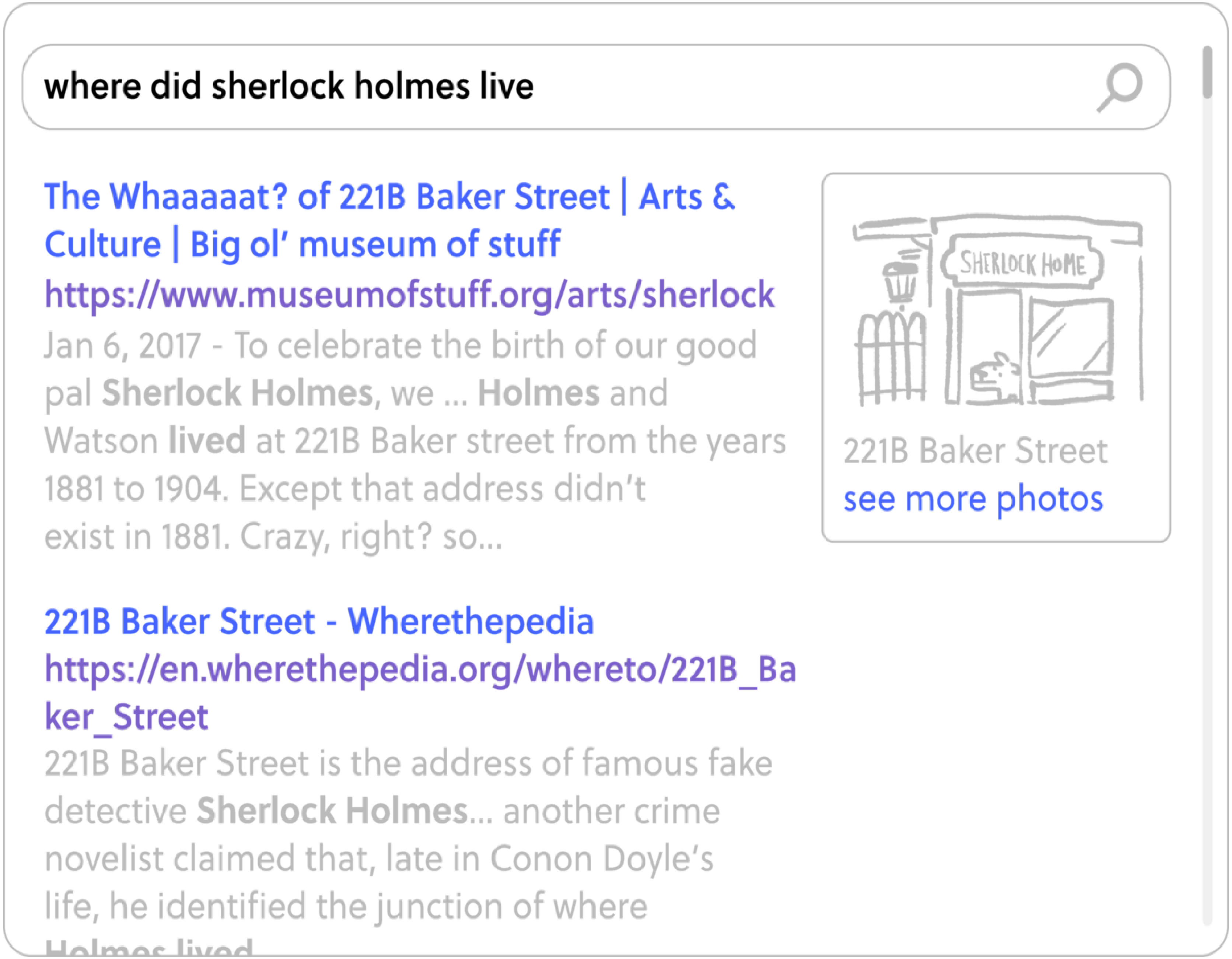
Searching The Whole World

Powered by @Brilliant App

Lecture prepared and delivered by

Dr Syed Khaldoon Khurshid



where did sherlock holmes live

The Whaaaaat? of 221B Baker Street | Arts & Culture | Big ol' museum of stuff
https://www.museumofstuff.org/arts/sherlock
Jan 6, 2017 - To celebrate the birth of our good pal **Sherlock Holmes**, we ... **Holmes** and Watson **lived** at 221B Baker street from the years 1881 to 1904. Except that address didn't exist in 1881. Crazy, right? so...

