords and topics.

4. Document Exploration Techniques

- **Hierarchical Document Exploration**
  - Documents are hierarchically clustered based on their topic similarity
  - The cluster results are shown in a tree view to guide the data navigation
- **Search and Query Based Approaches**
  - transforms search results into a visual representation to illustrate the insight of content relationships among documents.
  - Graph layout and projection-based approach are commonly used to represent the search and query results showing the relationships among documents or text snippets.

Text Visualization Browser

Text Visualization Browser is a Visual Survey of Text Visualization Techniques

http://textvis.lnu.se/

Selected Text Visualizations

1. Prefix Tag Clouds

- A **tag cloud** (or **word cloud**) is a visual representation of text data to quickly perceive the most prominent terms
- A tag is usually one word and the importance is visualized by the font size
- Major **limitation** is that they do not relate different word forms but treat every form as an individual tag
  - E.g. Network ← → Networks
  - ⇒ Waste of screen space, bad comprehension
- **Solution**: Prefix Tag Clouds address this limitation
- Main application area:
  - Former: Indexing method in community oriented websites
  - Nowadays: **Text summarization**

Source: Michael Burch et al., Prefix Tag Clouds 2013

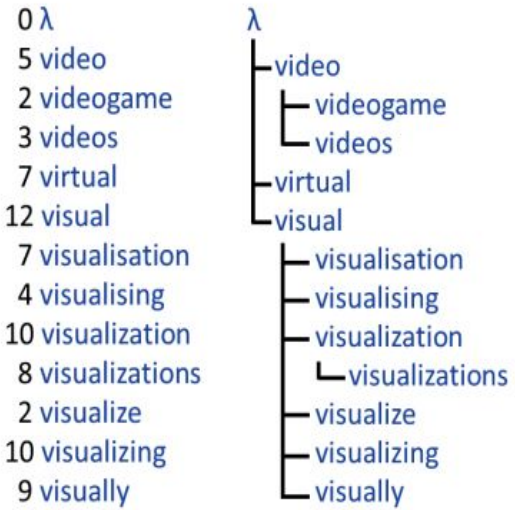
Tag cloud showing most frequent words of all publication titles containing the string “visual”

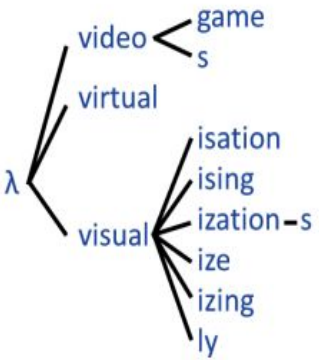
Prefix Tag Clouds (2) - Prefix tree

- A prefix tree is used to group different word forms together

Creation:

1. Adding the empty string λ, which serves as the root element
2. Ordering lexicographically
3. Generating the prefix tree from the ordered list



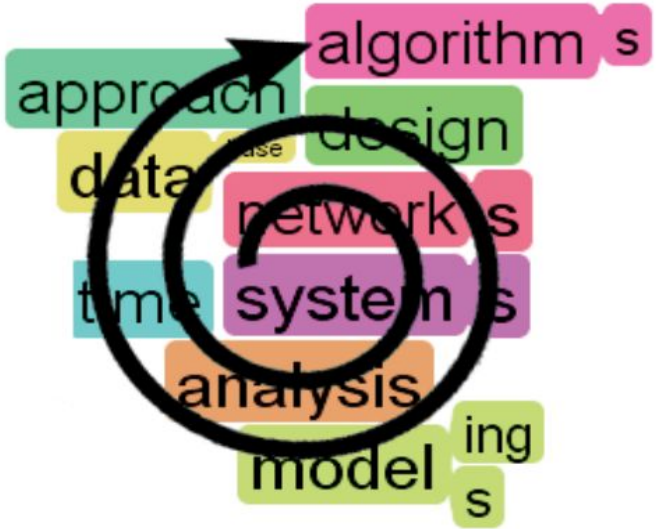


Prefix Tag Clouds (3) - Subtree Rendering and Cloud Generation

1. Splitting the tree into subtrees
2. Visualizing subtrees are in a circular tag cloud layout with the most frequent tags in the center and tags with decreasing frequencies towards the boundary

