Dealing with very large texts
By Gordon Rugg

Background: Search Visualiser is useful for finding words within large texts, and for seeing patterns and structures in large texts. It can handle a document half a million words long fairly comfortable, on the right settings (tiny squares and one document per screen). It's also useful for getting swift overviews of large numbers of texts.

This article was originally posted on October 4, 2012.

Sometimes a problem that's difficult to solve in one format is very easy to solve in another format.

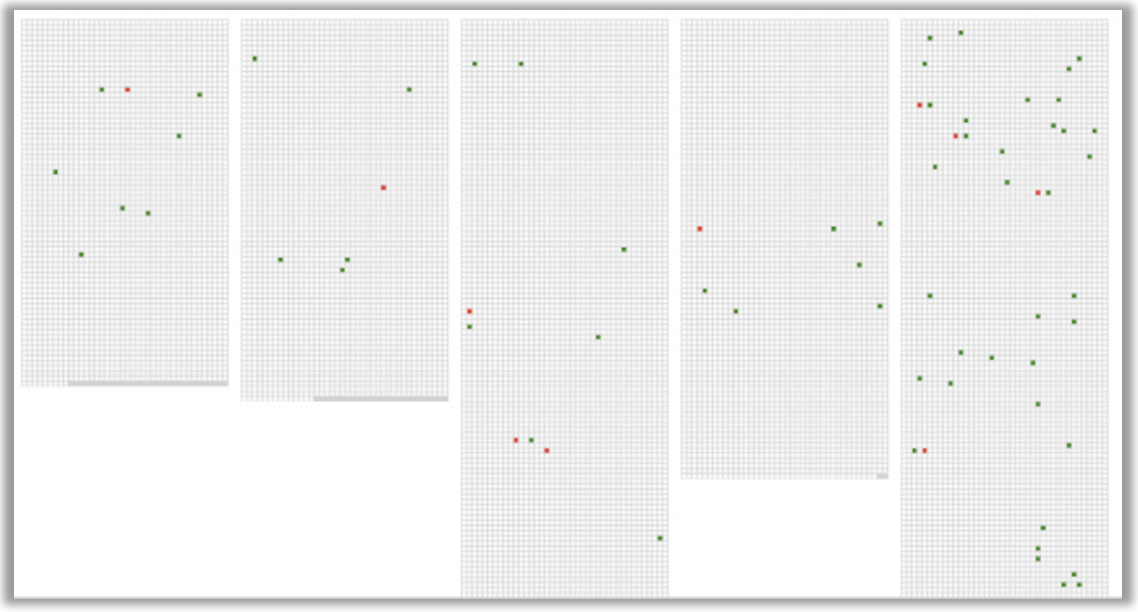
That's the core concept behind the Search Visualiser. Finding the key points in large quantities of text is very difficult when you're looking at the text in text format; it's like looking for the right needle in a pile of other needles. It's usually much easier to find those key points when the text is in a different format – in this case, visual.

Example: What are they saying about my product?

Here's an example. Imagine that you want a quick overview of what people on the Web are saying about your product.

The image below shows a visualization of documents containing the terms "bad" "good" and "value for money" to illustrate how you can do this.

