Google Ngram Viewer

Google Ngram Viewer: What is it?

Google's Ngram Viewer allows users to search the Google books database for word frequencies over a period of time. The books used are scanned in different languages from public libraries all over the world. Ngram Viewer therefore allows users to search word frequencies in any language over any period of time.

See https://books.google.com/ngrams/info

Culturomics

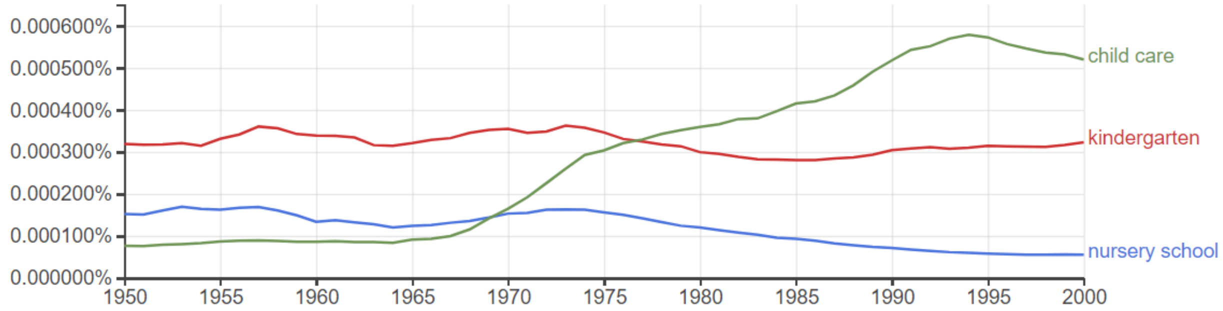
The use of the Google Ngram Viewer gave rise to a whole new field of studies called culturomics, for a while leading to the hope that that the millions of digitized books by Google and the Ngram Viewer tool would completely transform the way we study culture. But what is *culturomics*? "Culturomics is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts. Researchers data mine large digital archives to investigate cultural phenomena reflected in language and word usage."

On culturomics, see Michel et al. (2011), Michel and Lieberman Aiden (2011), Letcher (2011).

How does the Google Ngram Viewer work?

If you searched in Google Ngram Viewer the following three ngrams from 1950 to 2000: "nursery school" (a 2-gram or bigram), "kindergarten" (a 1-gram or unigram), and "child care" (another bigram), the Google Ngram Viewer will display the following graph where

1. The x-axis contains years in increments of 5 years
2. The y-axis shows the percentage of all the bigrams contained in the Google sample of books written in English and published in the United States, that are "nursery school" or "child care"? Of all the unigrams, what percentage of them are "kindergarten"?



N-GRAM- "statistical analysis of text or speech content to find n (a number) of some sort of item in the text."

Before Getting Started:

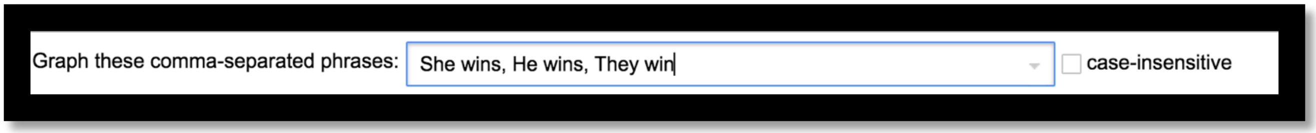
- Words in NGRAMs search are case sensitive (make sure to capitalize proper nouns)
    - If words are capitalized that should not be, it will only search for words in text where the word is capitalized
- Can analyze words or phrases
- Each item being searched should be separated by a comma (Ex: She wins, He wins, They win)
- The lower the "smoothing level", the more accurate the frequency lines

Step one:

- Have an idea how two or multiple words relate to one another and what a correlation in frequency may prove

