# Identifying plot structures in fiction with Search Visualiser

By Daniel Allington & Gordon Rugg

*Background:*
*This article describes how structures in fiction, such as stories within stories, can be identified via the distribution of textual features such as names of characters.*

*This article  was originally posted on October 4, 2012*

*Daniel Allington is based at The Open University, Milton Keynes, UK*

This working paper reports an exploratory study in the use of Search Visualiser as a tool for what Moretti (2000a) has called 'distant reading'. This term refers to literary analysis without actual reading – let alone close reading – of individual works, which Search Visualiser facilitates by providing a purely graphical representation of texts in which selected keywords appear as coloured cells in a grid. We set out to assess the viability of using Search Visualiser to study plot structure in novels by pinpointing mentions of the names of major characters. While the primary aim of distant reading is to study very large volumes of texts, including both canonical and non-canonical works, it is often helpful to test unfamiliar tools on familiar materials, so we chose to study the works of two very well-known authors: Jane Austen and Arthur Conan Doyle (who were, quite coincidentally, also the starting points taken by Moretti [2000b] in his study of genres).

In this study, our data comprise the Project Gutenberg e-texts of Arthur Conan Doyle's four Sherlock Holmes novels together with those of the four novels that Jane Austen published in her lifetime. These texts were chosen for reasons of availability in electronic form, and our approach in the current instance is to acknowledge rather than attempt to remedy their shortcomings as texts: they are not scholarly editions, but imperfect digitisations of historically unremarkable printings. We would suggest that they will suffice for the purposes of this working paper, which is primarily a technical proof of concept; the texts have therefore been taken directly from the Project Gutenberg website without further editing (thus, a blank space at the end of each visualization corresponds to the standard Project Gutenberg licence).

Here, the four novels are represented by Search Visualiser images showing distributions of the names of two main characters from each of the novels. These visualizations are compared and contrasted, leading to tentative conclusions about Austen and Doyle's different approaches to plot structure. Although these conclusions are highly provisional with regard to the specific authors under analysis, they argue for the viability of Search Visualiser as a tool in the study of plot structure.

## The images

We'll start with Jane Austen. In each of the four of the Austen novels under analysis, we searched for mentions of the female protagonist and the male protagonist. The images are shown at the foot of this article. In each case, the resulting visualization shows that the female protagonist (shown in red) is mentioned more frequently than the male protagonist (shown in green), and is in most cases mentioned throughout the whole of the text.

The exception to this pattern is *Emma*, in which there are two short sections (one about about a third of the way through and one close to the end) in which the male protagonist is not mentioned at all and the female protagonist is mentioned only rarely (*Pride and Prejudice* features a single similar section close to the beginning, though it is shorter). Throughout the remainder of *Emma*, mentions of both characters tend to be intermingled, but each of the remaining three Austen novels shows a different pattern. *Mansfield Park* begins with intermingled mentions of the female and male protagonists, but mention of the male protagonist becomes much less frequent from a point about three-fifths of the way through. *Pride and Prejudice* follows a slightly different pattern, with the disappearance of the male protagonist beginning just over three-quarters of the way through, and a return of the male protagonist towards the very end of the text. *Sense and Sensibility* begins with relatively scarce mentions of either the male or the female protagonist. Following this, mentions of both characters increase in frequency, with the female protagonist being mentioned frequently throughout while the male protagonist drops in and out of view: this gives the resulting visualization a banded structure, with sections in which mentions of the two are intermingled alternating with sections in which only the female protagonist is mentioned. As with *Pride and Prejudice*, the novel ends with a section in which mentions are intermingled, following an extended section in which the male protagonist was unmentioned by name. The above-described patterns are relatively subtle in comparison to those which the Search Visualiser has revealed for example in Shakespeare, but they appear to accord with observation: Austen's narratives revolve around her female protagonists, and not around their mysterious (and frequently absent) male suitors, a characteristic of the  courtship novel genre, whose focus was always *female* subjectivity (see Rogers, 1981).

Now we'll look at the Conan Doyle novels, in which we might expect to find very different patterns: the detective story did not come into existence until well after Austen's death, and was associated with a very different readership (a point to which we shall return below). In analyzing each of these works, we chose to search for mentions of the protagonist, Sherlock Holmes, and one other central character *apart from* the first-person narrator, John Watson. Conan Doyle's first Sherlock Holmes novel, *A Study in Scarlet*, shows a much more clearly pronounced pattern than the four novels examined above. To begin with, Holmes is frequently mentioned, and Drebber only rarely. Mentions of Drebber increase towards the approximate mid-point of the novel, from which point Holmes ceases to be mentioned altogether and Drebber is mentioned only sporadically. Mentions of Drebber again grow more frequent until the end of the novel, by which point Holmes has begun to be mentioned again, albeit more rarely than before. The second Sherlock Holmes novel, The Sign of the Four, shows a different pattern: at first only Holmes is mentioned, then Holmes and Thaddeus Sholto together, and then only Holmes again; after a few mentions of both together there is then a substantial chunk of the novel in which neither of these characters is mentioned, bisected by a narrow band of mentions of Holmes; finally, Holmes is mentioned a few last times before the very end. (A search for more than two characters would show that Thaddeus's father, Major John Sholto, is mentioned repeatedly in the section from which Holmes is absent.) In Conan Doyle's third Sherlock Holmes novel, *The Hound of the Baskervilles*, however, we see a return to the pronounced patterning of *A Study in Scarlet*. Stapleton, the second character that we chose to focus on in our analysis, is hardly mentioned in the first third of the story, in contrast to Holmes himself, who is mentioned frequently. This is followed by a substantial section in which Stapleton is mentioned more frequently than Holmes, a shorter section in which there is little mention of either, and a section in which mentions of the two characters are intermingled; towards the very end, however, only Stapleton is mentioned. In the last Sherlock Holmes novel, *The Valley of Fear*, for which we chose to visualise mentions of Holmes and Baldwin, there is an even stronger pattern: before the novel's approximate mid-point, there are mentions of Holmes but not Baldwin; then, after a brief intermingling, mentions of Holmes stop, and only Baldwin appears in the visualization until we reach a thin band of mentions of Holmes at the very end.