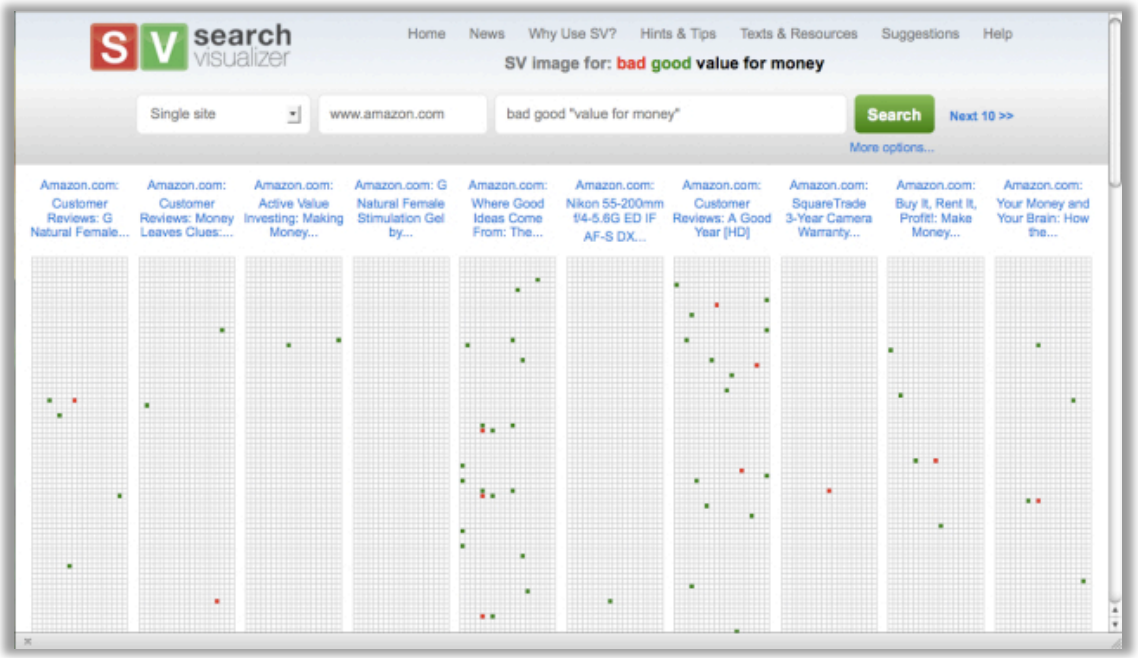
Each column shows a separate document; each square within a column represents a word in the document, starting at the top left, and working across and then down towards the bottom right, just the way that you would read the words in the document. The red squares show where the word "bad" occurs; the green squares show where the word "good" occurs, and the black squares show where the phrase "value for money" occurs.



You can see at a glance that there are a lot more green squares than red squares – the overall picture is positive. That's saved you reading several thousand words.

You can also see things that you wouldn't see from a statistical analysis of the same records. For instance, in the fourth record, all the mentions of "bad" and "good" are clustered in one section of the document.

That's an example of the principle behind the Search Visualiser. The screenshot below shows what a typical Search Visualiser search looks like, using the free online version of SV.



You're seeing ten records at a time. Your keywords are shown toward the top of the screen, colour-coded. The Search Visualiser is searching a single site, amazon.com. You also have the option of searching the entire Internet, and, if you use the commercial version, of also searching your hard drive or network. If you want to read the underlying record, then you just click on the blue link above the column.

This visualisation lets you scan very large amounts of text very swiftly. You have the option of choosing different square sizes, and of choosing different numbers of records to view per screen.

The example above shows a simple, basic search. It's easy to do more powerful searches. For instance, you can add synonyms to your search. You do that by separating each word in the list of synonyms using a comma without a space, as in the next example.

Finding the key data point

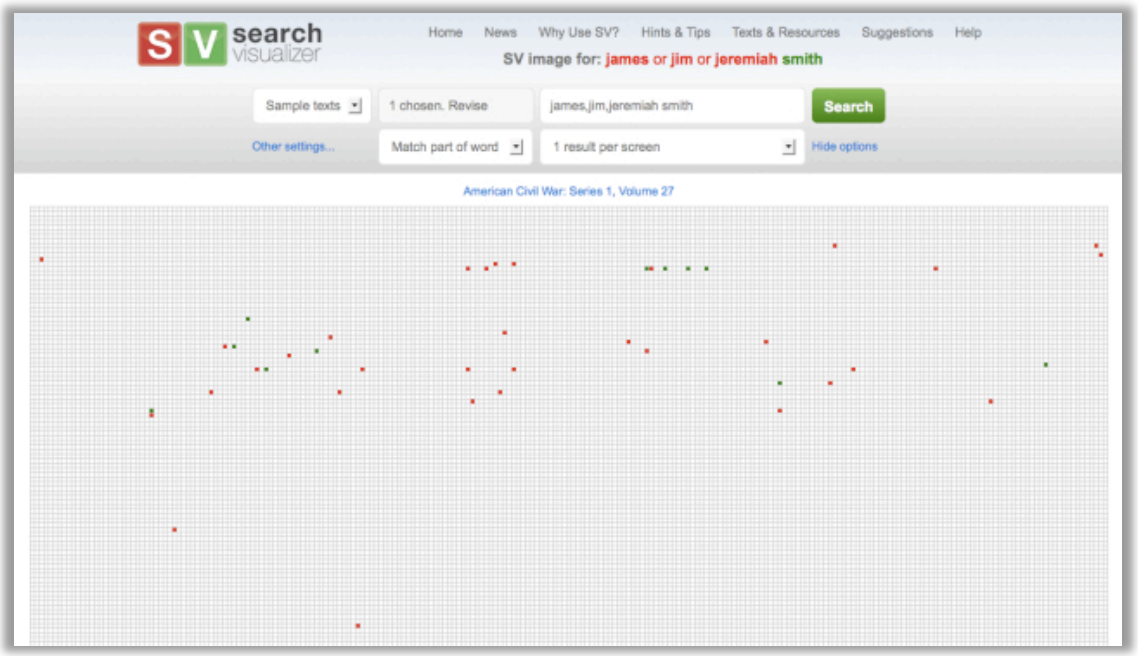
The previous example showed how you can get an overview of a large number of documents.

