Why visualize text?

Visualization goals

- **Understanding** - read a document
- **Summaries** - get the “gist” of a document
- **Clustering** - group together similar contents
- **Correlate** - compare patterns in text to other data

High-level tasks

- Find documents of interest in a collection
- Find documents similar to ones I want
- Identify the main themes or ideas of documents or collections
- See the history of changes in a document
- Find connections between documents
Interaction tool includes context of position and length of document

All details available through interaction

Search is built in

Tool is familiar, readily available and convenient
overview + detail display

syntactic structure is mapped to a color encoding
I don't know if you have had the same experience, but the snag I always come up against when I'm telling a story is this dashed difficult problem of where to begin it. It's a thing you don't want to go wrong over, because one false step and you're sunk. I mean, if you fool about too long at the start, trying to establish atmosphere, as they call it, and all that sort of rot, you fail to grip and the customers walk out on you.

Get off the mark, on the other hand, like a scalded cat, and your public is at a loss. It simply raises its eyebrows, and can't make out what you're talking about.

And in opening my report of the complex case of Gussie Fink-Nottle, Madeline Bassett, my Cousin Angela, my Aunt Dahlia, my Uncle Thomas, young Tuppy Glossop and the cook, Anatole, with the above spot of dialogue, I see that I have made the second of these two floaters.

I shall have to hark back a bit. And taking it for all in all and weighing this against that, I suppose the affair may be said to have had its inception, if inception is the word I want, with that visit of mine to Cannes. If I hadn't gone to Cannes, I shouldn't have met the Bassett or bought that white mess jacket, and Angela wouldn't have met her shark, and Aunt Dahlia wouldn't have played baccarat.

Yes, most decidedly, Cannes was the point d'appui.

Right ho, then. Let me marshal my facts.

LO, praise of the prowess of people-kings of spear-armed Danes, in days long sped, we have heard, and what honor the athelings won! Oft Scyld the Scefing from squadroned foes, from many a tribe, the mead-bench tore, awing the earls. Since erst he lay friendless, a foundling, fate repaid him: for he waxed under welkin, in wealth he throve, till before him the folk, both far and near, who house by the whale-path, heard his mandate, gave him gifts: a good king he! To him an heir was afterward born, a son in his halls, whom heaven sent to favor the folk, feeling their woe that erst they had lacked an earl for leader so long a while; the Lord endowed him, the Wielder of Wonder, with world's renown. Famed was this Beowulf: far flew the boast of him, son of Scyld, in the Scandian lands. So becomes it a youth to quit him well with his father's friends, by fee and gift, that to aid him, aged, in after days, come warriors willing, should war draw nigh, liegemen loyal: by lauded deeds shall an earl have honor in every clan.

YOU don't know about me without you have read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly—Tom's Aunt Polly, she is—and Mary, and the Widow Douglas is all told about in that book, which is mostly a true book, with some stretchers, as I said before.
1. **US GOVERNMENT INTELLIGENCE REPORT: 7 OCTOBER 2008 from MI-5**

MI-5 states that 100,000 bullets, six guns and military ware were among items police say they recovered after raiding a house in Narok, Kenya, belonging to Mr Thabit Othieno. The MI5 officers are investigating how sleeping bags, Nato-approved 7.62mm calibre ammunition, military desert boots, combat rain jackets, military fuel tanks, British Army inscribed Land Rover spare parts, military machetes and knives, landed in the hands of an unauthorised person.

3. **US GOVERNMENT TELEPHONE INTERCEPT: [October 10 2008]**

Call placed from a pre-paid cell phone (Caller 1) in an apartment complex in Barcelona, Venezuela to a pre-paid cell phone (Caller 2) in a business district Carabobo, Venezuela. Conversation took place in Spanish at about 1217, Barcelona local time.

Caller 1: Jorge’s place was raided last night. I think they [probably the police] were watching the place. They found the guns.

Caller 2: (interrupted) Shut up! Shut up! Just shut up! You can’t say that.

