Shredding the Great Books
- or -
Google Ngram Viewer Turns Snippets into Insight

ISC1057
Janet Peterson and John Burkardt
Computational Thinking
Fall Semester 2016

Computers have the ability to store enormous amounts of information.

But while it may seem good to have as big a pile of information as possible, as the pile gets bigger, it becomes increasingly difficult to find any particular item.

In the old days, helping people find information was the job of phone books, indexes in books, catalogs, card catalogs, and especially librarians.

Now new books can appear online as digital copies; old books can also be put online after some difficulty. How is this done? What tools are there for finding and accessing this information? What kind of problems can arise? What kinds of new investigations can be made now that so much data is available, and so many new tools can be used?

DIY
WORLD WAR II
PANDORA
KING STRAUB
MILLER'S
JO NESBO

In 2004, Google announced a plan for a universal electronic library,

This meant making an electronic copy of every book in existence.

To do this, they had know: *how many books are there?*

This question raises a new one: *What do we mean by a book?*

Do we include a comic book? A phone book? A high school yearbook? A copy of a magazine or newspaper? A map?

In our human experience, there are many situations in which we have to say, "I can't exactly define what I mean, but I know it when I see it." However, if we are going to deal with a computer, we do have to define our concepts exactly, even if the resulting definition does not always match our human expectations.

Thus, one aspect of computational thinking is:

Dealing with computers requires a precise definition of ideas.

Google considered the idea of counting every distinct bound volume as a separate book, but this would mean a hardback and paperback version of the same material would count twice, and that several works bound in a single volume would only count once.

Another possibilitity was to count a book only if it had an index number like the International Standard Book Number (ISBN) or Library of Congress catalog number, or WorldCat accession number (OCLC). However, these indexing systems also had problems; a series of books might have a single number and a single book might be assigned several numbers.

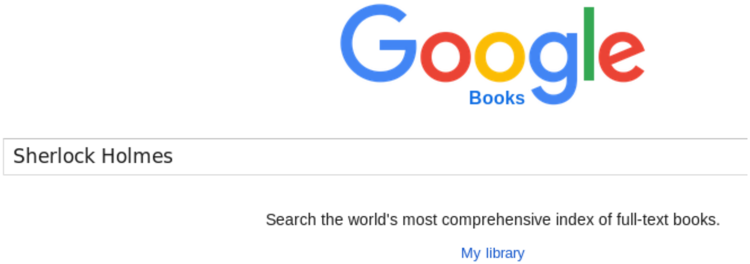
So Google decided to collect a billion raw catalog data records from 150 libraries and other book catalogers, and then tried to reduce these records down so that each one corresponded to a unique written work, ignoring whether it was paperback or hardback, or bound separately or with other works, and so on.

This resulted in an estimate of **129,864,880** distinct books.

Is 129,864,880 distinct books a large number?

- The FSU library has about 3,235,243 books;
- The British National Library has 13,950,000 books;
- The Harvard Library has about 18,900,000 books;
- The United States Library of Congress has 23,892,068 books;

These collections have been made over many years; the books vary in size; some are falling apart or are very fragile; others are in storage; there may be many copies or editions of essentially the same book.



Sherlock Holmes

Search the world's most comprehensive index of full-text books.

To create the electronic library, Google contacted university libraries and offered to “digitize” their collections. (We’ll need to examine this concept more carefully shortly.)

In exchange for access to each book, Google would share a copy of the electronic version with the library.

The result would be a huge database called Google Books.

Many libraries cooperated; they were glad to have old, decaying books preserved in this way, and they believed that users would have now have access to many more books than any single library could hold.

As of 2015, Google had scanned 25 million books, roughly 1/5 of their goal.

Although Google had negotiated with the libraries, it did not notify publishers and authors that it was scanning copyrighted works.

Librarians were eager to cooperate; they see their job as making books freely available, but are constrained by:

- limited budget;
- limited hours of operation;
- lack of space for storing books;
- difficulty of accessing rarely used books;
- the need to carefully index and catalog books;
- the need to restore books to the shelf upon return;
- danger of accidental or deliberate damage to books;
- book theft;
- slow operation of interlibrary loan;
- growing demands for computers and computer areas;

One important reason that librarians supported digitization of books was the fact their collections were crumbling. For hundreds of years, books, newspapers, and journals were printed on paper made from wood pulp that was bleached to appear white; but over time, the paper turns dark and brittle and begins to crumble into dust.

Some librarians had started trying to save their most precious books, but there was little money to deal with this new unexpected and slow-moving catastrophe.

The fact that Google would scan their books for free, and give them access to the scanned images, seemed like a miracle.

HarperCollins
THE BOOKSELLER INDUSTRY AWARDS PUBLISHER OF THE YEAR 2012
FRANCE
Books from
TURKEY IN ALL ITS COLOURS

Publishers objected to Google's plans. Publishers already felt that they lost sales because libraries had legal permission to share copies of books to one user at a time.

Google would be making all books available to everyone.

Publishers felt that Google was hiding behind non-profit libraries, and that Google was certain to earn money by selling adds that would appear alongside the scanned books.

Publishers didn't think that was wrong - they just wanted to be the ones who made that money!

They proposed the creation of an organization that would control the digitization of books, charging a license fee to every individual user, organization, or library, the way ASCAP does for music.

Authors, in turn, felt that publishers did not deserve to control new uses for the works they had written.

They insisted on filing a separate law suit, claiming that not only did Google not have the right to redistribute their works, but that the publishers also did not have this right.

Now the dispute potentially involved thousands of people, all with different motives and needs.

Each participant in the conflict, Google, librarians, publishers, and authors, could feel that they had right on their side; but each could also be accused of “really” only being interested in money.

