How we draw texts

A literature review on text visualization tools*

by Jaume Nualart

Canberra, July 2013

PhD student at Faculty of Arts and Design. University of Canberra

Contents

1. Introduction

1.1. About this research

1.1.1. Kind of research:

This literature review is part of a longer research project. It is the beginning of a PhD in Communication at the University of Canberra, Faculty of Arts and Design. The PhD is supervised by Assoc. Prof. Mitchell Whitelaw.

Australia and UK are pioneers in so-called *practice based and practice-led research* in several official Master and PhD degrees programs (Candy and Studios 2006), especially in a wide range of humanities studies; from creative writing to performance, exhibitions, cinema, art craft and painting (Webb [2012]).

The reasons for conducting a creative production research are:

1. The researcher has a scholarship with the National Information and Communication Technology of Australia (NICTA). The scholarship has a project agreement that defines a list of milestones and deliverables with the project's partner, UC (see B.1). The scholarship's milestones and deliverables include a report on state-of-the-art text visualization tools, software, and documentation.

2. The researcher is an experimental software developer who for the last decade has been coding free licensed software.

3. The researcher already has productive experience using the proposed methodology.

1.1.2. Proposed methodology

The proposed methodology has been applied by the author as a personal creative strategy for the first time in the extinct "Ludwig Boltzmann Institute. Center for media art research" (Linz. Austria) where he was a researcher of the visualization team in 2008-09. working with Dietmar Offenhuber, Moritz Stefaner, Evelyn Münster, Mar Canet & Sandor Herramhof. We presented several relevant data and visualization tools in the XXXth Ars Electronica exhibition (Lubwig Bolzman Institute [2008]).

The research starts with an inspiring tour through the history and trends of text visualization tools according to the aims listed in the next subsection (1.1.3).

Based on previous practical work, the proposed methodology is based on the following three principles:

1. Work on practical problems, involving challenges to be confronted in real scenarios. Focusing on solving or improving specific tasks clarifies the path to the solution.

Therefore there are possible collaborations with two research groups: the Machine Learning Research Group, from NICTA and specially with Principal Researcher Wray Buntine (NICTA [2013]); and the Lens.org project (cambia.org [2013]), a group that works with large amounts of texts from patents. Nowadays patent texts is an active field of research (Yang et al. [2008]). The text of patents is at once structured and unstructured. Structured because it follows quite a strict format, including: title, abstract, claims/subclaims tree, very detailed figures, inventors, applicants, patent owner, and one or more classifications codes. Unstructured because the texts are heavy legal texts with very long sentences in which it can be difficult to find dependencies and self references (Sheremetyeva, 2003).

2. Design visualization tools as a top-down task. Top-down approach starts with the big picture. It breaks down from there into smaller segments (Wikipedia, 2013b). This methodology advocates for a top-down design starting where the problem has been detected: the human level. Since there are many technical options for how to develop a piece of software, there is no need to care about it from the very beginning. The choice of a computer language or a software at the first stage can limit the freedom of creativity and, consequently, the result of the research project.

3. One task, one visualization. But: one dataset, multiple visualizations. Since data is the blood of the visualization's body, a good dataset can be used for many purposes, so each purpose may require a different visualization tool in order for each tool to serve the task effectively. Of course, one task can have several visualizations that can solve or help to solve it, but, in a real case, a specific visualization is finally chosen for a specific task. An example of this is the web site Manyeyes [http://www-958.ibm.com], that allows anyone to create a visualization from a dataset

We aim to develop these methodologies in greater depth as the project goes on.

1.1.3. Early research question: how can data visualization tools help people working with texts?

In this early work we are not presenting a concrete question that will drive this PhD research. Instead this document represents a starting point in the process of developing a question and shaping the work of the project The final question will form a part of the research proposal, to be presented after this document is completed.

Despite that, at this point, it is possible to define what we understand by the general afirmation: "help people working with data". Sometimes data visualization is the only way to show evidence. The classic example of this was demonstrated in 1973 by the statistician Francis Anscombe (see 1.1). Anscombe's example shows how even simple statistics can be confusing and, in some cases, only the visual representation of data can show significant features.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

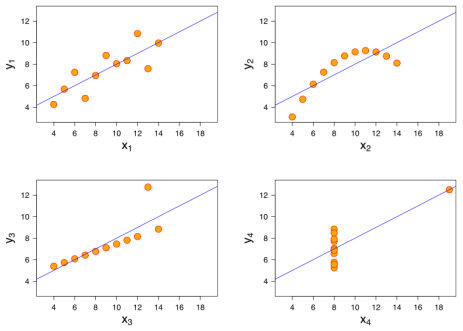

Figure 1.1.: Anscombe's quartet: all four sets are identical when examined using simple summary statistics, but vary considerably when graphed. [image from wikipedia http://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Anscombe%27s_quartet_3.svg/425px-Anscombe%27s_quartet_3.svg.png]

The aim of this research is to "help people working with data" using visualization techniques that can show patterns, behaviors and/or evidence of the reality represented by data, improving the way, the speed or the clarity with which the facts under the data are shown or discovered.

In this document we are not trying to include every single text visualization tool. This would be a never ending project. This is why, at this point, we can outline initial areas of focus. The points we have considered when analyzing text visualization tools for this study are:

- Individual document representation, in particular ways to extract meaning from texts based on writing style, document structure, and language register instead of pure statistics. We are interested in representing the meaning and salient features of texts because a convenient visualization of texts can speed up and/or improve our ability to select texts and manage the time to tackle them intellectually. The research output of fields like natural language processing, linguistics computing, and machine learning offer techniques for producing high quality data representing complex texts. We propose that combining these techniques with suitable visualization can improve the way we examine and understand texts.

- Representation and exploration of collections of texts. Ways to explore, and select texts, and navigating and analyzing collections of texts is a daily task for a lot of people working with computers and data and there is a lot of room for new tools and ideas. Information retrieval is a critical factor in an environment characterized by excess of information (Baeza-Yates et al., 1999). When a user conducts a search, the information retrieval systems normally respond with a list of results. In many cases, the presentation of those results play an important role in satisfying user information needs. A bad or inadequate presentation can hinder the satisfaction

of information needs (Baeza-Yates et al., 2011). Typically, information retrieval systems present the results of a query in flat, one dimensional lists. Usually, these lists are opaque in terms of order, i.e. the users do not know why the list has a particular order. To refine their search, the users have to interact again, normally by filtering the first output of results.We propose that new techniques for representing text collections - including search results - can contribute to improvements in navigation, exploration and retrieval.

1.2. Datavis as an academic field

1.2.1. How new is data visualization as an academic field?

here is no a consensus on the date of birth for data visualization. Probably the most well documented, popular, and complete time-line of data visualization is the "Milestones Project" where Michael Friendly and Daniel J. Denis (Friendly, M and Denis, D. J., 2001d) show visualizations from 6200 B.C. beginning with the "Konya town map" - considered the "oldest known map".

Tufte (Tufte and Graves-Morris 1983) in "The visual display of quantitative information" presents works from ancient Greeks and pictures in caves from neolithic age. Since there we can find data representation in every single culture and epoch.

Therefore data visualization, understood as extraction of data from reality and shaped in a two or three-dimensional abstract representation, can be considered as part of human culture.

Today data visualization is a consolidated academic field.(Strecker and International Development Research Centre (IDRC) [2012]). Below we present a short list of the main references for this relatively new field:

- Seven out of top the ten universities, in the Times Higher Education ranking (TSL Education Ltd. [2012]), have departments or research groups related to data visualization. Data visualization has developed in a wide range of departments, from computer science or statistics to linguistics or graphical design, and from chemistry or physics to genetics or history. Recently data visualization has emerged as a distinct field, with masters programs and departments dedicated to it (see table 1.1).

institution	rank 2012	Department/Course	URL
Harvard University	1	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
Massachusetts Institute of Technology	2	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
University of Cambridge	3	—	—
Stanford University	4	Stanford Vis Group	http://vis.stanford.edu/
University of California, Berkeley	5	VisualizationLab	http://vis.berkeley.edu/
University of Oxford	6	Visual Informatics Lab at Oxford	http://oxvii.wordpress.com/
Princeton University	7	PrincetonVisLab	http://www.princeton.edu/researchcomputing/vis-lab
University of Tokyo	8	—	—
University of California, Los Angeles	9	IDRE GIS and visualization	https://idre.ucla.edu/visualization
Yale University	10	—	—

Table 1.1.: Top universities and data visualization departments

- Conferences in the last five years ordered by number of participants and mainly dedicated to data visualization (see table 1.2).

Conference	Place	Topic	Participants	URL
NICAR 2013	USA	data journalism	149	http://ire.org/conferences/nicar-2013/
dd4d 2009	France	information visualization	52	http://www.dd4d.net
FutureEverything 2013	UK	Technology / society / art	52	http://futureeverything.org/
resonate 2013	UK	Creative code	44	http://www.thisisresonate.co.uk/resonate-13/
graphical web 2012	Switzerland	open web / datavis	38	http://www.graphicalweb.org/2012/
IEEEVis - VisWeek 2012	USA	information visualization	-	http://ieeevis.org/
EuroVis 2013	Germany	Computational Aesthetics	-	http://www.eurovis2013.de
Siggraph 2013	USA	computer graphics and interactive techniques	-	http://s2013.siggraph.org
OzViz 2012	Australia & NZ	workshops for visualisation practitioners, academics and researchers		http://www.ozviz2012.org/

Table 1.2.: Conferences dedicated mainly to visualization ordered by participants (source: Stefaner, M. [2013])

- A part from some dedicated journals, the most important contributions on data visualization are found in conference proceedings (see table 1.3).

Name	URL
Parsons Journal for Information Mapping	http://pjim.newschool.edu/issues/index.php
Journal of Visualization	http://springer.com/materials/mechanics/journal/12650
IEEE Transactions on Visualization and Computer Graphics (TVCG)	http://www.computer.org/portal/web/tvcg
Information Visualization	http://ivi.sagepub.com/
International Journal of Image Processing and Data Visualization (IJIPDV)	http://iartc.net/index.php/Visualization
IEEE Vis (Former Visweek)	http://ieeevis.org/
EuroVis	http://www.eurovis2013.de/
ACM CHI	http://chi2013.acm.org/
EG CGF	http://www.eg.org
IVS	http://www.graphicslink.co.uk/IV2013/

Table 1.3.: More important journals dedicated to data visualization

1.2.2. A multidisciplinary field

Data visualization, during its history, has had contributions from multiple fields. As a natural process, most visualization methods have been created or invented out of a necessity to address specific communication needs. Looking back to the so-called golden age of visualization (Chen et al. 2008) it seems that the creations with more impact are the ones with an intention of advocacy or a factual claim that, without a visualization tool, would not have been easy to demonstrate. This is the case of Cholera in London (Friendly, M and Denis, D. J., 2001a) or Playfair's UK exports (Friendly, M and Denis, D. J., 2001c) diagram or Minard's time line map (Friendly, M and Denis, D. J., 2001b, Tufte, E.). This necessity, along with creativity, has produced new ideas in a variety of fields and knowledge contexts. Figure 1.2 presents a word cloud of the professions of the creators of the most relevant visualization methods from 1765 to 1999.

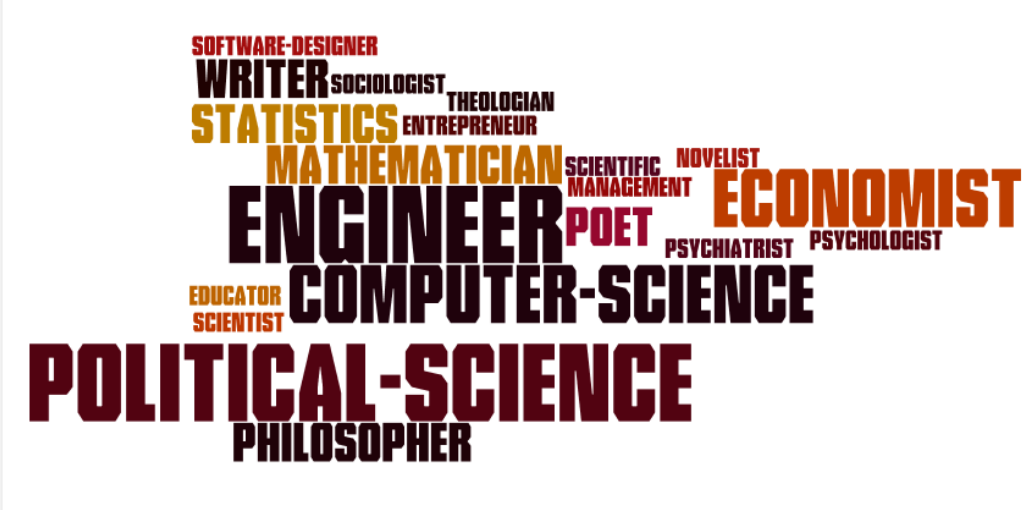

Figure 1.2.: From 1765 to 1999, a list of some of the most important visualization methods invented is: Timeline, Bar Chart, Pie chart, Flows in maps, Venn diagram, Histogram, Gantt Chart, Flowchart, Tagcloud, Social Networks, Boxplot, Star plot, Treemap, Headmap, Tagcloud, Sparkline. We present a word cloud of the professions of their inventors (made with Wordle).

Among others the authors include Joseph Prestley, defined as theologian, dissenting clergyman, natural philosopher, chemist, educator, and political theorist, who has been credited with the first publication of a time line in 1765. Also included are the famous visualization gurus William Playfair (pie chart, 1801), Charles Joseph Minard (flows in maps, 1869), John Tukey (boxplot, 1977) and Edward Tufte (sparkline, 1999). The structured data and the complete list of authors used to generate this word cloud can be found in B.2.1 on page 95

In fact, this multidisciplinary character is applicable even today, where most relevant names in the data visualization scene have very different backgrounds: from mathematics, statistics or pure computer sciences (Santiago Ortiz, Hilary Mason, Amanda Cox, Mike Bostock, Nathan Yau) to graphic design (John Maeda, Stefanie Posavec) or journalism (Alberto Cairo, David McCandless) and from academics (Ben Shneiderman, Jeff Heer, Andrew Vande Moere, Fernanda Viégas, Martin Wattenberg, Chris Harrison, Jeff Clark) to activist (Kim Rees, Jake Porway), freelance (Moritz Stefaner, Andy Kirk, Gregor Aisch) and corporate (Stephen Few, Robert Kosara, Manuel Lima) worlds. In coming years it is likely that new generations will come from the visualization departments in a growing percentage.

1.3. Data and text visualizations

1.3.1. The name of the field

There are several expressions that may refer to what data visualization means and there is not universal consent on their exact definitions. In scientific publications and books it is easy to find several expressions used, sometimes, indistinctly: data visualization, information visualization, infographics/information graphics, information design. Buzzwords and abbreviations such as datavis, dataviz, and infovis are also becoming popular (see Table 1.4 & Figure 1.3).

Examples of this are the titles of some books; is it really necessary the use of that variety of expressions? "Data Visualization: a successful design process" by Andy Kirk (Dec 26, 2012), Information Graphics by Sandra Rendgen (May 27, 2012), "Information Visualization: Beyond the Horizon" by Chaomei Chen (May 24, 2006), Infographics: The Power of Visual Storytelling by Jason Lankow, Josh Ritchie and Ross Crooks (Sep 4, 2012), The Information Design Handbook by Jenn Visocky O'Grady and Ken Visocky O'Grady (Sep 23, 2008).

Every author has a reason to use one or another expression, nevertheless, it doesn't seem there is a common criteria for it.

Google query	No. of results
infographics	18,600,000
data visualization	6,230,000
information design	2,230,000
information graphics	1,350,000
information visualization	996,000
infovis	699,000
datavis	573,000
data interfaces	375,000
dataviz	352,000

Table 1.4.: The names of visualization in Google

There is controversy over the distinctions between all these expressions (Kosara, R. [2011]). Given that their popularity is changing fast. we will probably need a bit more time to see which ones will stay and which ones will fall into oblivion.

In this work we will be using these expressions ordered in three conceptual groups as follows:

1. Data visualization, information visualization: our preferred expression is data visualization. We will also use information visualization as a general name for the field we work on. We also use the word representation as a synonym of visualization.

2. Infographics, information graphics: we consider this expression as a subfield of data visualization that, usually, is a static image and is not the output of a computer

code written specifically for this function. Digital infographics are usually created with image editors. As Cairo argues, the creation of infographics relies more on storytelling and narrative skills than on computer programming skills. (Cairo [2012]).

3. Information design, interface design, data interfaces: more and more data visualization tools are overlapping with digital interfaces. The digital interface is the permeable and thin medium that puts humans in contact with digital creations. The interface designer is the person who defines and innovate in the language, shape, accessibility and effectiveness of this layer in order to be used in one or more ways when a human interacts with it (ware_information). Nowadays, the evolution of digital interfaces and radical proposals is an active field (Doerk et al. [2011]).

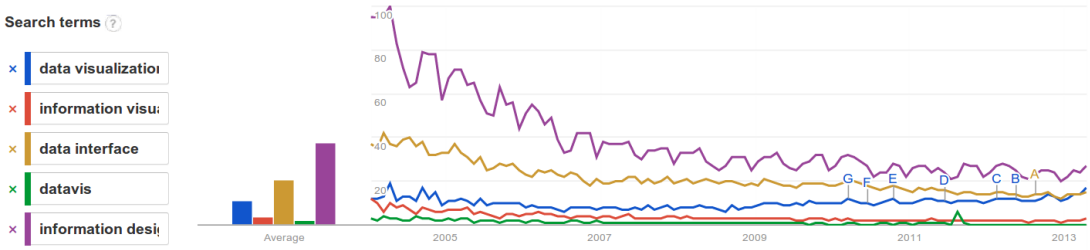

Figure 1.3.: Google trends in the use of the multiple names of visualization

1.3.2. The data types

Here we refer to data types as the formats that data can have in order to be processed for later representation. In the visualization process, data can be transformed several times before it takes on the desired or required format according to the visualization technique used.

Here again there is no a single classification of kinds of data or data-types. We are going to use Shneiderman's classification (1996) called Task by data Type Taxonomy (TTT) that divides data types in seven groups: 1-, 2-, 3-dimansional, Temporal, Multi-dimensional, Tree and Network.

- 1-dimensional data: textual documents, program source code, alphabetical lists, etc.
- 2-dimensional data: geographic maps, floorplants, newspaper layouts, 2-axes diagrams, etc.
- 3-dimansional data: real world objects such as molecules, bodies, buildings, etc.

- Temporal data: all kinds of timelines.
- Multi-dimensional data: most relational and statistical databases are conveniently manipulated as multi-dimensional in which items with n attributes become points in a n-dimensional space.
- Tree data: collections of hierarchical data where each item (except the root) links to a parent item. For example: genealogical data, file systems trees, genetic data, etc.
- Network data: any collection of items and their relationships. For example: social relationships, computer networks, etc.

1.3.3. Data as textual documents

This study focuses on visualization tools applied to textual documents. According to Shneiderman's classification, regular texts would be considered 1-dimensional data. A text is a sequential data that goes right-to-left or left-to-right and line by line, top to bottom. However a text can have multiple internal structures, e.g. according to morphology it can have paragraphs, sentences and words. According to the information structure, a text can be ordered by chapters, parts, sections, subsections, etc. If the text has a format like HTML, then it can be ordered by body, divs, paragraphs, etc. In those examples the text has a tree structure as a data type.

Sometimes, text visualization is not widely considered as a subfield of data visualization. Illinski (2013) asserts that text cannot be considered as a data type. Silić (2010) says that "unstructured text is not suitable for visualization". In fact, as mentioned above, most text visualizations transform the initial "unstructured" textual data into a new structured and reduced dataset. This new dataset is no longer a 1-dimension data type, but a categorical or a network dataset. And it can be represented with a wide range of tools not specific to natural text representation (Hearst, 2009, Grobelnik and Mladenic, 2002).

As we will show in chapter 2, most visualizations of texts do not represent raw data, that is the text as it is, instead they transform the text into smaller chunks of data, normally extracting a representative part of a text. Such a process is a data transformation process and it happens for example when a text is reduced to a list of words according to their frequency. In this case the method chosen to represent the data belong to a family of methods that suit this data type. In this review we will go through all the most referenced strategies to represent texts or collections of texts, paying special attention to the strategies to represent textual data as it is, as a regular text, with all its complexities, irregularities and rich abstractions.

The analysis of texts is a key field for text visualization. In the next subsection we give a brief commentary on this field and its relationship with this work.

1.3.4. Text analysis

Text analysis, almost synonymous with text mining, (Feldman and Sanger 2006) is an interdisciplinary field that includes information retrieval, data mining, machine learning, statistics, linguistics and natural language processing. According to M. Hearst (Hearst, M. 2003) in text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down. Text mining is a subfield of data mining whose typical applications, among others, are to analyze or compare literary texts, analyze biological and genomics data sequences or, more recently, discover patterns in customers behavior or discover fraud in the use of credit cards. Hearst differentiates this case from the information extraction operations, like the extraction of people's names, addresses or job skills. This later task can be done with 80% of accuracy, but the first task, the full interpretation of natural language by a computer program, again following Hearst, looks like will not be possible for "a very long time".

Following the classification of Keim (2008) texts can be analyzed on different levels of abstraction:

- Statistical level (e.g. frequencies of words, average sentence length, number of tokens or types, etc.)
- Structural level (structural components of a document, such as header, footer, title, abstract, etc.)
- Syntactical level (principles and rules for constructing sentences)
- Semantic level (linguistic meaning)
- Pragmatic level (meaning in context; consequence for actions)

In order to study visualization tools for texts it is as important to follow the literature of visualization tools as well as the literature of text analysis. The two fields are importantly interrelated. On one hand, the output of the text analysis can limit the possibilities of the visual presentation and interaction of the text itself.

1.3.5. Methods, tools and visualizations

Following the idea that data visualization is used in and builds from a wide range of points of view and fields of knowledge, we found it necessary to define three important concepts used to refer to the visualization artifacts and the produced representation of data. These are: methods, tools and tasks, and visualizations. Since there is no clear definition and use of these expressions we want to define how we are using them and how we define their relationships. This is turn will allow us to refer to the elements of visualizations with three conceptual views. These three concepts have a hierarchical

relationship: from a non-contextualized raw technique for representing data, what we call a method, to problem solving techniques based on specific data, what we call tools and tasks. To, finally, a contextualized visual discourse, i.e a complete visualization. This definitions can help us to emphasize or deconstruct parts of the visualization cases we are presenting.

The concepts can be represented in what we call the visualization pyramid (see 1.4).

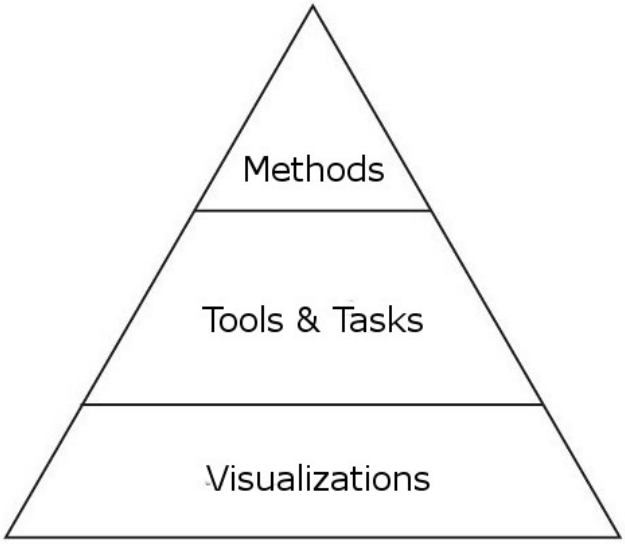

Figure 1.4.: No. methods < No. tools < No. visualizations

Methods

This is the smallest piece of data representation. Examples of visualization methods are: pie chart, bar chart, word cloud, Venn diagram, time line, etc. A method is a way to represent data and it does not include any context. A method is always integrated in a context. A context can have a dataset, an introductory text, a history, an interface, among other components.