This shows where Conan Doyle uses the device of the story within a story, where the main narrative gives way to an embedded narrative which is a substantial story in its own right. In *The Sign of the Four*, for example, the denouement takes the form of an extended oral narrative that is interrupted at its midpoint when Holmes hands the storyteller a drink (hence the narrow band of mentions of Holmes).  Austen also made use of embedded narratives, of course, but these were much shorter, and consequently do not show up in such an obvious way in the visualizations presented here.

## Discussion and conclusion

Austen and Conan Doyle wrote at opposite ends of the nineteenth century in very different genres, so it's not surprising that they used different devices. What surprised us was how vividly the different structures showed up visually in the images above, suggesting the potential of Search Visualiser for this kind of research.

As noted in the introduction, it is in distant reading that we see Search Visualiser's potential for literary study; as such, we would caution against using findings such as those of the preceding section to form qualitative judgements of authors' merits. It might be tempting to argue, for example, that Conan Doyle's novels exhibit a more strongly varied structure, and conclude from this that they were in some way superior to Austen's. But it would be equally tempting to argue that the greater obviousness of that structure in our visualizations reflects a cruder approach to narrative on his part: he wrote for a less refined and more masculine audience than Austen, after all. Evaluative arguments (or assumptions) of that nature have been made even by major proponents of distant reading (for example, Moretti chose to focus his first experiments on 'the rivals of Conan Doyle' rather than on 'the rivals of Austen' for the arbitrary and somewhat contentious reason that 'detective stories have the advantage of being a very simple genre' [2000b, p. 212]). However, we would suggest that they are a distraction from the business at hand, since no form of distant reading is likely to resolve them.

For us, the major advantage of a tool like Search Visualiser is that it enables texts to be compared at a glance, permitting the analyst to investigate the content and structure of those texts in an intuitive but non-evaluative way. This article is a simple demonstration of concept, showing how this approach to visualization can be used to reveal structural features within entire novels. In future work, we plan to look at larger samples of texts in more detail.

## References

Moretti, Franco (2000a). 'Conjectures on world literature'. *New Left Review* 1 (January-February). Available online at http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature

Moretti, Franco (2000b). 'The Slaughterhouse of Literature'. *Modern Language Quarterly* 61 (1): 207-227.

Rogers, Katharine M. (1981). 'Dreams and Nightmares: Male Characters in the Feminine Novel of the Eighteenth Century'. In: Janet Todd (ed.), *Men by Women*. New York: Holmes & Meier Publishers. pp. 9-24.

## Technical notes

The images below may take some time to load – they are high resolution images of eight entire novels. We used high resolution to reduce the risk of image distortion on some browsers. The original png file for an SV image of "Emma" – the longest of the novels – was under 200K in size.

The original files used were from Project Gutenberg, and were visualised using the "single site" option of SV. The relevant url is shown at the head of each visualisation.

If you wish to visualise those files yourself, you may need to remove the last part of the url shown, back to the final backslash – the "single site" option is affected by how the site in question is structured, and sometimes you'll need to truncate the url back to successive backslashes until you reach a point in the structure where SV can access the relevant files.

The images are shown "long and thin" with a comparatively small number of squares in each horizontal line; this configuration usually shows clusters of keyword occurrences more clearly than a "short and broad" configuration. For side-by-side comparison of the images, you should be able to save them as images using the "save as image" facility on your computer (usually right-click on the mouse, but systems vary) and then view them with software such as Preview. The images below are "copyleft" – you're welcome to use them for non-commercial purposes, provided that you cite this original source in any publication of them.

**The images**

Image 1: Mentions of “Emma” (in red) and “Mr Knightley” (in green) in *Emma*.

