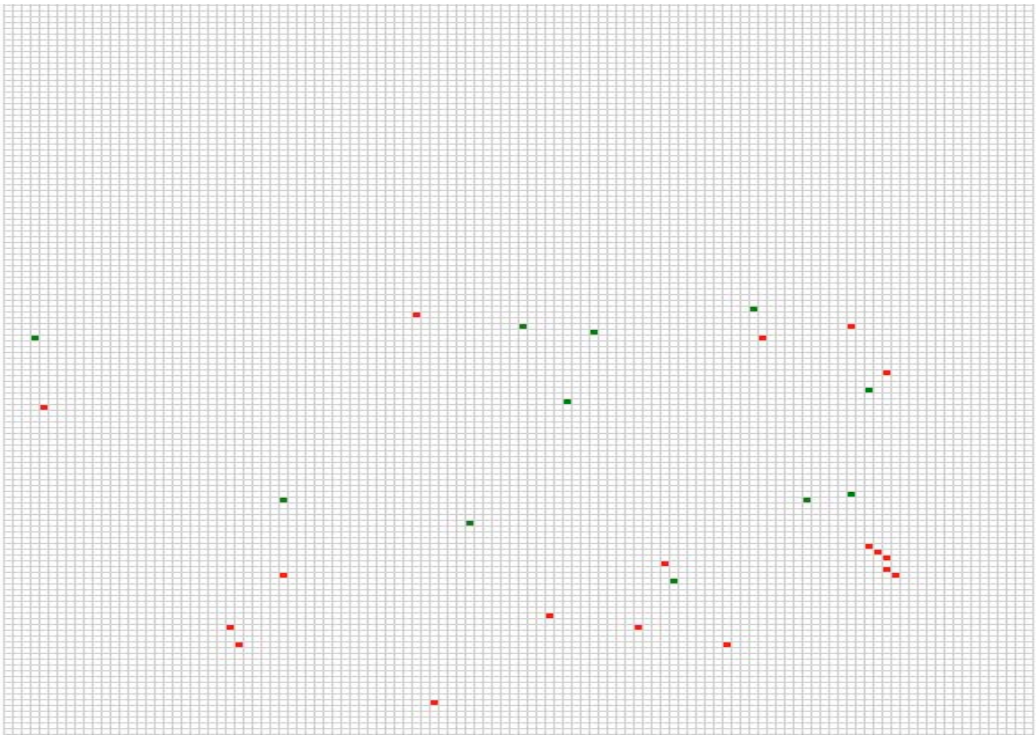
Mentions of War: Analyzing Large Quantities of Historical Text Visually
Gordon Rugg & Jo Hyde
February 1, 2012

A major problem for historians is source material. Often it's a problem because the source material is pitifully limited. Often, though, it's a problem because there's so much source material that the historian has difficulty seeing the wood beyond the trees. Take the official war records for the Battle of Gettysburg, for instance: the volume dealing with those few days runs to well over a thousand pages, containing over half a million words. How can anyone tease out patterns in so much text?

One solution is to look at the patterns that the words show, rather than looking at the words themselves. Imagine that you're looking at a document where someone has used color-coded highlighter on every occurrence of two keywords, but where the words themselves have been grayed out. All you can see is the distribution of the keywords. What can you tell from that? You can actually tell a lot from proximity, from frequency, and from pattern.

Here's an example. It shows part of the official war record for the Battle of Gettysburg. Every occurrence of the word "Buford" in this extract is shown with red highlighter; every occurrence of "Stuart" is shown with green highlighter.

***Buford* and *Stuart* in part of the Gettysburg volume**



The first half of this extract doesn't mention either name. Then both names are mentioned repeatedly, intermingled with each other. The closing section mentions only Buford.

What does this mean? The obvious, and correct, interpretation is that the intermingled section involves some sort of interaction between Buford and Stuart, followed by Stuart dropping out of the picture (literally as well as figuratively). This extract is in fact the part of the record which covers the capture of Stuart's camp at the end of the battle, with a flurry of cavalry action on both sides; the extract ends with coverage of Buford's subsequent actions.