⇒ Layout supports well the identification of popular tags

⇒ Good spatial utilization



Source: Michael Burch et al., Prefix Tag Clouds 2013

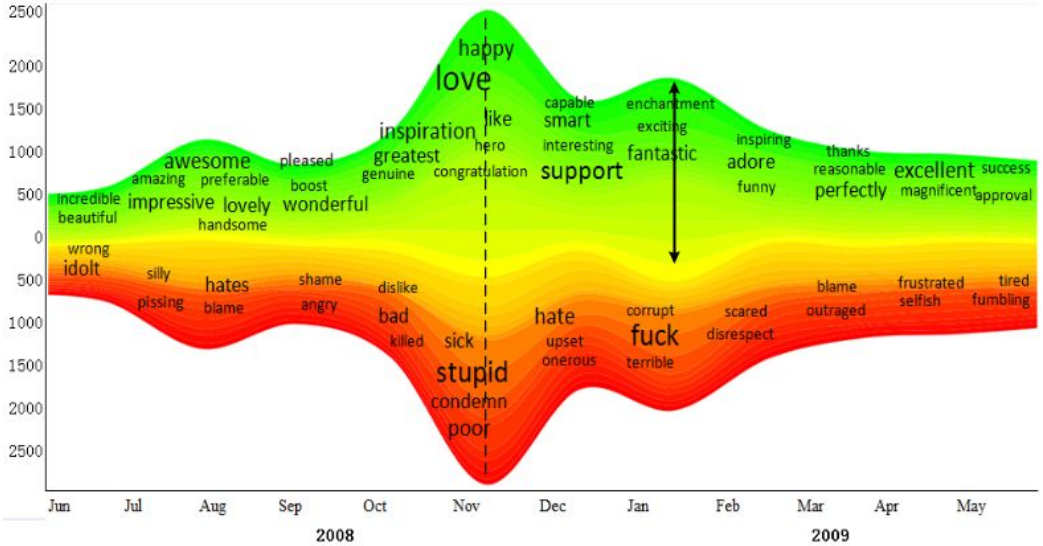
Spiral placement of the prefix subtrees

Prefix Tag Clouds (4) - Summary

- Tag cloud variant that makes use of prefix trees to deal with word forms
- Word forms are visualized by colour
- Easy to understand for the viewer
- Mainly applicable for text summarization

2. Visualizing Sentiment Evolvement for Tweet Events

- Motivation: rising interest in sentiment classification
- Challenge: large tweet streams stop people from reading the whole classified list to understand the insights
- ⇒ Usage of 2 co-trained classifiers
- ⇒ Usage of a “river” graph to visualize the intensity and evolvement of sentiment



Source: Shenghua Liu et al., Co-training and Visualizing Sentiment Evolvement for Tweet Events. 2013

Visualization of about “Obama Election”

Co-Training of Classifiers (1)

- Manually labeling a large number of tweets is a labor-intensive task
  - ⇒ Use of semi-supervised method to utilize the unlabeled tweets to boost performance
- Since tweets are extremely short, it is necessary to extract more features than the textual ones
  - ⇒ Co-training of 2 classifiers
    - Text-view vs. non-text view

Overview

1. Train 2 classifiers C1, C2 on a common set of labeled tweets L in a time period t1
2. Select the confident ones to augment the labeled set
3. We select the p positive ones and the n negative ones when the classifiers agree most (iteratively executed)
4. Finally we obtain weights for each tweet, which denote the probabilities that tweet t belongs to class positive, negative or neutral

Co-Training of Classifiers (2)

Text view based features

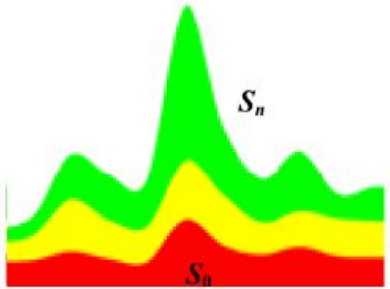
- Identification of sentiment word via sentiment dictionary
- Compute the textual feature for each sentiment word (e.g. via word net affect)

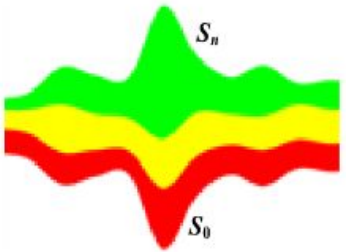
Non-text view based features

- Emoticons
  - E.g. :) :-) (>_<) etc.
- Temporal features
  - Classifying the post time into different hours, dates, day of week and months as temporal features
- Punctuation
  - Exclamation mark (!), question mark (?), ⇒ expression of the emotional intensity

