Searching for common names in large documents
By Gordon Rugg

Background: Finding common personal names in large documents or sets of documents is a notorious problem for most search engines, particularly if the person has one or more middle names, or if they are also known by a nickname. With Search Visualiser, it is easy to solve both these problems. This article describes how to do this, with worked examples from a very large document.

This article was originally posted on March 2, 2012

A classic problem in online search is finding someone who has a common first name and a common surname, like James Smith or Jane Jones. If you're using ordinary search engines, there are several options you can try, but they're all limited.

One is to use the specified phrase option, which tells the search engine to treat your chosen phrase as it if was a single word. A common way of showing this is to put your chosen phrase within inverted commas, e.g. "James Smith".

The trouble with this is that it will only show you records which have those two words immediately next to each other and in that order. It won't show you any cases of James F. Smith or of Smith, James.

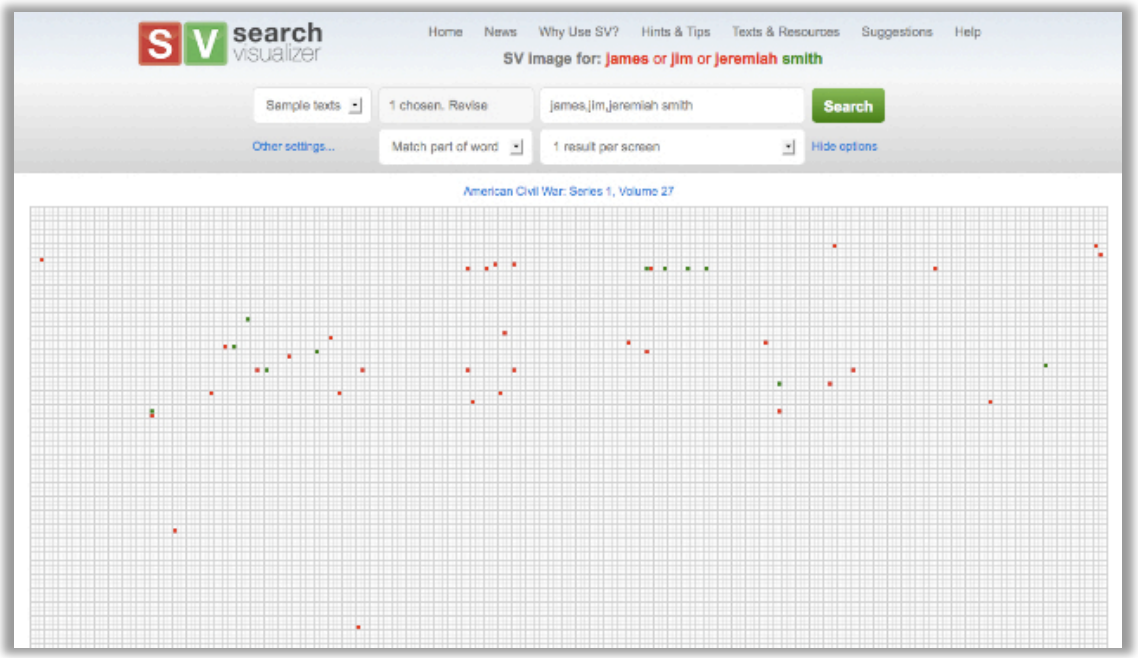
You can get round that problem by using what's known as proximity search – you tell the search engine to find cases where the two words are within a specified distance of each other, such as James and Smith within three words of each other. This would find cases such as James Edward Charles Windsor Smith. However, this usually involves getting into the advanced search options, and in practice most users are reluctant to do this. Quite a few search engines simply don't offer proximity search as an option.

Even if you are comfortable with doing proximity searches, there's a further complication with real-world use of common names. Common names are particularly likely to occur as nicknames, so someone called James on their birth certificate might actually be known as Jim, Jimmy or Jimbo. Some people are known by nicknames that have nothing to do with their first names, such as Dusty. As a further complication, many people use one of their middle names as their preferred first name.

So, if you're searching old records for an ancestor called James Smith on his birth certificate, but known to his friends as Jim and to his diving mates as Jeremiah, you're facing a challenge. On ordinary search engines, you can treat the nicknames and official name as OR options in the advanced search, but that will produce a huge number of false hits as varied as Jimmy Carter or the biblical Jeremiah. In principle, you could try a combination of OR search and proximity search, but that would be a non-trivial task for a professional information retrieval specialist. It would look something like this:

[James OR Jim OR Jeremiah] AND WITHIN THREE WORDS OF [Smith]

Here's what the same search looks like on Search Visualiser:



This screenshot shows the opening section of the official war record for the Battle of Gettysburg. That volume is over half a million words long, and contains a lot of mentions of people called James, Jim or Smith, plus quite a few mentions of people called Jeremiah.