221B Baker Street
see more photos

221B Baker Street - Wherethepedia
https://en.wherethepedia.org/whereto/221B_Baker_Street
221B Baker Street is the address of famous fake detective **Sherlock Holmes**... another crime novelist claimed that, late in Conon Doyle's life, he identified the junction of where **Holmes lived**

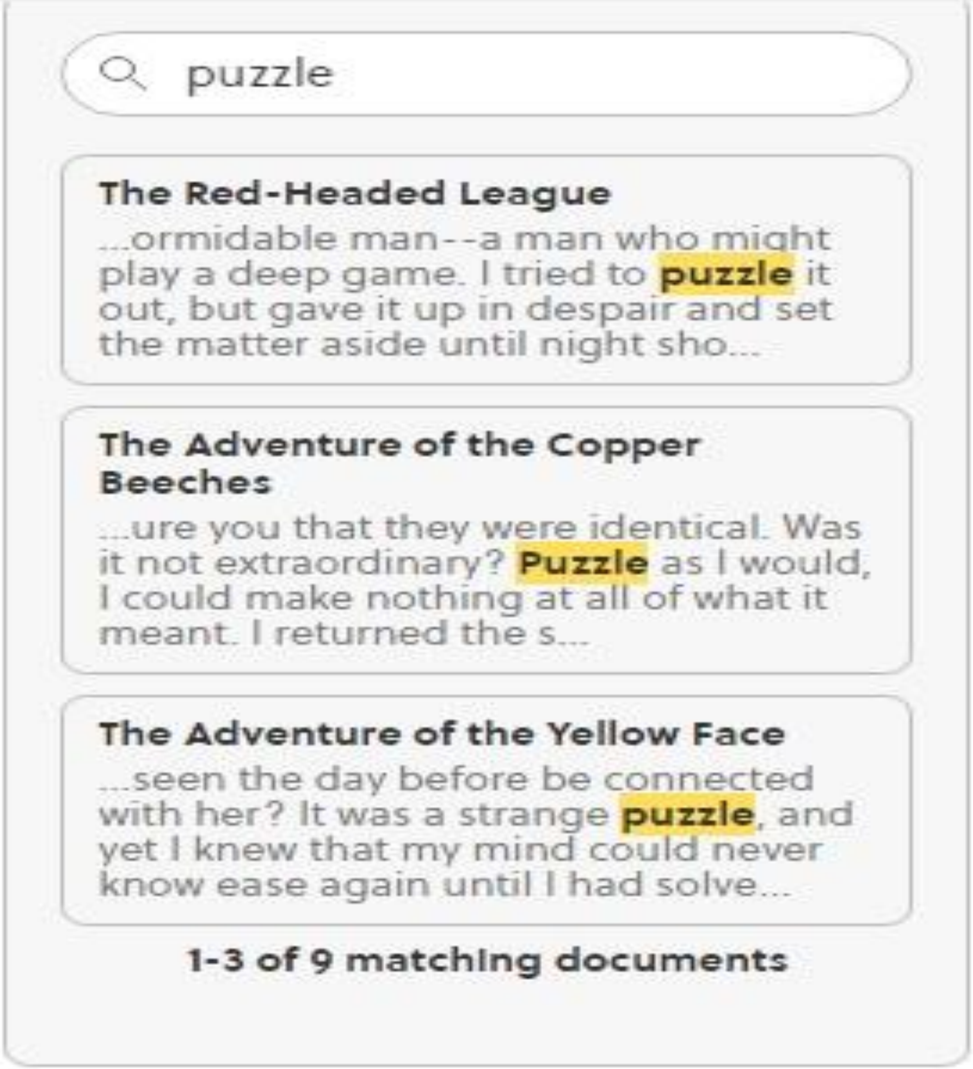
- When we type words into a search bar, computers in many different places spring to work in the blink of an eye. **Search engines give us a level of access to information that would have been breathtaking to humans a generation ago.**

- It’s hard to develop intuition about how search works. **Computers are very fast, but the amount of information collected on the internet is very big.**
- Should looking for all occurrences of the word "**puzzle**" in ten years of your email history be fast because computers are fast, or is it slow because ten years of email history is an awful lot of information to sort through?

- To find all the instances of a word in a document, **the simplest strategy is to read the document from beginning to end.**
- This would be tedious and slow for a human. But when books are **stored as text on a computer**, the approach can work.
- A computer can read through documents the same way a human does, letter by letter, word by word, from the beginning to the end. **Unlike a human, a computer can do this task with blazing speed and essentially perfect accuracy.**

Here’s a search engine that lets you query words from the Sherlock Holmes stories by Arthur Conan Doyle.