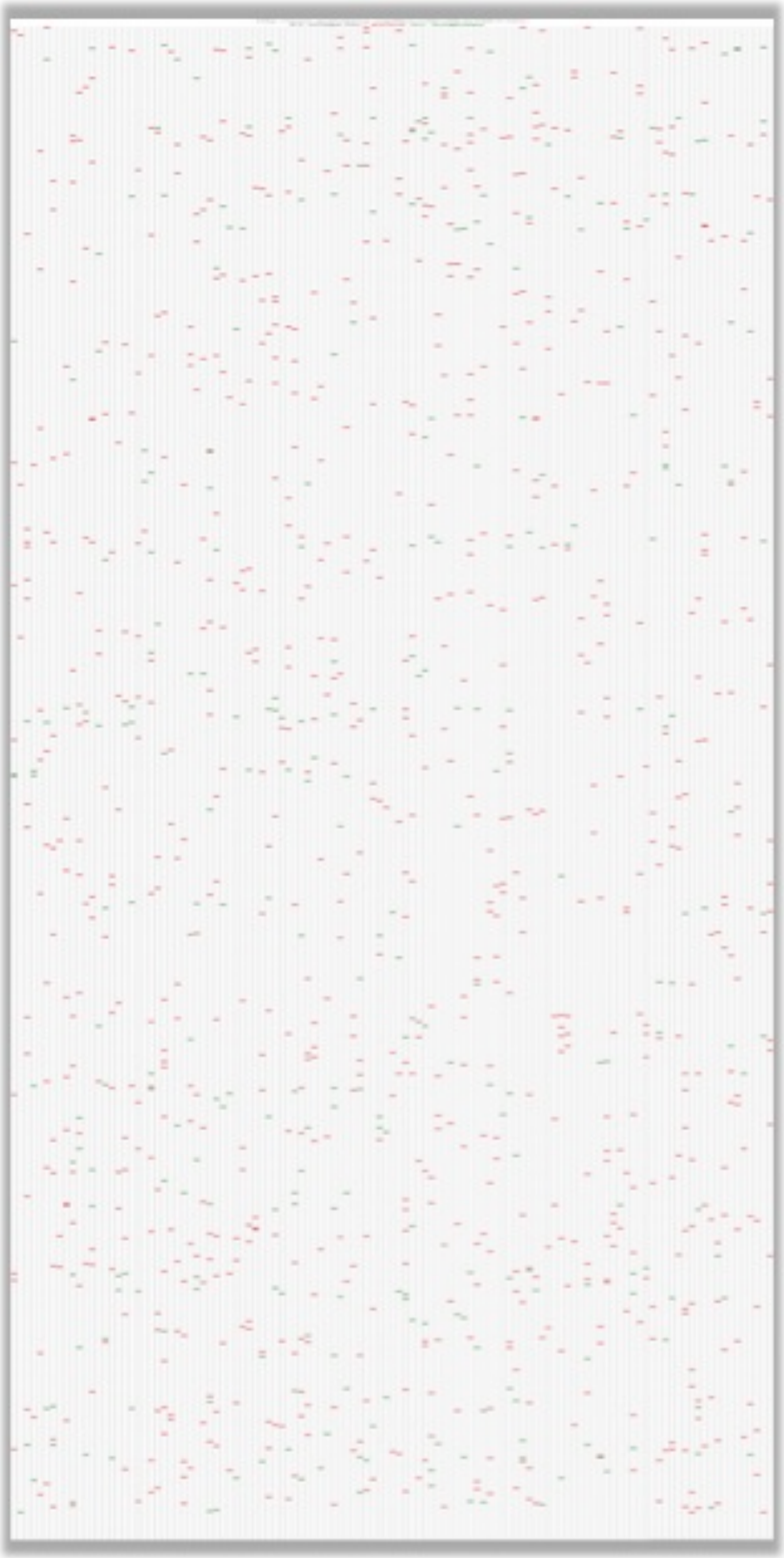


Image 2: Mentions of “Fanny” (in red) and “Edmund” (in green) in *Mansfield Park*.