Tools and tasks

When a dataset is injected in a method we get a tool that can perform one or more tasks in a particular context. A method can be used in different contexts and in different ways depending on the required tasks.

Visualization tool is the most used expression when referring to a data representation system.

Visualization and contexts

Finally, when one or more tools are used in a specific context, some properties need to be defined and adapted:

- Graphic components: scales, sizes, and colors.
- Texts: introduction, header, and legends.
- Data: format and injection.

A tool, under this point of view, can produce a lot of visualizations.

2. Literature review

2.1. Classification of tools

As mentioned in the introduction (1.1.3) this research focuses on visualization tools in which the data type is a 1-dimensional text, i.e., a natural text. If we follow typical text analysis that represents a text with a bag-of-words then we will quickly end up studying network representations and this is not the goal of this research.

However, as we also have mentioned, we are interested in text collection visualization and exploration tools, and this is a more common case of study with a variety of standard visualization methods, including network diagrams among others.

The ground level of classification of text visualization tools according to the type of data has two categories:

- Textual documents: this is the representation of single texts. We can understand text as a sequence of words ordered according to the hierarchy: document > paragraphs > sentences > other punctuation marks (“:”. “;”, “?”, “!”. “,”) > words > syllable and phonemes. In cases where the text is a book or other kind of structure, then, it is possible to have more granularities including: chapters > sections > sub sections > ... Also we assume the metadata of the text and other attached texts: title, authors, publishers, copyright notes, acknowledgement, dedication, preface, table of contents, forward, glossary, bibliography, index, etc.

- Text collections: groups of texts in which each item is a clearly differentiable entity. Usually when talking about collections of texts, we talk about texts that have some similarity, either in register, length, or structure. All the cases we have reviewed are collections of the same kind of texts. Heterogeneous collections of texts are also referenced in the literature (Meeks, E. 2011), especially for a representative analysis of a field of knowledge, in which cases the goal of the collection is to include the maximum variety of expressions and a wide vocabulary. In these cases the dataset is heterogeneous according to its structure and register.

Starting from these two very differentiated kinds of visualizations, we have added several subjective subdivisions to each case according to the goals of this research. The aims of the classification is to support the description and explanation of the reviewed cases, as well as to suggest key features of text visualization tools.

Single texts visualization

- Whole <-> Part
- Sequential <-> Non sequential
- Discourse structure <-> Syntactic structure
- Search
- Time

Text collections

- Items <-> Aggregations
- Landscape
- Search
- Time

2.1.1. For single texts:

Whole or part?

In some cases a part of the text is considered the essence of the text and it is used for the visualization instead of the whole text.

Nevertheless, there are representations where the whole text participates. These cases often represent the whole text implicitly. Examples of this are:

- The chapters of a book but not all its text.
- The representation of all the sentences of the text as colored lines.
- The verbs of a text, providing an impression of the style of the text.
- The characters of a novel and their appearance within the text.
- The places or dates present in the text.
- etc.

The cases in which the whole text is explicitly represented, for obvious reasons, are cases with short texts, e.g texts of songs, speeches, poems, etc.

In some cases, such as A.1.1 only certain words within the text are represented; nonetheless we classify this case as a whole text representation because the whole novel, chapter by chapter, is implicitly represented along the circle.

In cases in which the whole text is represented, even implicitly, as one main element of the visualization we have classified them as whole text visualization.

Does the visualization follow the text sequence?

Is the visualization following the same sequence as the original text does? By sequence we mean the same order as in the text. If yes, the case will be considered sequential, otherwise it will be called non-sequential.

For example, a typical case that does not follow the sequence of the original text is a word cloud (see 1.2).

Discourse structure or Syntactic structure?

Among others, a text can have two kinds of structure that we consider useful for our research. There is a structure that is completely subjective to the author's point of view, this is the so-called discourse structure. In linguistics, discourse is a broad concept. Here we are using it to refer to the parts of a text and the outline of a document: parts, chapters, sections, subsections, etc. The discourse structure is widely used when visualizing texts because it is an easy way to draw a representation of the text sequence.

The second structure that we consider is the syntactic structure. Syntactic can also refer to several concepts, here we are using it to refer to: sentences, phrases, and words as verbs, nouns, and morphemes. This is an objective structure and depends on the rules of linguistics rather than an author's decision. In text visualization, elements of this structure, such as sentences, are very common.

2.1.2. For text collections:

Are the items of the collection differentiated or represented as aggregations?

How is each item of the collection graphically represented? Is each text represented as a graphical entity, i.e., a point or a word or short sentence? Can the items in the visualization be counted, i.e., are they visually differentiated?

There are cases in which each item is not represented by a graphically separated entity, but as a colored zone for example. Another case is when the items are accumulated and shown as frequency distributions. When the items of the collection are not graphically distinct (visually countable) then we talk about an aggregation visualization instead of an items visualization.

Just data or data and landscape?

Is there any graphical context accompanying the items of the collection? Is there any other dataset, a part from the one coming from the text, also represented? Some cases

present the items immersed in a graphical environment, like a map. This context can be a real geographical map, a metaphor, or, for example, a conceptual landscape composed by words forming a second layer that complements the collection data layer in which every distance plays a role: item-item (similarity between documents), word-item (importance of a word in a document), word-word (similarity between words in the collection).

Scales and axes are not considered as landscape; nor are the elements of the interface in which the tool is embedded.

2.1.3. For both single texts and collections:

Properties applicable to single text and to texts collections visualizations are:

Is time involved?

Are the texts changing in time? There is a set of visualization tools that show the changes of a datasets over time. Most popular tools of this kind have been developed in computer science representing code evolution or in Wikipedia showing a number of aspects of article revisions.

We also include in this category visualizations in which the dataset changes over time, for example in a visualization of the latest news, the dataset grows over time.

Is the visualization a result of a search query?

The visualizations of the output of the information systems retrieval is well defined kind of visualization characterized by the changing number of represented items depending on the number of search results obtained. This is a growing visualization subfield related to the disciplines of information systems and information retrieval (Mann 2002, Hearst 2009).

Valid for short or for long datasets

It is rare that a visualization tool is independent of the size of the dataset that is represented. In cases in which the tool is clearly designed for a specific dataset size, the reader will find an explanation.

2.2. Analysis of tools

We have collected a list of tools classified using the proposed axes defined in 2.1. We expect to cover the most important aspects of text visualization. This review focuses on the specific ideas for text visualization, rather than the dataset and the contexts of each case.

We have reviewed 49 cases. The cases are listed in A. Each case includes: name of the project, author(s), year of publication, description of the original dataset, short

descriptions of the visualization methods used, description of the visualization, reference to related literature, the URL of the project, and a screen shot of the tool.

The cases are grouped into two sections and four subsections:

- Single text visualization (23 cases)
  - Whole text visualization (15 cases)
  - Part text visualization (8 cases)
- Text collection visualization (26 cases)
  - collection of items (16 cases)
  - collection of aggregations (10 cases)

For each subsection the cases are sorted by year of publication (descendant).

In order to assist the reader, this collection can be viewed using an online browser and visualizer based on one of the reviewed tools, called AREA (see 2.1).

To access it, please visit: `http://research.nualart.cat/textvistools/`

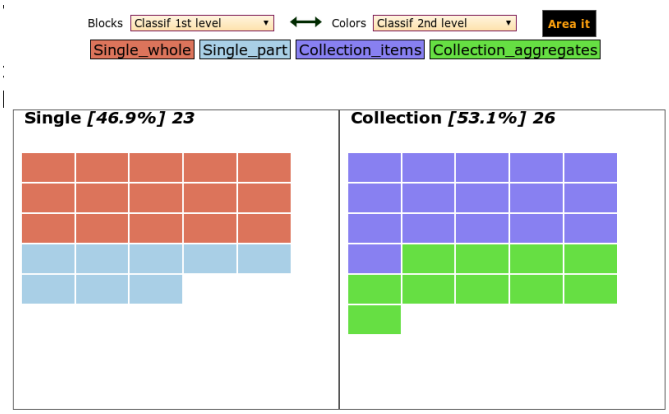

Figure 2.1.: The 49 reviewed cases visualized with a tool called AREA (screen shot)

A number of fields has been collected for each case: name of the tool, short name of the tool, author(s), year of publication, url for further information, original dataset , discipline related to the work, description of the visualization method, description of the tool, screen shot, thumbnail, Classification (Single or Collection), Classification (Single-Whole, Single-Part, Collection-Items, Collection-Aggregations), Classification (Time), Classification (Search), Classification (dataset small, dataset large, N/A).

...

2.2.1. Single text visualization

2.2.1.1. Whole text visualization (15 cases)

1) Literature - Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

2) Literature - Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

3) Literature - Poem Viewer by Katharine Coles et al. (2013)

4) Politics - State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

5) Literature - Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

6) Science/papers -On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

7) Science/papers - Texty by Jaume Nualart (2008)

8) Religion - Bible Cross-References by Chris Harrison (2008)

9) Literature - Literature fingerprint by Daniel A. Keim and Daniela Oelke (2007)

10) Wikipedia - History Flow by Fernanda Viégas and Martin Wattenberg (2003)

11) Literature - Colour-coded chronological sequencing by Joel Deshaye and Peter Stoicheff (2003)

12) Literature - 2-D display of time in the novel by Joel Deshaye (2003)

13) Literature - 3-D display of time in the novel by Joel Deshaye (2003)

14) Any - Arc diagram Wattenberg by Martin Wattenberg (2002)

15) Health - TileBars by Marti A. Hearst (1995)

Description:

- Number of cases: We found fifteen cases in the whole text visualization category.
- Years: The cases have been published from 1995 to 2013 (eighteen years).
- Authors: All the authors come from academic fields. The most prolific authors in this category are Jeff Clark and Joel Deshaye with three cases each, followed by Martin Wattenberg, two cases.
- About the datasets: Most of the text corpora used belong to literature (eight

cases). Most of authors use literature texts to demonstrate new visualization tools; especially well known texts, like classic novels.

- About the methods: All the cases except one (14 ▷ Arc diagram Wattenberg by Martin Wattenberg (2002)) use color as part of the visualization method. Five cases use methods that are bar chart derivatives (cases 4, 5, 6, 9 and 11). Three cases use curves connecting parts of the texts: two arcs and one radial diagrams (cases 1, 8 and 14).

Discussion:

It is not possible to establish a common method for whole text visualizations. As expected, all the cases present an axis that represents the whole text. In all the cases except in two, the text line is represented by a horizontal or a vertical line. The exceptions represent the text line as a circle (1 ▷ Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)) and as an iconification of a text in a page, (7 ▷ Texty, a visualization tool to aid selection of texts from search outputs by Jaume Nualart (2008)).

Since whole text visualization always includes an abstraction of the text that we call text line, a question arises: which part of the text is physically present in the reviewed whole text visualizations? It is surprising that most of the cases, nine out of fifteen, did not show a single word (cases: 4, 5, 6, 7, 8, 9, 10, 11, 15). Four cases show a small number of words (1, 2, 12, 13). Only two cases show all the text, cases 3 and 14.

The most common pattern is to show the occurrence of some feature - perhaps term, or topic, or cross-reference, or character - within the text as a whole (all the cases except 3, 12, 13 and 15). Except in Wattenberg's arc diagrams, occurrence is represented by same color.

Is interesting to observe how some cases represent very similar data in very different ways. This is the case of History Flow by Viégas and Wattenberg (A.1.1) and Favoured Traces by Fry (A.1.1): both show the same thing - document version history by document section - but the first is spatialized and the second animated.

We can also see similarities between Texty (A.1.1), and TileBars (A.1.1). Here Texty uses the same technique as TileBars, highlighting some words of the text inside a rectangular figure that represents the whole text.

We can also observe cases that are opposite or complementary one to another, like Wattenberg's Arc diagram (A.1.1), which shows repetition, and Hurst's novelty visualization (A.1.1), which shows only new strings, not repetitions

Literature and other complex texts, like political speeches (4 ▷ State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)) or the bible (8 ▷ Bible Cross-References by Chris Harrison (2008)) dominate the kind of corpora of this category (ten cases). In our opinion, this is surprising because these texts are complex, sometimes with a high level of abstraction and little formal structure. In our opinion, in order to introduce or test a new tool, it is a good idea to work with more structured and *easier* texts, with greater regularity in vocabulary, text length, discourse structure and language register. Literature, due to its inherent freedom of writing, does not need to follow any pattern or

rule that can help us to structure the unstructured. Examples of more structured texts are scientific papers, patent texts, health diagnostics, etc.

Nevertheless, depending on how the text is treated and processed, the nature of the text is not always a key point. For example, the work of Matthew Hurst (5 ▷ Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)) tracks the introduction of new terms along a text. This tool can be used by literature experts and, at the same time, it can be applied to any other text with results not related to the complexity of the texts because of the ubiquity of the method. Yet it would be interesting to apply this method to scientific papers in which the style is a lot more defined. Similar ideas are applicable to 1 ▷ Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013), and 4 ▷ State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011), and 9 ▷ Literature fingerprint by Daniel A. Keim and Daniela Oelke (2007).

2.2.1.2. Part text visualization

16) Literature - Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)

17) Any - Wordle by Jonathan Feinberg (2009)

18) Books - Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)

19) Literature - Phrase Nets by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

20) Google data - Word Spectrum: Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

21) Google data - Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

22) Literature/songs - Document Arc Diagrams by Jeff Clark (2007)

23) Any book - Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

Description:

- Number of cases: We found eight cases in the part text visualization category.
- Years: The cases have been published from 2005 to 2013 (eight years).
- Authors and datasets: Two cases by Jeff Clark (cases 16 and 22) and one by the creative couple M. Wattenberg and F.B. Viégas in collaboration with F. van Ham (case 19) use literary texts. The two cases by Chris Harrison use large bi-gram datasets published by Google. There is one case not dependent on the nature of the text: Wordle (case 17), the very popular word cloud method introduced by

J. Feinberg. Finally two interactive tools that allow large datasets are presented: Docuburst (case 18) and Gist icons (case 23).

- About the methods: In six of the eight cases (cases 16, 17, 18, 19, 22 and 23) the dataset is reduced to what is call a bag of words and only these words are present in the visualization. The cases 20 and 21 are representations of all bi-grams that two compared words share.

Discussion:

Part text visualization is a successful and popular way to visualize. This is probably because it is impressive that a long text can be effectively represented by a small set of words. Very simple statistical methods, like word frequency, can have an easy-to-understand result. A list of variously-sized-words is a direct way to communicate with any user, from beginner to expert. Most of the part text methods found online use statistical methods to extract the part from the whole.

We argue that the extraction of part of the corpora can be affected by the structure of the text corpora and its complexity. In the reviewed visualizations, half of the cases present unstructured text corpora, but the criteria to extract the part from the whole is very well defined: lists of verbs (16 ▷ Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)), words that are found in the text in an "X and Y"-pattern (18 ▷ Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)), and lists of words not included in a list of predefined empty words (21 ▷ Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)).

Of course, cases with extraction based on word or phrase functionality instead of pure statistics would be more affected by the nature of the text. We have been focusing on these cases because they look more interesting according to the goals of the research. Thus, there is the case of Novel Views: Les Misérables - Characteristic Verbs (case 16), which represents only verbs, and the case of Docuburst (case 18) which uses the *crowdsourced* lexical database Wordnet as a human-like backup. And again the cases of Phrase Net (case 19) and the two Google bi-gram cases (cases 20 and 21).

A common behavior we have found in part text visualization is that once a part of the text is extracted all the cases except one (22 ▷ Document Arc Diagrams by Jeff Clark (2007)) discard any reference to the original text sequence in the visualization. See the following point for a further development of this idea.

2.2.1.3. Other subcategories:

Sequential visualization

We found sixteen cases out of twenty-three in single text visualization in which the visualization keeps a similar sequence as the original text does. Seven of the sixteen visualize the sequence using a discourse structure, mainly chapters. The rest, nine cases use syntactic elements to represent the original sequence of the text, mainly words.

It is remarkable that only one case (22 ▷ Document Arc Diagrams by Jeff Clark (2007)) belonging to part text visualization follows the original text sequence. And, at the same time, all the whole text visualization cases are also sequential. It seems that sequentiality is intrinsic to whole text visualization. Whole text visualizations do not literally represent every word of the text, but a graphical metaphor of the whole text: a text line. This text line can represent a discourse structure of the text or a syntactic one, in any case, graphically there is a line or an area that represents the length of the text.

Sequentiality in visualization allows the reader to go back and forward between the visualization as in the text. In the case of a long text, a book (nine out of sixteen cases) the visualization can act as a map or a guide to the text.

Non sequential visualization

As non sequential visualization we have found five cases: three of them are clouds of words (cases 17, 20 and 21). One is network-like (19 ▷ Phrase Nets by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)) and one is an all-verbs-in-text representation (16 ▷ Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)).

Discourse structure in the visualization

1) (365 chapters) - Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

2) (5 volumes) - Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

5) (48 volume/book) - Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

6) (15 chapters) - On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

8) (chapters/books) - Bible Cross-References by Chris Harrison (2008)

11) (narrative structure) - Colour-coded chronological sequencing by Joel Deshaye and Peter Stoicheff (2003)

12) (4 sections) - 2-D display of time in the novel by Joel Deshaye (2003)

13) (sections + narrative + chronological times) - 3-D display of time in the novel by Joel Deshaye (2003)

All eight cases found that follow a discourse structure of the text in the visualization are sequential visualizations. There are no cases in which the discourse structure appears unordered regarding the text. This is not surprising because in the cases in which a text is

divided in chapters and each chapter is represented as an entity, we have considered them as collection visualizations (e.g. 4 ▷ State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)). This is why all the cases in this section represent the parts of the text ordered and aligned in a line or a curve. Out the eight cases, five represent chapters or sections of a book, two represent volumes, and there is one special case (11 ▷ Colour-coded chronological sequencing by Joel Deshaye and Peter Stoicheff (2003)) in which the text is divided in colors according to purely narrative topics and scenes. This is the only case we could find that uses discourse structure elements deeper than chapter, sections, books and volumes. Probably any deeper method, such as, for example, narrative topics, would require manual text line segmentation.

Syntactic structure in the visualization

3) (sentence analysis) - Poem Viewer by Katharine Coles et al. (2013)

16) (verbs) - Novel Views:Les Misérables - Characteristic Verbs by J Clark(2013) (NOTE: this is the only one in non sequential)

4) (words->sentences) - State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

7) (words) - Texty by Jaume Nualart (2008)

18) (words hyponymy) - Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)

9) (sentences) - Literature fingerprint by Daniel A. Keim and Daniela Oelke (2007)

22) (words/phases) - Document Arc Diagrams by Jeff Clark (2007)

23) (words) - Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

The second half of the sequential cases, eight, use *intrinsic text elements*, like groups of words (four cases, cases 7, 18, 22 and 23), verbs (one case, 16), sentences (two cases, 4 and 9), and a complete text analysis (case 3). Syntactic structure requires parsing of the text word by word, sometimes using a database of lexical or semantic list of words, sometimes parsing sentences and paragraphs. Syntactic structure visualization is less dependent on the nature of the text in the sense that the methodology is agnostic to the complexity of the text. Usually software automatically extracts or marks the chosen syntactic elements.

Search results visualization

18) Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)

23) Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

15) TileBars by Marti A. Hearst (1995)

The three cases under search results visualization were presented as web applications and therefore interactive. The user could query the visualization tool and get a unique representation for each search. The three cases we are presenting are no longer on line. Docuburst (case 18) is a Prefuse application that can be downloaded (Collins et al. [2009]).

TileBars is a classic case (according to Google scholar, cited 625 times) by a classic author on visualization and interfaces of search engines, Marti Hearst. Docubusrt and Gist Icon are interactive radial visualization, in fact, Gist Icons is one of the references and main influences in the Docuburst development, as explained in Docuburst paper.

Search result visualization tools are not widely implemented in information retrieval systems and most of the result outputs are 1-dimensional lists of itemized texts. (Nualart and Perez-Montoro). These three cases have in common that they are applied to large datasets and, starting with a search query, they show an improved search output with the aim of helping the user in the process of reading and filtering the results. In the three cases it seems a key task is to distinguish between similar items. TileBars searches PubMed, more than twenty million papers. Docuburst uses WordNet lexical database (155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs) to classify the visualized text. And Gist Icons examples represent, among other datasets: the complete set of approximately 7 million USPTO patents and Enron email data set comprised of 500,000 emails.

Under the collection of texts category (2.2.2) we present nine cases of visualization tools applied to search. results.

Dataset changing over time

6) On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

10) History Flow by Fernanda Viégas and Martin Wattenberg (2003)

We present two cases in which the visualization tools can be used to understand or follow the evolution over time of the represented text. The dynamic text visualization demonstrates that data visualization can be almost the only way to solve some tasks and

not only one more way of pure data advocation. For example it is really difficult to show how an entry from the Wikipedia evolves over time according to editors participation (10 ▷ History Flow by Fernanda Viégas and Martin Wattenberg (2003)). History flow is a clear solution to solve this problem and it helps to understand part of the complex collaborative process that Wikipedia is.

In the second case (6 ▷ On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)) the animated visualization demonstrate how changed the theories of Darwin along the editions of his Origin of Species. In Ben Fry's words: "The first English edition was approximately 150,000 words and the sixth is a much larger 190,000 words. In the changes are refinements and shifts in ideas — whether increasing the weight of a statement, adding details, or even a change in the idea itself."

2.2.2. Text collections

2.2.2.1. Items visualization

24) Literature (NOTE: it converts single text in collection) - Novel Views: Les Misérables - Segment Word Clouds by Jeff Clark (2013)

25) Literature - Grimm's Fairy Tale Network by Jeff Clark (2013)

26) Twitter - Spot by Jeff Clark (2012)

27) Science - Word storm by Quim Castella and Charles Sutton (2012)

28) Literature - Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

29) Wikipedia - Notabilia by D. Taraborelli, G. L. Ciampaglia and M. Stefaner (2010)

30) Media art - X by Y by Moritz Stefaner (2009)

31) Search engine - Search Clock by Chris Harrison (2008)

32) Online media - Digg Rings by Chris Harrison (2008)

33) Science - Royal Society Archive by Chris Harrison (2008)

34) Wikipedia - WikiViz: Visualizing Wikipedia by Chris Harrison (2007)

35) Visualization - Area by Jaume Nualart (2007)

36) Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

37) Search engines - Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001)

38) Search engines - Touchgraph by TouchGraph, LLC. (2001)

39) Internet - HotSauce by Ramanathan V. Guha (1996)

Description:

- Number of cases: We found sixteen cases in the items visualization category.
- Years: The cases have been published from 1996 to 2013 (seventeen years).
- Authors: The most prolific authors in this category are Chris Harrison (cases 13, 32, 33 and 34) and Jeff Clark (cases 24, 25 and 26), followed by Moritz Stefaner with two cases (cases 29 and 30).

- Disciplines and datasets: It is remarkable that nine cases are datasets from the Internet: Wikipedia (cases 29, 34 and 36), search engines (cases 31, 37 and 38), twitter (case 26), on line media (case 32), web pages (case 39) . In this category there are only three cases that use literary texts (cases 24, 25 and 28). Finally two cases represent scientific papers (cases 27 and 33), one case represents media art datasets (case 30) and one represents non specific collections (case 35).