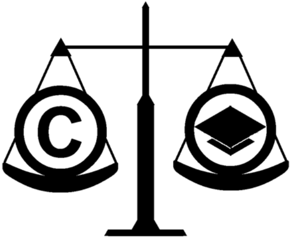
Google books

The Authors Guild filed a federal lawsuit in New York against Google in September 2005, which was followed by a lawsuit from the Association of American Publishers.

In 2008, a settlement was proposed by representatives of authors, publishers, and Google, which included:

- Google would pay $60 to the author of every book that was scanned;
- Google would pay $125 million to recompense copyright owners, and to fund a Book Rights Registry that would continue to distribute royalties to owners;
- Google would set up free portals in 4,000 colleges and universities;

In March 2011, this proposed settlement was rejected by the court.



As users, we might think Google Books is wonderful, free access to everything. But here are arguments made against it:

- Google is not a public library, it is a company. We tolerate libraries sharing books, but they don't do this for a profit;
- Google is worse than a pirate - the difference between Google Books and a pirate web sites is that pirates distribute music others have copied, but Google actually makes the illegal copies first, and then distributes them;
- Google is a monopoly - Google will control how the information is used;
- Censorship: Google may be pressured by various governments to suppress certain literature. As a profit-making company, it may yield to such pressure;
- Orphans: books whose copyright owners can't be determined are called orphans. Google doesn't pay anything to scan and display such books. Why should Google get this windfall?
- Privacy: Google will be able to tell what books people read; collecting, using, and even selling such data may violate privacy.

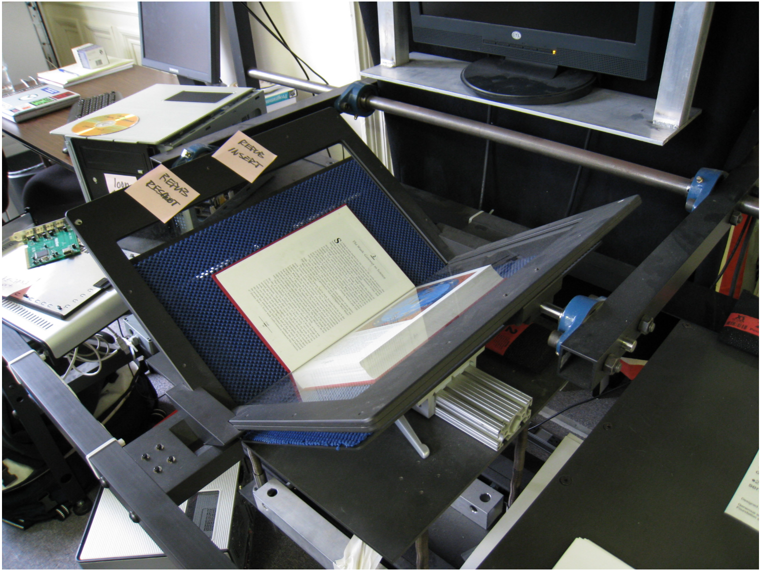
Google also tried to reduce the controversy with other actions:

- For any copyrighted work, Google would only allow users to see small portions or “snippets”, not the whole thing;
- Google would allow any publisher to withdraw all their books entirely from the database;
- Google would allow any author to withdraw their book, or allow only snippets;
- Google would include a “buy this book” link along with a search result.

The Authors Guild continued their lawsuit, but in 2013 it was dismissed by the US Circuit Court,

In October 2015, their appeal to the Second US Circuit Court of Appeals was dismissed.

In December 2015, the Authors Guild appealed to the Supreme Court, and this appeal was rejected in April 2016.

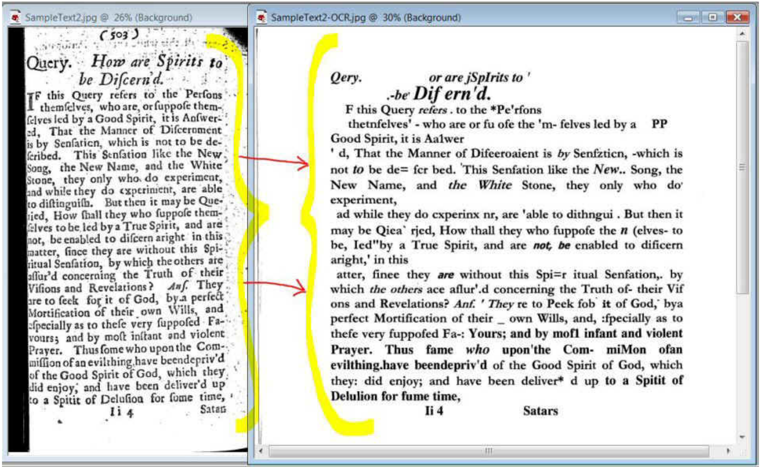

REPUB INSERT
REPUB RESHOOT

Scanning a book is essentially taking a photograph of each page.

In order to control cost, the operator only has to place the book into a cradle, after which the book scanner can turn pages automatically.

The quality of the resulting image depends on the strength of the scanner light, the physical state of the book, the printing style, the scanner resolution, and many other factors.

The resulting image may be too faint to read, or the text on the next page may show through, or because of physical faults the page may have been partially turned down.


SampleText2.jpg @ 26% (Background)
SampleText2-OCR.jpg @ 30% (Background)

(503)

Query. *How are Spirits to be Diſcern'd.*

IF this Query refers to the Perſons themſelves, who are, or ſuppoſe themſelves led by a Good Spirit, it is Anſwered, That the Manner of Diſceroment is by Senſaticn, which is not to be deſcribed. This Senſation like the New Song, the New Name, and the White Stone, they only who do experiment, and while they do experiment, are able to diſtinguiſh. But then it may be Queried, How ſhall they who ſuppoſe themſelves to be led by a True Spirit, and are not, be enabled to diſcern aright in this matter, ſince they are without this Spiritual Senſation, by which the others are aſſur'd concerning the Truth of their Viſions and Revelations? *Anſ.* They are to ſeek for it of God, by a perfect Mortification of their own Wills, and eſpecially as to theſe very ſuppoſed Favours; and by moſt inſtant and violent Prayer. Thus ſome who upon the Commiſſion of an evilthing have beendepriv'd of the Good Spirit of God, which they did enjoy, and have been deliver'd up to a Spitit of Deluſion for ſome time,

Ii 4 Satan

Qery. or are jSpIrits to '

.-be Dif ern'd.