Caller 1: Sorry, and we need a new source for the, uh, car parts.

Caller 2: I think I have a source. My friend has a connection to a guy who is selling parts. I’m going to contact him today or tomorrow.

Caller 1: When do you think we can get them?

Caller 2: I don’t know. Just don’t talk to anybody. I’ll call you on this phone when we have anything. Don’t call me and don’t do anything.

[END]

1. **News Article: Bangkok Post, March 2, 2008**

Thai authorities seized an Ilyushin IL-76 aircraft carrying tons of weapons from North Korea during a refueling stop in Bangkok, a government official said. The pilot told Thai authorities the aircraft was headed to Sri Lanka, but its final destination was unknown, according to a spokesman for the Thai prime minister. It contained about 35 tons of weapons, including rocket-propelled grenades, shoulder-launched rockets and tubes that may be missile components, the spokesman said. The plane, which was detained Saturday, had five people onboard -- four from Ukraine and one from Belarus. They will appear in court Monday on charges related to illegal weapons smuggling, the spokesman said.

1. **REPORT DATE: 9 October 2008 [provided to CIA by Pakistani Criminal Investigation Unit, Karachi Division]**

NOTE: Surveillance report on the activities of Maulana Haq Bukhari, suspected to be a top leader within the Karachi faction of Lashkar-e-Jhangvi. Bukhari is frequently accompanied by Akram Basra, who acts as a driver and bodyguard. Additional information provided by police informants.

23 July 2008 – A delivery was made to a house in Lyari Town (a constituent town of Karachi) in a house believed to be used by Bukhari. The delivery was made by a two men in street clothing (as opposed to a uniform) who arrived in a white van, license LHR 6354, with single blue stripe on each side. The delivery consisted of three medium boxes (requiring two hands to move) and a small box (handsized). The boxes appeared to be heavy. One large box was square, the other rectangular. The small box was rectangular. It is unknown if Bukhari was home at the time of the delivery.

8 August 2008 – An unknown man visited the Lyari Town house where Bukhari is believed to stay. He arrived at 1615, and was let into the house immediately. Loud voices could be heard for a few moments, then they subsided. About fifteen minutes later a silver Mercedes left the rear of the house with what appeared to be three occupants. Due to the tinted glass it was impossible to identify the occupants of the vehicle. The house was surveilled for the next five hours; however no one came or went.

16 September 2008 – Bukhari was followed from his Lyari Town house to an apartment about 1.5 kilometers southeast. He entered the building and stayed there for about two hours. When he left he returned to his Lyari Town house at which time the observer believed he’d been discovered and left the area.

23 September 2008 – Bukhari and probably Basra are reported to have visited a house in the Katchi Abadis Old Settlement, on 835T Longhi Street. The informant, a vendor with business in the area, was passing by and noticed Bukhari and one other person enter the building at about 1430. The informant knew Bukhari by sight, but had never met him.
Visualization Pipeline

1. Raw Data
2. Data tables
3. Visual structures
4. Visualization

Data transformations, visual mappings, view transformations, user interaction

Insight!
Why is text visualization hard?

Text is high dimensional
(10,000+ unique words in a document)

Some words have ordering
January, February, March...
one, two, three...

Words have meaning, relationships and context

On St. Patrick’s Day, the local bar serves green beer.

The pub downtown has been known to make their pints chartreuse on Saint Paddy’s.

Patrick walked around the corner of Green St. and Day St., right into this bar sticking out of the wall and knocked himself senseless.
Extracting structure

**metadata** - data about the data
  who wrote it, when, what collection is it from, how long is it, format, keywords

**lexical level** - break the document into *tokens*
  characters, words, n-grams

**syntactic level** - identify the function of tokens
  identify the part of speech or named entities (people, places, money, dates)

**semantic level** - extract meaning from the text
  identify relationships between entities, themes, sentiment, etc..
History flow

Viégas et al., “Studying Cooperation and Conflict between Authors with history flow Visualizations”
Jigsaw: Calendar and timeline views
To identify research topics, we used standard, internally developed topic clustering technology. The statistical model underlying the code is called a mixture model [5]. The technology was originally developed for site administrators to help build and maintain category hierarchies. The text-clustering component suggests a set of categories when no explicit structure exists. We used titles, references, and keywords in the clustering process. A standard list of stop words, months of the year, journal and proceeding titles, and version and page numbers were removed from influencing the cluster results.