Typed "puzzle" into the search bar below:



puzzle

The Red-Headed League

...ormidable man--a man who might play a deep game. I tried to **puzzle** it out, but gave it up in despair and set the matter aside until night sho...

The Adventure of the Copper Beeches

...ure you that they were identical. Was it not extraordinary? **Puzzle** as I would, I could make nothing at all of what it meant. I returned the s...

The Adventure of the Yellow Face

...seen the day before be connected with her? It was a strange **puzzle**, and yet I knew that my mind could never know ease again until I had solve...

1-3 of 9 matching documents

Based on the results, does Arthur Conan Doyle use "puzzle" as a verb, a noun, or both?

Correct answer: **Both as a noun and a verb**

Explanation

- When you search for **"puzzle"**, only the first three of 9 results are shown.
- Nevertheless, you can see that “**puzzle**” is used both as a noun (as in The Adventure of the Yellow Face: “It was a strange puzzle”) and as a verb (as in The Red-Headed League: “I tried to puzzle it out”).

- A reasonably designed computer program on a personal computer takes about a **second** to "read" any one of these documents to find all instances of a word or phrase:
  - the complete works of Shakespeare
  - Leo Tolstoy’s epic novel War and Peace
  - the Bible
  - the collected stories of Sherlock Holmes.
- **How long should you expect it to take a computer to read *all four* of those documents from beginning to end to count all instances of a word or phrase?**

Correct answer: **About four seconds**

Explanation

- This wasn't a trick question! If a computer needs one second to work through each document, it will need four seconds to work through four documents.
- **If you were expecting a trick question, you may have answered "About a second," because you imagined that a computer with multiple processors (or "cores") could use parallelism to search all four documents simultaneously.**

There are two reasons why that's not the right answer:

- A computer with several processors could have split each individual document into several pieces, one piece per processor, and searched different parts of the document simultaneously. So the estimate of a second per document might already account for using all available processors.
- In reality, due to the way that computer memory works, the process of reading a document from beginning to end to search for a word or phrase is not limited by the speed of the processor. Instead, it's limited by how fast information stored on the hard drive can be moved around inside of the computer to a place where the processor can use it.

- Computers can read the Bible from beginning to end faster than a human can turn a page. When we say the computer is “reading,” **we mean that it is doing something similar to what a human does: processing the text letter by letter, word by word, from front to back.**
- So unlike humans, computers can quickly perform search by reading straight through — the entire Bible takes about a second.

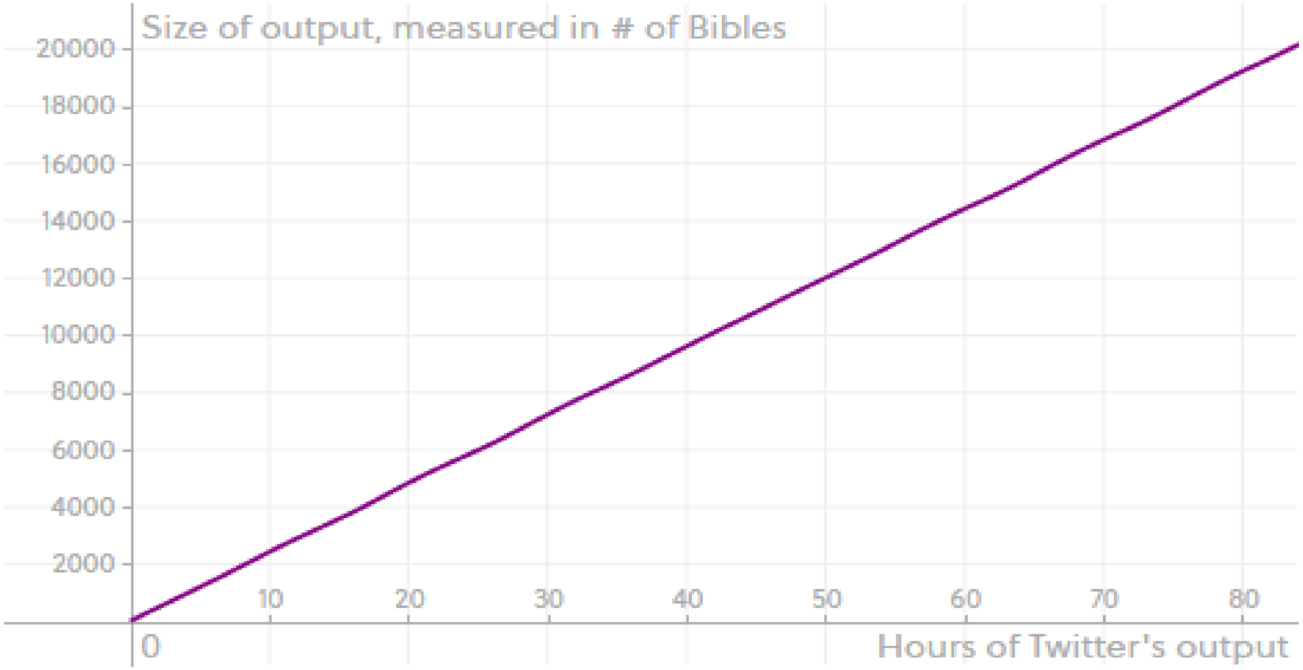
- According to Wikipedia, there are at least 450 different English translations of the Bible.
- How long would it take for a computer to "read" all 450 translations in order to find out how many times each one uses a certain word?