F this Query *refers* . to the *Pe'rfons

thetnfelves' - who are or fu ofe the 'm- felves led by a PP

Good Spirit, it is Aalwer

' d, That the Manner of Difeeroaient is *by* Senfzticn, -which is not ***to*** be de= fcr bed. 'This Senfation like the *New*.. Song, the New Name, and *the White* Stone, they only who do' experiment,

ad while they do experinx nr, are 'able to dithngui . But then it may be Qiea` rjed, How thall they who fuppofe the ***n*** (elves- to be, Ied"by a True Spirit, and are ***not, be*** enabled to dificern aright,' in this

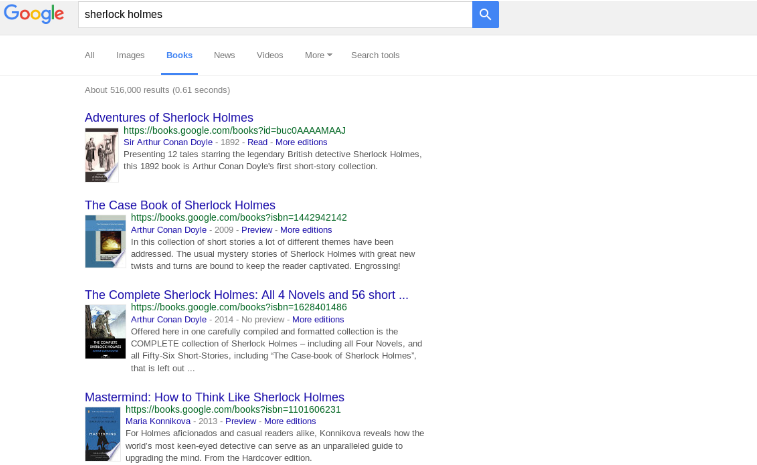
atter, finee they ***are*** without this Spi=r itual Senfation,. by which *the others* ace aflur'.d concerning the Truth of- their Vif ons and Revelations? *Anf. ' They* re to Peek fob' it of God,' bya perfect Mortification of their _ own Wills, and, :fpecially as to thefe very fuppofed Fa-: **Yours; and by mofl infant and violent Prayer. Thus fame *who* upon'the Com- miMon ofan evilthing.have beendepriv'd** of the Good Spirit of God, which they: did enjoy; and have been deliver* d up **to a Spitit of Delulion for fume time,**

Ii 4 Satars

In order for the information to be recognized, it is necessary to use optical character recognition (OCR), which looks at the photographic image, and attempts to recognize the original text.

For a sharp, clean printing of a modern book, OCR is very successful. However, OCR can become unreliable for books that are damaged, or printed too lightly or darkly, that use unusual fonts, or were printed hundreds of years ago.

A difficult book to scan results in a garbled OCR text which can only be fixed by human intervention. However, this would massively increase the cost of Google Books, and so in most cases, the OCR text is left as it is.



sherlock holmes

About 516,000 results (0.61 seconds)

Adventures of Sherlock Holmes
https://books.google.com/books?id=buc0AAAAMAAJ
Sir Arthur Conan Doyle - 1892 - Read - More editions
Presenting 12 tales starring the legendary British detective Sherlock Holmes, this 1892 book is Arthur Conan Doyle's first short-story collection.

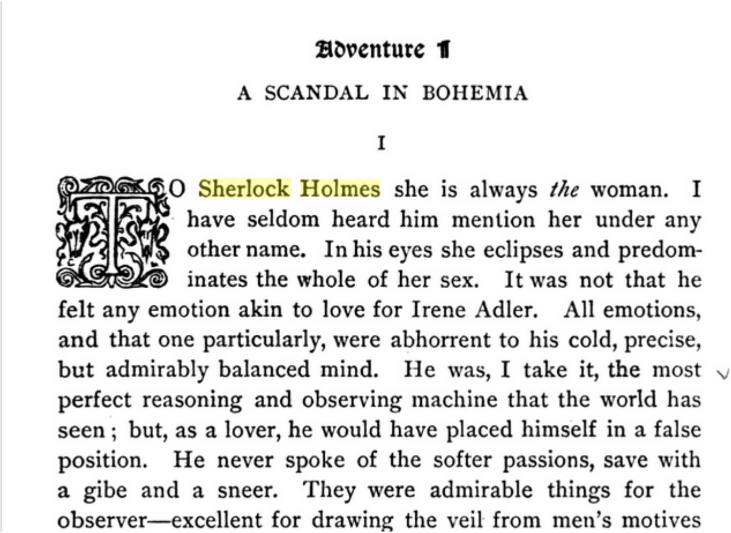
The Case Book of Sherlock Holmes
https://books.google.com/books?isbn=1442942142
Arthur Conan Doyle - 2009 - Preview - More editions
In this collection of short stories a lot of different themes have been addressed. The usual mystery stories of Sherlock Holmes with great new twists and turns are bound to keep the reader captivated. Engrossing!

The Complete Sherlock Holmes: All 4 Novels and 56 short ...
https://books.google.com/books?isbn=1628401486
Arthur Conan Doyle - 2014 - No preview - More editions
Offered here in one carefully compiled and formatted collection is the COMPLETE collection of Sherlock Holmes – including all Four Novels, and all Fifty-Six Short-Stories, including "The Case-book of Sherlock Holmes", that is left out ...

