Google's Ngram Viewer

Jeffrey Barlow
Pacific University

Follow this and additional works at: http://commons.pacificu.edu/inter11

Recommended Citation
Google’s Ngram Viewer

Rights
Terms of use for work posted in CommonKnowledge.
Google’s Ngram Viewer

Editorial by Jeffrey Barlow

Google has brought out applications and various cloud-related services so very quickly that it is sometimes quite easy to miss them. Google’s Ngram Viewer is such an application from late 2010.

Ngram is simultaneously a service, a site, an application, and a search device—all summed up in the name “Viewer”. It surveys the many millions of books in the humanities scanned and indexed to date by Google. The Viewer permits the user to scan for strings of up to five words found in any or all of the works in a search process largely defined by the user, and in a variety of languages, too.

Here is Google’s explanation of the Ngram Viewer:

“Since 2004, Google has digitized more than 15 million books worldwide. The datasets we’re making available today to further humanities research are based on a subset of that corpus, weighing in at 500 billion words from 5.2 million books in Chinese, English, French, German, Russian, and Spanish. The datasets contain phrases of up to five words with counts of how often they occurred in each year.”

There are many potential uses for such a search device. Students of intellectual or cultural history, for example, are often going to be interested in the occurrence and distribution of particular terms or ideas over long time periods.

For example, we are currently assisting a senior history major with her thesis research. The topic is submarines and the policies adopted in an attempt to control them via treaty agreements.

One of the student’s major arguments after initial research is that the submarine was quickly associated with criminal behavior, as it began to develop in the late nineteenth and early twentieth centuries, because it violated Romantic traditions of warfare. These were basically related to chivalry and held that...

1 http://booksearch.blogspot.com/2010/12/find-out-whats-in-word-or-five-with.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+blogspot%2FCjSP+%28Book+Search%29
See also a very positive description of the Ngrams at: http://online.wsj.com/article/SB10001424052748704073804576023741849922006.html
Yet more detailed analysis can be found at: http://www.sciencemag.org/content/early/2010/12/15/science.1199644
gentlemen did not read each other’s mail, nor would they kill civilians or attack from ambush as did submarines, etc.

Furthermore, the student believes, the technological development of the submarine eventually outran attempts to craft appropriate policy, and by World War II, war had grown so terrible that the submarine seemed no worse than many other hitherto inhuman weapons. In short, by 1945 the Romantic vision of warfare was pretty much dead.

These are not, of course, easy arguments to prove or demonstrate. Most arguments in intellectual history usually must rest on evidence drawn from one or more influential writers, and always leave upon questions as to how typical or widespread that particular perspective in fact was. With the Ngram Viewer, we have an opportunity to discuss the occurrence of any given search string on the basis of a large sample.

In an effort to improve our own understanding of these issues, and to learn to use Ngram, we turned to it. I first ran a search on the word “submarine” in Ngram’s English-language corpus. Here were the results: ² [following page]

---

² [following page]

Figure 1 Ngram Graph of the Word "submarine"
This graph [previous page] nicely demonstrates two important points of which the student was already aware: concern about submarine warfare peaked in the early period of the First World War, and again in the Second World War, but the use of the term was apparently much more frequent in the first period. This is possible support for the student’s position that the submarine had pretty much lost its criminal aura by the end of World War II.

However, we also learn some of the limitations of an Ngram search. When we click on the graph we can go to the list of books where the term was used. Here we find out that there are various meanings of *submarine*, which includes not only the naval weapon, but also the undersea telegraph cable. So we must narrow our search to see if the first simple conclusion—that apparently there were more works published which discussed submarines in WW I than in WW II—is accurate or that it may be some statistical freak resulting from searching a word used in a variety of contexts.

So now we run the search “submarine warfare.” Here are the results:  

[following page]

---


176
Figure 2 Ngram Graph of Words “submarine warfare”
We have now filtered out the references, we assume, to everything but submarine warfare. We see that our initial conclusions seem to still stand. Though there are some fluctuations in the percentage of books which utilized the term, the respective peaks are still visible.

This new search is interesting however, because there is a bump in usage between 1812 and 1820. This is very early for submarine warfare, and we want to see what is causing the bump. We refine the search at the above link by going into the frame 1800-1917 at the foot of the graph. We then further refine the search in the next page, by undertaking a more sophisticated search that enables us to sort the documents by years published.

This leads us to a number of books from the period of 1812-1820, most referring to the proposal of Robert F. Fulton to develop submarines for use by the British government against Napoleon. This, incidentally, is something of which neither myself nor the student doing the actual work was aware.

Now we want to take advantage of the language filter on Ngram to see if there are differences between German and English language sources. This is because, as has long been known, Germany made good use of the submarine earlier than any other power, and it was precisely the degree to which it violated Romantic notions of war that caused Woodrow Wilson, among others, to condemn Germany for its use. Germany, in short, may have been much less prejudiced toward submarine warfare than were the Allies, particularly America.

Although German is not one of the pitifully few languages I can use, I eventually come up with the search term *Unterseebootskrieg* and in an Ngram arrives at a page with quite similar graphs to our English language search found here.

[following page]

---


5 [http://ngrams.googlelabs.com/graph?content=Unterseebootskrieg&year_start=1800&year_end=2000&corpus=8&smoothing=3](http://ngrams.googlelabs.com/graph?content=Unterseebootskrieg&year_start=1800&year_end=2000&corpus=8&smoothing=3) (Note: The capitalization of the 'U' in "Unterseebootskrieg" is important when performing an ngram search)
Figure 3 Ngram Graph of the Word “Unterseebootskrieg”
There are useful starts for our search, but we still need to see if the terms *submarine* and *criminal* are often linked in the periods we are studying. We arrive at this graph:  


---

6 http://ngrams.googlelabs.com/graph?content=submarine%2C+criminal&year_start=1800&year_end=2000&corpus=0&smoothing=3
Figure 4: Ngram Graph of the Word "submarine" and "criminal"
Here we begin to learn a bit more. One of the student’s assumptions seems to be holding up rather well. In English at least, the submarine seems to be less linked with the notion of “criminal” in the Second World War than in the First.

This is by no means a safe conclusion without much additional research, but it has shown some of the power of the Ngram viewer, and in addition our searches have led us to a number of useful sources.

There are some systematic issues of which users attempting major generalizations should be aware. There is an amusing analysis of the occurrence of the F word---or is it the S word? ---found here.  

However, there are also some distinct limitations in the device. Changes in spelling over time can be critical, and it is also difficult to be fully confident with the thoroughness of any given search.

We think that Ngram is a very powerful search tool and that it should be investigated by any one interested in changes in words or concepts over time. For those wishing to do high-level searches, Google permits the user to download discrete data sets and to manipulate them locally. So, like any Internet search, Ngram is only a beginning. It is also a great deal of fun. Once again, thanks, Google!


8 For somewhat contrarian views go to: http://nataliacecire.blogspot.com/2010/12/google-books-ngrams-and-number-of-words.html