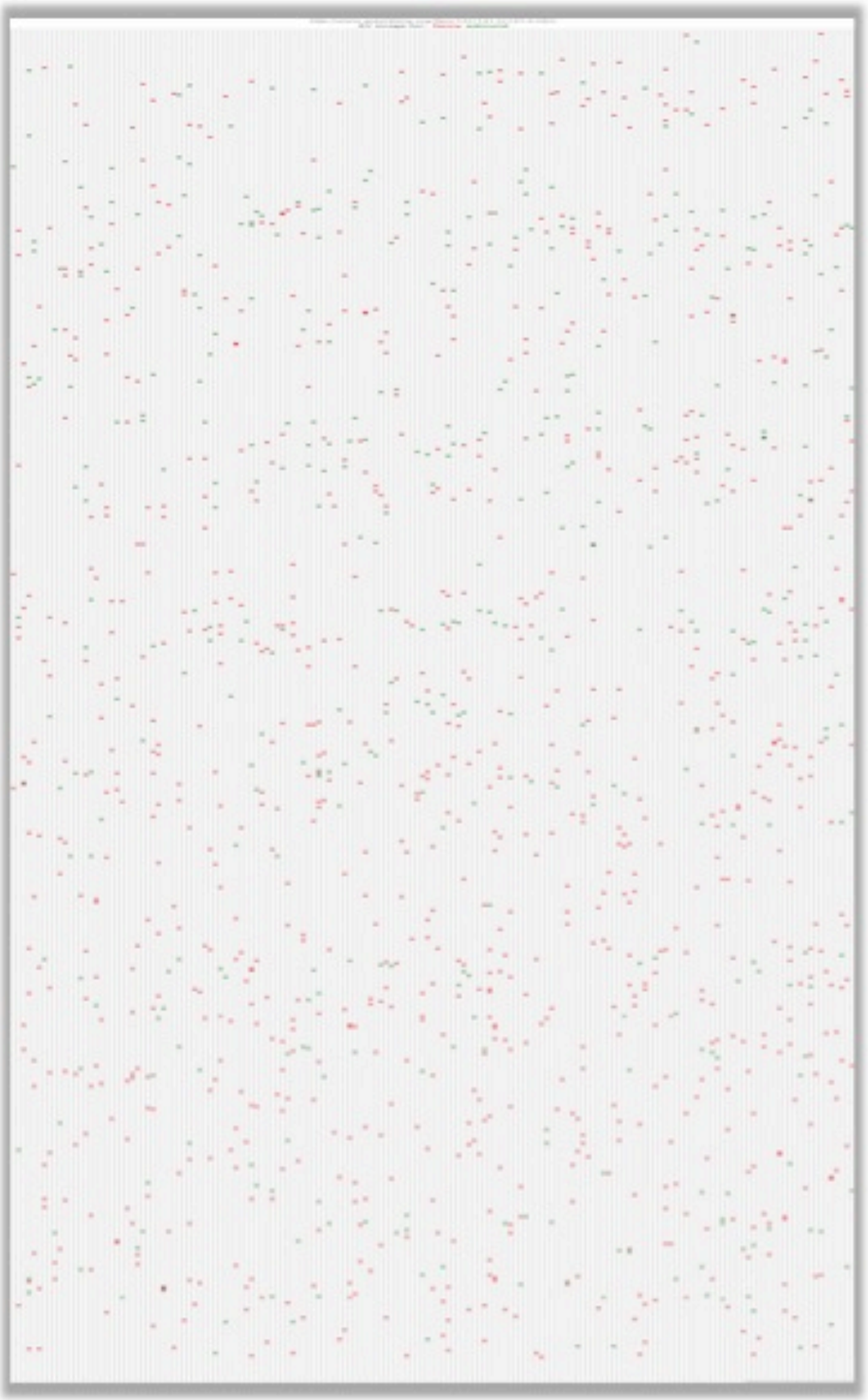


Image 3: Mentions of “Elizabeth” (in red) and “Darcy” (in green) in *Pride and Prejudice*.