Mastermind: How to Think Like Sherlock Holmes
https://books.google.com/books?isbn=1101606231
Maria Konnikova - 2013 - Preview - More editions
For Holmes aficionados and casual readers alike, Konnikova reveals how the world's most keen-eyed detective can serve as an unparalleled guide to upgrading the mind. From the Hardcover edition.

If you go to Google Books and type in “Sherlock Holmes”, you will see a list of all the books in the database in which this name occurs. Just as with Google Search, an attempt is made to put the best matches first. One of the first items is “The Adventures of Sherlock Holmes”, with the address
https://books.google.com/books?id=buc0AAAAMAAJ.

You can select this book, and you will be able to “page through it”, with the additional feature that all the pages on which the words “Sherlock Holmes” appears will have a bookmark allowing you to quickly jump there.



Adventure I

A SCANDAL IN BOHEMIA

I

TO Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise, but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen; but, as a lover, he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer—excellent for drawing the veil from men's motives

Here is the actual first page of the Sherlock Holmes adventure titled A Scandal in Bohemia, after being scanned into Google Books.

You probably don't realize that this page, which looks very clean and readable, presents challenges to an OCR translator.

The heading **Adventure 1** is printed in an unusual font called "Blackletter".

Notice that the "T" in *To Sherlock Holmes she is always the woman* is displayed in a fancy style suggestive of an old handwritten manuscript.

There is a hyphenated word split between text lines 3 and 4. Will the OCR know what to do with this?

There is a pencil mark to the right of the word *most* in text line 7.

Note the phrase *observer–excellent* which you automatically realize represents two separate words, though the dash seems to join them.

On this single page we see many potential problems.

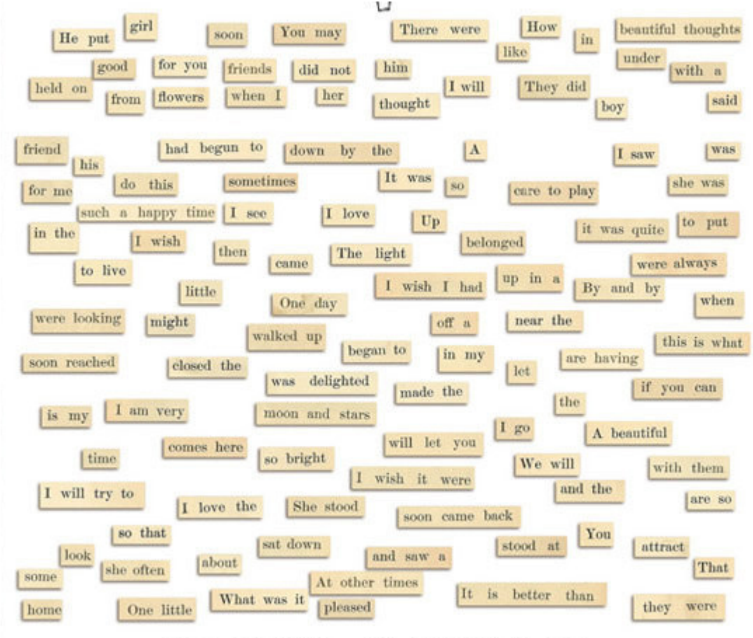
Adventure IA SCANDAL IN BOHEMIA

O Sherlock Holmes she is always the woman I have seldom Heard him mention her under any other name in His eyes she eclipses and predominates the whole of her sex It was not that I felt any emotion akin to love for Irene Adler All emotions and That One Particularly Were abhorrent to His cold required but admirably balanced mind I was I take it the most perfect reasoning and observing machine That the world has seen but as a lover I would have Placed himself in a false position I never spoke of the softer passions save with a gibe and a sneer They Were admirable things for the observer excellent for drawing the veil from men’s motives

It's only fair to say that this is actually a very good copy of the text. But it is by no means perfect, and there are a number of interesting and even serious problems:

- The two line heading has merged into a single line;
- The initial fancy T has disappeared;
- All the punctuation has disappeared;
- Many letters have been randomly capitalized;
- Several occurrences of he have become I;
- The word precise became required!

The word changes are particularly disturbing, since they change the meaning, and are harder to notice.


He put
girl
soon
You may
There were
How
in
beautiful thoughts
like
under
good
for you
friends
did not
him
with a
held on
from
flowers
when I
her
thought
I will
They did
boy
said
friend
had begun to
down by the
A
I saw
was
his
do this
sometimes
It was
so
care to play
she was
for me
such a happy time
I see
I love
Up
it was quite
to put
in the
I wish
then
belonged
The light
came
were always
to live
up in a
I wish I had
little
By and by
when
One day
were looking
might
off a
near the
walked up
this is what
began to
in my
soon reached
closed the
are having
let
was delighted
made the
if you can
the
is my
I am very
moon and stars
I go
A beautiful
will let you
time
comes here
so bright
We will
with them
I wish it were
and the
I will try to
are so
I love the
She stood
soon came back
You
so that
sat down
stood at
attract
look
about
and saw a
That
she often
some
At other times
It is better than
What was it
home
One little
pleased
they were

The raw database of Google Books is not suitable for computer analysis, because it consists of a sequence of photographs of book pages, that is, they are really pictures, not texts!