Discussion:

The big difference between single text visualizations and collection visualizations is the nature of the text. In collections most of the texts are not literature and are accessible on line. Probably the nature of the text is less important when the goal of the representation is the collection instead of the text itself.

Item visualizations use methods that are independent from the nature of the items. Once the collections of texts are itemized then the dataset can be considered a general case of data visualization and not *pure* case of text visualization. This is why in this category, to generalize, the methods we find are well known and used in other visualization fields. Following this reasoning we find six network visualizations (cases 25, 28, 34, 37, 38 and 39), three time lines (cases 31, 32 and 33), and three cases that use also time lines but allowing categorization-based grouping (cases 26, 30 and 35).

Finally, there are four cases that, in our opinion, are really specific to text visualization. Two of them are dedicated to item comparison: 24 ▷ Novel Views: Les Misérables - Segment Word Clouds by Jeff Clark (2013) and 27 ▷ Word storm by Quim Castella and Charles Sutton (2012). Segment Word Clouds is noteworthy to say that it transforms a single text in a text collection. It represents the chapters of *Les Misérables* as word cloud items. That makes it easy to compare them. It also uses color to spot words that are newly prominent in the text.

Word Storm is a reinvention of the word cloud, more concretely is a variation of Wordle (case 17) that makes word clouds comparable. This is done by assigning a fixed position to each word. This simple idea makes it visually easy to compare word clouds while maintaining the usual word cloud features.

Finally, the most original cases that deserve to be mentioned specifically are Notabilia (case 29) and Chromograms (case 36). Notabilia is a very specific design that shows the evolution of the Wikipedia discussions called "Article for deletion". These discussions are sometimes close to "flame wars" due to the controversy over the simple existence of some articles. Notabilia represents the evolution and the final decision of the one hundred longest discussions. The visualization is done by Moritz Stefaner and is an interactive bushtree that highlights branches when moused over. The shape of the branches tells you about the nature of the discussion: cyclic, straight, or never-ending.

Chromograms is also a work based on Wikipedia data. It analyzes the comments of the editors for each edition of a Wikipedia entry. Visually it produces color coded stripes that in a short area easily tell you about years of editions of Wikipedia entries.

2.2.2.2. Aggregation visualization

40) Literature - Grimm's Fairy Tale Metrics by Jeff Clark (2013)

41) Topic Models - Termite by J. Chuang, C.D. Manning, J. Heer (2012)

42) Wikipedia - Pediameter by Müller-Birn, Benedix and Hantke (2011)

43) Google suggestions - Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

44) Google ngrams - Web Trigrams: Visualizing Google's Tri-Gram Data by Chris Harrison (2008)

45) Political speech - FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007)

46) Online News - Newsmap by Marcos Weskamp (2004)

47) Email conversation - TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48) Search engine - WebBook by S.K. Card, G.G. Robertson, and W. York (1996)

49) Any texts - Dotplot Applications by Jonathan Helfman (1994)

Description:

- Number of cases: We found ten cases in the aggregations visualization category.
- Years: The cases have been published from 1994 to 2013 (nineteen years).
- Authors and datasets: Only Fernanda B. Viégas has participation in two of the ten cases presented in this category (cases 43 and 47). The rest of authors only participate in one case each. The nature of the texts is very similar to the items visualizations category. Five cases from corpora that can be found on line [Wikipedia, case 42; Google, cases 43 and 44; online news, case 46; search engine results, case 48]. As *standard* unstructured texts there is one from literature (40 ▷ Grimm's Fairy Tale Metrics by Jeff Clark (2013)), one from political speeches (45 ▷ FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007)), and one from a one year email conversations between two people (47 ▷ TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)). Finally there are two cases that are quite special: 41 ▷ Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012) and 49 ▷ Dotplot Applications by Jonathan Helfman (1994). All the cases are discussed below.

Discussion:

Aggregation visualization is the category with a clearly biggest variability of methods. The ten methods in this category, apart from representing collections, only have in common that they are not representing specific items.

Due to the very special cases, we will comment the cases one by one:

40 ▷ Grimm's Fairy Tale Metrics by Jeff Clark (2013) is a matrix (a table-like) visualization that allows row sorting by clicking columns. The columns defines in a quantitative order thirteen measures related to the sixty two Grimm's fairy tales. It is a powerful tool to understand, compare and go through all the stories.

41 ▷ Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012) is a case that represents an intermediary dataset called topic models (Wikipedia [2013c]). Topic model is a more "clever" way to get a bag-of-words from a text than the typical word frequency statistical analysis. Termite is not visualizing texts but comparing parts of texts. It is a tool to compare topic models.

42 ▷ Pediameter by Müller-Birn, Benedix and Hantke (2011) is an specific interface that uses bar charts to show Wikipedia editions in real time. It is remarkable that the project used a device called Arduino (Wikipedia [2013a]) to detect editions and transcribe them to a physical indicator, merging digital and material worlds.

43 ▷ Web Seer by Fernanda Viégas & Martin Wattenberg (2009) is also an specific visualization method that shows most popular search queries based on Google suggestions. The tool allows comparing queries representing the suggestions in trees and connecting the matching branches. The simplicity of this case contrasts with the its power of communication: fast and easy to understand.

44 ▷ Web Trigrams: Visualizing Google's Tri-Gram Data by Chris Harrison (2008) uses a similar way to represent than Web Seer. It uses the huge dataset of Google n-grams. It represents and compares three words sentences (tri-grams).

45 ▷ FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007) is an interactive interface dashboard-styled that allows text comparison. The central representation uses a visual representation of frequent concepts similar to Texty (case 7) and TileBars (case 15). It allows text browsing and shows line graphs of frequent words along the texts.

46 ▷ Newsmap by Marcos Weskamp (2004) uses the treemap technique to propose a new way of real-time news reading and monitoring based on on-line Google news feeds. It is totally customizable in terms of news topic, country, published time. It allows also news search. This tool is available on line for free use.

47 ▷ TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006) is an experiment that developed a very specific interface allowing to follow and analyze the evolution of email correspondence between two people over the course of one year. It represents the words that characterize each of the writers and its evolution over time.

48 ▷ WebBook by Stuart K. Card, George G. Robertson, and William York (1996) is a surprising application at its time, 1996. It transformed a search engine results list in a multimedia (at the time text and images, basically) mash up using the metaphor of a

book. This was a pure collection of texts (web pages) visualization that presented the results as aggregations of text and images.

Finally we present 49 ▷ Dotplot Applications by Jonathan Helfman (1994) a great idea of visualization that has multiple uses. It is comparable to 14 ▷ Arc diagram Wattenberg by Martin Wattenberg (2002). Main use of dotplots is text comparison, including multi-language comparison, text version comparison and programming code comparison.

2.2.2.3. Other subcategories

Landscape as a second data layer

40) Grimm's Fairy Tale Metrics by Jeff Clark (2013)

26) Spot by Jeff Clark (2012)

28) Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

33) Royal Society Archive by Chris Harrison (2008)

47) TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

37) Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001)

38) Touchgraph by TouchGraph, LLC. (2001)

49) Dotplot Applications by Jonathan Helfman (1994)

The typical idea of landscape data is a network visualization with two layers of data, as is the case in 28 ▷ Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011). The first layer is the list of Marcel Proust texts represented as items, and the second layer is a network of topic models of the texts. The positions of the nodes of both layers are optimized. Proximity means more related nodes. This definition of landscape can be found also in the defunct cases of search engine results: 37 ▷ Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001) and 38 ▷ Touchgraph by TouchGraph, LLC. (2001).

The rest of the cases classified under this category show collections of texts in combination with more data. This is the case in Dotplot which represents the coincidence or not of strings in various texts, and in Grimm's Fairy Tale Metrics that combines a list of texts in the rows with several parameters in the columns. These parameters are not directly part of the text, but recalculated features related to the text, like the length, the lexical diversity and the presence of different groups of words that represent entities (for example: body -> hand, head, heart, eyes and foot) in each tale.

A third kind of landscape is based on representing timed metadata, these are the cases of 26 ▷ Spot by Jeff Clark (2012), 33 ▷ Royal Society Archive by Chris Harrison (2008),

and 47 ▷ TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006).

A common feature that landscape visualizations have is the capacity to compare a collection of texts simultaneously with a second parameter. The limitation of these visualizations is the number of items represented; large numbers create problems with overlapping items.

Search results visualization

26) Spot by Jeff Clark (2012)

43) Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

35) Area by Jaume Nualart (2007)

45) FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007)

47) TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

46) Newsmap by Marcos Weskamp (2004)

37) Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001)

38) Touchgraph by TouchGraph, LLC. (2001)

48) WebBook by S.K. Card, G.G. Robertson, and W. York (1996)

Compared to single text visualization, it is remarkable that collections visualizations have a lot more cases with search capacities (three cases versus nine). Common sense suggests that when presenting a collection of texts one natural feature of the tool may be a way to select part of the collection according to some criteria, i.e. filter and search features.

All the cases in this category allow search queries and output a unique visualization for each query. All the cases includes a search box and a search button.

Dataset changing over time

42) Pediameter by Müller-Birn, Benedix and Hantke (2011)

29) Notabilia by D. Taraborelli, G. L. Ciampaglia and M. Stefaner (2010)

36) Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

46) Newsmap by Marcos Weskamp (2004)

The four cases we include in this category allow the user to follow the evolution of the texts in the collection over time. Only one is designed to be real-time (46 ▷ Newsmap by Marcos Weskamp (2004)), but potentially all of them can show the collection for a specific date and time.

One difficulty for a tool that represents collections of texts changing over time is the access to an updated feed or an accessible API. Probably for this reason three out of the four use Wikipedia data and the other one uses Google news. In all the cases these are on line sources that have long allowed public access to their feeds.

2.3. Conclusions

This document will continue to be revised during the PhD research. This version is the first beta version (1.0b) and it will be a key tool in shaping future research.

At the time that these conclusions are being written, the practice part of the research has already begun. The first goal is to develop ideas to visualize collections of papers from scientific conferences. It will involve collection visualization as well as single text visualization. The aim is to develop two to three views or representations of the conference to help provide researchers with rich information on the contents of the conference. The idea involves show what the conference is about as well as which topics are not present. This literature review is the starting point. For the moment we are inspired by the cases 41 ▷ Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012) and 40 ▷ Grimm's Fairy Tale Metrics by Jeff Clark (2013) for the conference overview. For one-document-view we are influenced by 15 ▷ TileBars: Visualization of Term Distribution Information in Full Text Information Access by Marti A. Hearst (1995) and 7 ▷ Texty, a visualization tool to aid selection of texts from search outputs by Jaume Nualart (2008).

Regarding the literature review we are sure that we have not covered all the existing ideas related to text visualization. Certainly because data and text visualization is a relatively new field there is no single consolidated list of dedicated publications and sources. This presents a challenge in making a comprehensive survey of work in the field. Some of the cases we are presenting have been found in very specific publications; for example Joel Deshaye and Peter Stoicheff and their works on Faulkner's visualizations (cases 11, 12 and 13). Reading Stoicheff's notes it can be seen that they developed the tools just to assist in a very specific study of William Faulkner's narrative time lines. There are no references to applications of these interesting ideas to other texts, suggesting that more works remain hidden in the depths of other fields.

Text visualization, as we argue in this document, can be considered a subfield of data visualization. Yet, the boundaries of the field are sometimes not clearly defined. This is the case of 31 ▷ Search Clock by Chris Harrison (2008), in which the text corpora is an enormous dataset of search engine queries. Can this dataset be considered a collection of texts if each of them is, in most cases, is only one or two words long? Is there a minimum length of a text to be considered as a text? We decided to treat this case as a collection of texts, short ones, but, ultimately, texts.

One important decision in this review has been the classification of the cases found. Since there are few papers that review text visualization tools, we referenced the classic data visualization reviews [Shneiderman, 1996] as well as new ones (Collins et al. 2009). In these cases the classifications were based on tasks that the visualization tool can solve rather than on the explicit aspects of the visualization. This is why we decided to propose our own classification that, while far from perfect, is hopefully a useful approach to a classification based on visual features.

This is the list of insights and gaps we have found so far:

- Single text visualization cases are applied mainly to literature. Literature is an unstructured text; apart from complex combinations of words it can have high levels of human abstraction and freedom of structures and experimentation. It might be more effective to apply visualization techniques to other kinds of texts with a more formal register and/or predefined style, such as legal texts, scientific papers, template based texts and communications, etc.

- We have found only one single/part text visualization case which is sequential (22 ▷ Document Arc Diagrams by Jeff Clark (2007)). Most part text visualizations extract the essence of the text based on some criteria and the original sequence of the text is lost. Since sequential visualization tools have some advantages, it seems there is room to develop part visualization tools that maintain the original text sequence.

- Collection visualizations use methods that can be found in data visualization in general. This idea invites experimentation in bringing more standard data visualization methods and tools to the specific text visualization subfield.

- Collections of aggregations is the category that has developed more specific designs and ideas. More work needs to be done in order to find some patterns in this kind of visualization.

Open questions related to nature of texts to be faced during this research include:

- How can we extract the human part of a text?

- In some ways, text to bag-of-words is a *machinification* of the text, a simplification. What are we loosing in this transformations?

- When the intrinsic human part of a text matters? In which tasks can be important to extract it and show it?

These are really big questions that quickly go outside the scope of the project, as it is defined here. To talk about many of these things we would need to go in to textual analysis in literature, as well as computational linguistics, digital humanities among other fields. As these fields are related but not directly belonging to data visualization, we assume that this research project will require investigation in these fields in order to

contextualize and understand the tasks and challenges in which data visualization can help.

And finally:

- Why is it that most of reviewed cases older than five years are no longer on line? If these tools are no longer (or never have been) in use, we should question their effectiveness. We have not investigated how many are part of commercial software products and how many, after being published, are just forgotten. In any case it is questionable why some great ideas did not become new standards. Should this research study or contribute to fighting this tendency? This project aims to produce applications that will be adopted in some field or to solve some task for some group of users; but as these cases show, adoption is a significant challenge.

We hope to answer these and other questions in the long and challenging journey of this practical research project.

Bibliography

Gary J Anglin, Hossein Vaez, and Kathryn L Cunningham. Visual representations and learning: The role of static and animated graphics. *Handbook of research on educational communications and technology*, 2:865–916, 2004.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Ricardo Baeza-Yates, AndreiZ. Broder, and Yoelle Maarek. The new frontier of web search technology: Seven challenges. In Stefano Ceri and Marco Brambilla, editors, *Search Computing*, volume 6585 of *Lecture Notes in Computer Science*, pages 3–9. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-19667-6. URL `http://dx.doi.org/10.1007/978-3-642-19668-3_1`.

Alberto Cairo. *The Functional Art: An introduction to information graphics and visualization*. New Riders, 2012.

cambia.org. The lens (beta). http://www.lens.org/lens/, 2013. URL `http://www.lens.org/lens/`.

Linda Candy and Cognition Studios. Practice based research : A guide practice and research. 2006.

Chun-houh Chen, Wolfgang Haerdle, and Antony Unwin. *Handbook of data visualization*. Springer, 2008.

Christopher Collins, Sheelagh Carpendale, and Gerald Penn. Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009.

Marian Doerk, Sheelagh Carpendale, and Carey Williamson. The information flaneur: a fresh look at information seeking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1215–1224, 2011. URL `http://dl.acm.org/citation.cfm?id=1979124`.

Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2006.

Friendly, M and Denis, D. J. Dotmap of desease. john snow (1855) - milestones in the history of thematic cartography, statistical graphics, and data visualization. http://www.datavis.ca/milestones/index.php?group=1850%2B&mid=ms126, 2001a. URL `http://www.datavis.ca/milestones/index.php?group=1850%2B&mid=ms126`.

Friendly, M and Denis, D. J. Flow map.Charles joseph minard (1869) - milestones in the history of thematic cartography, statistical graphics, and data visualization. http://www.datavis.ca/milestones/index.php?group=1850%2B&mid=ms135, 2001b. URL http://www.datavis.ca/milestones/index.php?group=1850%2B&mid=ms135.

Friendly, M and Denis, D. J. Playfair line graph: chart of national debt. william playfair (1786) - milestones in the history of thematic cartography, statistical graphics, and data visualization. http://www.datavis.ca/milestones/index.php?group=1700s%2B&mid=ms80, 2001c. URL http://www.datavis.ca/milestones/index.php?group=1700s%2B&mid=ms80.

Friendly, M and Denis, D. J. Milestones in the history of thematic cartography, statistical graphics, and data visualization. http://www.datavis.ca/milestones/, 2001d. URL http://www.datavis.ca/milestones/.

M. Grobelnik and D. Mladenic. Efficient visualization of large text corpora. In *Proceedings of the seventh seminar. Dubrovnik, Croatia*, 2002. URL http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf.

Marti a Hearst. Search user interfaces. *Search User Interfaces*, 54(Ch 1):404, November 2009. ISSN 00010782. doi: 10.1145/2018396.2018414. URL http://searchuserinterfaces.com/book/.

Hearst, M. What is text mining? http://people.ischool.berkeley.edu/~hearst/text-mining.html, 2003. URL http://people.ischool.berkeley.edu/~hearst/text-mining.html.

Daniel A Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler. Visual analytics: Combining automated discovery with interactive visualizations. In *Discovery Science*, pages 2–14, 2008.

Kosara, R. The many names of visualization | eagereyes. http://eagereyes.org/blog/2011/the-many-names-of-visualization, 2011. URL http://eagereyes.org/blog/2011/the-many-names-of-visualization.

W Howard Levie and Richard Lentz. Effects of text illustrations: A review of research. *ECTJ*, 30(4):195–232, 1982.

Lubwig Bolzman Institute. LBI visualization team. http://vis.mediaartresearch.at/webarchive/public/view/mid:4, 2008. URL http://vis.mediaartresearch.at/webarchive/public/view/mid:4.

Thomas M Mann. Visualization of search results from the world wide web. 2002.

Meeks, E. Documents | digital humanities specialist. https://dhs.stanford.edu/comprehending-the-digital-humanities/documents/, 2011. URL https://dhs.stanford.edu/comprehending-the-digital-humanities/documents/.

NICTA. NICTA - BuntineW. http://www.nicta.com.au/people/buntinew, 2013. URL http://www.nicta.com.au/people/buntinew.

Noah Iliinsky. *Choosing visual properties for successful visualizations*. s IBM Software - Business Analytics, 2013. URL http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03323usen/YTW03323USEN.PDF.

J. Nualart and M Perez-Montoro. Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2), jun . ISSN 1368-1613.

Svetlana Sheremetyeva. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20*, PATENT '03, pages 66–73, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119303.1119311. URL http://dx.doi.org/10.3115/1119303.1119311.

Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, 1996.

Artur Silic and Bojana Dalbelo Basic. Visualization of text streams: A survey. In Rossitza Setchi, Ivan Jordanov, RobertJ. Howlett, and LakhmiC. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6277 of *Lecture Notes in Computer Science*, pages 31–43. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15389-1. doi: 10.1007/978-3-642-15390-7_4. URL http://dx.doi.org/10.1007/978-3-642-15390-7_4.

Stefaner, M. Gender balance visualization. http://moritz.stefaner.eu/projects/gender-balance/#NUM/NUM, 2013. URL http://moritz.stefaner.eu/projects/gender-balance/#NUM/NUM.

Jacqueline Strecker and International Development Research Centre (IDRC). *Data visualization in review: summary*. IDRC, Ottawa, ON, 2012.

TSL Education Ltd. World university rankings 2012-2013 - times higher education. http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking, 2012. URL http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking.

Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

Tufte, E. Minard's figurative map of hannibal's war. http://www.edwardtufte.com/tufte/minard-hannibal. URL http://www.edwardtufte.com/tufte/minard-hannibal.

Jen Webb. The logic of practice? art, the academy, and fish out of water. *Journal TEXT*, (14):1–16, 2012. URL http://www.textjournal.com.au/speciss/issue14/Webb.pdf.

Wikipedia. Arduino — Wikipedia, the free encyclopedia, 2013a. URL `http://en.wikipedia.org/wiki/Arduino`. [Online; accessed 12-June-2013].

Wikipedia. Top-down and bottom-up design — Wikipedia, the free encyclopedia, 2013b. URL `http://en.wikipedia.org/wiki/Top-down_and_bottom-up_design`. [Online; accessed 27-June-2013].

Wikipedia. Topic model — Wikipedia, the free encyclopedia, 2013c. URL `http://en.wikipedia.org/wiki/Topic_model`. [Online; accessed 12-June-2013].

YunYun Yang, Lucy Akers, Thomas Klose, and Cynthia Barcelon Yang. Text mining and visualization tools - impressions of emerging capabilities. *World Patent Information*, 30(4):280–293, December 2008. ISSN 01722190. doi: 10.1016/j.wpi.2008.01.007. URL `http://linkinghub.elsevier.com/retrieve/pii/S0172219008000094`.

A. APPENDIX A

A.1. Single text visualization

A.1.1. Whole text visualization

1 ▷ Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

- data: Novel "Les Misérables"
- method: Radial text-line
- description: "A word used in multiple places in a text can be interpreted as a connection between those locations. Depending on the word itself the connection could be in terms of character, setting, activity, mood, or other aspects of the text."

[http://neoformix.com/2013/NovelViews.html]

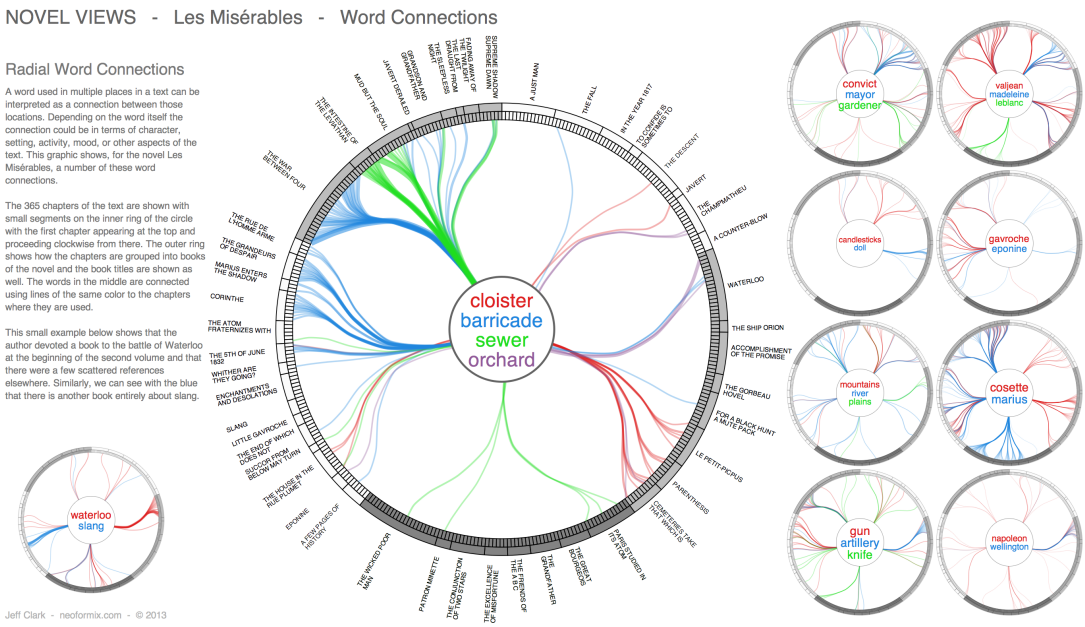

Figure A.1.: Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

2 ▷ Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

- data: Novel "Les Misérables".
- method: Horizontal multi text-line.
- description: "it shows where the names of the primary characters are mentioned within the text. Click on any of these images to see larger versions."