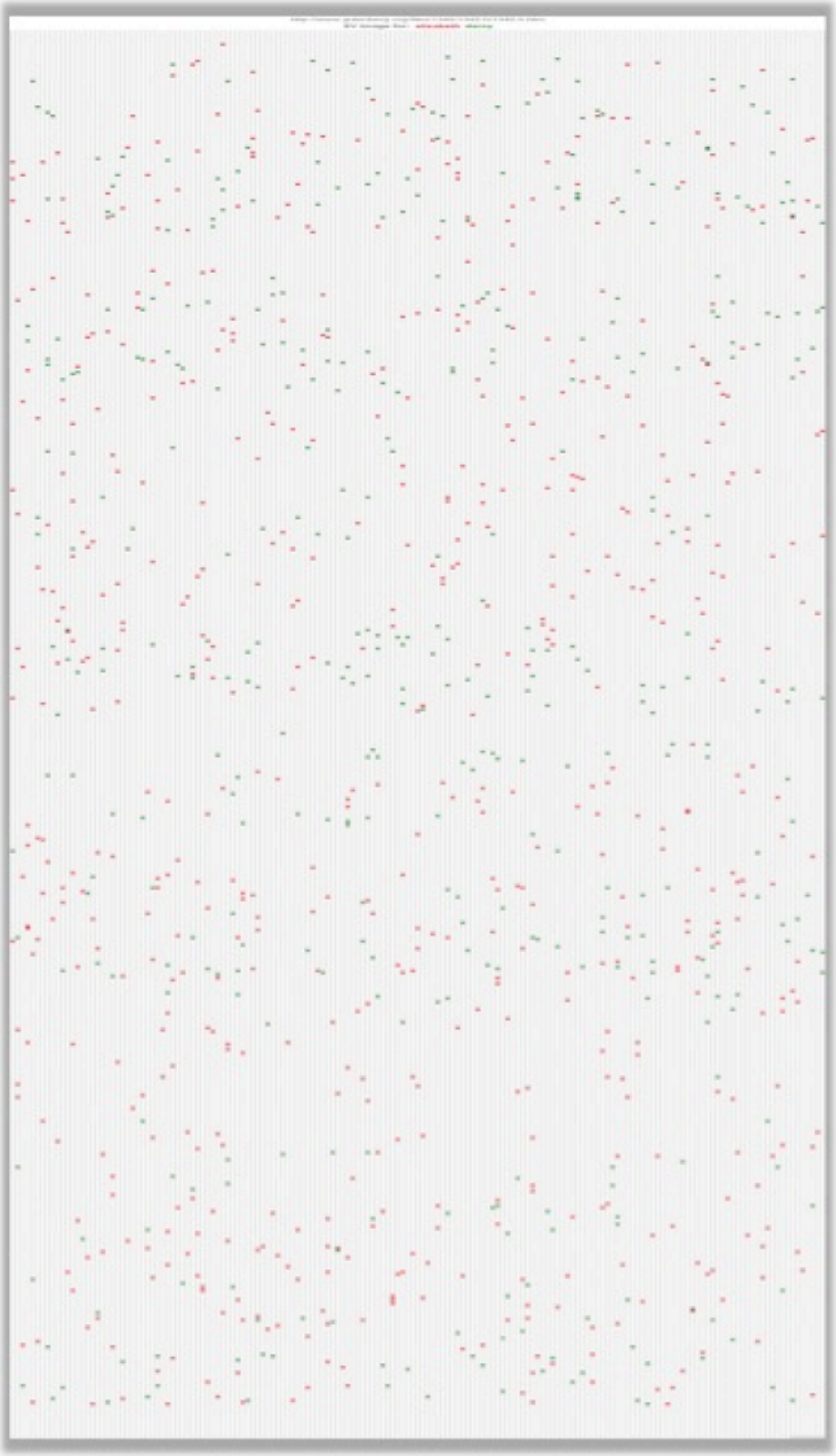


Image 4: Mentions of “Elinor” (in red) and “Edward” or “Ferrars” (in green) in *Sense and Sensibility*.

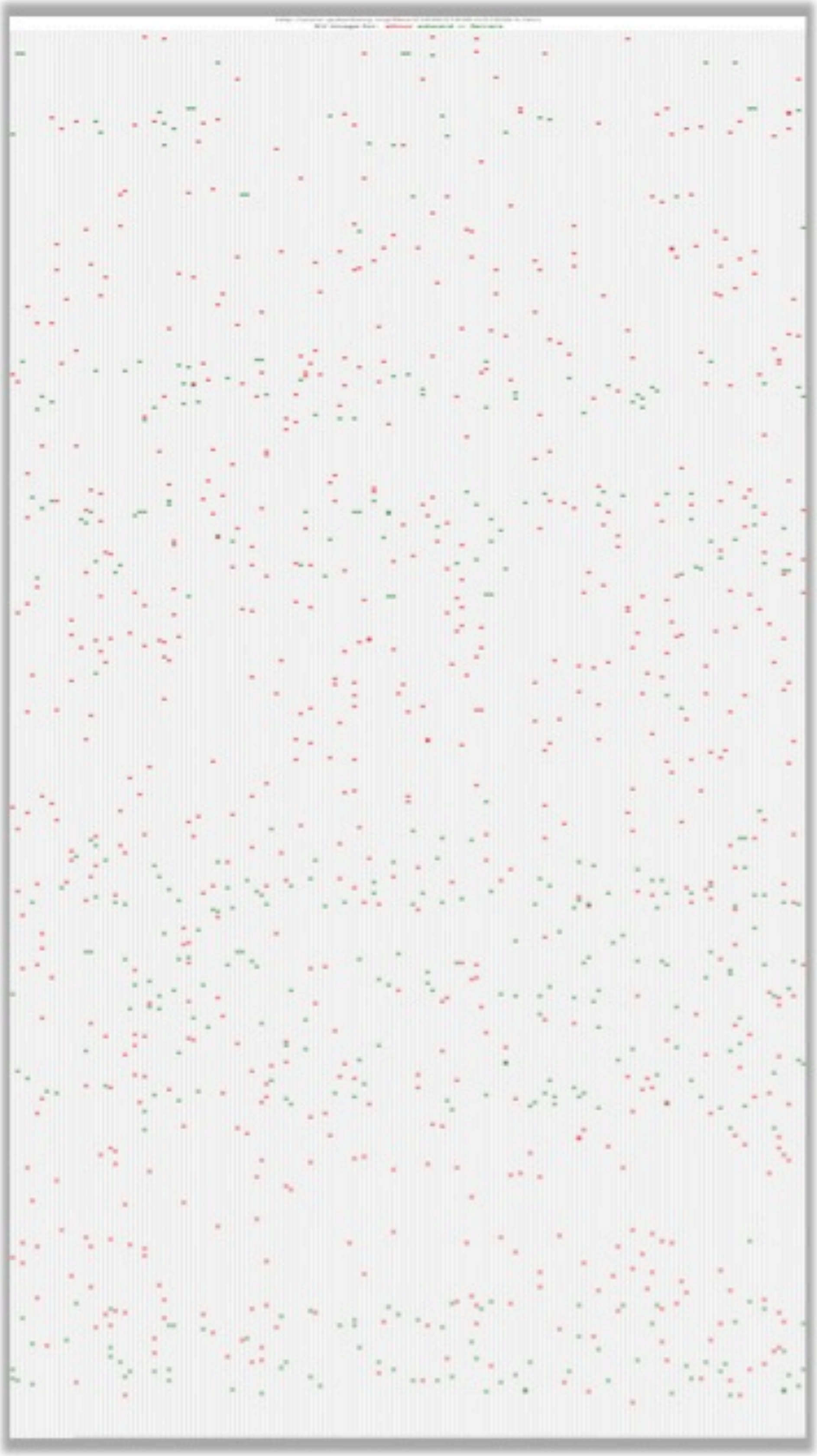


Image 5: Mentions of “Holmes” (in red) and “Drebber” (in green) in *A Study in Scarlet*.