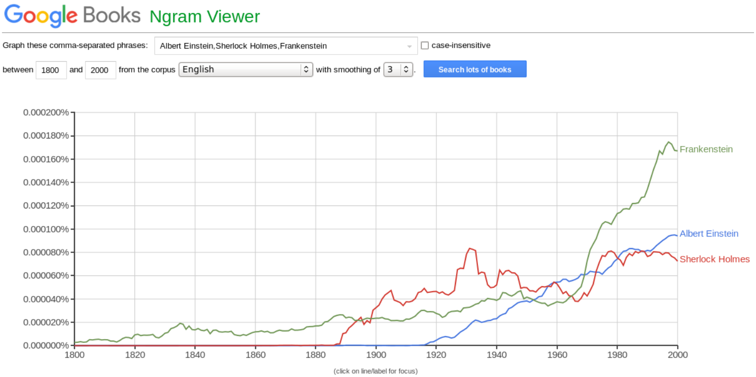
Google wanted to make the data computer accessible by applying OCR.

For copyright reasons, it was not possible to create OCR versions of the full text of the books.

Instead, they took each book and recorded all the individual words, pairs of words, triples of consecutive words and so on, up to strings of 5 words.

You can think of this as cutting a book up into thousands of scraps of paper. Each scrap contains a few words, and is labeled by the title of the book, language, and date of publication.

By cutting books up in this way, we've lost a lot of information. Can we do anything interesting with what is left?



Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Cutting a book up into single words creates so-called “one-grams”, “unigrams”, or “monograms”. Cutting it up into pairs of words creates bigrams, and similarly we have trigrams, quadgrams and quintgrams. That is as far as the cutting goes.

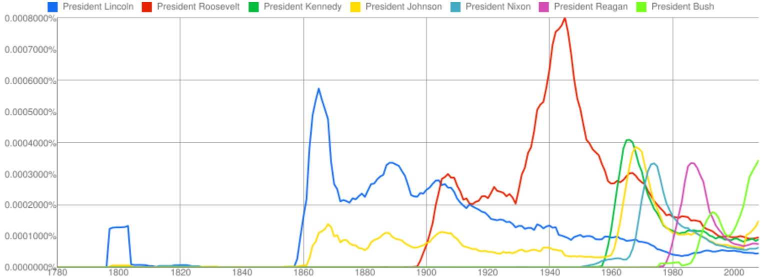
The Google Ngram viewer is at http://ngrams.googlelabs.com

When you start it, you see the results of a sample query, namely Albert Einstein, Sherlock Holmes, Frankenstein.

Each comma-separated item results in a line shown on a common plot. The horizontal axis records the years, and the vertical axis the relative frequency.

It’s important to keep an eye on that vertical axis, since its scale will be different for every plot.

The Ngram viewer requires that a word (ngram) show up at least 40 times in order to be searchable.


President Lincoln
President Roosevelt
President Kennedy
President Johnson
President Nixon
President Reagan
President Bush
0.0008000%
0.0007000%
0.0006000%
0.0005000%
0.0004000%
0.0003000%
0.0002000%
0.0001000%
0.0000000%
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000

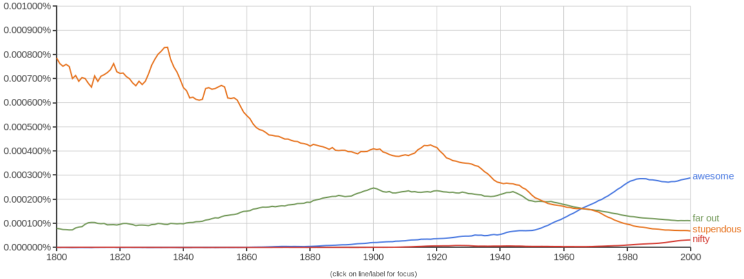
We can search on presidents Lincoln, Roosevelt, Kennedy, ...

We include the word "president" in each search string, otherwise we would find references to many other people with the same name.

Notice the double humps for presidents Roosevelt and Bush.

Notice the false information at the year 1800. Lincoln wasn't even born until 1808, and what we see here is the result of books about Lincoln for which an incorrect publication date was entered.

So we not only have OCR errors to worry about, we also have mistakes made by the catalogers!


0.001000%
0.000900%
0.000800%
0.000700%
0.000600%
0.000500%
0.000400%
0.000300%
0.000200%
0.000100%
0.000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
awesome
far out
stupendous
nifty
(click on line/label for focus)

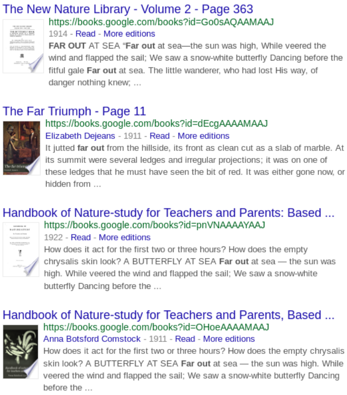
When did people stop saying “far out” and start saying “awesome”?

We can type in a few such words and try to judge what is happening.

It certainly seems like ”nifty” and ”awesome” are newcomers compared to ”stupendous” and ”far out”. It seems as though ”stupendous” is rapidly dropping, while ”nifty” and ”awesome” are rising and ”far out” is holding more or less steady.

Search in Google Books:

1800 - 1929	1930 - 1987	1988 - 1992	1993 - 1997	1998 - 2000	awesome	English
1800 - 1914	1915 - 1990	1991 - 1994	1995 - 1997	1998 - 2000	nifty	English
1800 - 1834	1835 - 1910	1911 - 1924	1925 - 1975	1976 - 2000	far out	English
1800 - 1810	1811 - 1821	1822 - 1858	1859 - 1939	1940 - 2000	stupendous	English



The New Nature Library - Volume 2 - Page 363
https://books.google.com/books?id=Go0sAQAAMAAJ
1914 - Read - More editions
FAR OUT AT SEA "**Far out** at sea—the sun was high, While veered the wind and flapped the sail; We saw a snow-white butterfly Dancing before the fitful gale **Far out** at sea. The little wanderer, who had lost His way, of danger nothing knew; ...