In this essay, we describe some of the ways in which this approach can be used by amateur and professional historians to investigate large quantities of text.

The software we have used for the examples is the Search Visualizer, available for free at www.searchvisualizer.com

The text we have used for the examples is from the official war records from the American Civil War. The full records are available for free download at http://digital.library.cornell.edu/m/moawar/waro.html

The full set consists of 49 substantial volumes for the land campaign, plus another set for the naval campaign. Three of these volumes are freely available on the Search Visualizer website, in the "sample texts" section (accessed via the "entire Web/single site/sample texts" option on the search bar). The first of these is Volume 1, covering the start of the war; the second is the volume dealing with Gettysburg; the third is Volume 49, dealing with the end of the land war. We chose Gettysburg partially because of its historical significance, but also because of its size – this volume is over half a million words long, so it's a significant challenge to the historian.

Our opening illustration showed the underlying principle for visualizing text: the experienced reader is well able to interpret patterns and structures in a text, and can start making sense of a visualization immediately. For instance, a Civil War historian looking at the distributions of mentions of "cavalry" in the Gettysburg volume would expect to see a lot of mentions toward the start of the image, when the two sides were coming into contact, and a lot of mentions at the end, when the Confederate army withdrew from the field, harrassed by Union cavalry; they would not expect to see so many mentions in the middle section, which was mainly fought by infantry and artillery. This is exactly what the visualization shows. It's possible to fit this visualization, for half a million words, onto a single sheet of typing paper. The image below shows this visualization for the entire Gettysburg volume.

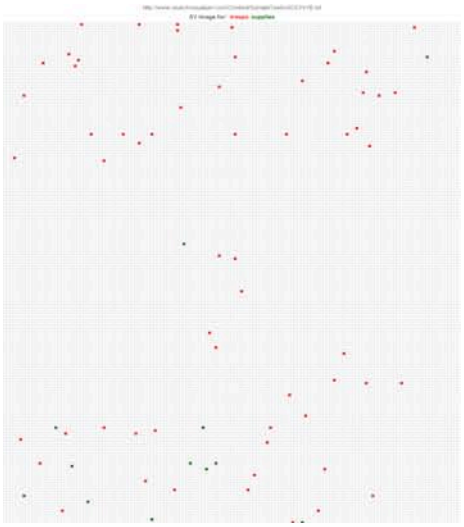
This is obviously an extreme example, but it demonstrates what can be done using this approach. The software produces visualizations such as this within seconds, making this a quick, simple way to start making sense of substantial documents. In the next part of this essay, we describe some ways of examining documents in more detail, and of spotting significant patterns within them.

Mentions of “cavalry” in the Gettysburg volume

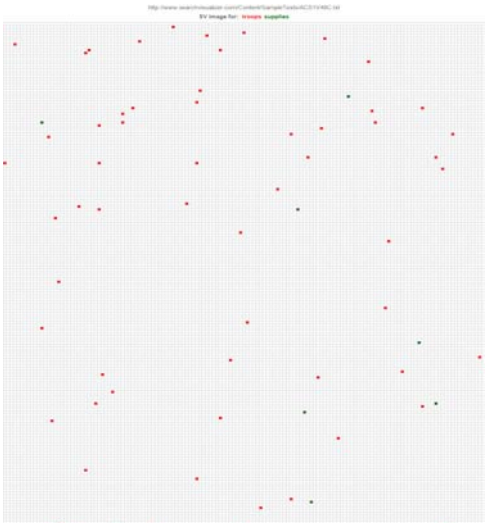
Comparing texts to each other: frequency of mentions

Here's what happens when you visualize mentions of "troops" and "supplies". The Union example is from Volume 1 of the war records, and the Confederate example is from Volume 49.