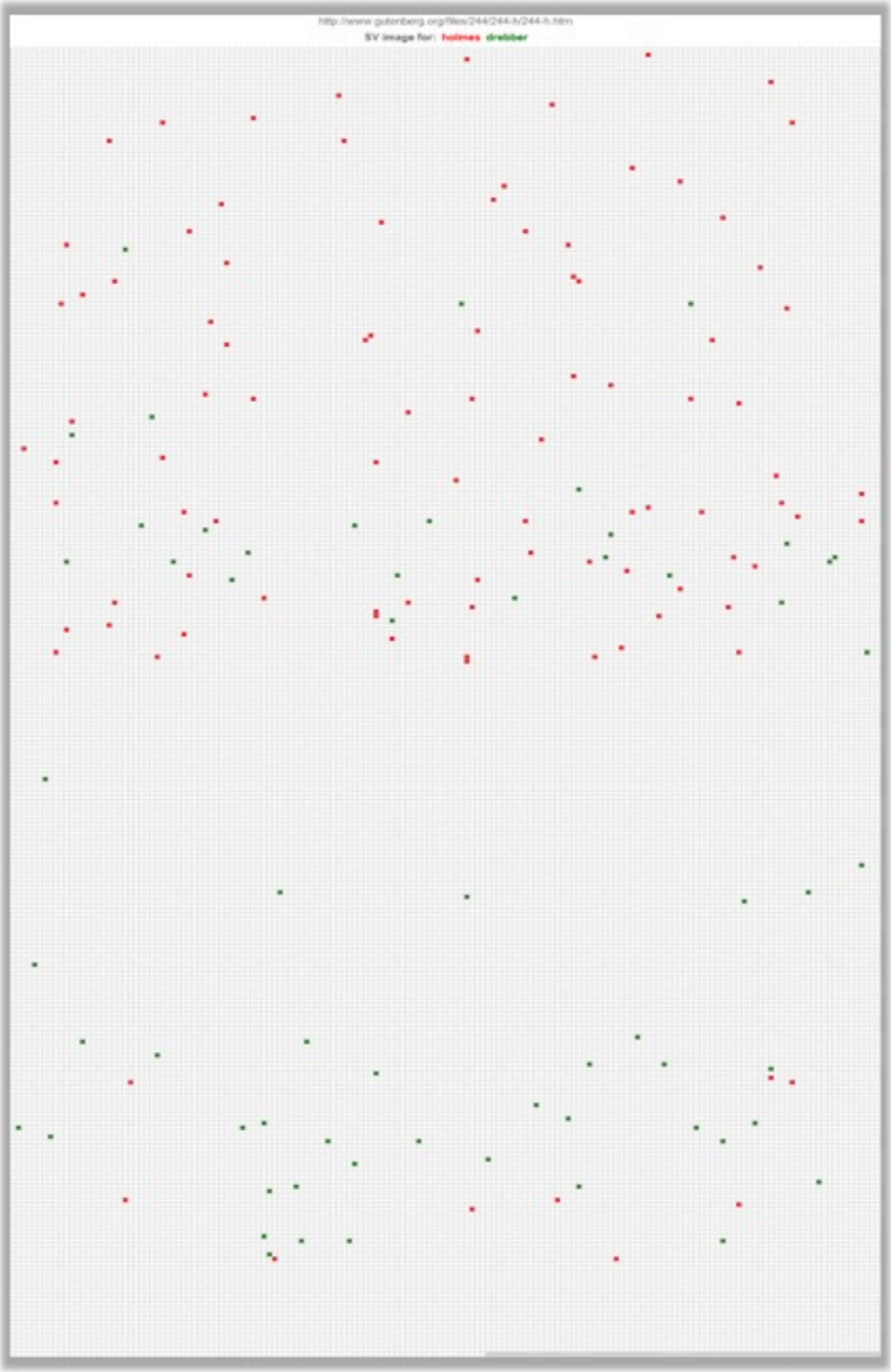


Image 6: Mentions of “Holmes” (in red) and “Thaddeus” i.e. Thaddeus Sholto (in green) in *The Sign of the Four*.