The Far Triumph - Page 11
https://books.google.com/books?id=dEcgAAAAMAAJ
Elizabeth Dejeans - 1911 - Read - More editions
It jutted **far out** from the hillside, its front as clean cut as a slab of marble. At its summit were several ledges and irregular projections; it was on one of these ledges that he must have seen the bit of red. It was either gone now, or hidden from ...

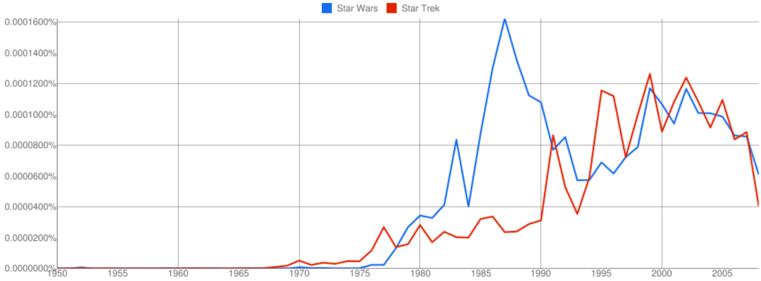
Handbook of Nature-study for Teachers and Parents: Based ...
https://books.google.com/books?id=pnVNAAAAYAAJ
1922 - Read - More editions
How does it act for the first two or three hours? How does the empty chrysalis skin look? A BUTTERFLY AT SEA **Far out** at sea — the sun was high. While veered the wind and flapped the sail; We saw a snow-white butterfly Dancing before the ...

Handbook of Nature-study for Teachers and Parents, Based ...
https://books.google.com/books?id=OHoeAAAAMAAJ
Anna Botsford Comstock - 1911 - Read - More editions
How does it act for the first two or three hours? How does the empty chrysalis skin look? A BUTTERFLY AT SEA **Far out** at sea — the sun was high. While veered the wind and flapped the sail; We saw a snow-white butterfly Dancing before the ...

We should not be too sure about the results for “far out”.

After all, this phrase has another meaning, simply “not nearby”.

The Ngram Viewer allows us, for any of the search words, to examine the books in which the search word occurred. Here is part of what we see if we ask to look at books from 1911-1925 using “far out”.


🟦 Star Wars
🟥 Star Trek
0.0001600%
0.0001400%
0.0001200%
0.0001000%
0.0000800%
0.0000600%
0.0000400%
0.0000200%
0.0000000%
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005

We can ask Ngram Viewer to give us the history of the usage of the terms “Star Wars” and “Star Trek”.

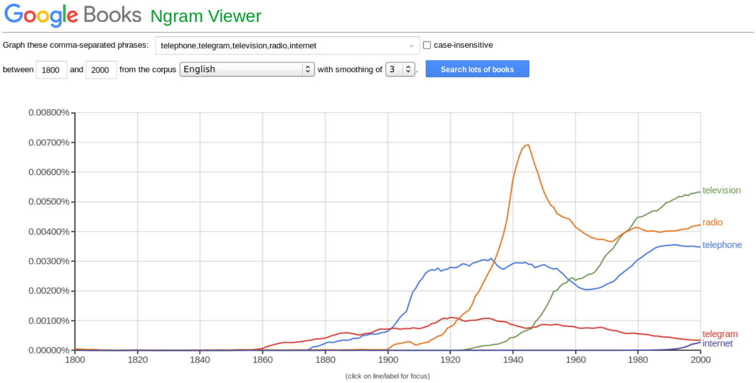
We expect to see both terms set to zero until quite recently.

It’s interesting to look at who’s “winning” from year to year.

Also, Star Trek has peaks around 1978, 1980, 1985, 1991, 1995, 1999, 2002, 2004. Star Trek movies were released in 1979, 1982, 1984, 1986, 1989, 1991, 1994, 1996, 1998, 2002...

Star Wars shows peaks at around 1980, 1983, 1987, 1999 and 2002. Movies were released in 1977, 1980, 1983, 1999, and 2002... Notice that the Ngram Viewer seems to have completely ignored the first movie!

However, note that if there are less than 40 mentions of a topic in a given year, this is actually rounded down to zero.



Google Books Ngram Viewer

Graph these comma-separated phrases: telephone,telegram,television,radio,internet

□ case-insensitive

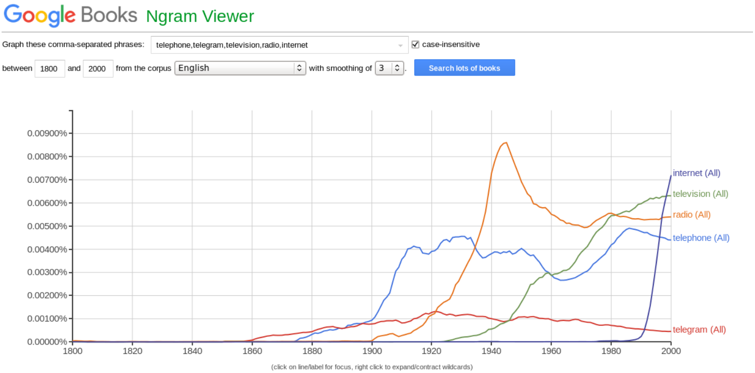
between 1800 and 2000 from the corpus English with smoothing of 3.

We can ask Ngram Viewer to compare the relative usage of the words “telegram”, “telephone”, “television”, “radio” and “internet.”

Some things make sense:

- telegram pops up first, and decreases over time
- telephone is next
- radio zooms up into the 1940’s and then drops to a plateau
- television shows up early, in the 1920’s, and goes up and up

but what’s going on with the internet? Why are its numbers so poor?