Union records, early in the war



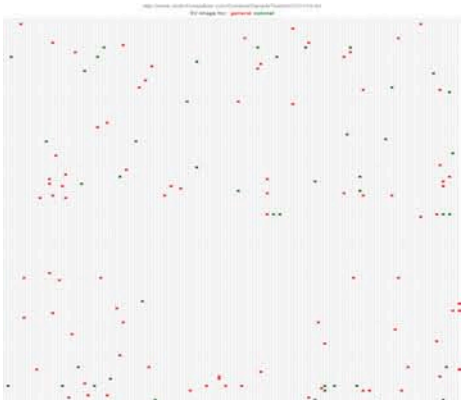
Confederate records, late in the war



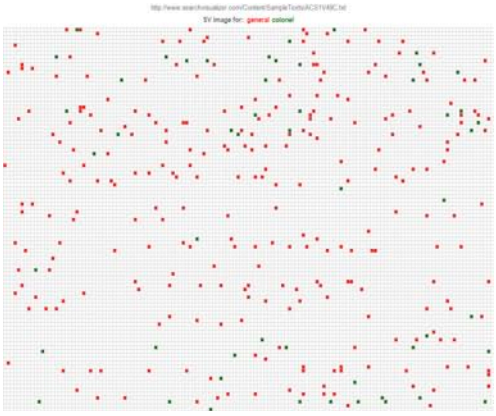
There's some difference between them in terms of number of mentions of troops and supplies, but not an enormous amount, although they're from opposite ends of the war, and from two opposing armies.

If we now look at the mentions of "general" and "colonel" from the same documents we see a very different pattern.

Union records, early in the war



Confederate records, late in the war



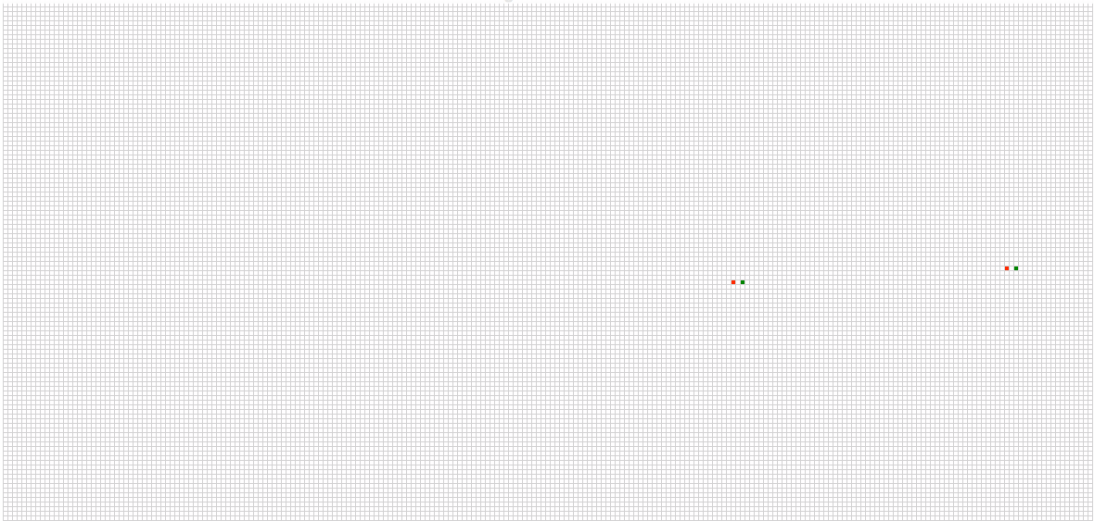
Why is there such a striking difference between the two sets of documents? That's a good question, and one that specialists in this field can doubtless answer. For the amateur historian, this is a good example of how the visualization allows you to run a

quick check on something that strikes you when reading a text, to see whether there really is something going on.

Sometimes, as in the example above, a question jumps out from the data. You're reading a document, and you notice something about it that you weren't expecting. On other occasions, you'll be reading a document because you already have a question that you're trying to answer. Sometimes that question is a personal one – you might be trying to find out, for instance, whether one of your ancestors fought at a particular battle. Sometimes the question is on a much larger scale. Our next two examples illustrate these two types of question.