Sentiment Visualization

- Define density function ρ_i as the distribution of the number of tweets belonging to sentiment class i
- The upper boundary of the bottom curve function of the graph S_0 is the lower boundary of the curve of graph S_1
- Further visualization aspects:
  - Graph geometry, layer ordering, coloring, sentiment word labels





Source: Shenghua Liu et al., Co-training and Visualizing Sentiment Evolvement for Tweet Events. 2013

Visualizing Article Similarities in Wikipedia

- Visualization providing insight about similarities among Wikipedia articles in terms of structure as well as content
- Usage of Vector Space Model incl.:
  - Weighting via Tf-Idf
  - Stemming - *e.g. "fishing", "fished", "fisher" ⇒ "fish"*
  - Stop word removal - *e.g. is, that, at etc.*

⇒ Representation of documents as high-dimensional **vector**

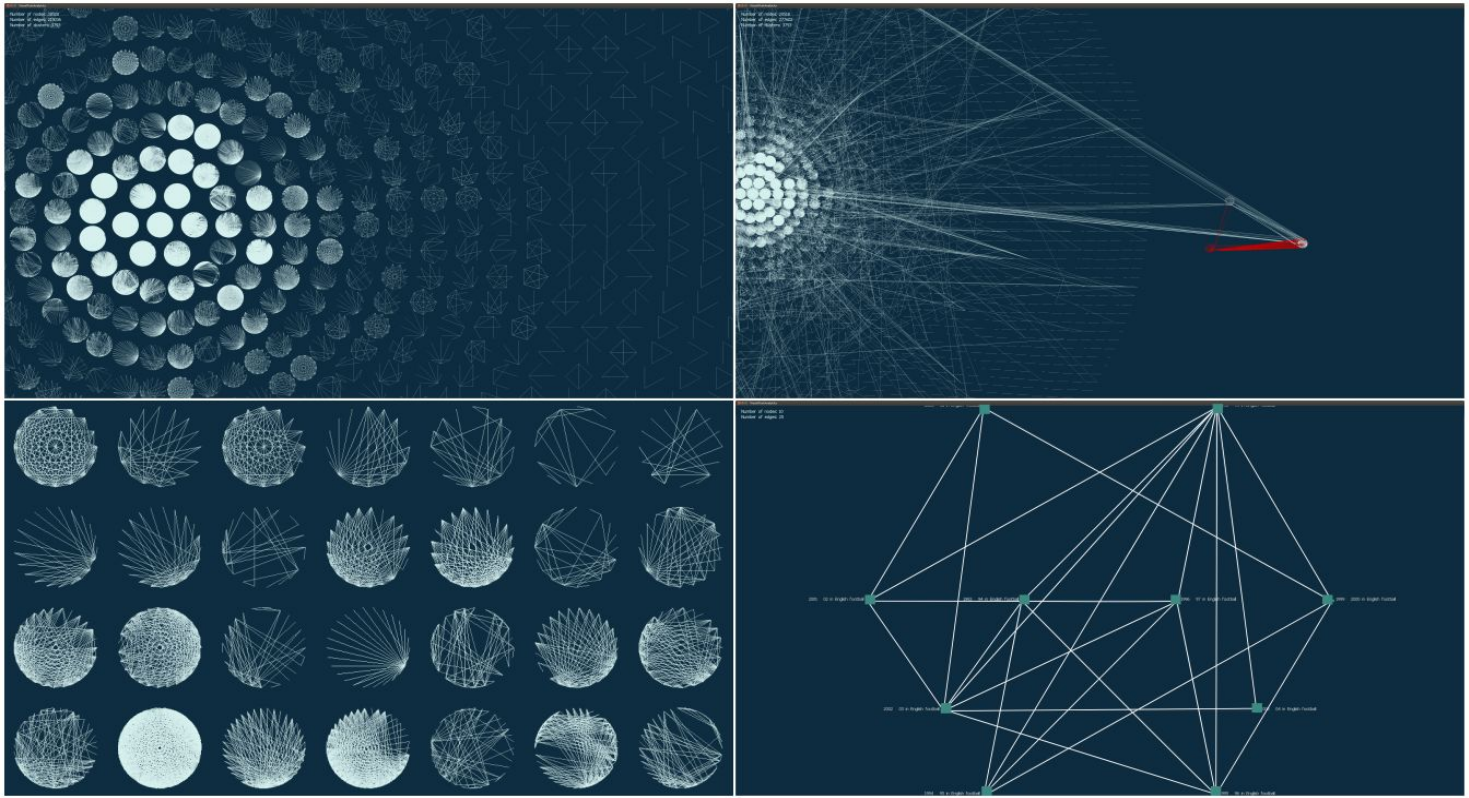
- Data was gathered and processed via a **pairwise comparison** of all Wikipedia articles
  - Cosine-Similarity measure used for semantic similarity estimation