Google Books Ngram Viewer

Graph these comma-separated phrases: telephone,telegram,television,radio,internet ☑ case-insensitive

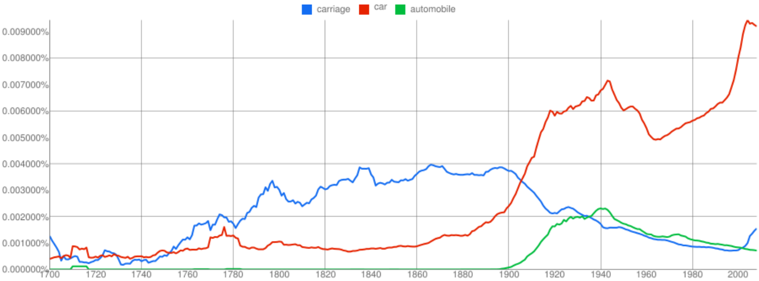
between 1800 and 2000 from the corpus English with smoothing of 3.

(click on line/label for focus, right click to expand/contract wildcards)

Ngram Viewer is case sensitive unless we tell it not to be.

That means that if we search for **internet** then **Internet** will not count.

Simply by checking the case-insensitive box, we can repeat our search, but now look at the result. Apparently, “radio” is rarely spelled “Radio”, but almost all mentions of the internet spell it as “Internet”!


carriage
car
automobile
0.009000%
0.008000%
0.007000%
0.006000%
0.005000%
0.004000%
0.003000%
0.002000%
0.001000%
0.000000%
1700
1720
1740
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000

How about the words “carriage”, “automobile” and “car”?

We might suspect that “carriage” would disappear around 1900, just as “automobile” and “car” suddenly pop up.

The actual graph is more complicated. “carriage” doesn’t want to die out...but then again, “carriage” has other meanings besides something a horse pulls.

But did people in the 1800’s actually have cars? I don’t think so! Did they call a carriage a car back then? Seems unlikely! What is going on?

Specification of a patent for a Pendulous Rail-road Car
Isabel Trevithoe, a poem by C.A.R., 1879
A Key to the Classical Pronunciation of Greek, 1830,
including Car-nus, Car-nu'tes, Car-pa'si-a, ...
CAR, verb, to cover the cop
the case of Habeas Corpus in the 3d of Car. 1
(the third year of the reign of King Charles I)
Carcase (car-case), n. dead body, body,
[Enter Caratach] Car: ‘‘Now, what's the matter?’’
instructions for making a flying car, in which a man may sit
Rahla turned the car on and started to back out of the
parking place *(1881? No way! Mistaken date!)*

Unfortunately, when we look at the data, we see some very peculiar things!

First of all, it's true that a train consists of railroad cars, so there were things called cars back in the 1800's.

But a poem, written by a person with initials C.A.R., is listed as a hit.

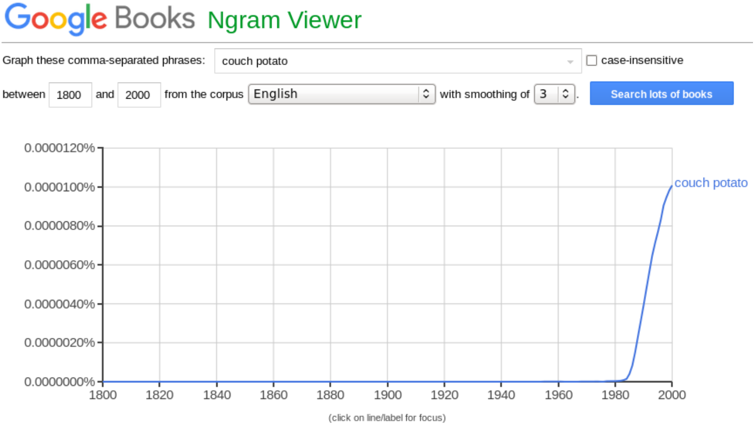
Also, a list of Greek words, hyphenated, produces a string of pseudo "car" hits.

An OCR mistake reads "CAP" as "CAR"

English laws are dated by the king's name, in Latin, abbreviated, so Charles I is Car. I

In plays, a character's name is abbreviated, so Caratach become "Car:".

And worst of all, a book published in 1981 was mistakenly listed as 1881, and so we have people driving around in a car back then!



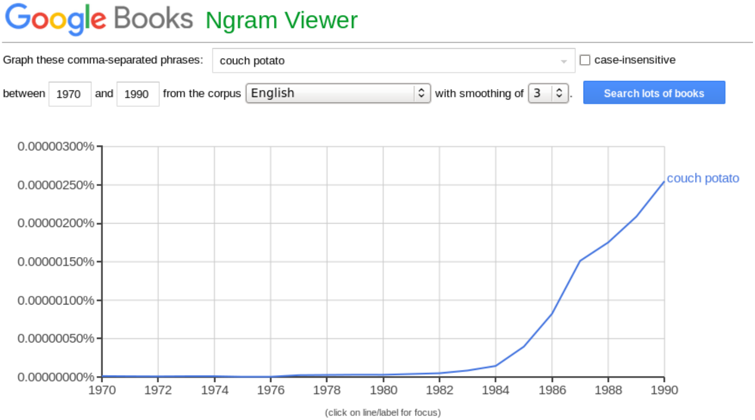
Graph these comma-separated phrases: couch potato □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

We are all familiar with the phrase “couch potato”, but there was a time when no one had heard that phrase.

It had to have an origin. Can we estimate when this happened?

We can use Ngram viewer to search for the origin of the phrase “couch potato”, or at least to search all the published literature in its database.