The Search Visualiser is also good for finding the right piece of specific information within a large number of documents, or within a single large document.

Here's an example. Suppose you're searching a very large file for mentions of a James Smith. It's a common name, and both the names "James" and "Smith" are individually common. An added complication is that there may be a middle name, or a middle initial, which would cause problems for the "search for this phrase" type of search on most search engines. There's also the issue of nicknames.

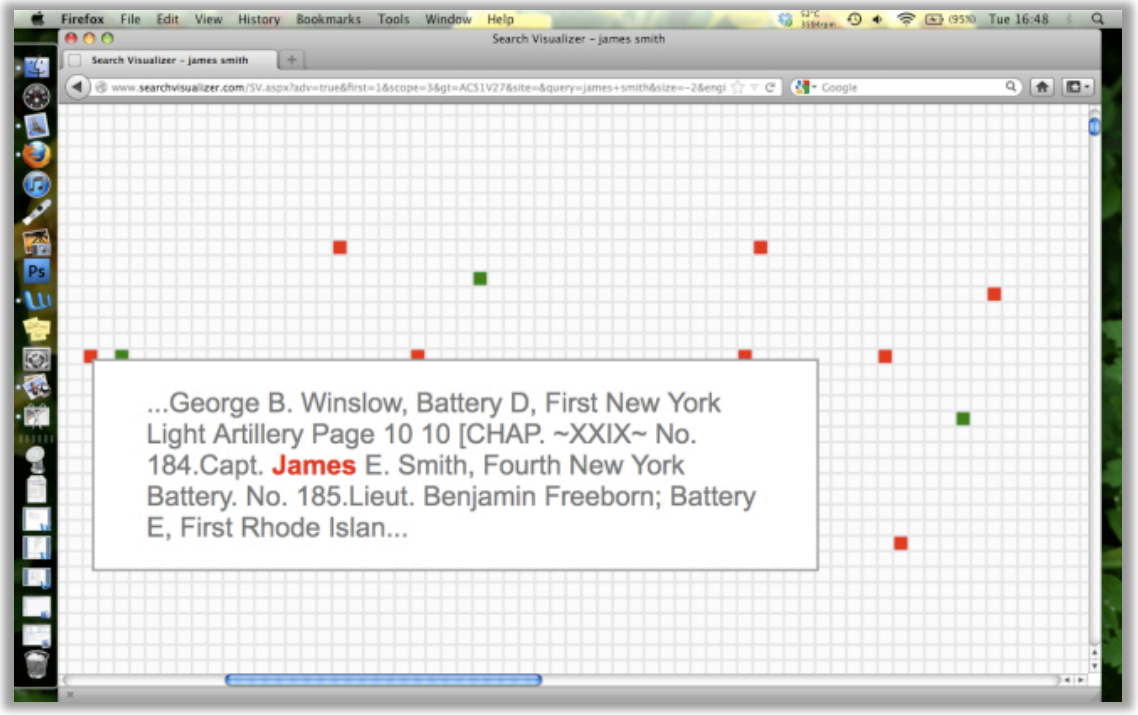
Some search engines have a thesaurus of common nicknames, which they add automatically to your search, but that doesn't help if the nickname is unusual, like a James being nicknamed Jeremiah.