Five InfoVis and 22 CHI clusters emerged from using the clustering tool. We used PaperLens in the process of manually naming each cluster by investigating papers and authors in the cluster. For the CHI data, some topics were divided into several clusters, which we combined into one cluster, but we did not move individual papers into other clusters. This resulted in some papers being placed in odd clusters but is typical of any clustering solution. We ended up with the 15 CHI clusters shown in Figure 1.

Figure 1. PaperLens tightly couples views across papers, authors, and references and consists of 6 main parts: (a) Popularity of Topic (b) Selected Authors (c) Author List (d) Degrees of Separation Links (e) Paper List (f) Year by Year Top 10 Cited Papers/Authors.
Authorlines

Viégas and Smith, "Newsgroup Crowds and Authorlines: Visualizing the Activity of Individuals in Conversational Cyberspace"
text: ROM 9:5 Whose are the fathers, and of whom as concerning the flesh Christ came, who is over all, God blessed for ever. Amen.

/tmp/words22058:
## Vector Space Model / Bag of Words

<table>
<thead>
<tr>
<th></th>
<th>Book 1</th>
<th>Book 2</th>
<th>Book 3</th>
<th>Book 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sir</td>
<td>481</td>
<td>79</td>
<td>118</td>
<td>12</td>
</tr>
<tr>
<td>about</td>
<td>316</td>
<td>175</td>
<td>192</td>
<td>61</td>
</tr>
<tr>
<td>jeeves</td>
<td>313</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>255</td>
<td>96</td>
<td>139</td>
<td>80</td>
</tr>
<tr>
<td>holmes</td>
<td>2</td>
<td>459</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>your</td>
<td>210</td>
<td>406</td>
<td>282</td>
<td>92</td>
</tr>
<tr>
<td>man</td>
<td>126</td>
<td>288</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>could</td>
<td>166</td>
<td>286</td>
<td>185</td>
<td>105</td>
</tr>
<tr>
<td>doctor</td>
<td>4</td>
<td>34</td>
<td>554</td>
<td>0</td>
</tr>
<tr>
<td>romana</td>
<td>0</td>
<td>0</td>
<td>419</td>
<td>0</td>
</tr>
<tr>
<td>k9</td>
<td>0</td>
<td>0</td>
<td>347</td>
<td>0</td>
</tr>
<tr>
<td>page</td>
<td>3</td>
<td>9</td>
<td>327</td>
<td>0</td>
</tr>
<tr>
<td>wizard</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>255</td>
</tr>
<tr>
<td>dorothy</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>238</td>
</tr>
<tr>
<td>jim</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>152</td>
</tr>
<tr>
<td>little</td>
<td>86</td>
<td>269</td>
<td>50</td>
<td>146</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Weighting the vector-space

**Term Frequency**

\[ TF_{td}(word) = \text{count}(word) \text{ in } d \]

**Term Frequency by Inverse Document Frequency**

\[ TF.IDF_{td}(word) = TF_{td}(word) \times \log \frac{N}{Df(word)} \]

Df() is # of documents containing the word, N is the number of documents
Right ho, Jeeves
Sir Yes Yes Yes Bertie Angela Aunt Jeeves
Brazil passes Bill on Internet Privacy

GOP Candidate to Obama: ‘Man up’ and meet me on the basketball court
Parallel Tag Clouds

Collins, “Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora”
might have been observed fumbling with the tie, shuffling the hat, and behaving in all other respects in her presence like the complete dumb

home
at the battle of agarour .
in your corner .
guise .

is
her .
men cipher and that the only member of the household with brains and resources is jeeves .