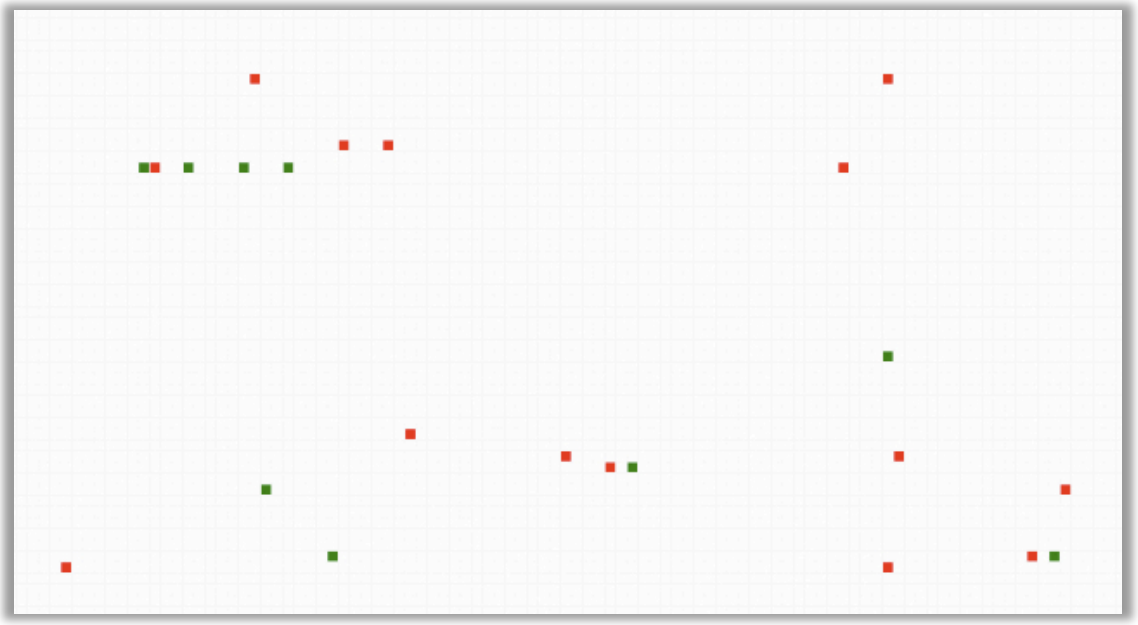
Imagine that someone has gone through that document highlighting occurrences of James or Jim or Jeremiah in red, and occurrences of Smith in green.

Where there's a red square followed immediately by a green square, then it shows one of the three J names immediately followed by "Smith".

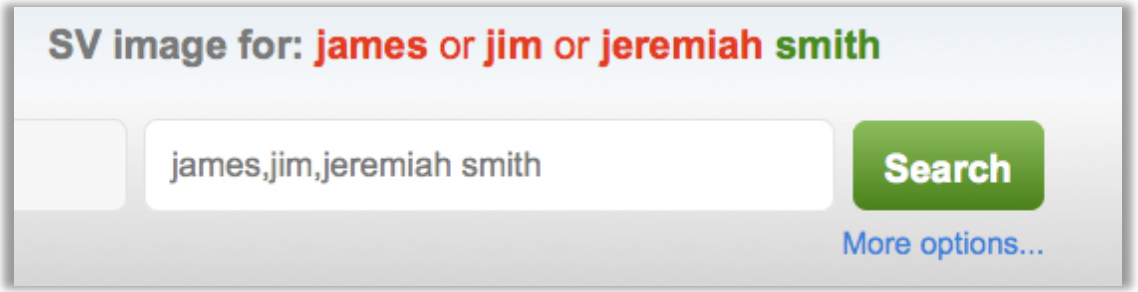
Where there's a small gap between a red square and a green square, it's showing a middle name or middle initial, such as "James F. Smith".

There's also a case of a green square followed by a red square, which is a "Smith, James" in an index at the start of the document.

You can easily see where there's a real hit, and where there's a hit on only one of the names.



So how do you do this within Search Visualiser?



If you look closely at the search bar in the screenshot, you see that the three versions of the first name are right next to each other, separated by commas, without spaces after the commas. That's how you tell SV that you want to treat those three words as synonyms of each other.

There's then a space, followed by the name Smith. That's telling SV that you want to treat Smith as a separate keyword.

You can have more than one cluster of synonyms if you want: for instance, if your ancestor used more than one spelling of "Smith" for their surname, then you could search for James,Jim,Jeremiah Smith,Smyth,Smythe

This would appear in the SV command section as

James,Jim,Jeremiah Smith,Smyth,Smythe

so that you can keep track of which words are being treated as synonyms of each other.

Notes

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