For the first example, we searched for mentions of Denis Burke in the Gettysburg volume. Burke was a captain with the 88th New York Infantry. He appears four times in the Gettysburg volume. The illustration below shows two of those mentions.

Proximity: Mentions of Denis Burke in the Gettysburg volume

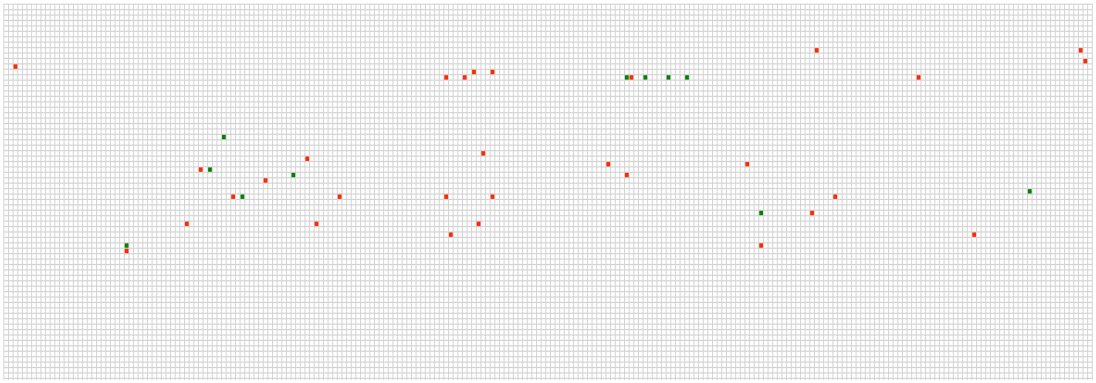


In principle, it should be easy to track down records about an officer who fought in a significant battle. In reality, it can be difficult to do this using traditional methods, for various reasons.

Take the name itself, for instance. When you look closely at the visualization, you notice that there's a gap between the red dot representing "Denis" and the green dot representing "Burke". The reason is that Burke appears in the records as "Denis F. Burke". A simple online search for the phrase "Denis Burke" in inverted commas (the usual convention for finding a specific phrase) would probably not register "Denis F. Burke" as a hit. An online search for the two words *Denis* and *Burke* without inverted commas would get round this problem, but at the cost of finding false positives – there were several other soldiers called Burke at Gettysburg. That isn't too much of a problem with a name like "Burke" which is relatively uncommon. If you're trying to find the right Smith, on the other hand, then a feature like the one above can be invaluable.

The next illustration shows what happens when you search the same volume for words *James* and *Smith*. It's a close-up of the opening part of the volume.

Early mentions of *James* (red) and *Smith* (green) in the Gettysburg volume



This illustration shows that there are numerous Smiths and numerous Jameses, and various permutations on *James Smith*, including *Smith, James*, and some instances of the two names separated by an initial, such as a *James E. Smith* and a *James J. Smith*.

The next illustration shows what happens if you are trying to find out about a bigger question: In this case, the Confederate land forces' view of the navy in the closing stages of the war. It shows occurences of the words *navy* and *naval* in the Confederate records from the closing stages of the war. The two words are being treated as synonyms, and both occur as a red dot.

Mentions of navy and naval in the Confederate records at the end of the war

One immediately striking feature is how few the mentions are, although the collected documents run to some three hundred pages. Most of those mentions are at the beginning and the end of the text. Why is the navy not being mentioned in between?

When you go down into the underlying text, one possible explanation leaps out. One of the last mentions of the navy in the first part of the text runs as follows. (Capitalization etc preserved as in the orginal.)

"MERiDIAN, Miss., February 4, 1865.
His Excellency the PRESIDENT:
The navy at Mobile is a farce. Its vessels are continually tied up at the wharf; never in co-operation with the army. The payment of its expenses is a waste of money. I send by mail a communication, giving my reasons for these expressions.
R. TAYLOR,
Lieutenant- General."

It is little surprise that subsequent mentions of the navy are few, and occur mainly as references to the navy yards. The closing section of the volume relates to the arrangements for the navy after the war's end.