pretty hard chap to outgeneral .
not accustomed to this gluttonous appetite for his society .
at his best .
seldom baffled for more than the nonce .

vocabulary considerably
that, while no one sees his flesh and blood with a keener and more remorselessly critical eye, he is nevertheless a man who delights

but there are limits, and sharply defined limits, at that.

when the sun has gone to bed and all the bunnies come out to have their little suppers ?

very thin .
a shoulder run through the frame.

lay propped up against the pillows had become a better, stronger, finer being .

had added to note that the dear old chap was a bit simmered up .

walked away and left him at it .

eventually toppled off at the door of longham manor was a very different barham from the gay and insouciant , boulevardier , of bond street and piccadilly .

working himself in your affairs .

brain all the way down , but it continued to beat me what could have caused the outbreak of hostilies, and I blunged my foot

could scarcely pass this

be said to have breathed freely .

across the face with a wet dish - rag

 pride : inducing arrogance

and it ' s no good looking like that .

* well . I wasn ' t going to , of course

* a wooster can stand a good deal , but he cannot stand having his name bandied in a public place.

likes to have his story readly .

? " I

! " I

* I take it you believe in love at first sight ?

* she said again , and from the tone of her voice, I could see that I had got her going .

I was in no mood for bantering words .

standing motionless in the shadows

was getting his eight hours .

best self tell you that he is always at his shiniest and most level - headed in moments of peril .

from being dragged in the mine and avoiding wide publicity of the wrong sort ?

would do that ?

* would perform the task admirably .

way

bedtime

she squashed among the profusely in the standing - room , only section , but at least, I felt , I was out of

can stand a good deal , but he cannot stand having his name bandied in a public place .

* * * I don ' t know when I have had a nastier shock .

seen this scripture - knowledge prize at a kids ' school we were at together , and you know what he ' s like .

along these lines , and you catch him at his best .

heart had long been hers and was waiting ready to be scooped up in demand , had decided to take up her option . I should

of the old yucks - and - fanciful days - the emotional , free - speaking girl who had so often risen in her stumps .
lay propped up against the pillows had become a better, stronger, finer Bertram.

had failed to note that the dear old chap was a bit steamed up.

wooster who

walked away and left him at it.

eventually toppled off at the door of Kingham Manor was a very different Bertram from the gay and insouciant _boulev
Abstract

We present a new technique, the phrase net, for generating visual overviews of unstructured text. A phrase net displays a graph whose nodes are words and whose edges indicate that two words are linked by a user-specified relation. These relations may be defined either at the syntactic or lexical level; different relations often produce very different perspectives on the same text. Taken together, these perspectives often provide an illuminating visual overview of the key concepts and relations in a document or set of documents.

Index Terms—Text visualization, tag cloud, natural language processing, semantic net.

1 INTRODUCTION

Scholars have long dreamed of turning text—from books to entire libraries—into maps. A visual perspective on a text has many potential uses. A map can serve as a summary and provide a jumping-off point for close reading. Mapping techniques may also be used to compare multiple texts, whether books by different authors or speeches by different politicians.

Unfortunately, anyone who sets out to map a book quickly runs into two problems. The first is purely conceptual: most interesting texts are large (a typical novel has more than 100,000 words) so some kind of analysis and summarization is necessary to make them amenable to visualization. The key issue is to define an effective unit of analysis; such a unit could range from letters to words to general concepts and ideas. Due to the current state of natural language processing, choosing the right unit involves a tradeoff between reliability and validity. On one end of the scale, computers can reliably pick out the individual words in a book, leaving the task of putting the words together to a human. On the other end, programs that aim to extract high-level meaning from text—say a semantic network of people and their relations—face significant error rates and are easily misunderstood by their users.

The second problem in the visual display of text involves legibility. In most visualizations, one wants to use spatial position as a meaningful variable. Yet a readable set of words obeys spatial constraints on alignment, grouping, and type size. The conflict between positioning and legibility can lead to displays that are hard to read or where spatial position is essentially random.