[http://neoformix.com/2013/NovelViews.html]

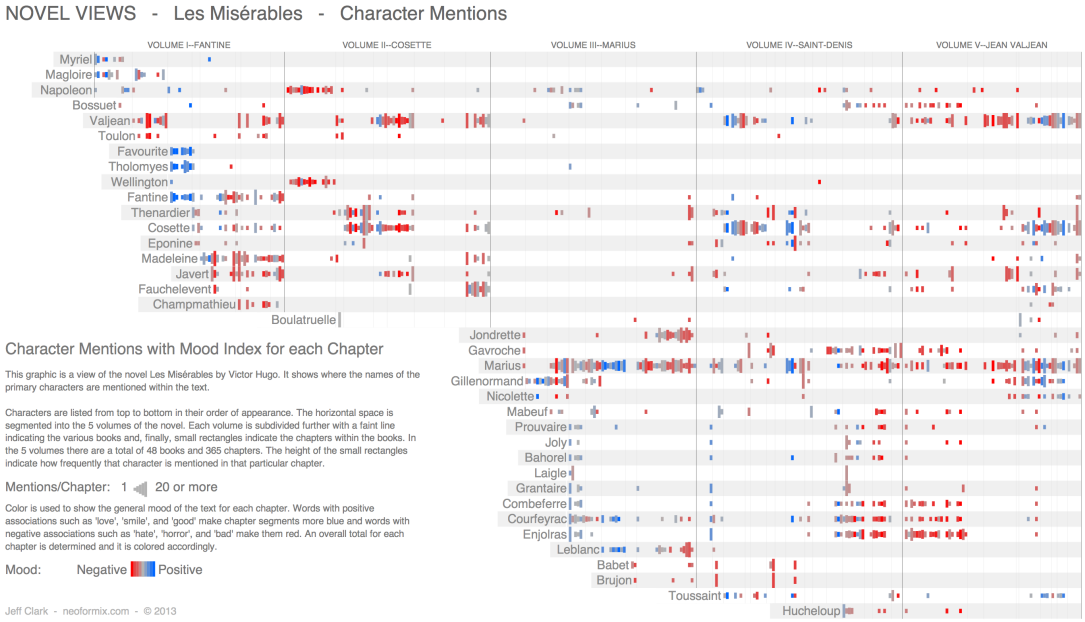

Figure A.2.: Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

3 ▷ Poem Viewer - (on process) 2013 project Imagery Lens for Visualizing Text Corpora - Oxford e-Research Centre at the University of Oxford by Katharine Coles, Min Chen, Alfie Abdul-Rahman, Chris Johnson, Julie Lein, Eamonn Maguire, Miriah Meyer, and Martin Wynne. (2013)

- data: Poems.
- method: Advanced Text-line.
- description: Poem Viewer is an experimental service for the exploration and analysis of poetry through visualization. Poem Viewer is part of an on-going research project and is currently a work in progress.

[http://ovii.oerc.ox.ac.uk/PoemVis/]

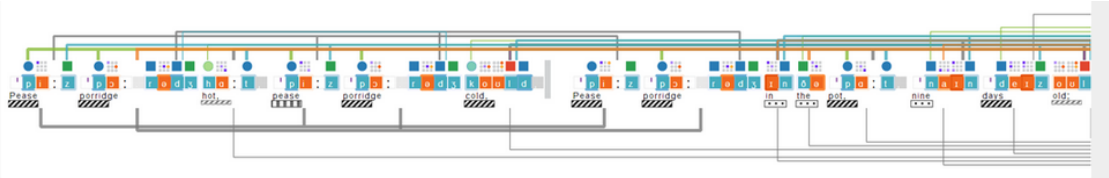

Figure A.3.: Poem Viewer - (on process) 2013 project Imagery Lens for Visualizing Text Corpora - Oxford e-Research Centre at the University of Oxford by Katharine Coles, Min Chen, Alfie Abdul-Rahman, Chris Johnson, Julie Lein, Eamonn Maguire, Miriah Meyer, and Martin Wynne. (2013)

4 ▷ State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

- data: Speech to text. Obama's State of the Union speech 2011 [N/A]
- method: Sentence Bar Diagrams.
- description: "First we have two Sentence Bar Diagrams for the speeches from 2010 and 2011. Sentence Bar diagrams use color coding to show the topic of the various sentences in the text and bar length to show how long the sentences are."

[http://neoformix.com/2011/SOTU2011.html]

Figure A.4.: State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

5 ▷ Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

- data: Literature
- method: Original pixel-block method
- description: Tracking the introduction of new terms in a novel. In the visualization each column represents a chapter and each small block a paragraph of text. The color of the block indicates the % of new words. It is not clear what is the utility of this tool; an specialist in linguistics could say how this visualization can be used.

[http://datamining.typepad.com/data_mining/2011/09/visualizing-lexical-novelty-in-literature.html]

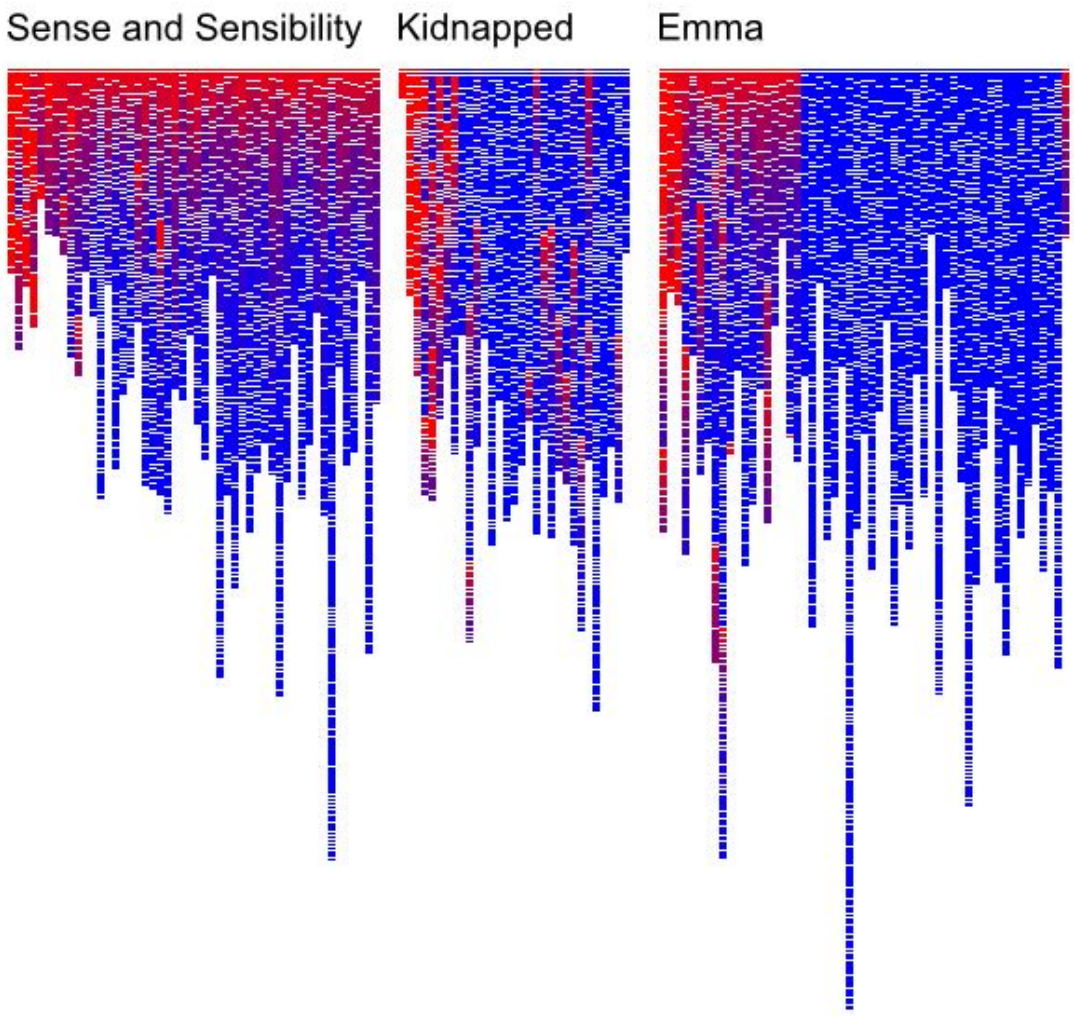

Figure A.5.: Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

6 ▷ On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

- data: All the versions of Darwin's On the Origin of Species Book.
- method: Matrix of pixel-blocks representing all the text in chunks. Colors are used to show each edition additions.
- description: We often think of scientific ideas, such as Darwin's theory of evolution, as fixed notions that are accepted as finished. In fact, Darwin's On the Origin of Species evolved over the course of several editions he wrote, edited, and updated during his lifetime. The first English edition was approximately 150,000 words and the sixth is a much larger 190,000 words. In the changes are refinements and shifts in ideas whether increasing the weight of a statement, adding details, or even a change in the idea itself.

[http://benfry.com/traces/]

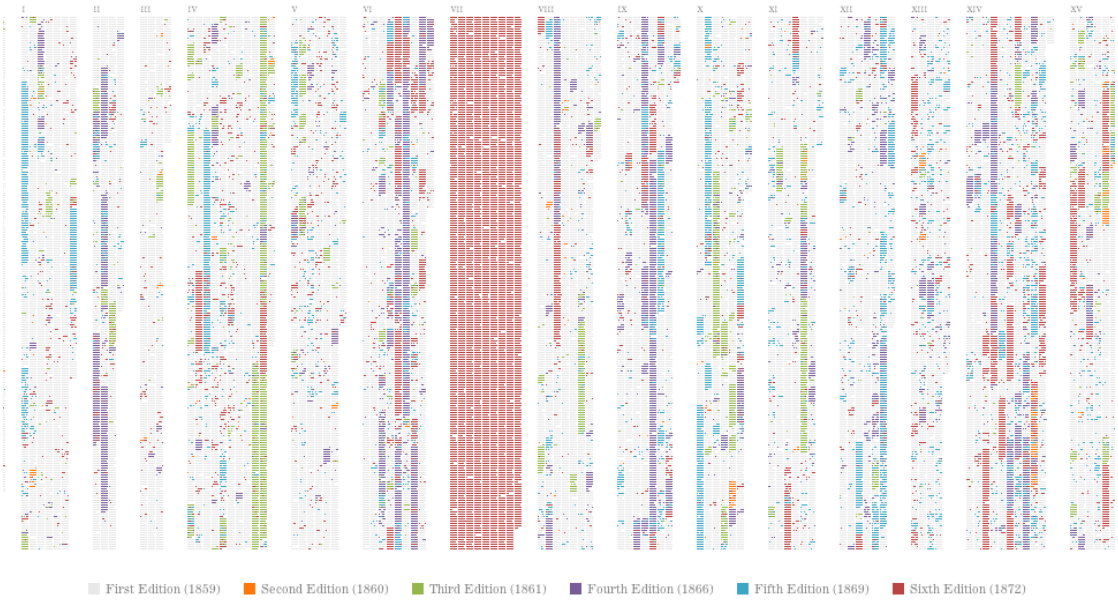

Figure A.6.: On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

7 ▷ Texty, a visualization tool to aid selection of texts from search outputs by Jaume Nualart (2008)

- data: Ars Electronica Jury Statements (from 1987 to 2007)
- method: Iconic representation of a text
- description: a Texty is an image, an icon that represents the physical distribution of keywords of a text as a flat image. These keywords are grouped in vocabularies, to each of which a color is linked. Texty reveals, the structure, conceptual density and subject matter of a text.
- paper: *Nualart, J. Pérez-Montoro, M (2013). "Texty, a visualization tool to aid selection of texts from search outputs" Information Research.*

[http://vis.mediaartresearch.at/textass/texty.php]

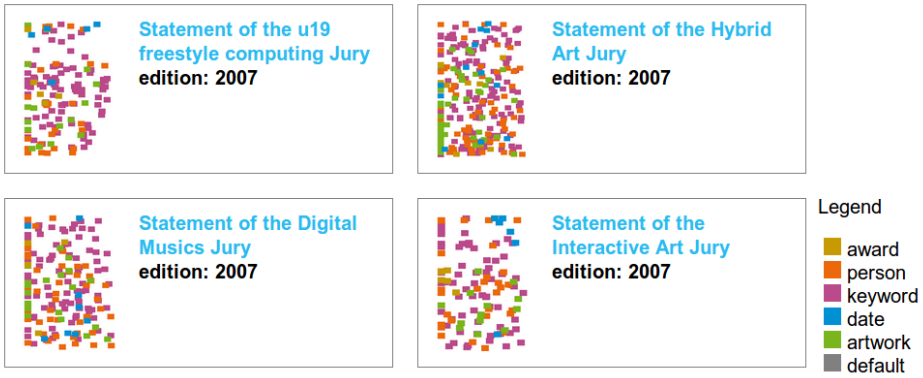

Figure A.7.: Detail of four textys showing texts and the legend from Ars Electronica Jury Statements (2008)

8 ▷ Bible Cross-References by Chris Harrison (2008)

- data: Bible [Large dataset]
- method: Arc graph + bar graph
- description: "The bar graph that runs along the bottom represents all of the chapters in the Bible. Books alternate in color between white and light gray. The length of each bar denotes the number of verses in the chapter. Each of the 63,779 cross references found in the Bible is depicted by a single arc - the color corresponds to the distance between the two chapters, creating a rainbow-like effect."

[http://www.chrisharrison.net/index.php/Visualizations/BibleViz]

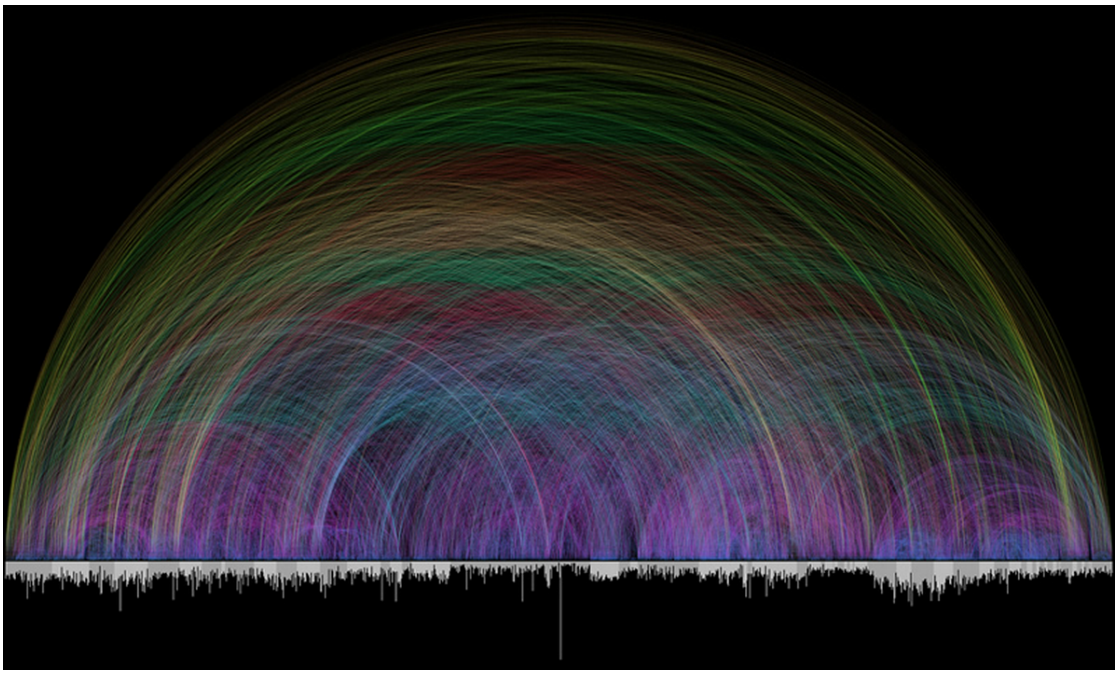

Figure A.8.: Bible Cross-References by Chris Harrison (2008)

9 ▷ Literature fingerprint by Daniel A. Keim and Daniela Oelke (2007)

- data: Any text, from middle to long
- method: Colored squares representing elements of a text in a sequencial way.
- description: Colored squares representing elements of a text in a sequencial way. Description/Comments: the authors develop visualization techniques to show the results of some linguistic analysis that represent a literature fingerprint, demonstrating text authorship.
- paper: *Keim D. A. & Oelke D. (2007). Literature Fingerprinting: A New Method for Visual Literary Analysis. In: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (VAST '07). IEEE Computer Society, Washington, DC, USA, 115-122. DOI=10.1109/VAST.2007.4389004 http://dx.doi.org/10.1109/VAST.2007.4389004*

▷ [http://bib.dbvis.de/uploadedFiles/71.pdf]

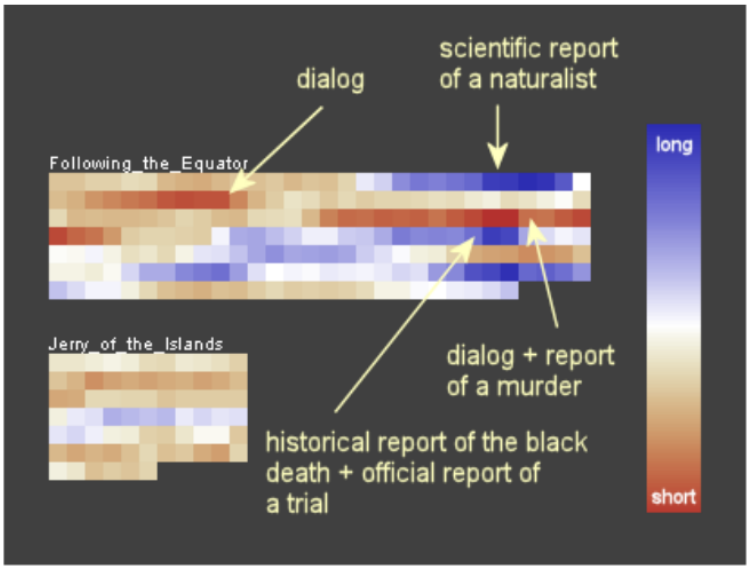

Figure 3: The figure shows the fingerprints of two novels that almost have the same average sentence length. In the detailed view, the different structure of the two novels is revealed. The inhomogeneity of the travelogue *Following the Equator* can be explained with the alternation of dialogs, narrative parts and quoted documents.

Figure A.9.: Sentence length visualization

10 ▷ History Flow by Fernanda Viégas and Martin Wattenberg (2003)

- data: Wikipedia [Small dataset]
- method: Flow chart time-line
- description: "the history flow application charts the evolution of a document as it is edited by many people using a very simple visualization technique. All text segments contributed by the same editor are marked with a unique color. The diagram shows how some editors produce long lasting content, while others don't. With enough editors contributing to an article, almost every paragraph or even sentence gets modified in the long run."
- paper: *Viégas, F. B., Wattenberg, M., & Dave, K. (2004, April). Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 575-582). ACM. [http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf]*

[http://hint.fm/projects/historyflow/]

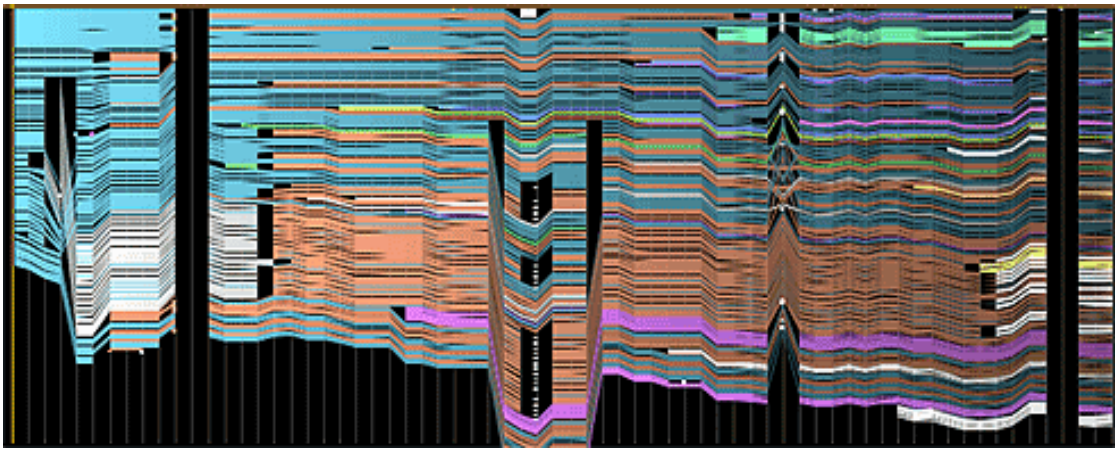

Figure A.10.: History Flow by Fernanda Viégas and Martin Wattenberg researchers at IBM's Visual Communication Lab (2003)

11 ▷ Colour-coded chronological sequencing by Joel Deshaye and Peter Stoicheff (2003)

- data: The Sound and the Fury by William Faulkner
- method: Timeline
- description: Case collected by Peter Stoicheff: "This visualization (contributed by Joel Deshaye and Peter Stoicheff) shows a compressed version of the colour-coded "April Seventh, 1928" narrative on the left. The middle bar extracts the narrative of Benjy's present day. The right bar extracts the flashbacks to Caddy's wedding. Although the section seems randomly ordered, within it the present and each flashback reside independently as coherent, chronological sequences."
- paper: *Stoicheff, R.P. "Faulkner's Foreign Levy: Macbeth ,The Sound and the Fury, and Writerhood." The Sound and the Fury: a Hypertext Edition. Ed. Stoicheff, Muri, Deshaye, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <http://www.usask.ca/english/faulkner>*

[http://drc.usask.ca/projects/faulkner/main/benjy_spectrum.htm]

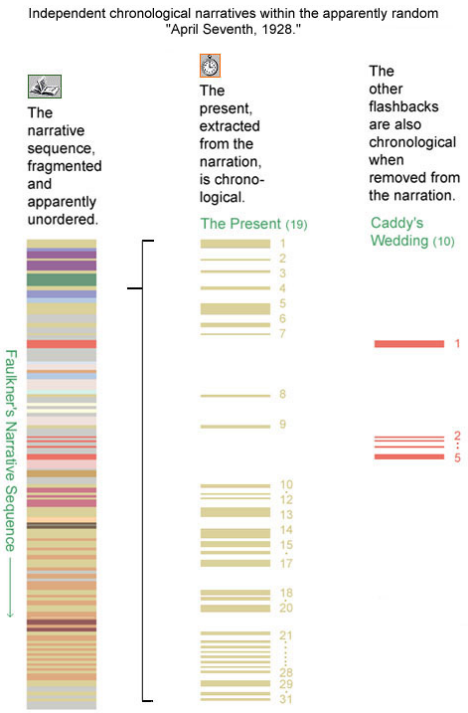

Figure A.11.: Colour-coded chronological sequencing of William Faulkner's novel The Sound and the Fury, by Joel Deshaye and Peter Stoicheff (2003)

12 ▷ 2-D display of time in the novel by Joel Deshaye (2003)

- data: The Sound and the Fury by William Faulkner
- method: Customized 2-D diagram
- description: Case collected by Peter Stoicheff: "This graph (contributed by Joel Deshaye) shows two dimensions of time in The Sound and the Fury: chronological time, and Faulkner's re-ordering of chronological time into the text's narrative sequence. Faulkner scrambles the chronology not only by flashbacks, but also by the non-linear sequence of the novel's sections. Through this graph it becomes clear that the novel follows a conventional in medias res structure, whose existence is otherwise obscured by more local narrative complexities."
- paper: *Stoicheff, R.P. "Faulkner's Foreign Levy: Macbeth ,The Sound and the Fury, and Writerhood." The Sound and the Fury: a Hypertext Edition. Ed. Stoicheff, Muri, Deshaye, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <http://www.usask.ca/english/faulkner>*