The examples above show how visualizing text in this way makes it possible to see patterns in the text, and to find relevant sections of text, within a single language. There are other advantages in using this approach. One is that the images are derived from text, but are pure images, without any text visible in them. This means that you can compare visualizations for texts in different languages, provided that the keywords in each case are equivalent to each other.

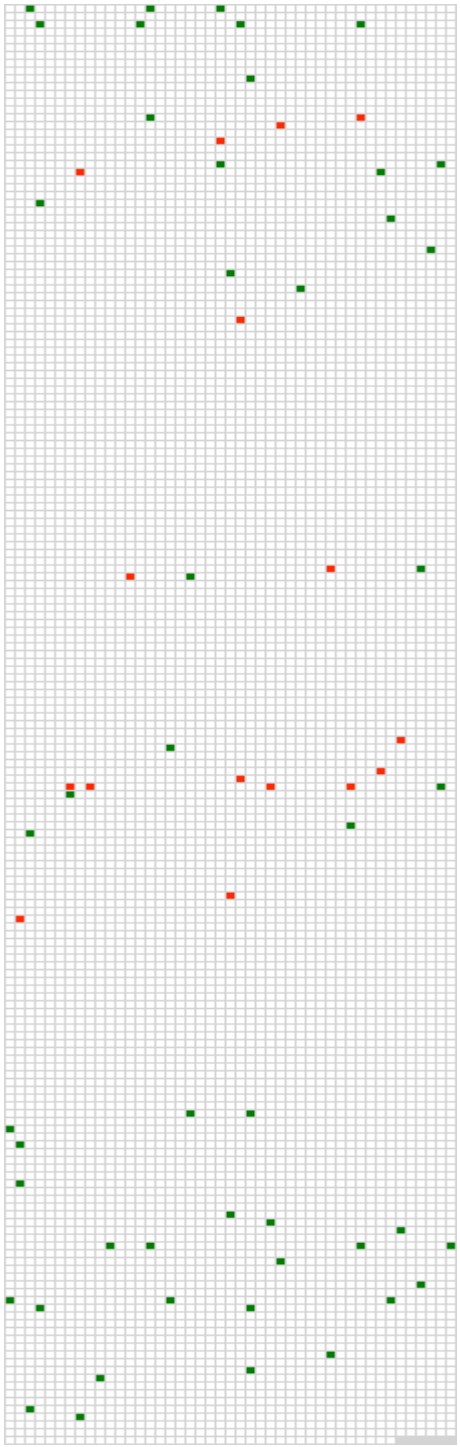
This can be very useful to historians dealing with records in different languages, as a way of seeing which themes are mentioned, and how prominent those themes are, and how they are structured. For instance, if you are looking at documents relating to the same war, in the languages of the two different sides, do they both place the same emphasis on the same topics, or do they have very different perceptions of the same events?

The next section demonstrates this with texts in German and English. For consistency with the rest of this essay, we've used a German and an English text about Gettysburg, but the same principle could obviously be applied to other texts.