The image below shows how you can handle this with Search Visualiser. It's part of an SV search for mentions of James Smith in a very large document – over 600,000 words.



If you switch SV to its most powerful settings, then you can see about a couple of hundred thousand words per screen. In the screenshot above, you can see that the search is using the synonyms James, Jim and Jeremiah (all shown in red), plus the name Smith (shown in green). There are several places where you see a red square next to a green square. Several of them turn out to be a James Smith with a middle initial. One is a "Smith, James". There are also a lot of other people called James (or Jim or Jeremiah) and a fair few more called Smith.

If you want to look at the context of a keyword, you can click on the column to bring up an interactive enlarged version of the record. When you hover over a coloured square in the interactive version, you see the text surrounding that square, with the keyword highlighted in its corresponding colour.

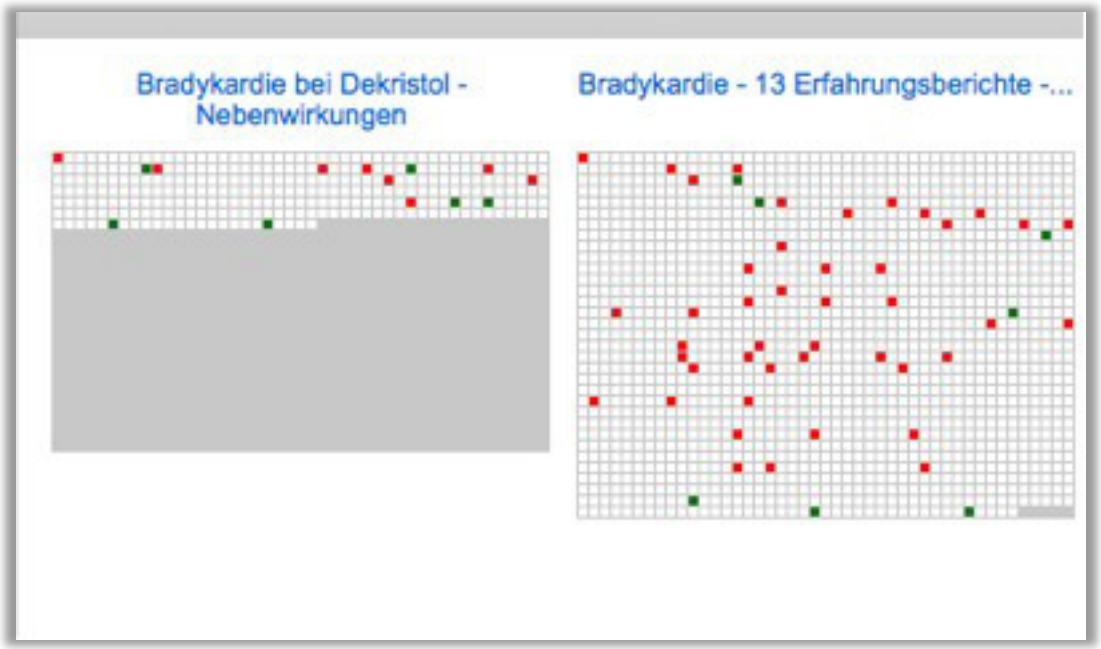


If you decide that you want to go to the underlying document, you just click on any of the white squares.

Other uses

Our blog has articles about various other ways of using Search Visualiser. Some of these are very different from other search engines. For instance, you can use Search Visualiser to find relevant records in a language that you don't speak. If you're trying to keep up with what's going on in other countries, that can be very useful.

For instance, imagine that you're a pharmaceutical researcher, wanting to know what work is being done in Germany on medication for bradycardia (slow heart rate). You can use an online dictionary to find the German terms for bradycardia and medication, then enter them into SV as your keywords. Here's an example of what you might find.



You can see that the second record is longer than the first, and contains much more about bradycardia (red squares) than about medication (green squares). You can start making sensible judgments about what's going on in each of these records, and about which of them would be worth running through translation software, even without speaking a word of German.

Notes:

The Search Visualiser is available for online use, free, at:
www.searchvisualiser.com