[http://drc.usask.ca/projects/faulkner/main/sf_2d_timegraph.htm]

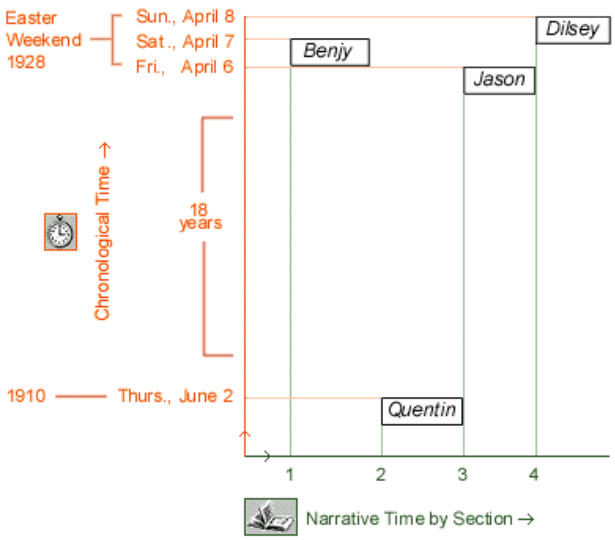

Figure A.12.: 2-D display of time of William Faulkner's novel The Sound and the Fury, by Joel Deshaye and Peter Stoicheff (2003)

13 ▷ 3-D display of time in the novel by Joel Deshaye (2003)

- data: The Sound and the Fury by William Faulkner
- method: Highly customized 3D diagram
- description: Case collected by Peter Stoicheff: "This graph (contributed by Joel Deshaye) represents time in three different dimensions. First, there is the sequence of four sections in the novel - the narrative time or récit (shown in green). Second, there is the chronology, l'histoire (shown in orange), beginning with the earliest recollected date in the novel - Damuddy's death in 1898. These two dimensions show how the conventional view of linear time can be disrupted by a fictional narrative. Third, there is a representation of the proportion of memories of and flashbacks to the past in each section (shown in blue). This last dimension shows how the novel progresses from an emphasis on the past toward an emphasis on the present."
- paper: *Stoicheff, R.P. "Faulkner's Foreign Levy: Macbeth ,The Sound and the Fury, and Writerhood." The Sound and the Fury: a Hypertext Edition. Ed. Stoicheff, Muri, Deshaye, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <http://www.usask.ca/english/faulkner>*

[http://drc.usask.ca/projects/faulkner/main/sf_3d_timegraph.htm]

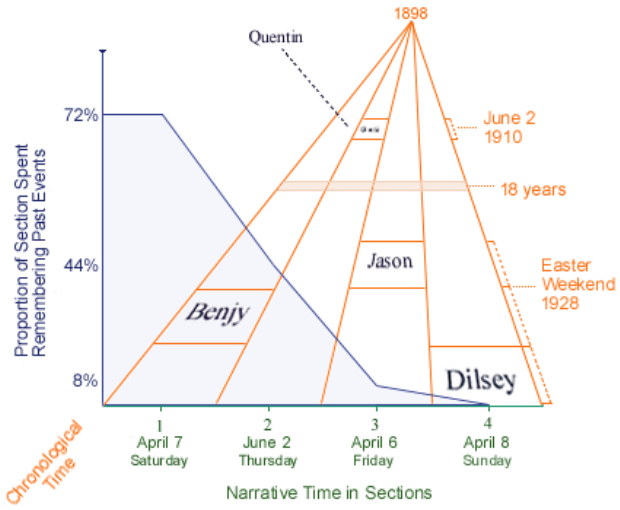

Figure A.13.: 3-D display of timeof William Faulkner's novel The Sound and the Fury, by Joel Deshaye and Peter Stoicheff (2003)

14 ▷ Arc diagram Wattenberg by Martin Wattenberg (2002)

- data: Any text or string [Small dataset]
- method: Arc diagram
- description: Arc diagram is capable of representing complex patterns of repetition in string data. Arc diagrams improve over previous methods such as dotplots because they scale efficiently for strings that contain many instances of the same subsequence.
- paper: *Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on (pp. 110-116). IEEE.*

▷ [http://domino.watson.ibm.com/cambridge/research.nsf/58bac2a2a6b05a1285256b30005b3953/e2a83c4986332d4785256ca7006cb621/$FILE/TR2002-11.pdf]

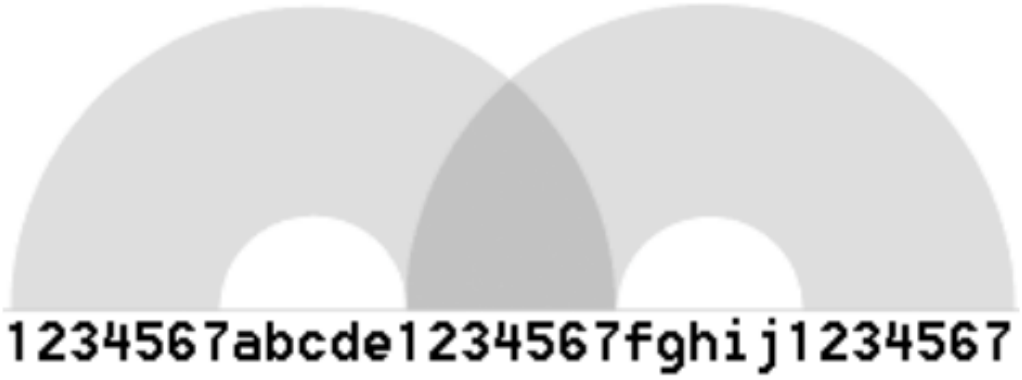

Figure A.14.: Arc diagram by Martin Wattenberg (2002)

15 ▷ TileBars: Visualization of Term Distribution Information in Full Text Information Access by Marti A. Hearst (1995)

- data: search results from information retrieval systems. Applied to PubMed results.
- method: iconic representation of a text
- description: "TileBars, which provides a compact and informative iconic representation of the documents' contents with respect to the query terms. The goal is to simultaneously indicate: the relative length of the document, the frequency of the term sets in the document, and the distribution of the term sets with respect to the document and to each other. Each large rectangle indicates a document, and each square within the document represents a TextTile. The darker the tile, the more frequent the term (white indicates 0, black indicates 8 or more instances)"
- paper: Hearst, M. 1995. "TileBars: visualization of term distribution information in full text information access." Proceedings of the SIGCHI conference on Human http://dl.acm.org/citation.cfm?id=223912 (March 26, 2013).

[http://dl.acm.org/citation.cfm?id=223912]

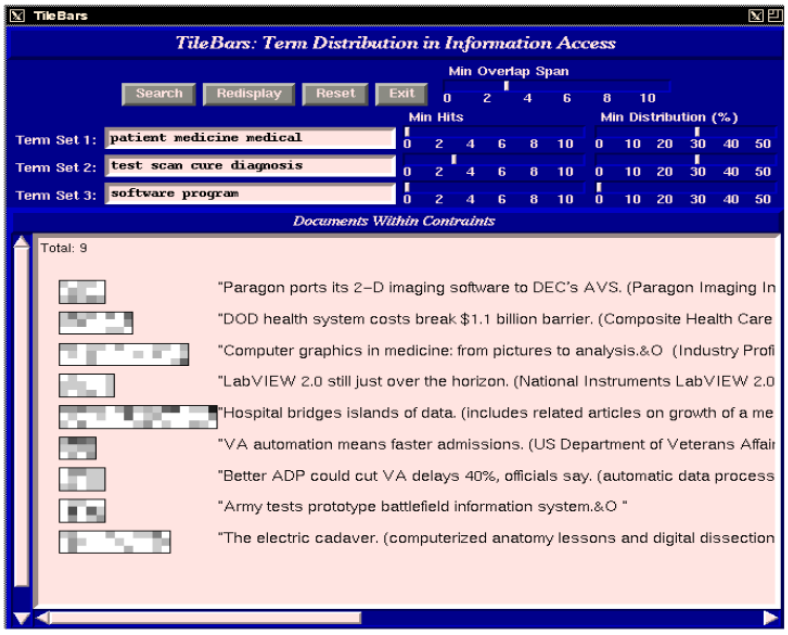

Figure A.15.: TileBar search on (patient medicine medical AND test scan cure diagnosis AND software program) with stricter distribution constaints.

A.1.2. Part text visualization

16 ▷ Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)

- data: Novel "Les Misérables" [N/A]
- method: Tables + condensed bar graphs
- description: "the verbs used together with character names in a novel can provide a glimpse into the personalities and actions of that character. For the primary people in the novel Les Misérables this graphic illustrates their characteristic verbs."

[http://neoformix.com/2013/NovelViews.html]

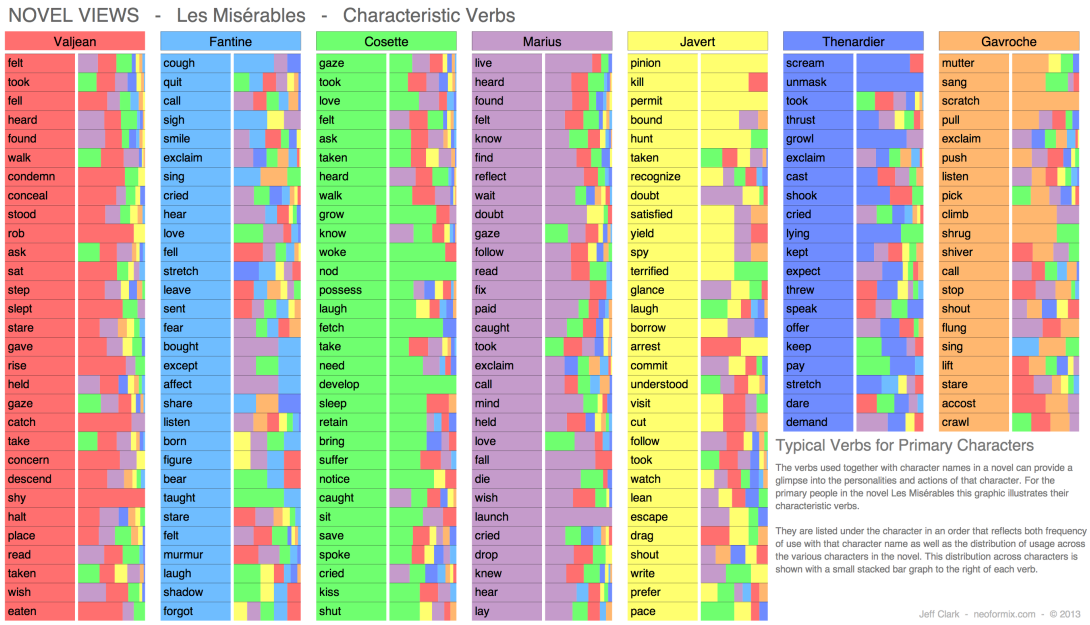

Figure A.16.: Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)

17 ▷ Wordle by Jonathan Feinberg (2009)

- data: any text
- method: Word cloud
- description: Wordle is a toy for generating "word clouds" from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text.
- paper: *Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with wordle. Visualization and Computer Graphics, IEEE Transactions on, 15(6), 1137-1144.*

▷ [http://cyber-kap.blogspot.com.au/2011/04/top-10-sites-for-creating-word-clouds.html]

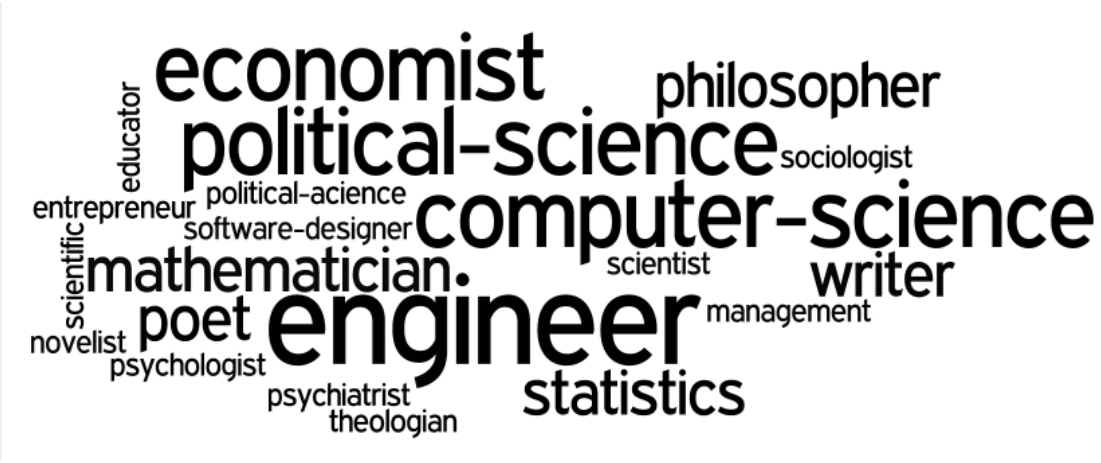

Figure A.17.: Wordle showing the professions of inventors of data visualization tools

18 ▷ Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)

- data: any text, specially used with books. It is a visualization of Wordnet lexical database.
- method: Radial, space-filling layout of hyponymy (IS-A relation)
- description: DocuBurst is the first visualization of document content which takes advantage of the human-created structure in lexical databases. The authors used an accepted design paradigm to generate visualizations which improve the usability and utility of WordNet as the backbone for document content visualization. A radial, space-filling layout of hyponymy (IS-A relation) is presented with interactive techniques of zoom, filter, and details-on-demand for the task of document visualization. The techniques can be generalized to multiple documents.
- paper: *C. Collins, S. Carpendale, and G. Penn, "DocuBurst: Visualizing Document Content Using Language Structure," Computer Graphics Forum (Proc. of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis)), vol. 28, iss. 3, pp. 1039-1046, 2009.*

▷ [http://vialab.science.uoit.ca/portfolio/docuburst-visualizing-document-content-using-language-structure]

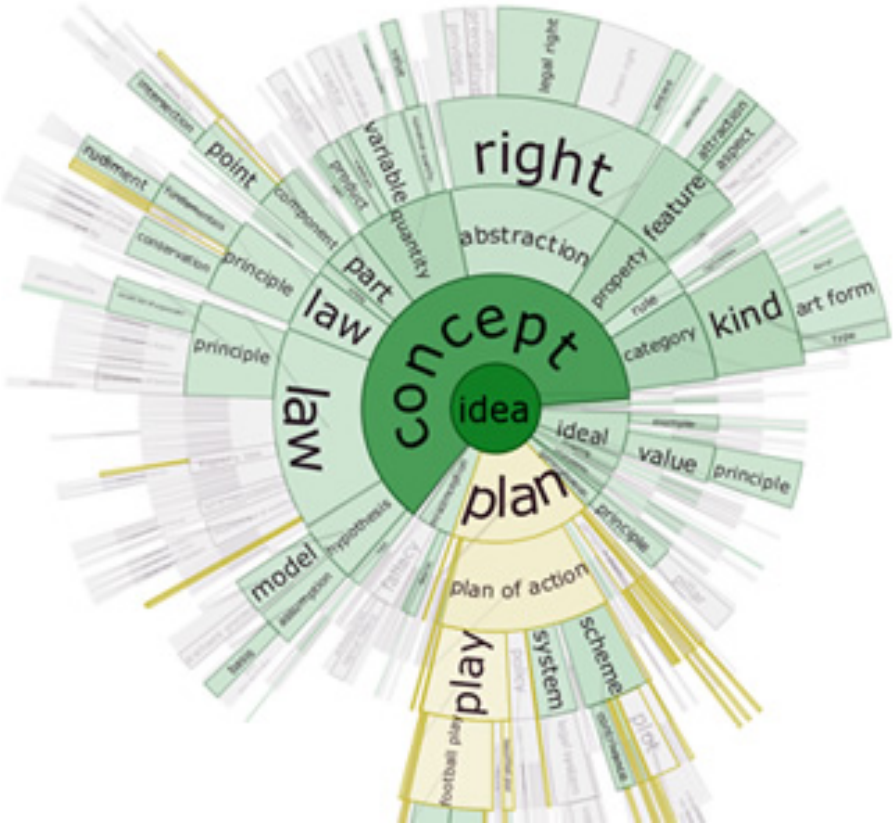

Figure A.18.: Screenshot of docuburst interactive interface.

19 ▷ Phrase Nets by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

- data: Any text. Jane Austen's novel "Pride and Prejudice." [N/A]
- method: Network of sized-words, some connected
- description: "Phrase Nets use a simple form of pattern matching to provide multiple views of the concepts contained a book, speech, or poem. The image below is a word graph made from Jane Austen's novel "Pride and Prejudice." The program has drawn a network of words, where two words are connected if they appear together in a phrase of the form "X and Y":"
- paper: *Van Ham, F., Wattenberg, M., & Viégas, F. B. (2009). Mapping text with phrase nets. Visualization and Computer Graphics, IEEE Transactions on, 15(6), 1169-1176.*

[http://hint.fm/projects/phrasenet/]

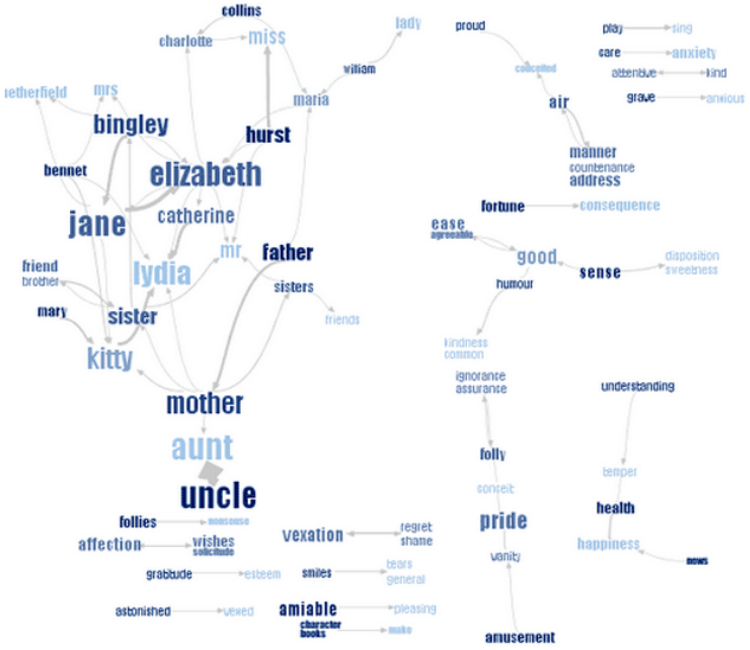

Figure A.19.: Phrase Nets of Jane Austen's novel "Pride and Prejudice." by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

20 ▷ Word Spectrum: Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

- data: The huge Google bi-gram dataset [Large dataset]
- method: Word spectrum
- description: "Using Google's enormous bigram dataset, I produced a series of visualizations that explore word associations. Each visualization pits two primary terms against each other. Then, the use frequency of words that follow these two terms are analyzed. For example, "war memorial" occurs 531,205 times, while "peace memorial" occurs only 25,699. A position for each word is generated by looking at the ratio of the two frequencies. If they are equal, the word is placed in the middle of the scale. However, if there is a imbalance in the uses, the word is drawn towards the more frequently related term. This process is repeated for thousands of other word combinations, creating a spectrum of word associations. Font size is based on a inverse power function (uniquely set for each visualization, so you can't compare across pieces). Vertical positioning is random."

[http://www.chrisharrison.net/index.php/Visualizations/WordSpectrum]

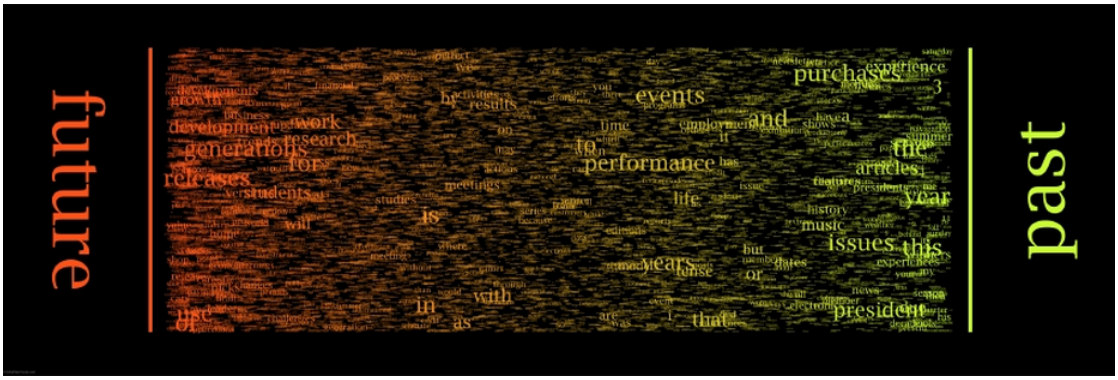

Figure A.20.: ▷ Word Spectrum: Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

21 ▷ Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

- data: The huge Google bi-gram dataset [Large dataset]
- method: Words rays
- description: "words are bucketed into one of 25 different rays. Each of these represent a different tendency of use (ranging from 0 to 100% in 4% intervals). Words are sorted by decreasing frequency within each ray. I render as many words as can fit onto the canvas. There is a nice visual analogy at play - the "lean" of each ray represents the strength of the tendency towards one of the two terms. As in the word spectrum visualization, font size is based on a inverse power function (uniquely set for each visualization, so you can't compare across pieces). Common words (a, the, for, as, etc.) are not shown."

[http://www.chrisharrison.net/index.php/Visualizations/WordAssociations]

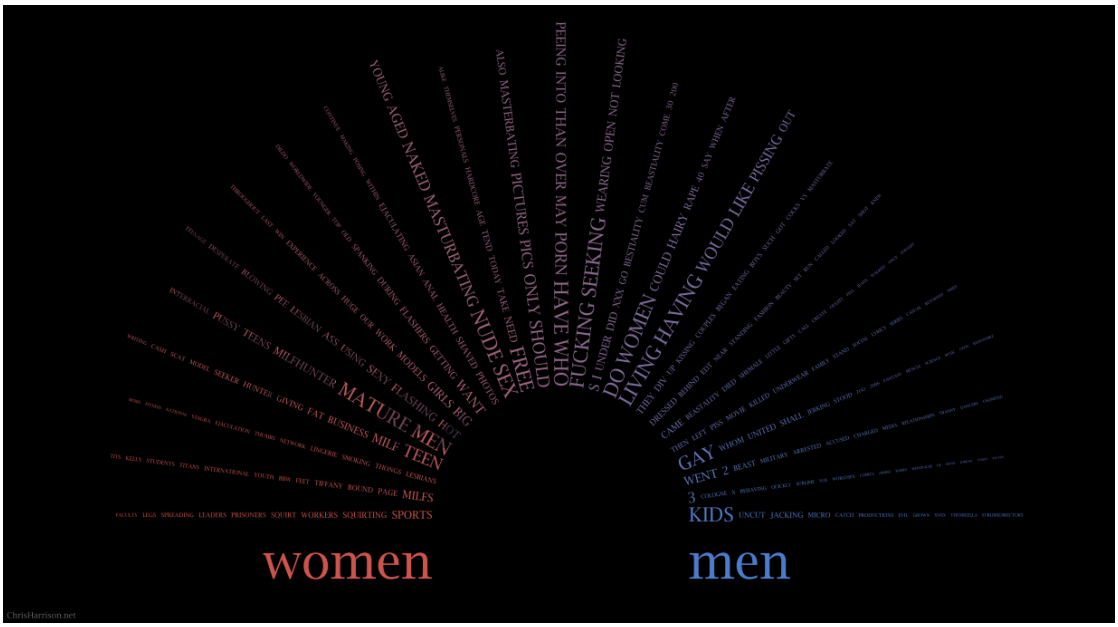

Figure A.21.: Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

22 ▷ Document Arc Diagrams by Jeff Clark (2007)

- data: any text
- method: Arc diagram
- description: Document Arc Diagrams illustrate the similarity structure within a text document by drawing arcs connecting segments of a document that share similar vocabulary.