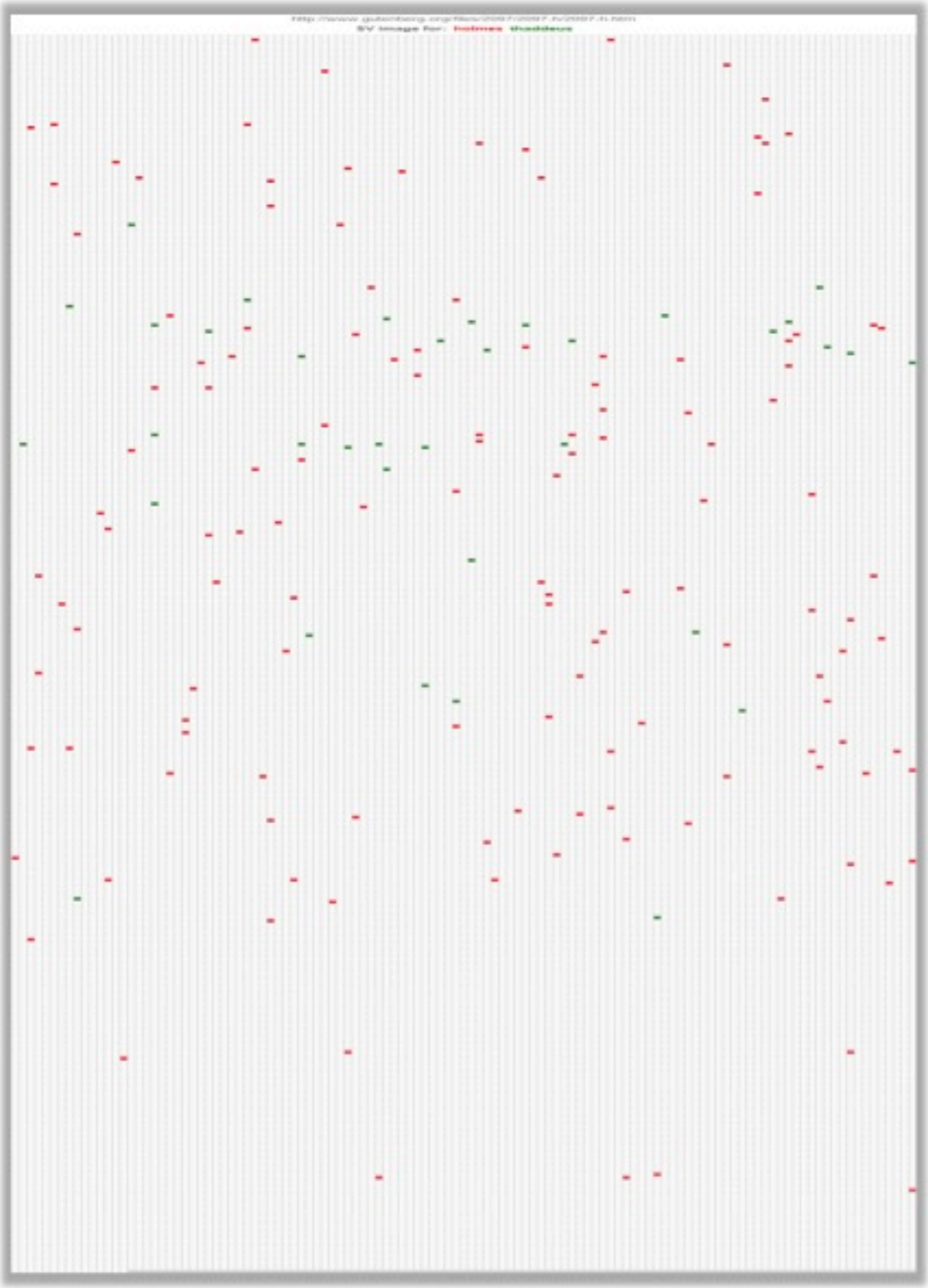


Image 7: Mentions of “Holmes” (in red) and “Stapleton” (in green) in *The Hound of the Baskervilles*.