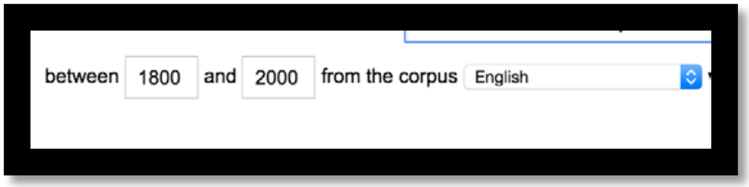
Step two:

- Type searches of interest into bar, next to where it says "Graph these comma-separated phrases"
- If you would like search to be case insensitive, check the box next to the search bar
- If you would like to put emphasis on one of the words being searched in relation to the other, place an asterisk next to word and a multiplication factor, enclose both word and factor in parenthesis (She wins*2)


Graph these comma-separated phrases: She wins, He wins, They win □ case-insensitive

Step three:

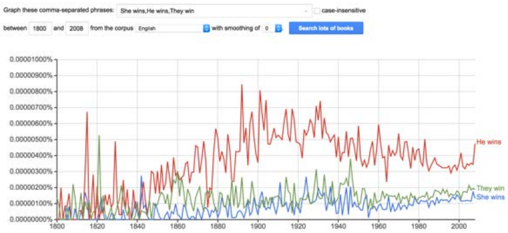
- Specify what years you would like the search to operate between
- Select a language


between 1800 and 2000 from the corpus English

Step four:

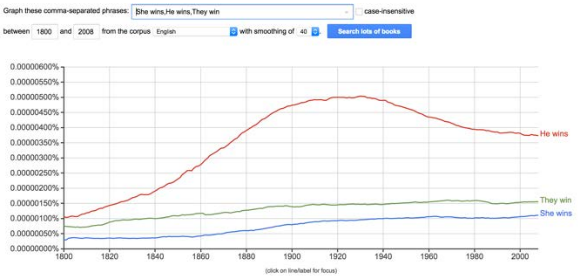
- Change the "smoothing"

- *Smoothing* is how jagged the lines created are
- The lower the smoothing, the more accurate but harder to read



Graph these comma-separated phrases: She wins,He wins,They win □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 0



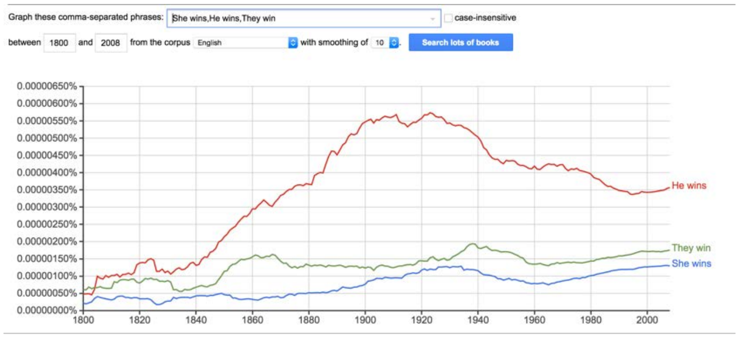
Graph these comma-separated phrases: She wins,He wins,They win □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 40

Step Five:

- Click "Search lots of books"
- Evaluate results!

Frequency



Graph these comma-separated phrases: She wins,He wins,They win □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 10

Year

Step Six:

- Make Conclusions!
- What are my data telling me?
- Use findings as evidence in research or as ideas for new patterns

References

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 14 January 2011, Vol. 331, pp. 176-182.

Michel, Jean-Baptiste and Erez Lieberman Aiden. 2011. "What we learned from 5 million books". https://www.ted.com/talks/what_we_learned_from_5_million_books?language=en

Letcher, David W. 2011. "Cultoromics: A New Way to See Temporal Changes in the Prevalence of Words and Phrases." *American Institute of Higher Education 6th International Conference Proceedings*. Vol. 4, No.1, pp. 228-236.

More TIPS

TIPS_NLP_Ngram (word & character).pdf

TIPS_NLP_Ngram and Word Co-Occurrence Viewer.pdf