In this paper we introduce a new text mapping technique, the phrase net, which seeks a balance both in analysis and display. Our unit of analysis is a "phrase," i.e., a particular relationship between words that can be defined using either simple pattern matching or syntactic analysis. This unit provides a higher level of analysis than individual words, but is easily understood by users and does not require unreliable artificial intelligence. Our visual displays use a standard graph layout engine that has been modified to ensure that text is readable via constraints on alignment and grouping.

In addition to describing the design and implementation of the phrase net, we also provide a series of sample use cases. Some of these were derived from our own exploration. To find others, we deployed a simplified version of the phrase net on the Many Eyes project.

van Ham et al. “Mapping Text with Phrase Nets”
On the other hand, applying different expressions to the same text can reveal a series of interrelated conceptual networks. The phrase nets of Jane Austen’s novel *Pride and Prejudice* in Figure 5 illustrate this. Matching “X and Y” shows a network of concepts and people. The main characters appear neatly organized in two clusters: Jane, Elizabeth, Lydia, Kitty, Catherine and Mr. Bingley form a central cluster, whereas “mother,” “aunt,” and “uncle” keep some distance. Positive attributes such as “sense,” “disposition,” “humor,” “kindness” cluster together while less flattering qualities such as “pride,” “conceit,” “vanity,” “folly,” and “ignorance” form a group of their own. Perhaps most interesting, to those familiar with the novel, is that “Darcy” does not appear in the network—in a certain sense he is the most solitary major character.

If we analyze the same text with the pattern “X at Y” we obtain an entirely different network that reveals the set of locations inhabited by the characters in the novel and the events that take place at those locations. In a sense, the user can direct exploration towards a particular dimension of the text by intelligently choosing the pattern to match.

Figure 1 shows the result of another targeted pattern. Here we have analyzed the whole Bible using the pattern “X begat Y”, a specific formulation from the King James Bible indicating a parent-child relationship. The resulting graph illustrates the lengthy genealogies that are recorded by many different books in the Bible. The network also uncovers a number of defining aspects of these lineages, such as the importance of Abraham.

### 4.2 Regular expressions and matching
The patterns we have shown so far are of the form “X <connector> Y”, where the connector is either a separate word or a phrase. However, regular expressions also allow us to specify patterns that match for specific pre- and postfixes to X and Y. Previously, the authors worked with a humanities scholar to analyze a set of 7,000 British novel titles between 1740 and 1850—indeed, much of the motivation behind building phrase nets comes from this collaboration. This scholar was interested in how the use of simple syntactic constructions such as “X of the Y” reflected changes in literary style over the centuries.
Don et al., “Discovering interesting usage patterns in text collections: integrating text mining with visualization”
Google nGram Viewer

Graph these comma-separated phrases: data visualization, information visualization

between 1800 and 2000 from the corpus English with smoothing of 3.

Search lots of books

https://books.google.com/ngrams
Literature fingerprinting

Hapax Legomena
word that appears only once

Keim, Oelke, “Literature Fingerprinting: A New Method for Visual Literary Analysis”
Literature fingerprinting

Simpson’s Index
probability that a token belongs to the set

Keim, Oelke, “Literature Fingerprinting: A New Method for Visual Literary Analysis”
Literature fingerprinting

Keim, Oelke, “Literature Fingerprinting: A New Method for Visual Literary Analysis”
The state of our union is ... dumber:
How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union

2 December 1823
James Monroe delivers an address that would later become known as the Monroe Doctrine, a key tenet of US foreign policy for almost two centuries.

2 December 1913
Woodrow Wilson’s speech follows more than 100 years of the address being delivered to Congress as a written submission. His precedent is (mostly) followed to this day.

3 January 1934
Franklin D. Roosevelt changes the name of the address from the President’s Annual Message to Congress to State of the Union.