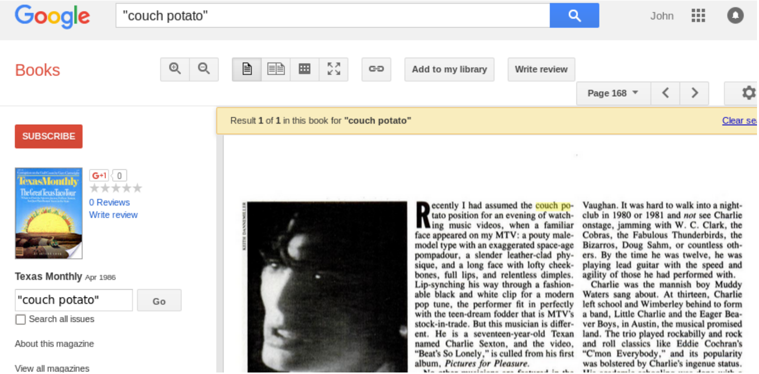
Graph these comma-separated phrases: couch potato

☐ case-insensitive

between 1970 and 1990 from the corpus English with smoothing of 3.

Our first search showed a dramatic growth in usage, but this can make it hard to see the origin, since the usage would typically be very low early on.

By narrowing the range of years, we can get a better clue as to when the phrase was starting to become popular.



Result **1** of **1** in this book for **"couch potato"**

Recently I had assumed the couch potato position for an evening of watching music videos, when a familiar face appeared on my MTV: a pouty male-model type with an exaggerated space-age pompadour, a slender leather-clad physique, and a long face with lofty cheekbones, full lips, and relentless dimples. Lip-synching his way through a fashionable black and white clip for a modern pop tune, the performer fit in perfectly with the teen-dream fodder that is MTV's stock-in-trade. But this musician is different. He is a seventeen-year-old Texan named Charlie Sexton, and the video, "Beat's So Lonely," is culled from his first album, *Pictures for Pleasure*.

Vaughan. It was hard to walk into a nightclub in 1980 or 1981 and *not* see Charlie onstage, jamming with W. C. Clark, the Cobras, the Fabulous Thunderbirds, the Bizarros, Doug Sahm, or countless others. By the time he was twelve, he was playing lead guitar with the speed and agility of those he had performed with.

Charlie was the mannish boy Muddy Waters sang about. At thirteen, Charlie left school and Wimberley behind to form a band, Little Charlie and the Eager Beaver Boys, in Austin, the musical promised land. The trio played rockabilly and rock and roll classics like Eddie Cochran's "C'mon Everybody," and its popularity was bolstered by Charlie's ingenue status.

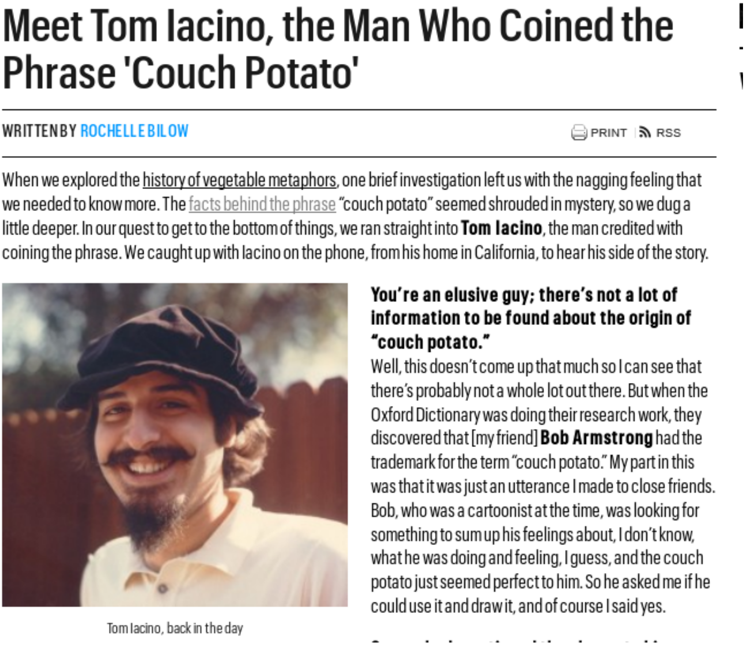
When we go into the book search, the closest thing we can find is an article in Texas Monthly, for April 1986.

Reading the text gives us some assurance that the phrase is being used in the way we expect.

Even though Ngram Viewer seems to show usage before 1980, the 1986 reference is the earliest I could spot in the database.

Because this is a magazine, we have confidence that the date is given correctly.

If it was a book, we would want to go to the first couple pages and look at the copyright page, because Google Books is full of incorrect date information!



Meet Tom Iacino, the Man Who Coined the Phrase 'Couch Potato'

WRITTEN BY ROCHELLE BILOW

PRINT | RSS

When we explored the history of vegetable metaphors, one brief investigation left us with the nagging feeling that we needed to know more. The facts behind the phrase "couch potato" seemed shrouded in mystery, so we dug a little deeper. In our quest to get to the bottom of things, we ran straight into **Tom Iacino**, the man credited with coining the phrase. We caught up with Iacino on the phone, from his home in California, to hear his side of the story.

Tom Iacino, back in the day

You're an elusive guy; there's not a lot of information to be found about the origin of "couch potato."

Well, this doesn't come up that much so I can see that there's probably not a whole lot out there. But when the Oxford Dictionary was doing their research work, they discovered that [my friend] **Bob Armstrong** had the trademark for the term "couch potato." My part in this was that it was just an utterance I made to close friends. Bob, who was a cartoonist at the time, was looking for something to sum up his feelings about, I don't know, what he was doing and feeling, I guess, and the couch potato just seemed perfect to him. So he asked me if he could use it and draw it, and of course I said yes.

Of course, another way to search is to just use a browser, and in this case, it turns out we can find some information about a man who claims to have invented the phrase back in the 1970s.

So Internet browsers can be a quicker way to some kind of facts...on the other had, we never would have been able to trace the growth in popularity of the phrase, or all its occurrences, without using Ngram Viewer!