Four Thousand Years Of Search Engines

Powered by @Brilliant App

Lecture prepared and delivered by

Dr Syed Khaldoon Khurshid

- As long as humans have been writing things down, humans have needed help r**emembering where things were written down**.
- In modern-day Iraq, archeologists have found a four-thousand-year-old tablet, the most ancient library catalog that we know about. The tablet is a listing of 62 other documents. Presumably, those other documents were all on other tablets, and **the catalog helped potential readers quickly see what documents were available.**
- Even ancient librarians knew that searching is difficult, **but the same techniques employed by these ancient tablet catalogs still find use in modern search engines.** All it takes is a few Sumerian insights to get around the intimidating amount of time it takes to search all of Twitter...

- Whether carved on tablets, inscribed in scrolls, or bound in books, humans continued using book listings to organize libraries for thousands of years. Thomas Jefferson's sale of 6,487 books to the United States government in 1815, which kickstarted the Library of Congress, **came with a catalog, a bound volume that organized the other books by topic**.
- It was around the same time that large libraries around the world started experimenting with using card catalogs to **replace their bound catalogs**. Information about each book in a library was **written on a separate index card** — the French used the backs of playing cards!

- The cards were then kept in some order: for example, in alphabetical order by the author's name.
- **What is an advantage of using a card catalog instead of a bound library catalog?**

Correct answer: **Card catalogs are easy to revise when books are added or removed**