Correct answer: **About eight minutes**

Explanation

- Since the computer reads one document per second, it takes 450 seconds to read all the translations from beginning to end. That's 7.5 minutes, or about eight minutes.

- You can't just rely on the blinding speed of computers when there is lots of text to search through. It takes about ***15 seconds*** for the users of Twitter to produce text equal in length to the Bible.



- If a computer can read through a book the size of the Bible in a second, how many hours of Twitter output could that single computer read through in an hour of computing?

Correct answer: **About 15 hours of Twitter's output**

Explanation

- There are two ways to solve this question. One is to observe that the computer can read text 15 times faster than Twitter can generate it, so with one hour the computer can read 15 hours of Twitter's output.
- The other option is to calculate the number of Bible-length texts that the computer can process an hour. This is 60·60=3600 Bible-length texts.
- Looking at where the graph above reaches 3600 Bible-length texts, this is closest to 15 hours of Twitter's output.

- The output from Twitter is a huge, but comprehensible, amount of information. It still takes weeks or months of Twitter’s output to match the amount of text contained in the Library of Congress's 39 million physical books.
- How long would you expect it to take for a single computer to search once through the entire Library of Congress, if the whole Library of Congress is about a month of tweets?

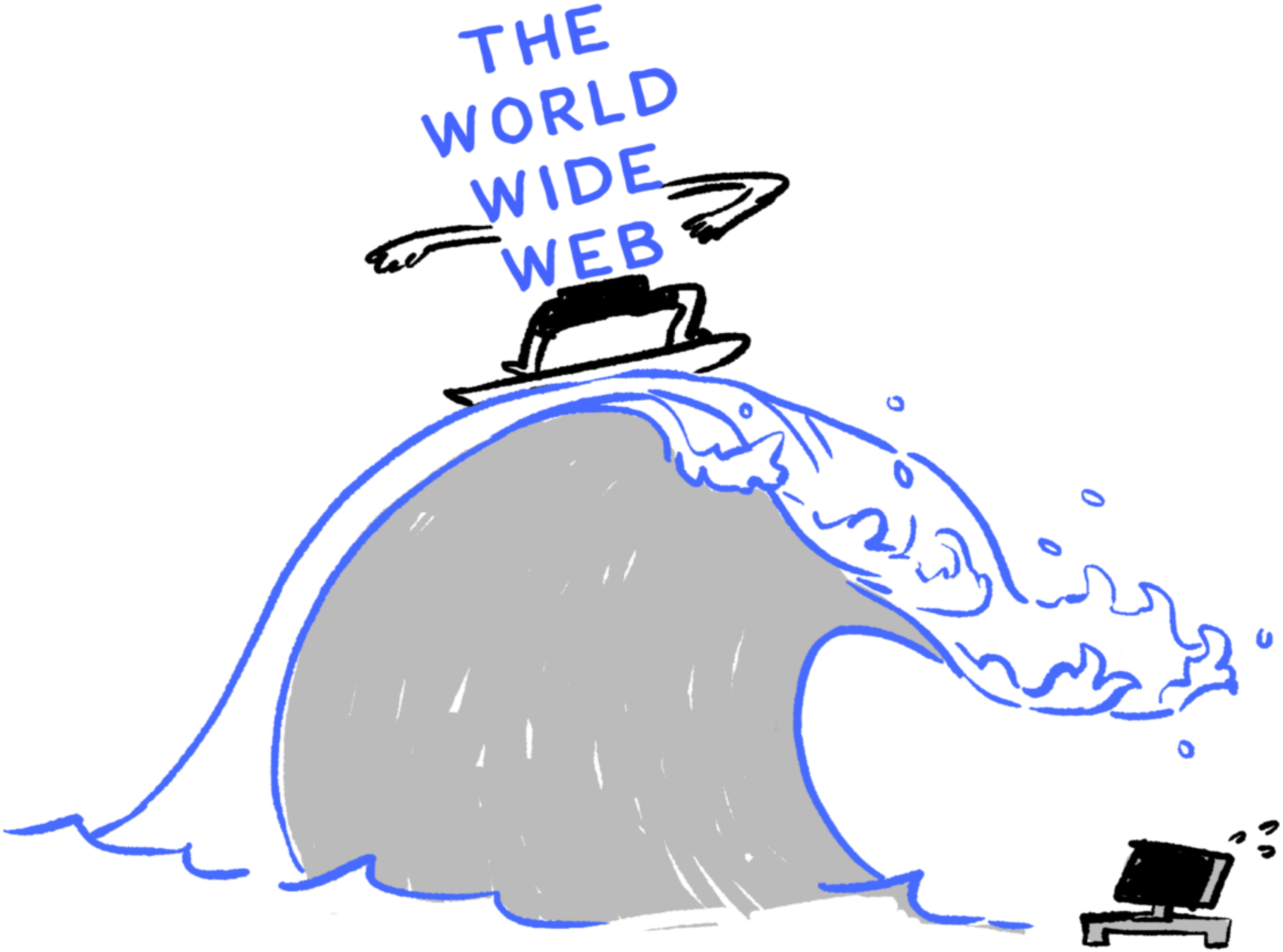
Correct answer: Hours or days

Explanation

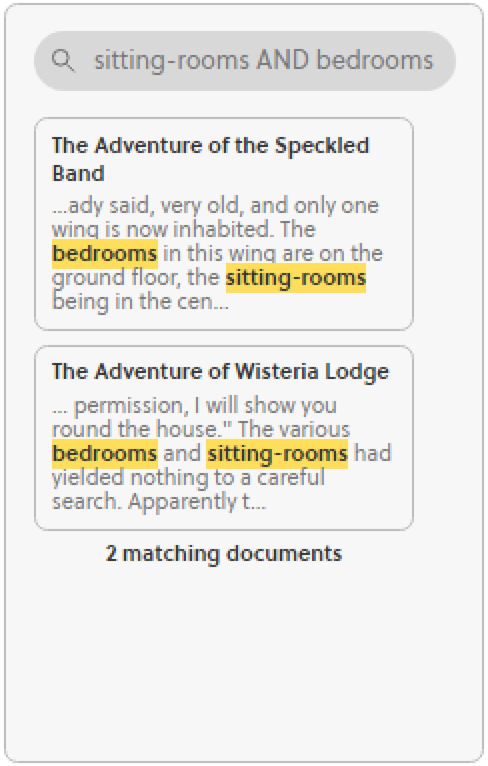
- There's more than one way to estimate the answer.
- In the previous question, you saw that a computer could read half a day of tweets in an hour. So the answer is more than "hours" but less than "months."
- If you remember that the assumed computer program can process tweets 15 times faster than Twitter can generate them, then a month of tweets — approximately 30 days — is about two days of computation time.