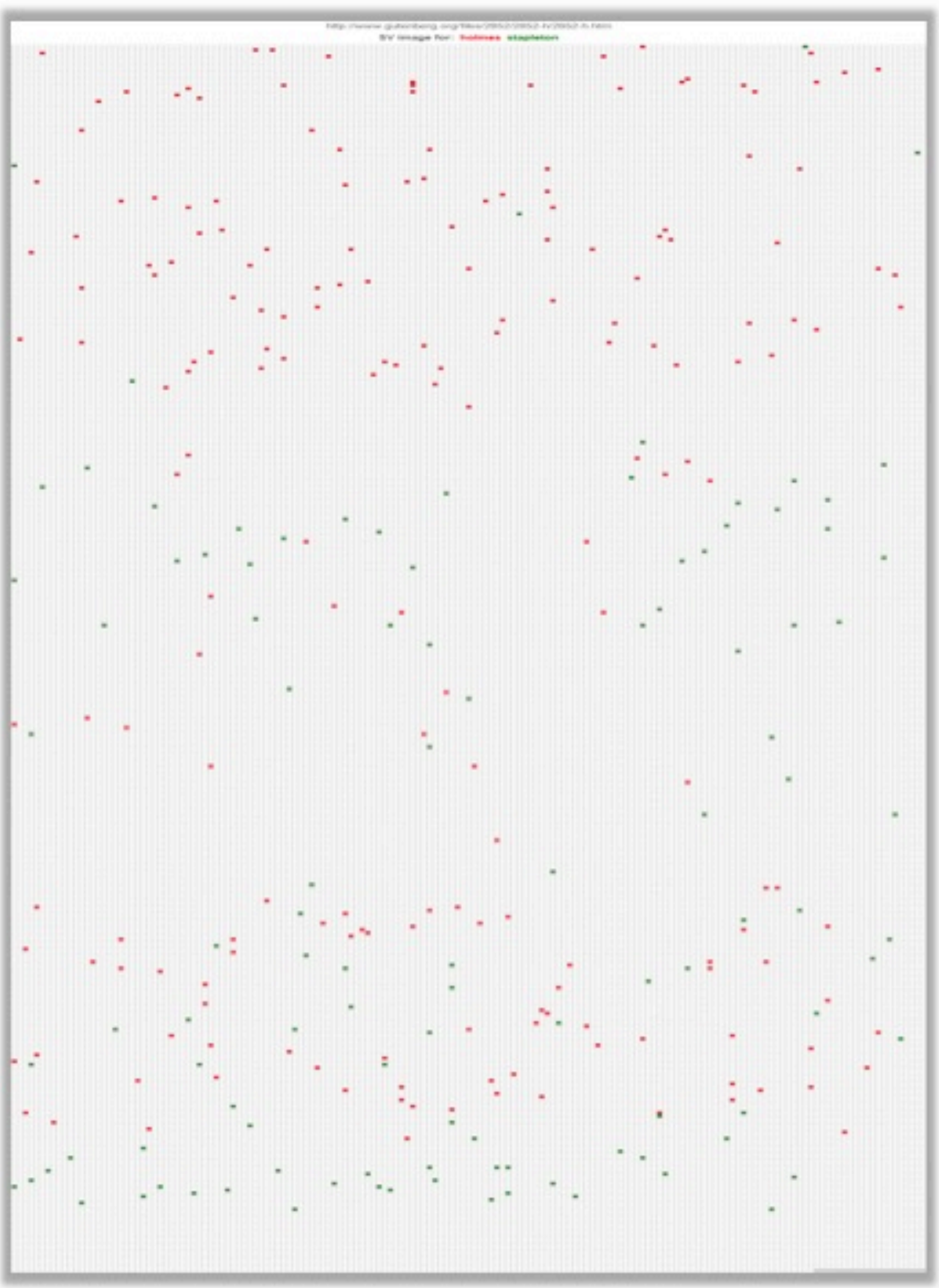
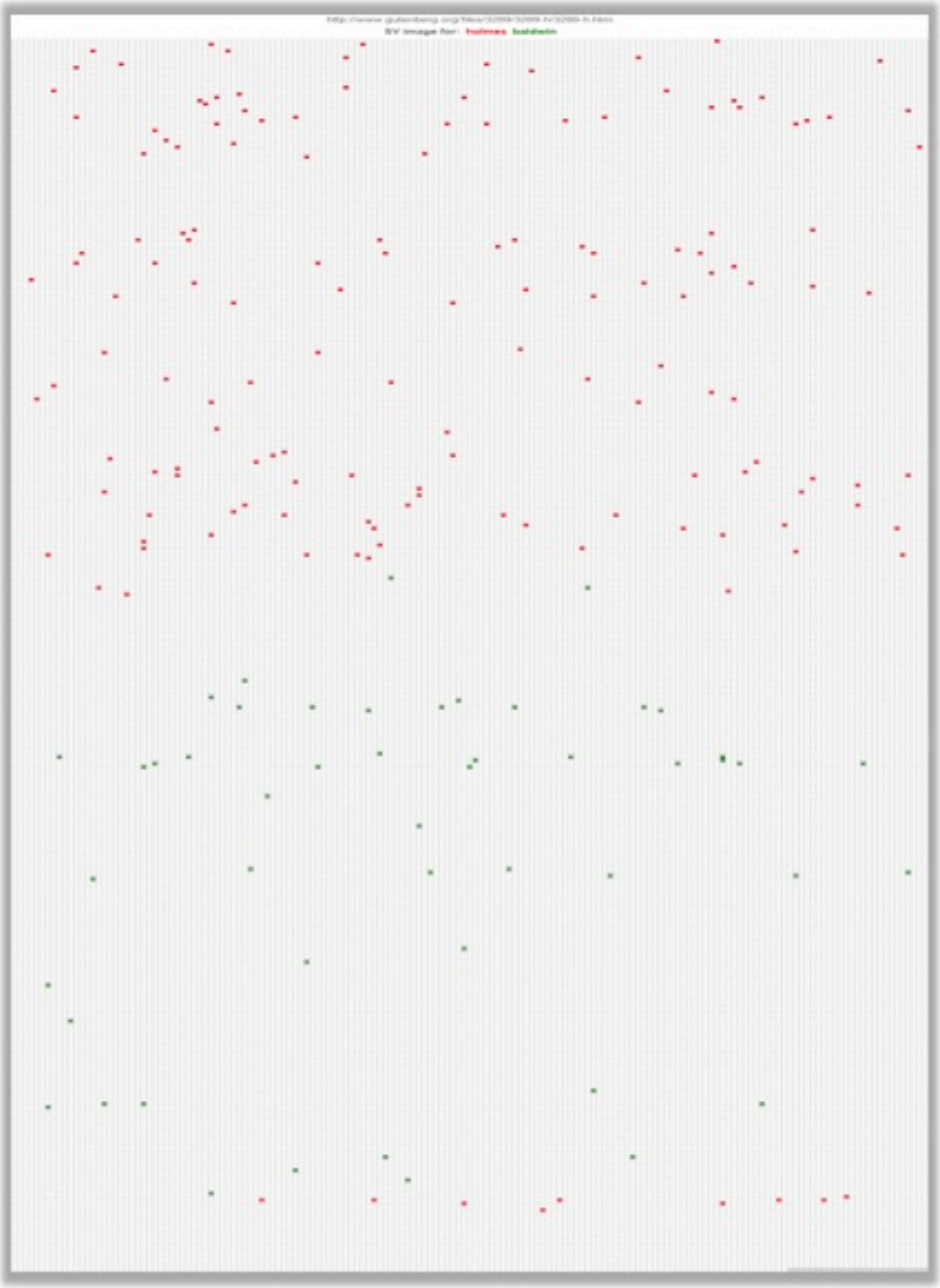


Image 8: Mentions of “Holmes” (in red) and “Baldwin” (in green) in *The Valley of Fear*.



**Notes:**

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