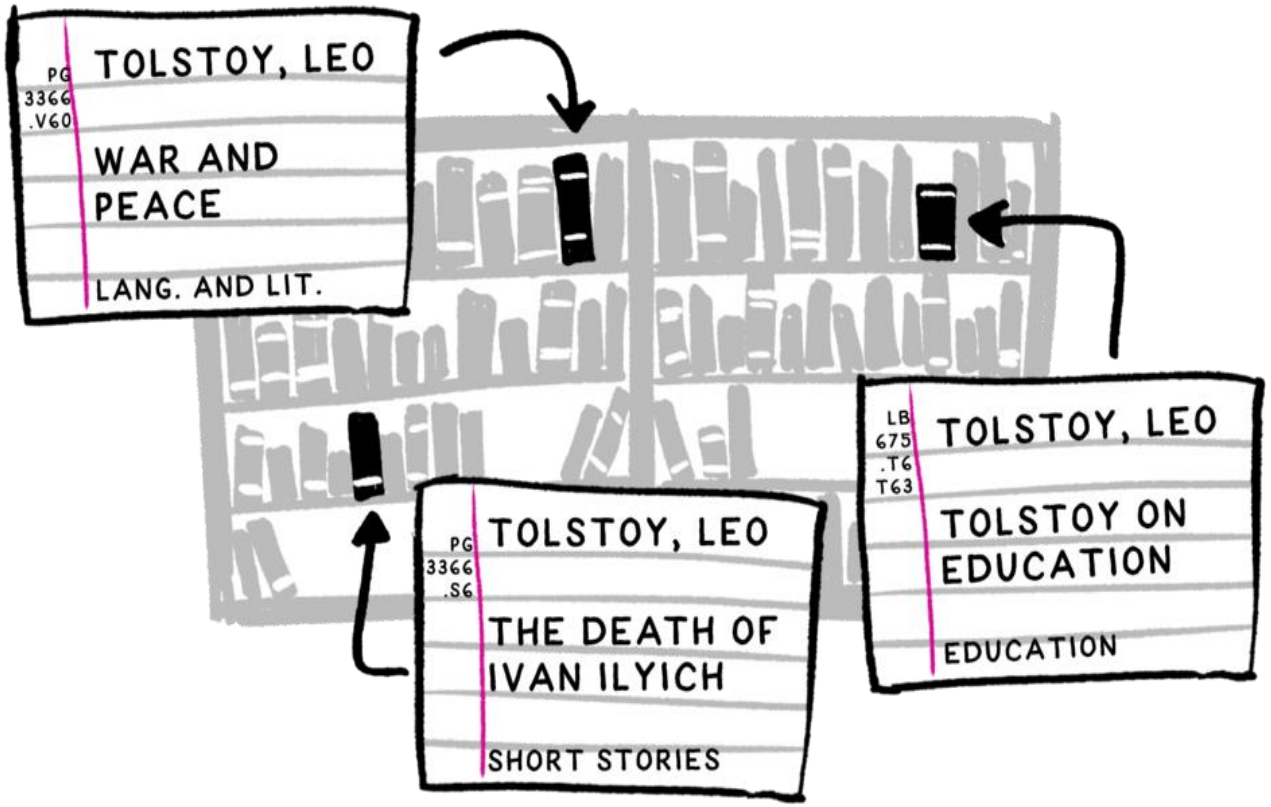
Explanation

- While harder to move from one place to another, **card catalogs have the advantage of being easy to modify: you can add a new card to both the library and the catalog without worrying about whether you'll have space to add the new record to the bound volume**. This was a real problem for libraries in Jefferson's time!

- The Jefferson purchase instantly broadened the scope of the collection... Throughout the first half of the nineteenth century, the Library [of Congress] continued to grow in both stature and reputation. However, the catalog woefully failed to keep up. The **Library produced a new bound catalog every ten years**, which was doomed to obsolescence as soon as it was printed.

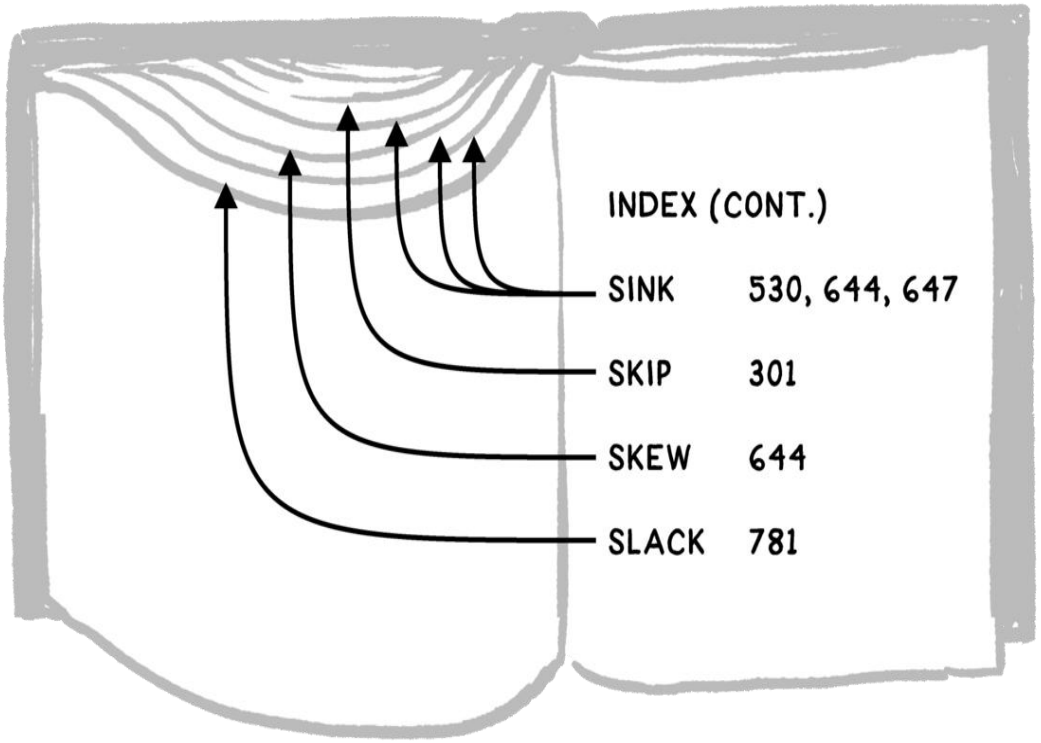
  – Ref: Excerpt from *The card catalog: books, cards, and literary treasures*, by the Library of Congress.