- The simple **read-from-front-to-back** approach will stop working at the scale of large libraries. A few days isn't an extraordinary amount of time, but you probably don't want to run a computer for several days just to figure out where a particular word appears in the Library of Congress.
- We expect search engines to respond in **fractions of a second**, even when faced with the fire hose of text they have to deal with.

- Even if you have multiple modern computers searching multiple parts of the text, you’d need a ridiculously large number of computers to achieve search-engine-like speeds.
- The simple approach to reading a book —letter by letter and word by word — lets a computer almost instantly search a book. That approach just won't work anymore when you want to search the entire web.
- **A large search engine like Google must keep track of hundreds of trillions of web pages and respond to millions of search queries every hour.**


THE
WORLD
WIDE
WEB

- **Modern search engines and our Sherlock search engine are built on a fundamentally different approach.**
- This foundation also allows you to search in more powerful ways, like searching for exact phrases or using **AND** and **OR** in queries. In coming lectures, you'll build up the skills you need to build a search engine by solving puzzles with Sherlock Holmes.



sitting-rooms AND bedrooms

The Adventure of the Speckled Band

...ady said, very old, and only one wing is now inhabited. The **bedrooms** in this wing are on the ground floor, the **sitting-rooms** being in the cen...

The Adventure of Wisteria Lodge

... permission, I will show you round the house." The various **bedrooms** and **sitting-rooms** had yielded nothing to a careful search. Apparently t...

2 matching documents

- **The ideas behind modern search engines are not new. In the next exploration, you’ll see that the basis of modern search has been around for hundreds — even thousands — of years.**

Any
Questions