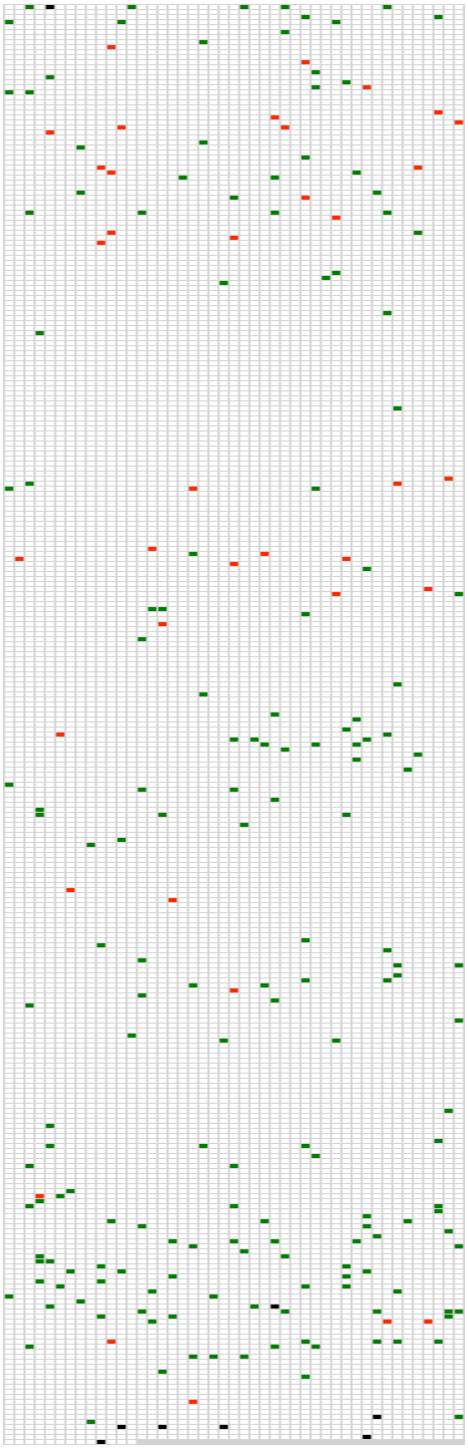
Visualization across languages

Here are the images for the German-language wikipedia entry on Gettysburg, with *kavallerie* (cavalry) in red and *Gettysburg* in green, and the English-language wikipedia equivalent, with *cavalry* in red and *Gettysburg* in green. (The black dots represent mentions of *wikipedia*.)

Kavellerie Gettysburg



cavalry Gettysburg



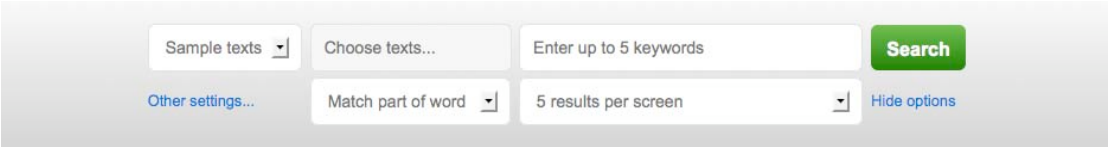
Conclusion

Visualizing texts makes it possible to handle large quantities of material swiftly and easily. Using this approach makes it possible to see structures within a text, and to do other things such as comparing the number of times that two different texts mention the same topic. It also makes it possible to answer questions which are difficult to answer using ordinary text search facilities.

Practical notes

The Search Visualizer is available at www.searchvisualizer.com without cost. We have put three volumes of the official war records of the American Civil War on the site, for use without charge and within the terms of the original copyright holders. To access these records, choose “Sample texts” from the option on the left of the search bar, and then click on the documents you want. For convenience, we have split two of the volumes into shorter thematic sections, but we have left the Gettysburg volume as-is, because it deals with a single event.

The SV search bar, showing location of the “Sample texts” option



The SV offers other search options. You can point it at a website of your choice, to search only records from that site. You can also use it to search the entire Internet.

The “results per screen” option is useful for dealing with records of different sizes. If you’re doing an Internet search for comparatively small records, then 5 or 10 records per screen allows you to search a lot of records quickly. If you’re working with very large records, then using 1 record per screen is more tractable. If you’re comparing two documents directly against each other, then 2 documents per screen is useful.

If you want to look at a visualization in more detail, then you can left-click on it, to bring up an enlarged, interactive version. This enlarged version lets you see the text surrounding a keyword when you hover over it.
You can save visualizations as images by right-clicking on them, and then choosing the destination folder where you want to save them.

If you want to get to the underlying document itself, then you can either click on the enlarged visualization, or simply click on the blue hyperlink above the relevant record.

The SV is protected by patent, and the terms and conditions specify that visualizations from it cannot be sold on commercially (e.g. on mugs and T-shirts). You are welcome to use images from the SV for non-commercial purposes.

We hope you will find this approach interesting and useful for your research.