16 January 1981
Jimmy Carter’s final address, delivered as a written message, is the longest ever state of the union.

http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level
Relative n-gram signatures

base document: *burroughs* warlord of mars

https://web.cs.dal.ca/~jankowsk/ngram_signatures/
ABSTRACT

We present Themail, 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 one's correspondence with an individual and how they change over the period of the relationship.

This paper describes the interface and content-parsing algorithms in Themail. It also presents the results from a user study where two main interaction modes with the visualization emerged: exploration of "big picture" trends and themes in email (haystack mode) and more detailed exploration (needle mode). Finally, the paper discusses the limitations of the content parsing approach in Themail and the implications for further research on email content visualization.

Author Keywords
Email archive, visualization, content

ACM Classification Keywords
H.5 Information Interfaces and Presentation; H.5.2 User Interfaces

INTRODUCTION

Email users tend to save the overwhelming majority of messages they receive [4]. In fact, email storage and retrieval have, early on, been identified by researchers as two of the main uses of this communication technology [18, 19]. It is not clear, however, why users save such large amounts of messages.

Figure 1: Screen shot of Themail showing a user's email exchange with a friend during 18 months.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Review Spotlight

- Expensive or cheap?
- "long wait" or "no wait"?

**Mentioned 63 times**

possess sage of the halos wisdom, and know in advance sushi zone only accepts cash and the waits will be long and arduous.

yes, its a long wait, learn the master of zen if you want to eat here.

we came here early to try to avoid the long wait most people here talk about.
A senior police officer was killed near Cairo on Wednesday in a bomb attack claimed by a militant group, while security forces stormed a hideout used by another Islamist organisation near Alexandria in a raid that left an officer and a militant dead. Militant violence has spiralled since last July, when the army toppled elected head of state Mohamed Mursi and the authorities launched a fierce crackdown on his supporters in the Muslim Brotherhood and other Islamist sympathisers. The attacks underline lingering instability in Egypt ahead of a presidential election in May that Abdel Fattah al-Sisi, the former army chief who deposed Mursi, is expected to win. The prime minister said the state was in "a fierce war" on terror. The police officer killed near Cairo was named as Brigadier General Ahmed Zaki. State media said he was killed outside his home in 6th of October City, 32 km (20 miles) outside Cairo, when a bomb placed under his car went off. Two conscript policemen were wounded in the bombing. A militant group called Alnad Mstr, or Soldiers of Egypt, said it carried out the attack in a statement posted on a Facebook account in its name that has carried past statements. The post included a photo of a man said to be Zaki on his way to his vehicle, describing him as "the criminal brigadier general in the (security) force for killing protesters". Mursi's removal from power last summer after mass protests against his rule tipped Egypt into the worst internal strife of its modern history. Hundreds of his supporters were killed by security forces as they broke up their protest camps. Militant attacks since then have killed around 500 people, mostly policemen and soldiers. The threat has been compounded by a flow of weapons from neighbouring Libya. The Interior Ministry said the hideout targeted by police at dawn on Wednesday near Alexandria was used by Ansar Bayt al-Maqdis, or Supporters of Jerusalem, the group behind some of the deadliest attacks of the last nine months. The militants had opened fire on the security forces as they arrived at the hideout in Borg El Arab, some 45 km (28 miles) south-west of Alexandria. The police officer killed in the raid was named as First Lieutenant Ahmed Saad and the dead militant as Hassan Abdel Aal, a 25-year old from the Nile Delta province of Dakahlia. Two other militants were detained, the ministry spokesman, Hany Abdel Latif, said in a televised statement. Footage broadcast on state TV appeared to show the body of a militant on the ground. The militants were "among the dangerous elements of the terrorist group Ansar Bayt al-Maqdis, which was planning to target police and military facilities and the security forces", the ministry said. The police seized weapons including explosive belts, automatic weapons, hand grenades and ammunition.

http://uk.reuters.com/article/2014/04/23/uk-egypt-violence-idUKBREA3M1KO20140423

Stanford NLP Named Entity Tagger
Jigsaw
Sentiment analysis

http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
In-spire: Theme view [classic]
In-Spire: Theme view