▷ [http://www.neoformix.com/2007/DocumentArcDiagrams.html]

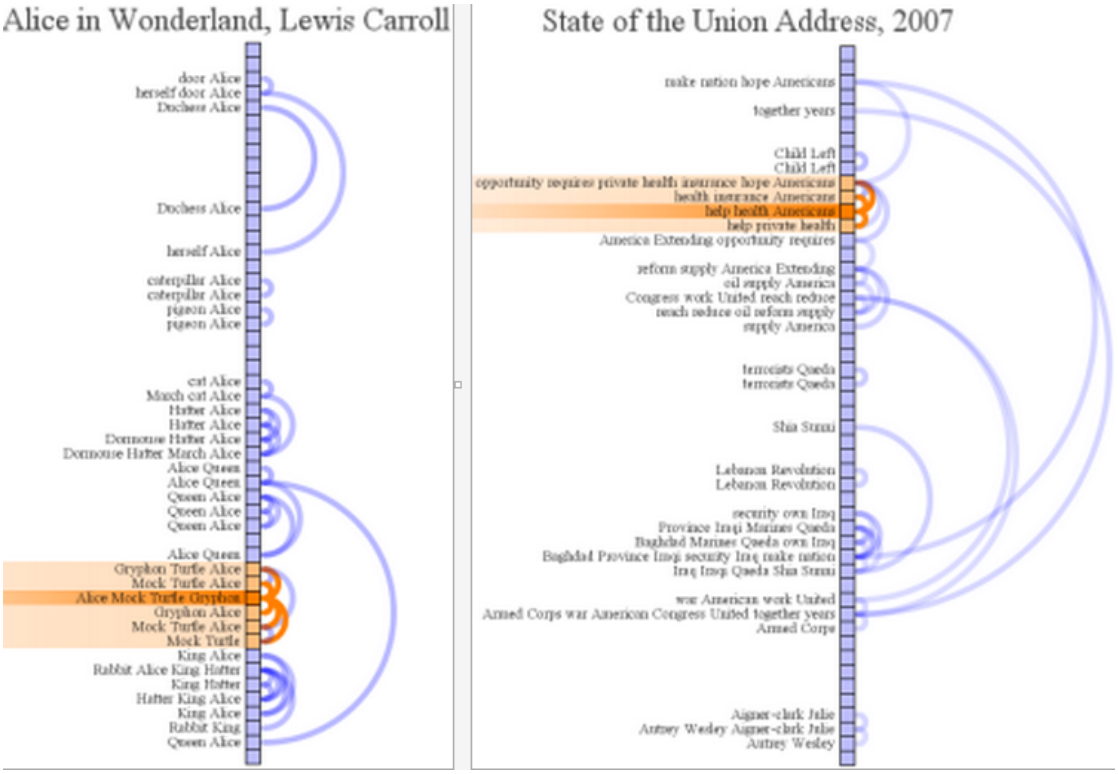

Figure A.22.: Jeff Clark version of arc diagrams

23 ▷ Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

- data: any body of literature: 1) The complete set of approximately 7 million USPTO patents; 2) Enron email data set comprised of 500,000 emails; 3) A collection of computer generated speech transcripts from MIT Media Lab Symposia.
- method: interactive radial histogram
- description: The shape contains the semantic profile of a single document where the peaks and valleys are defined by the relatedness of words or concepts to that document.
- paper: *Decamp P., Frid-Jimenez A., Guiness J., Roy D.: Gist icons: Seeing meaning in large bodies of literature. In Proc. of IEEE Symp. on Information Visualization, Poster Session (Oct. 2005).*

▷ [http://media.mit.edu/cogmac/publications/IEEEIcons.pdf]

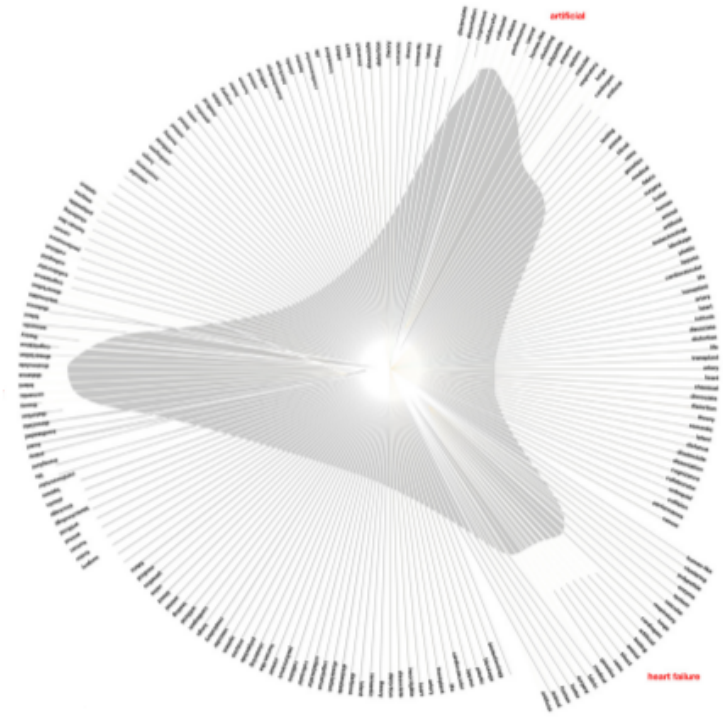

Figure A.23.: Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

A.2. Texts collection visualizations

A.2.1. Collections of items visualizations

24 ▷ Novel Views: Les Misérables - Segment Word Clouds by Jeff Clark (2013)

- data: Novel "Les Misérables" [N/A]
- method: Word cloud
- description: "a series of small word clouds for each book within the novel are shown."

[http://neoformix.com/2013/NovelViews.html]

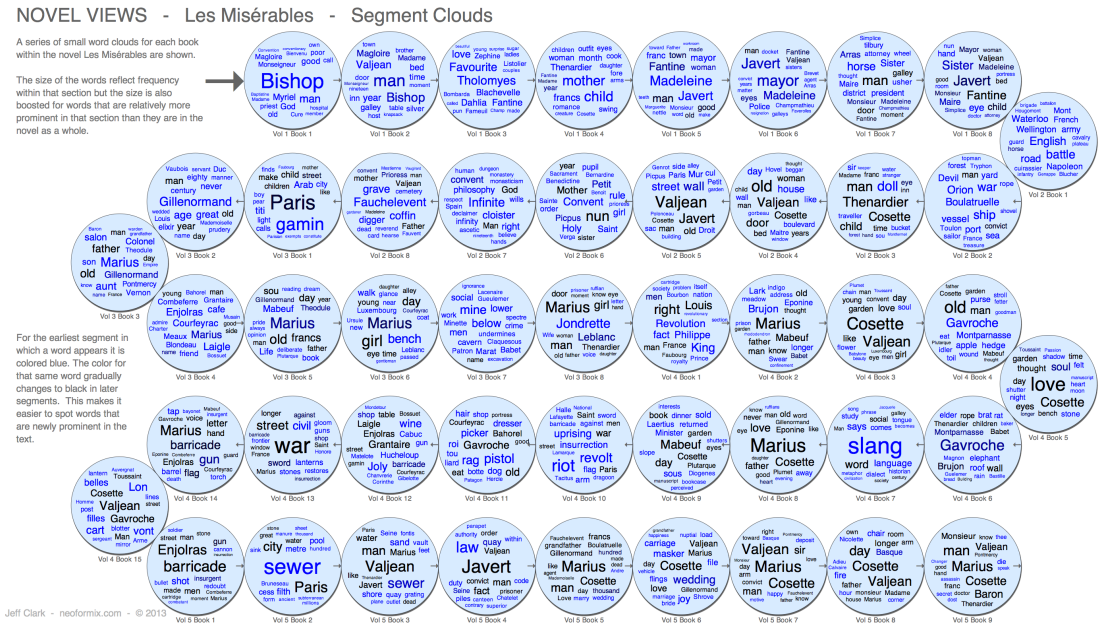

Figure A.24.: Novel Views: Les Misérables - Segment Word Clouds by Jeff Clark (2013)

25 ▷ Grimm's Fairy Tale Network by Jeff Clark (2013)

- data: 62 stories of the Grimms's Fairy Tales [Small dataset]
- method: 2-dimension networks
- description: "the graphic below is a simple network showing which stories are connected through the use of a common vocabulary. There are three different strengths of connection shown and I've tried to minimize the usual 'hairball' nature of these types of diagrams by only showing the top three connections for a story."

[http://neoformix.com/2013/GrimmNetwork.html]

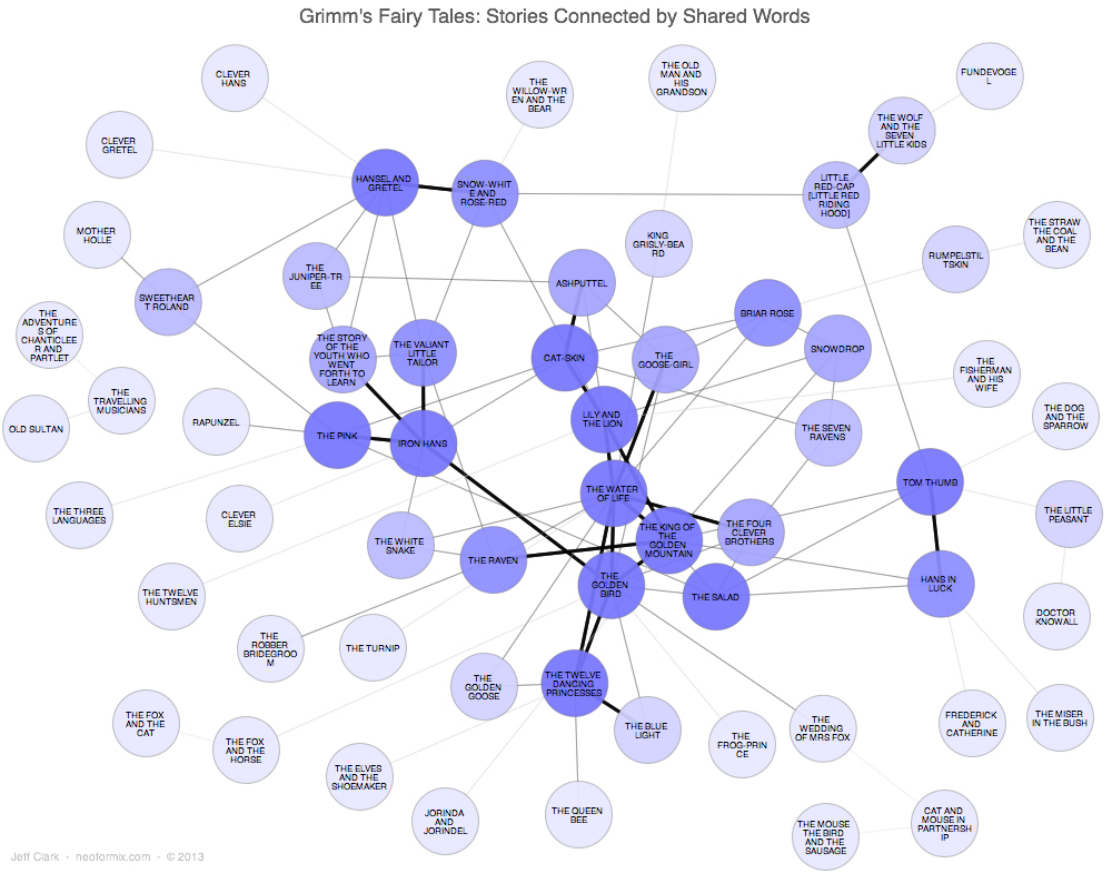

Figure A.25.: Grimm's Fairy Tale Network by Jeff Clark (2013)

26 ▷ Spot by Jeff Clark (2012)

- data: Twitter [Small dataset]
- method: Multi visualization of tweets based on a search query.
- description: "Spot is an interactive real-time Twitter visualization that uses a particle metaphor to represent tweets. The tweet particles are called spots and get organized in various configurations to illustrate information about the topic of interest."

[http://neoformix.com/2012/IntroducingSpot.html]

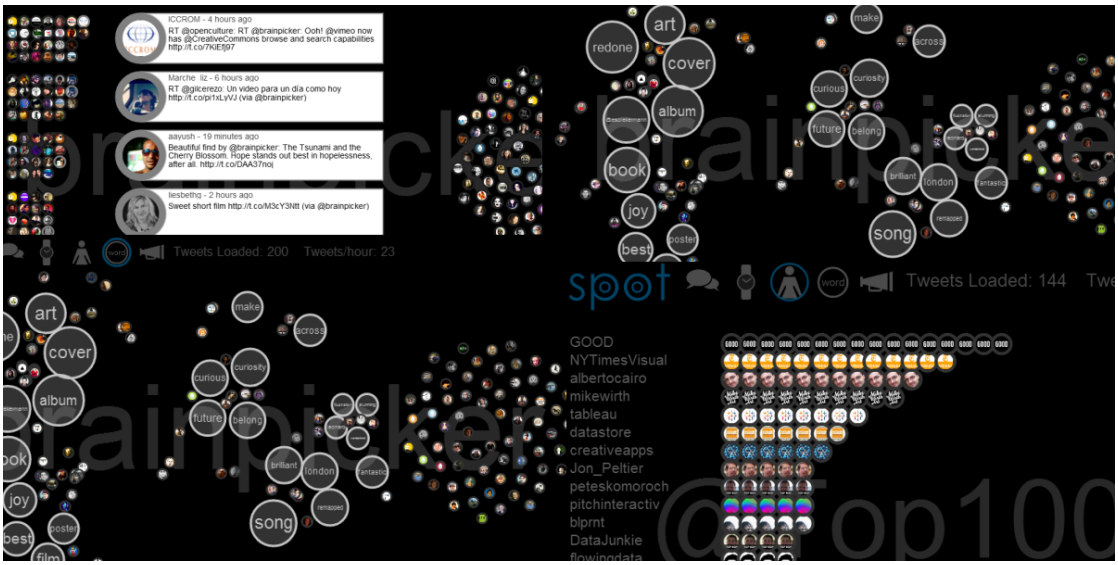

Figure A.26.: Spot by Jeff Clark (2012)

27 ▷ Word storm by Quim Castella and Charles Sutton (2012)

- data: Papers from ICML - International Conference on Machine Learning June 26July 1, 2012 Edinburgh, Scotland [Small dataset]
- method: Word cloud with fixed word position
- description: visualization of the conference session in the form of a word storm, which is a group of word clouds. The clouds are arranged so that if the same word appears in two clouds, it is in the same position. This is intended to make it easier to see the difference between clouds.
- paper: *Castella, Q., & Sutton, C. (2013). Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. arXiv preprint arXiv:1301.0503.*

[http://icml.cc/2012/whatson-all/]

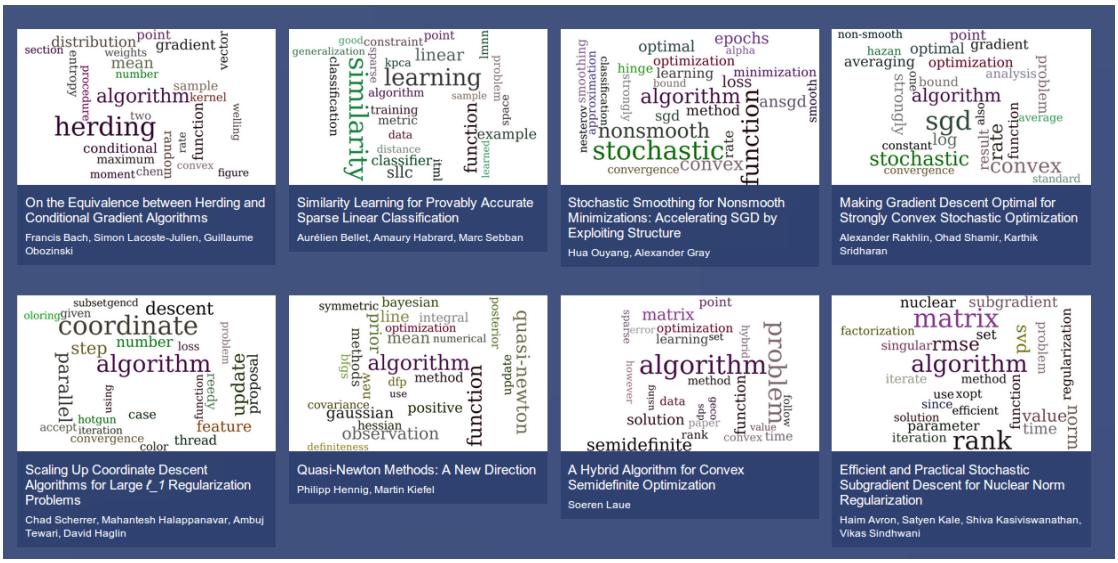

Figure A.27.: Word storm by Quim Castella and Charles Sutton (2012)

28 ▷ Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

- data: Marcel Proust texts [Large dataset]
- method: Topic model network
- description: "this is document-topic network representation. This visualization shows the relationships abd topics of a collection of documents. Visually this is represented by the distances between documents, topics and documents-topic."

[https://dhs.stanford.edu/algorithmic-literacy/topic-networks-in-proust/]

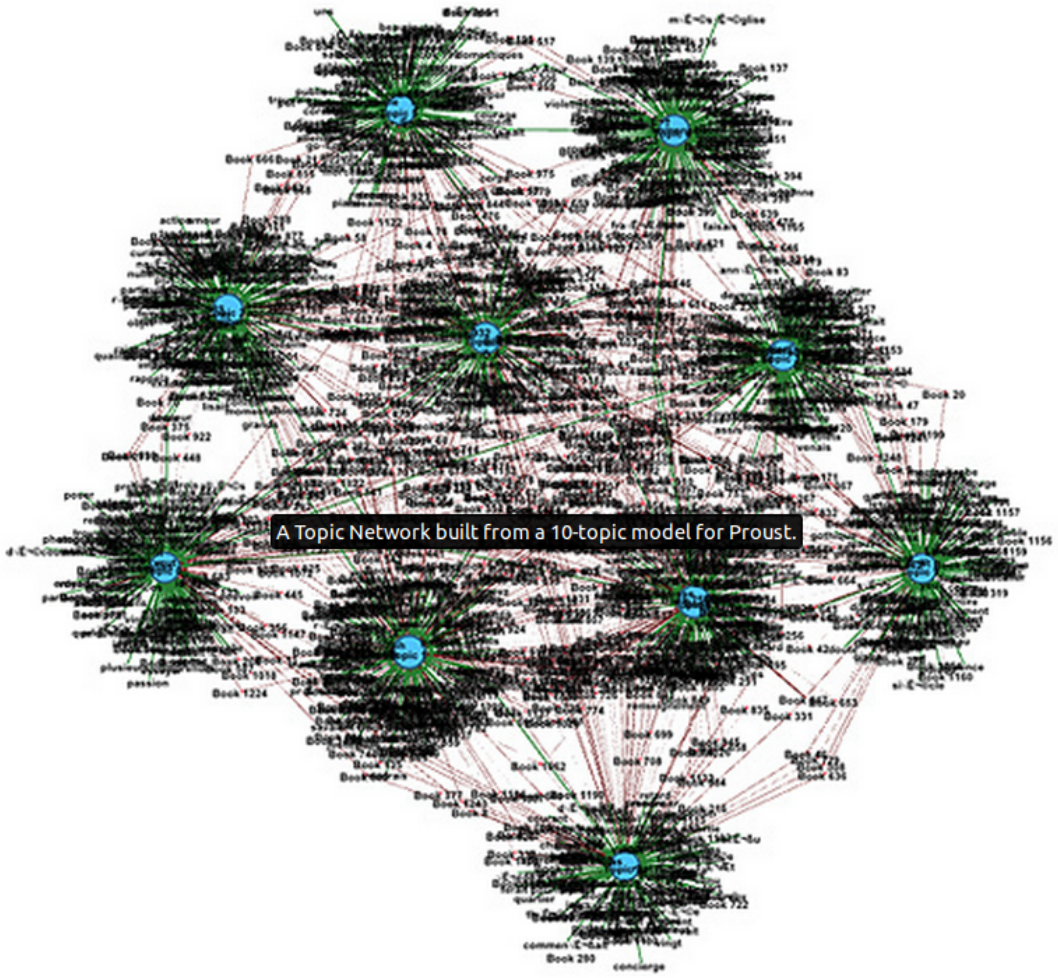

Figure A.28.: Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

29 ▷ Notabilia. 100 Longest Article for Deletion [AfD] discussions on Wikipedia by Dario Taraborelli, Giovanni Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization). (2010)

- data: 100 Longest Article for Deletion [AfD] of Wikipedia [Small dataset]
- method: Original. Bouquet of edition lines
- description: "visualization of debate deletion of entries in Wikipedia. Each time a user joins an AfD discussion and recommends to keep, merge, or redirect the article a green segment leaning towards the left is added. Each time a user recommends to delete the article a red segment leaning towards the right is added. As the discussion progresses, the length of the segments as well as the angle slowly decay."