Similarity Graph

- Structure:
  - Documents as Vertices
  - Similarities as Edges
- Edges are only drawn if similarity exceeds a defined threshold

⇒ Resulting graph consists of a large number of graph components

Visualizing Article Similarities in Wikipedia



Source: Riehmann et al., Visualizing Article Similarities in Wikipedia. Poster Abstracts of the Eurographics Conference on Visualization, 2016

Visualizing Article Similarities in Wikipedia

- In general intuitive visualization, but doesn’t scale very good
  - => Very large graph might be hard to explore
- Too static at the moment:
  - Changing the threshold leads to huge graph shifts
  - What is the right threshold then?
- The used similarity measures tend to focus more on structural similarity such as subgraphs of cities, countries, years, events, etc.
- Possibilities to improve the similarity recognition of the running text:
  - Data cleaning
  - Similarity measure weighting
  - Applying different similarity measures concurrently

3. Storylines

- Visual exploration and analysis in latent semantic spaces
- To explore and study a body of unstructured text without prior knowledge of its thematic structure.
- Integrates
  - latent semantic indexing
  - natural language processing
  - and social network analysis.
- The contributions of the work include
  - providing an intuitive and directly accessible representation of a latent semantic space derived from the text corpus
  - an integrated process for identifying salient lines of stories
  - coordinated visualizations across a spectrum of perspectives in terms of people, locations, and events involved in each story line
- Approach to explicitly represent the latent semantic dimensions in order to track the main topics in source data

Storylines - How it works

1. Story generation
    - visualize the entire dataset and provide an overview for exploratory analysis.
    - Information Seeking Mantra: overviews first, filter, and details on demand.
2. Social network analysis

4. Lexichrome

- Color evokes a response, describes an object, and shapes a culture
- Colors tell stories about those who exploit them, and invite emotional responses from those who see them.
- Emotional color associations are widely used in brand design

Coca Cola’s company mission statement results in a mix of colors, led by white, with their company signature color red being sixth Positive terms such as happiness, accountable, and passion all associate with red and appear in the Coca-Cola brand statement.

Source: http://www.coca-cola.com/global/glp.html

Lexichrome

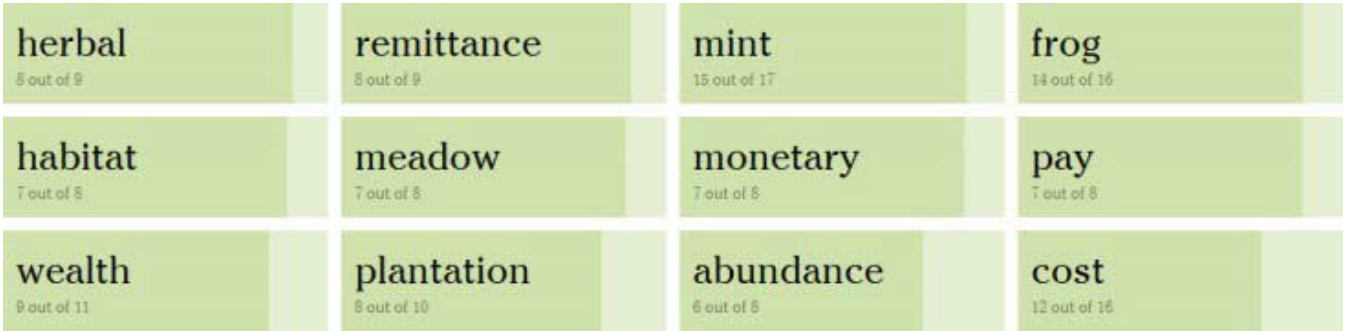
- An interactive culmination of numerous visualization techniques that seeks to bridge the gap between lexical semantics and popular notion of colors.
- Comprehensive lexicon of word-color associations, and further contextualize the implications of this dataset by “fingerprinting” their own texts.
- Applicable to a range of academic and creative personnel, including literary linguists, corporate brand managers, and writers;

Lexichrome Features

Visitors can explore the application's database in two ways:

- Interacting with an eleven-color palette
- searching for a particular word

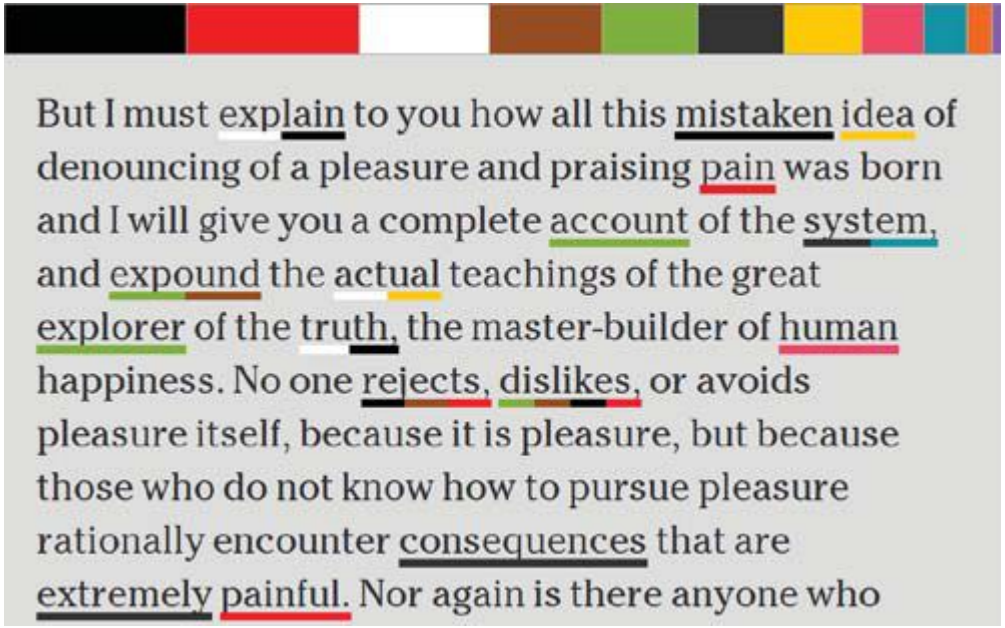
Source: Lexichrome: Examining Word-Color Associations with Visualization, Chris K. Kim, Christopher Collins



Source:Lexichrome: Examining Word-Color Associations with Visualization, Chris K. Kim, Christopher Collins

Lexichrome

Lexichrome has an ability to process user-provided texts and extract the relevant colors.



"Chromatic fingerprint" of a sample text based on word-color associations

Source:Lexichrome: Examining Word-Color Associations with Visualization, Chris K. Kim, Christopher Collins

Conclusion / Summary

- Combine the efficiency of computers with the intuition of humans
- Many approaches for text visualization but all approaches solve specific problems
  - None is perfect for general purpose
  - Different solutions depending upon the nature of data and output format
- Text visualizer browser is a nice tool that gives an overview and links to the all details of current and past research in the field.
- The usefulness of some tools is questionable especially in case of huge data sets, e.g. very large graphs
- Works needs to be done to combine all research into a single tool for better understanding and output

Thanks for your attention

References

- Michael Burch, Steffen Lohmann, Daniel Pompe, and Daniel Weiskopf. Prefix Tag Clouds. Proceedings of the International Conference on Information Visualisation (IV), pp. 45-50, 2013. - http://www.vis.uni-stuttgart.de/uploads/tx_vispublications/PrefixTagClouds-IV2013.pdf
- Shenghua Liu, Wenjun Zhu, Ning Xu, Fangtao Li, Xue-qi Cheng, Yue Liu, and Yuanzhuo Wang. Co-training and Visualizing Sentiment Evolvement for Tweet Events. Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion), pp. 105-106, 2013 - http://dx.doi.org/10.1145/2487788.2487836
- Patrick Riehmann, Martin Potthast, Henning Gruendl, Johannes Kiesel, Dean Jürges, Giuliano Castiglia, Bagrat Ter-Akopyan, and Bernd Froehlich. Visualizing Article Similarities in Wikipedia. Poster Abstracts of the Eurographics Conference on Visualization (EuroVis), pp. 69-71, 2016 -http://www.uni-weimar.de/medien/webis/publications/papers/potthast_2016a.pdf
- Lexichrome: Examining Word-Color Associations with Visualization, Chris K. Kim, Christopher Collins
- Storylines: Visual exploration and analysis in latent semantic spaces, Weizhong Zhu, Chaomei Chen, College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA, Science direct, Computers & Graphics 31 (2007) 338–349
- Coca cola logo has been taken from CocaCola website: http://www.coca-cola.com/global/glp.html