[http://notabilia.net/]

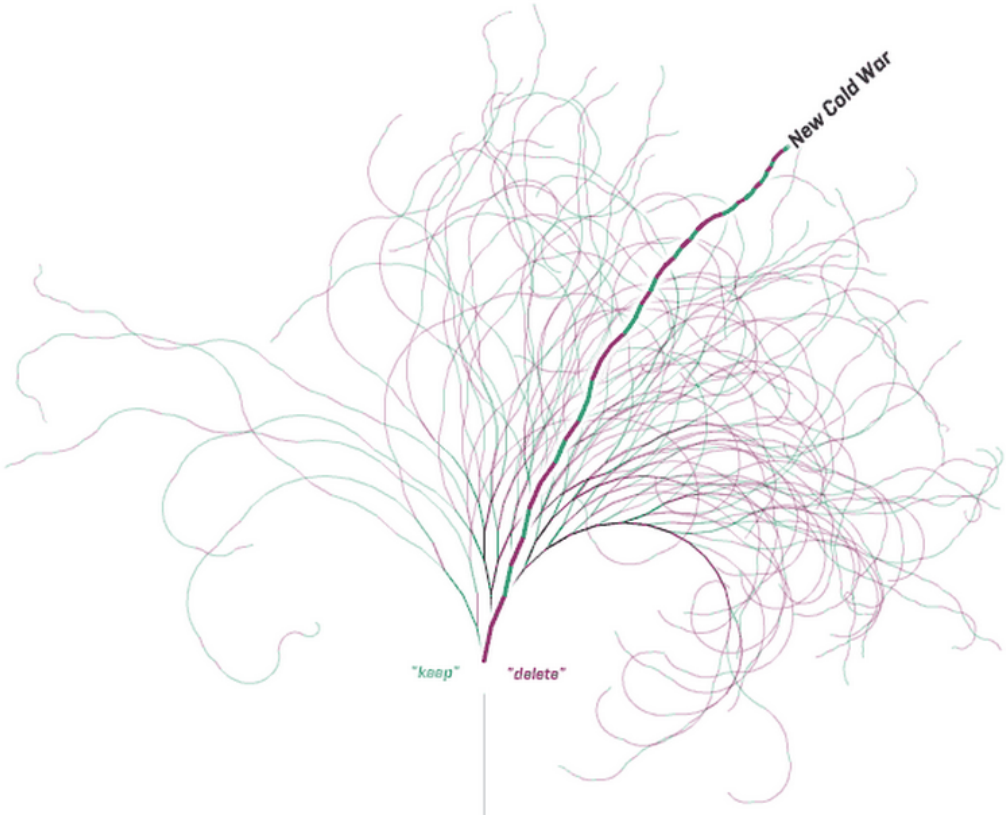

Figure A.29.: Notabilia. 100 Longest Article for Deletion [AfD] discussions on Wikipedia by Dario Taraborelli, Giovanni Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization). (2010)

30 ▷ X by Y by Moritz Stefaner (2009)

- data: Almost 40.000 submissions to the Prix Ars Electronica, from the early beginnings in 1987 up to 2009 [Large dataset]
- method: Visual spots in groups and colors.
- description: "X by Y visualizes all submissions to the Prix Ars Electronica, from the early beginnings in 1987 up to 2009. The goal is to characterize the "ars world" in quantitative terms. A series of diagrams groups and juxtaposes the submissions by years, categories, prizes and countries. The graphics are composed of little dots (each representing a single submission) to provide a visual scale for the statistical statements and thematize the relation of the totality and the individual."
- Book reference: *Stefaner, M., Ferré, S., Perugini, S., Koren, J., & Zhang, Y. (2009). User interface design. In Dynamic Taxonomies and Faceted Search (pp. 75-112). Springer Berlin Heidelberg.*

[http://moritz.stefaner.eu/projects/x-by-y/]

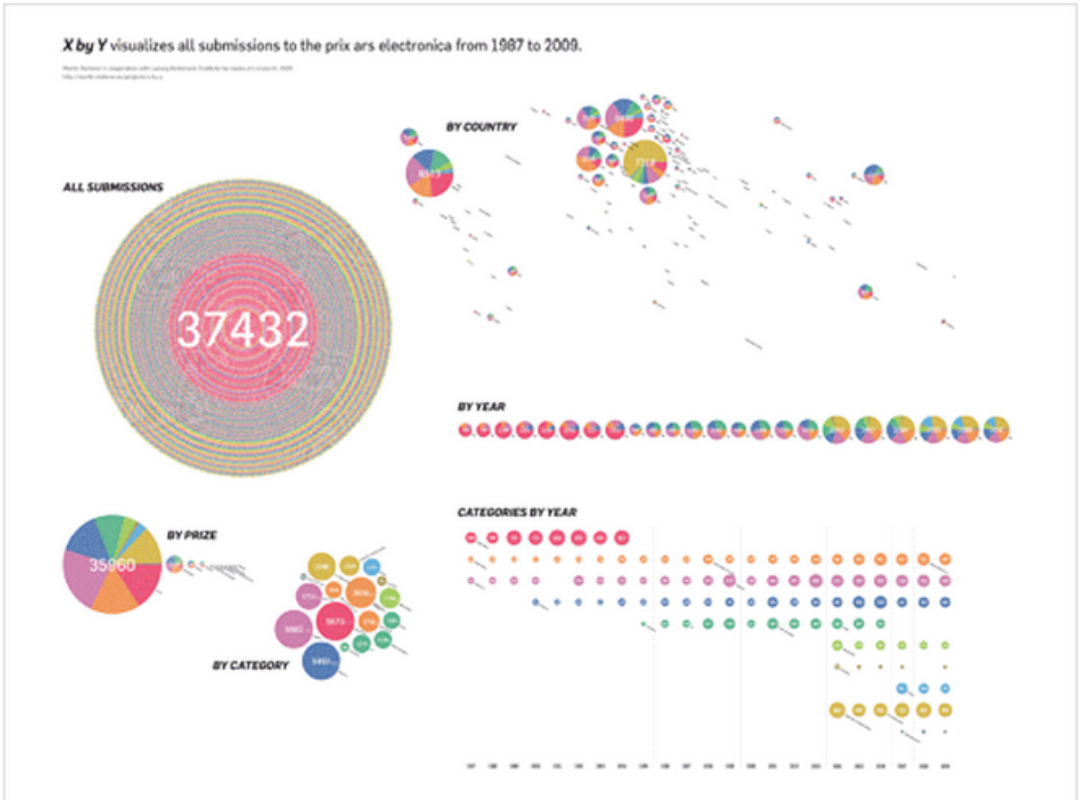

Figure A.30.: X by Y by Moritz Stefaner (2009)

31 ▷ Search Clock by Chris Harrison (2008)

- data: Search engine queries [Large dataset]
- method: Clock with queries
- description: "I was curious to see if data from search engines would support my anecdotal observations. I built a simple clock-like visualization that displays the top search terms over a 24-hour period. Displaying search terms in a cyclical layout (like a clock) allows continuous examination of trends that would otherwise be broken up. The data I had access to was both large and noisy. In response, I combined hourly data into week or year averages."

[http://www.chrisharrison.net/index.php/Visualizations/SearchClock]

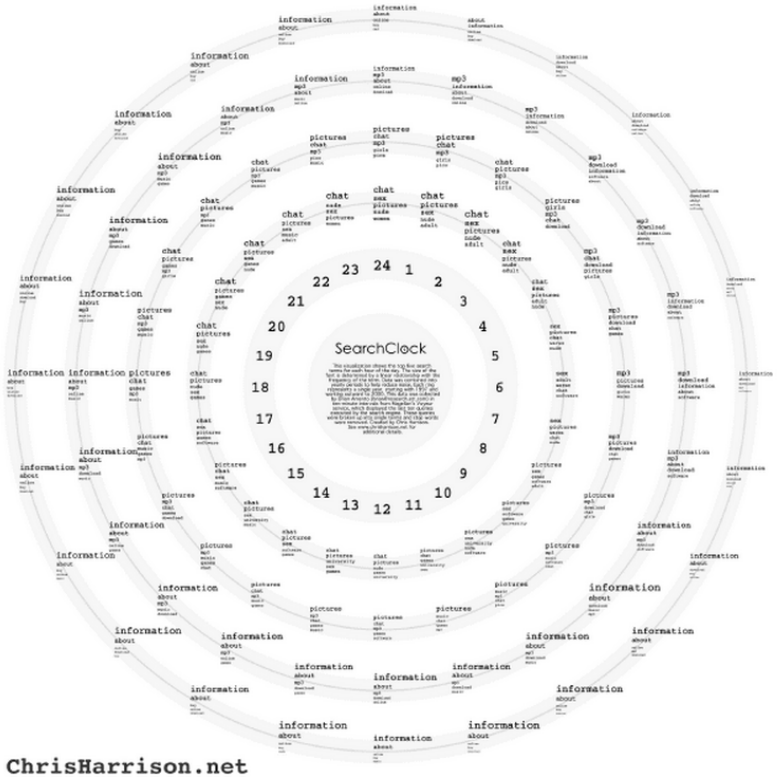

Figure A.31.: Search Clock by Chris Harrison (2008)

32 ▷ Digg Rings by Chris Harrison (2008)

- data: Digg data - top 10 most digg stories of the day for May 24, 2007 to May 23, 2008 [Large dataset]
- method: Tree-ring-like visualizations
- description: "using the Digg API, I grabbed the top 10 most-dugg stories of the day (by midnight) for the past year - May 24, 2007 to May 23, 2008. I then rendered a series of tree-ring-like visualizations (moving outwards in time). Rings are colored according to Digg's eight top-level categorizations (see key at bottom of page). Ring thickness is linearly proportional to the number of diggs the story received. I also made a pair of visualizations using Digg's entire archive, which goes back to December 1, 2004."

[http://www.chrisharrison.net/index.php/Visualizations/DiggRings]

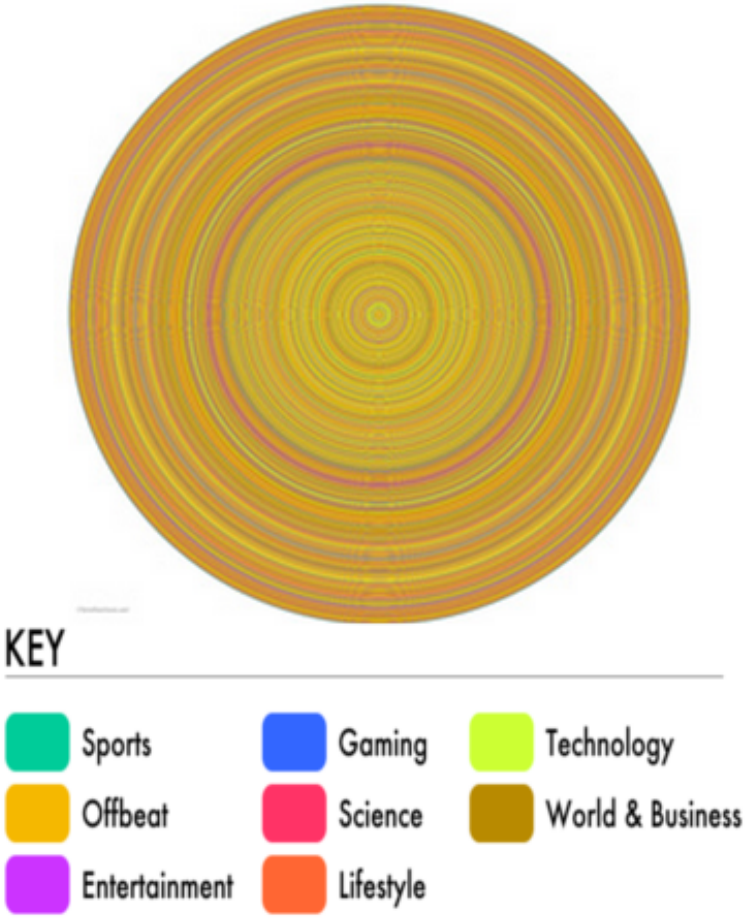

Figure A.32.: Digg Rings by Chris Harrison (2008)

33 ▷ Royal Society Archive by Chris Harrison (2008)

- data: titles of papers of the Royal Society Archive (1665-2005) [Large dataset]
- method: time-line
- description: "The Royal Society recently provided access to an archive of papers published in the scientific academy's prestigious journals. Some 25 thousand scholarly works are represented, which date from 1665 to 2005. Many notable scientific minds are represented, including Isaac Newton, Michael Faraday and Charles Darwin. This interesting data set was ripe for some visual tinkering. The database I used was put together by Brian Amento and Mike Yang of AT&T Labs."

[http://www.chrisharrison.net/index.php/Visualizations/RoyalSociety]

Figure A.33.: Detail of the visualization Royal Society Archive by Chris Harrison (2008)

34 ▷ WikiViz: Visualizing Wikipedia by Chris Harrison (2007)

- data: Wikipedia [Large dataset]
- method: network
- description: "Wikipedia is an interesting dataset for visualization. As an encyclopedia, it's articles span millions of topics. Being a human edited entity, connections between topics are diverse, interesting, and sometimes perplexing - five hops takes you from subatomic particles to Snoop Dog. Wikipedia is revealing in how humans organize data and how interconnected seemingly unrelated topics can be."

[http://www.chrisharrison.net/index.php/Visualizations/WikiViz]

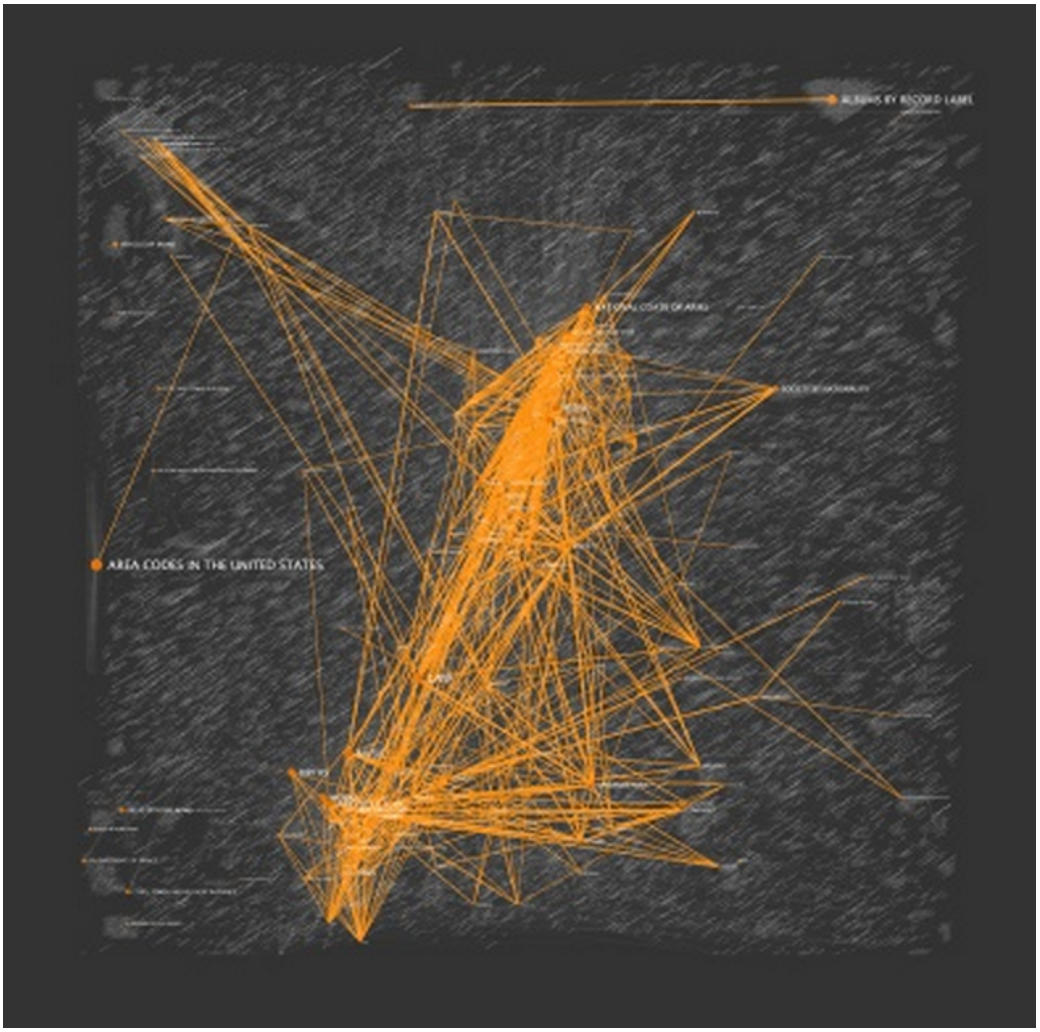

Figure A.34.: WikiViz: Visualizing Wikipedia by Chris Harrison (2007)

35 ▷ Area by Jaume Nualart (2007)

- data: Any collection of texts with parameterizable meta-data to discrete values [Small dataset]
- method: Database representation and browsing tool.
- description: AREA represents the whole dataset, it gives a size-overview of the data, is a data browser, is a data visualization, is a whole data understanding, is interactive, data can be filtered, some time using Area teaches about the main characteristics of the represented data.

[http://nualart.com/area2]

Figure A.35.: Area by Jaume Nualart (2007)

36 ▷ Visualizing Activity on Wikipedia with Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

- data: Wikipedia [Large dataset]
- method: Chromograms
- description: a chromogram describes the unique edit pattern of a Wikipedian over time by categorizing and color coding edits.Chromograms for distinct persons (and bots!) can be markedly dissimilar.
- Wattenberg, *M., Viégas, F. B., & Hollenbach, K. (2007). Visualizing activity on* Wikipedia *with chromograms. In Human-Computer Interaction–INTERACT 2007 (pp. 272-287). Springer Berlin Heidelberg.*

[http://pensivepuffin.com/dwmcphd/syllabi/info447_wi12/readings/wk09-Organizing/wattenberg.Chromo

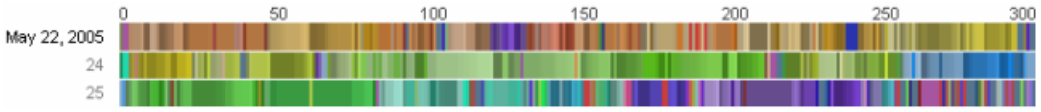

Figure A.36.: Visualizing Activity on Wikipedia with Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

37 ▷ Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001)

- data: Internet web pages [Small dataset]
- method: Search results as a map of web pages with tagged edges
- description: "Kartoo and Ujiko have been the more advanced search engine interfaces for a wide use. The resultsa were presented as a map of web pages with tagged edges. Kartoo (later Ujiko) was norm in 2001 and shut down in 2010."

[http://en.Wikipedia.org/wiki/Kartoo]

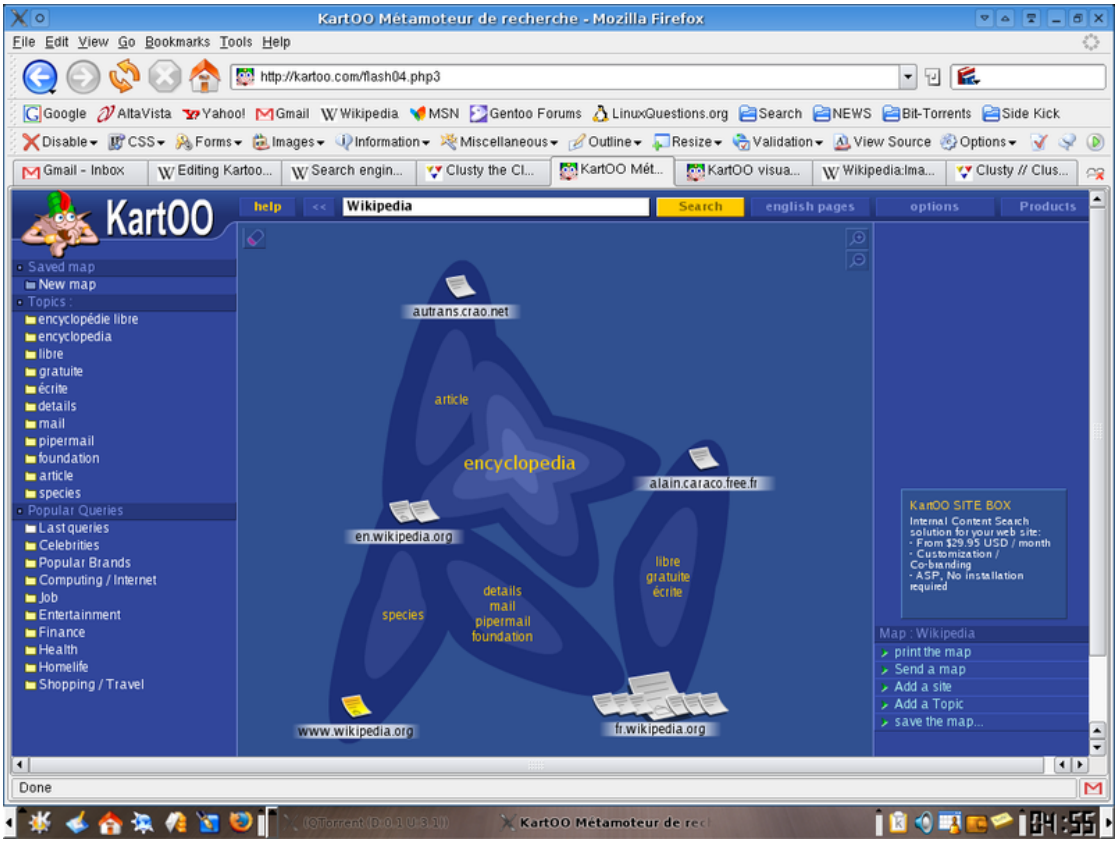

Figure A.37.: Kartoo/Ujiko by Laurent Baleydier and Nicholas Baleydier (2001)

38 ▷ Touchgraph by TouchGraph, LLC. (2001)

- data: Search engine results [Large dataset]
- method: Network visualization of search results
- description: TouchGraph was founded in 2001 with the creation of the original visual browser for Google. Since then, millions of people have used TouchGraph's tools to discover the relationships contained in Google, Amazon, Wikis, and other popular information sources.

[http://www.touchgraph.com/seo]

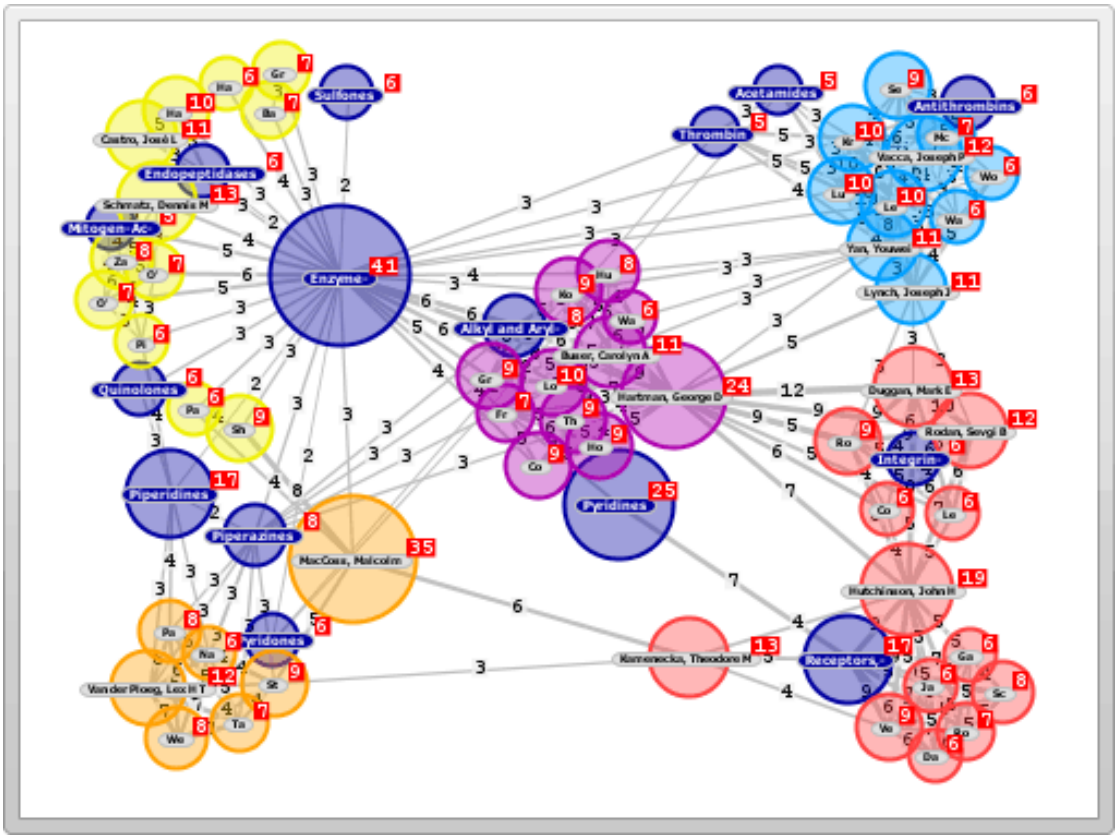

Figure A.38.: Touchgraph by TouchGraph, LLC. (2001)

39 ▷ HotSauce by Ramanathan V. Guha (1996)

- data: Websites relationships written in Meta Content Framework (MCF), a precursor of RDF. [Small dataset]
- method: 3D interactive network
- description: Ramanathan V. Guha: "HotSauce worked as a plug-in to an existing browser so that when a hyperlink to a MCF-enabled website was selected the user was dropped into a first-person perspective view of the Web. It was a videogame view with Web pages floating as brightly colored blocks in an infinite black space, something like the view from a starship cockpit navigating through some strange asteroid field. It was easy to fly into and around the space, using the mouse to guide the direction of flight and holding down buttons to go forwards and backwards. A page could be accessed by simply double-clicking on the relevant block. A 3D immersive environment where I could move around an Antarcti.ca like space, moving things around, interacting with others in that space. And I should be able to do this wherever I am!"

[http://mappa.mundi.net/maps/maps_018/]

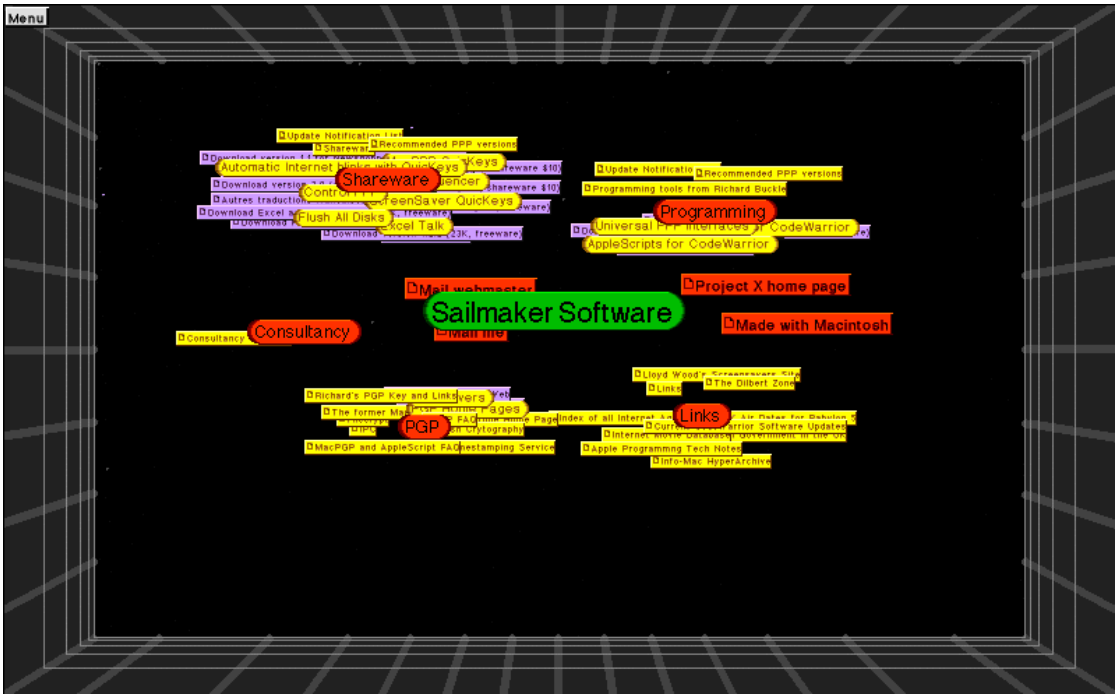

Figure A.39.: HotSauce by Ramanathan V. Guha (1996)

A.2.2. Collections of aggregations visualization

40 ▷ Grimm's Fairy Tale Metrics by Jeff Clark (2013)

- data: 62 stories of the Grimms's Fairy Tales [Small dataset]
- method: Sortable matrix with links to the dataset
- description: very complete metric analysis of the 62 stories of the Grimms brothers. This is a high quality example of a tool for linguistics analysis.

[http://neoformix.com/2013/GrimmStoryMetrics.html]

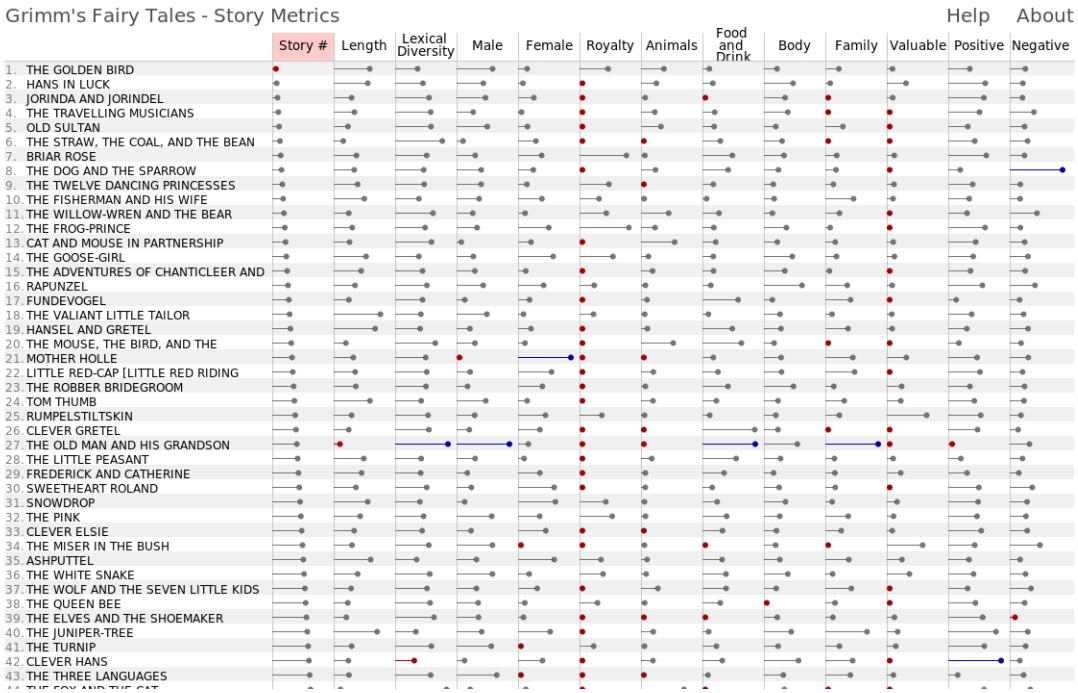

Figure A.40.: Grimm's Fairy Tale Metrics by Jeff Clark (2013)

41 ▷ Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012)

- data: any text corpora [Large dataset]
- method: Matrix/tabular view
- description: it is matrix view to support the assessment of topical term distributions and enable the comparison of latent topics. And, in general, it is a visualization of topic models and their terms distribution.
- paper: *Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 74-77). ACM.*

[http://vis.stanford.edu/papers/termite]

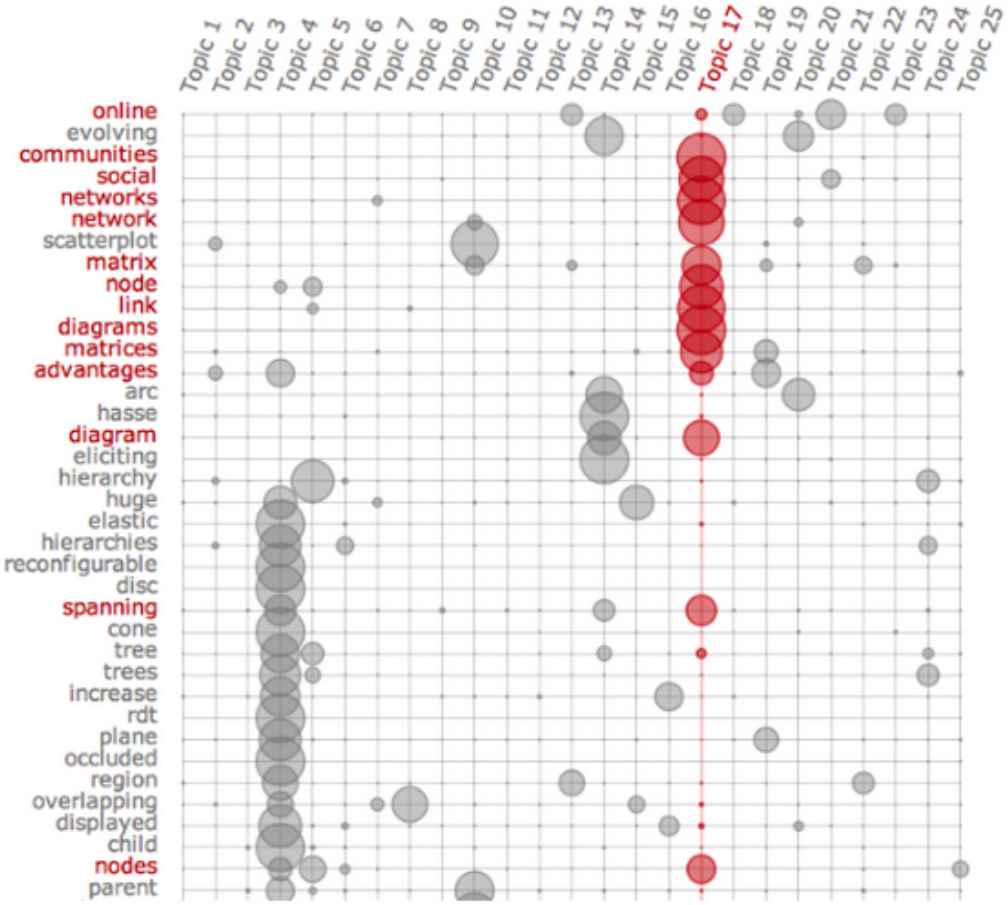

Figure A.41.: Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012)

42 ▷ Pediameter by Müller-Birn, Benedix and Hantke (2011)

- data: Wikipedia [N/A]
- method: time-line + arduino indicator
- description: a live visualization of Wikipedia Edits using pixel block in a time-line and an arduino indicator

[http://l3q.de/pediameter/]

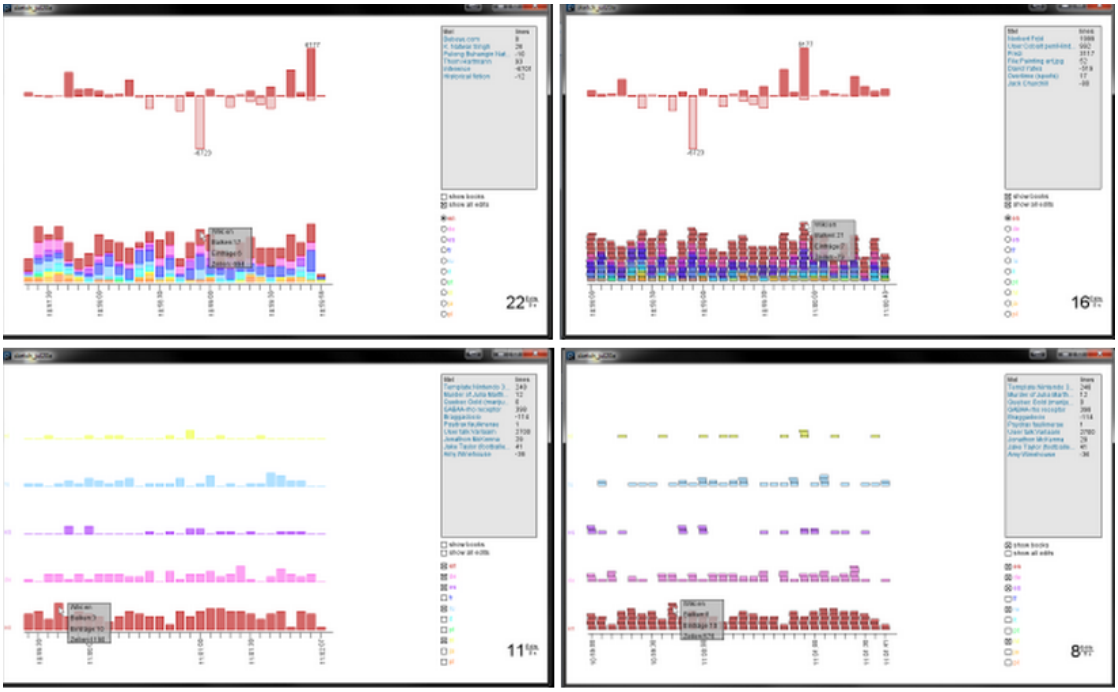

Figure A.42.: Pediameter by Müller-Birn, Benedix and Hantke (2011)

43 ▷ Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

- data: Google search suggestions [Large dataset]
- method: Connected trees
- description: a visualization that lets you compare two Google suggest queries.

[http://hint.fm/seer/]

Figure A.43.: Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

44 ▷ Web Trigrams: Visualizing Google's Tri-Gram Data by Chris Harrison (2008)

- data: Google n-gram dataset (2006) [large dataset]
- method: specific. Connected trees
- description: "As soon as I got my hands on the data, I quickly got to work on some straight forward visualizations. The first type compares two sets of trigrams, each starting with a different word. One visualization compares 'He' with 'She', while the other uses 'I' and 'You'. In the case of the 'He' vs. 'She', the top 120 trigrams for each were identified. The frequencies of the second word in the trigrams were combined and sorted, and rendered in decreasing frequency-of-use order. A similar process was used to create a ranking for the third (and final) word in the trigrams. Words are sized according to the square root of their use frequencies. The color-coded lines act like paths (a tree structure), enumerating all of the trigrams. The process was identical for the 'I' and 'You' version, except that only the top 75 trigrams were used."

[http://www.chrisharrison.net/index.php/Visualizations/WebTrigrams]

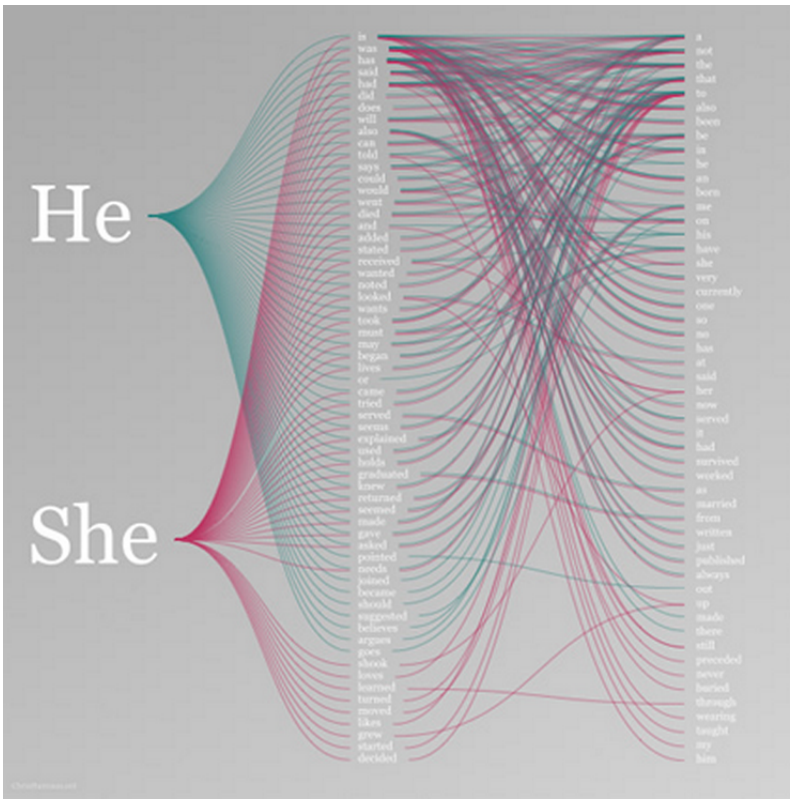

Figure A.44.: Differences in how the he and she subjects are used.

45 ▷ FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007)

- data: President Bush's eight annual speeches (2001-2007)
- method: rich interface for search, visualize and explore texts
- description: FeatureLens visualizes a text collection at several levels of granularity and enables users to explore interesting text patterns. The current implementation focuses on frequent itemsets of n-grams, as they capture the repetition of exact or similar expressions in the collection. Users can find meaningful co-occurrences of text patterns by visualizing them within and across documents in the collection. This also permits users to identify the temporal evolution of usage such as increasing, decreasing or sudden appearance of text patterns. The interface could be used to explore other text features as well.
- paper: *Don A., Zheleva E., Gregory M., Tarkan S., Auvil L., Clement T., Shneiderman B., Plaisant C.: Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In Proc. of the Conf. on Information and Knowledge Management (2007).*

[http://hcil2.cs.umd.edu/trs/2007-08/2007-08.pdf]

Figure A.45.: Screenshot of the FeatureLens interface.

46 ▷ Newsmap by Marcos Weskamp (2004)

- data: Google news [Small dataset]
- method: Treemap
- description: "Newsmap is an application that visually reflects the constantly changing landscape of the Google News news aggregator. Google News automatically groups news tories with similar content and places them based on algorithmic results into clusters. In Newsmap, the size of each cell is determined by the amount of related articles that exist inside each news cluster that the Google News Aggregator presents. In that way users can quickly identify which news stories have been given the most coverage, viewing the map by region, topic or time"

[http://marumushi.com/projects/newsmap]

Figure A.46.: Newsmap by Marcos Weskamp (2004)

47 ▷ TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

- data: Collection of emails conversation. [Small dataset]
- method: Speciphic interface. Time line with columns of words
- description: "a visualization that portrays relationships using the interaction histories preserved in email archives. Using the content of exchanged messages, it shows the words that characterize ones correspondence with an individual and how they change over the period of the relationship."
- paper: *Viégas, F. B., Golder, S., & Donath, J. (2006, April). Visualizing email content: portraying relationships from conversational histories. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 979-988). ACM.*

[http://smg.media.mit.edu/papers/Viegas/themail/viegas_themail.pdf]

Figure A.47.: TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48 ▷ WebBook by Stuart K. Card, George G. Robertson, and William York (1996)

- data: Search engine results [Small dataset]
- method: 3D interactive book of HTML pages
- description: WebBook was a dynamic multimedia contents constructor from lists of Internet pages. It is very inspiring the reading of the 1996 paper to understand the history of the world wide web.
- paper: *Card, S. K., Robertson, G. G., & York, W. (1996, April). The WebBook and the Web Forager: an information workspace for the World-Wide Web. In Proceedings of the SIGCHI conference on Human factors in computing systems: common ground (pp. 111-ff). ACM.*

[http://www.sigchi.org/chi96/proceedings/papers/Card/skc1txt.html]

Figure A.48.: WebBook by Stuart K. Card, George G. Robertson, and William York (1996)

49 ▷ Dotplot Applications by Jonathan Helfman (1994)

- data: Any text [Large dataset]
- method: Dotplot
- description: a bit minimalistic and very effective method to check similarity in large amount of texts, including multilanguage texts or computer code.
- paper: *Helfman, J. (1996). Dotplot patterns: a literal look at pattern languages. Theory and Practice of Object Systems, 2(1), 31-41.*

[http://imagebeat.com/index.php?id=17]

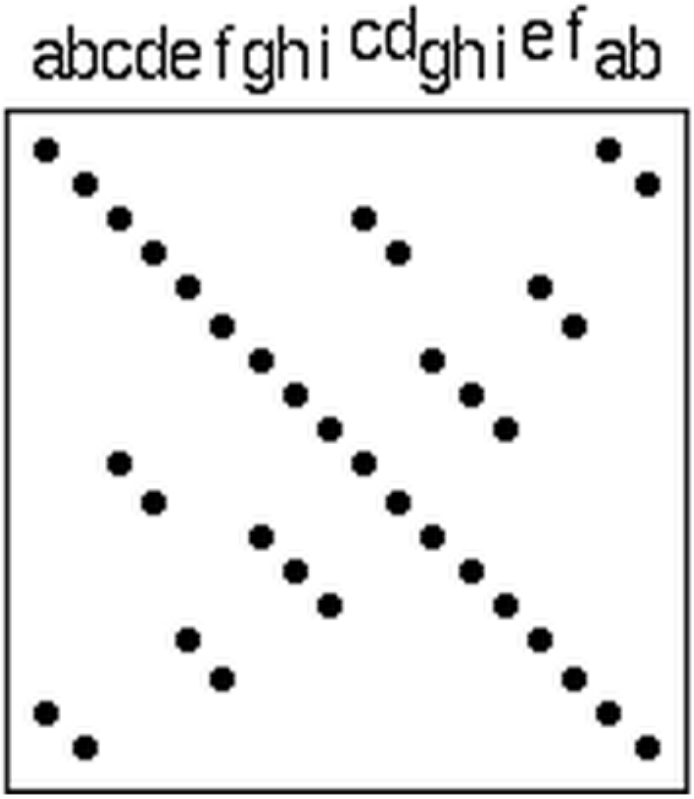

Figure A.49.: Dotplot Applications by Jonathan Helfman (1994)

B. APPENDIX B

B.1. NICTA scholarship: Milestone and agreement document

SCHEDULE 4 TO PROJECT AGREEMENT: MILESTONES AND DELIVERABLES

NICTA Milestones / Deliverables

No.	Deliverable / Milestones	Delivery Date
1.	nil	

Project Partner Milestones / Deliverables

UC will undertake the work and deliver the results through a series of reports and related software releases which will include source code and documentation.

The software is to be developed to a professional standard and is to include documentation to a level sufficient to allow a third party to be able to modify the software, as well as appropriate revision control and coding guidelines.

The parties will hold formal liaison meetings at the commencement of the project and from time to time thereafter. More precise standards for the software will be agreed in writing by the parties at the first formal liaison meeting. Detailed work plans for the subsequent 6 months will also be agreed in writing at each formal liaison meeting

No.	Milestones	Deliverable	Delivery
1.	Report and evaluation of existing state of the art on large corpus text visualisation	Report	1/9/2012
2.	Report on browser-based visualisation frameworks, feasibility of text visualisation in browser	Report	1/3/2013
3.	Report on initial text visualisation prototypes - rich overview	Report & Software	1/9/2013
4.	Report on text visualisation prototypes - visualisation of topic / document space	Report & Software	1/3/2014
5.	Report on technical implementation / integration of visualisation techniques	Report & Software	1/9/2014
6.	Final report - research outcomes and implications	Report & Software	1/3/2015

Figure B.1.: NICTA scholarship: Milestone and agreement. The dates have been relocated, starting Jan 1st, 2013.

B.2. Data for the word cloud in Figure 1.1: 1.2

B.2.1. Timeline data of some visualization tools

The data has been extracted manually from wikipedia.

Data format: Year of publication, name of the method, author(s), professional field and/or skills

- 1765 Timeline - Joseph Priestley [theologian, Dissenting clergyman, natural philosopher, chemist, educator, and political theorist]
- 1786 Bar Chart - book The Commercial and Political Atlas, by William Playfair (1759-1823).
- 1801 Pie chart - William Playfair's Statistical Breviary of 1801 [engineer and political economist]
- 1869 Flows in maps - Charles Joseph Minard [civil engineer]
- 1880 Venn diagram - by John Venn (1834–1923) [philosopher]
- 1891 Histogram - Karl Pearson, [Mathematician]
- 1896 Gantt Chart - by Karol Adamiecki, who called it a harmonogram. [economist, engineer]
- 1921 Flowchart - Frank Gilbreth [scientific management]
- 1923 Tagcloud ERUTARETTIL – André Breton [write, poet] and Robert Desnos [poet], proto-surrealist magazine Littérature. It weighs the importance of certain writers for the nascent surrealists.
- 1934 Social Networks - Jacob Moreno [psychiatrist and psychosociologist]
- 1977 Boxplot - John Tukey [mathematician]
- 1983 Star plot - John M. Chambers [Computer science, statistics]
- 1990 Treemap - University of Maryland, College Park professor Ben Shneiderman in the early 1990s [Computer Science]
- 1991 Headmap - Cormac Kinney [entrepreneur and software designer]
- 1992-1995 Tagcloud - Douglas Coupland's Microserfs (1995). [novelist]
- 1999 Sparkline - Edward Tufte [political science, statistics, and computer science]

B.2.2. Text used to create the word cloud using Wordle

[http://www.wordle.net/create]

theologian philosopher scientist educator political-science engineer political-science economist engineer political-science economist engineer philosopher mathematician economist engineer scientific management writer poet poet psychiatrist psychologist sociologist mathematician computer-science statistics computer-science entrepreneur software-designer novelist writer political-science statistics computer-science