- **Card catalogs** take work to build and maintain, but they make searching for the right book in a library much, **much easier**, at least if you happen to want to search by the same characteristic that was used to organize the cards.



- However, card catalogs **take a really big shortcut:** they limit the ways you can search for information. **If you want to find all the books by a particular author, they are great**. If you want to find all the books published in “South Asia”, or all the books that contain the word "puppy," a card catalog is unlikely to be set up in a way that can help you.

- **The *index* of a book has a similar role to a card catalog. An index is for helping you search for a page or passage within a book, rather than finding a book within a larger collection of books.**



- A book's index **only contains topics the publisher or author found to be worth including**. The index of a book on calculus is likely to contain the word 'integral', whereas the index of a book on world history will not contain it, *even though* the word **may occur** in the text: "An integral part of understanding history...".

- Imagine that you have a copy of *The Complete Annotated Sherlock Holmes*, and that this book has an index.
- You think that Conan Doyle wrote the phrase "You can't enjoy a puzzle unless you know its rules," but the word "puzzle" doesn't appear in the book's index.

What can you conclude?

- Just because a word like "puzzle" doesn't appear in a book's index, that doesn't mean the word didn't appear in the book!
- A word like "puzzle" might appear in a book's index, but that's up to the author or publisher.
- Generally speaking, a book's index will only reliably contain proper nouns like people's names.

- **A book's index is limited to the words that someone else declared were the most important words.** One superpower of modern search engines is that they allow you to search for *any* word that appears *anywhere* in a document.
- There was a piece of pre-internet technology that served a similar purpose: a ***concordance***. A concordance was like a "search-engine-in-a-book" for a single document or work.
- A concordance takes *every single word* that appears anywhere in a document, puts those words in alphabetical order, and lists every place in the document where those words appear.

Puttock. an owl, a puttock,
or a herring without a roe. Trois. and Cres. v
I chose an eagle, And did
avoid a puttock Cymbeline i
Puzzle. Your prescence needs
must puzzle Antony Ant. and Cleo iii
Puzzled. More puzzled than
the Egyptians in their fog. T. Night iv
Puzzles the will And makes us
rather bear those ills we have Hamlet iii
Pygmalion. Is there none of

- The complete Shakespeare concordance pictured above shows that the word "puzzle" appears in the play "Antony and Cleopatra" (Act III) and that "puzzles" appears in Hamlet (Act III).
- Does the Shakespeare play "Romeo and Juliet" contain the word "puzzles"?

Correct answer: **"Romeo and Juliet" definitely does not contain the word "puzzles."**

Explanation

- Because a concordance exhaustively lists every word along with each of its appearances, you can use it to decide that the exact word "puzzle" only appears once in the entire works of Shakespeare, in "Antony and Cleopatra." It does not appear in "Romeo and Juliet."

- To build a concordance without a computer, scholars needed to take *every single word* that appeared anywhere in the document, put those words in alphabetical order, and list every place in the document where those words appeared, along with their context.
- It was only worthwhile to make a concordance when people *really, really cared* about answering this question:
- "Where did I see that word before?"
- Unsurprisingly, this was only done for works of particular importance like religious texts. Hundreds of monks worked together to create the first concordance of the Bible in the thirteenth century.

- Concordances require an astounding **upfront investment to create**.
- Once a concordance exists, it acts like a simple search engine. A concordance only **lets you search for a single word at a time, and only within a single document**: the Bible, the works of Shakespeare, or the collected stories of Sherlock Holmes. But within those limitations, **it allows scholars to answer questions about word usage in seconds that would otherwise have taken weeks or months**.

- A computer can search Shakespeare by reading the stories, front to back, in about a second.
- But as we move to more internet-sized chunks of information, this approach falls flat on its face — Google is not able to handle the fire hose of internet content by reading it straight through.
- Much like the monks and their concordance, Google keeps structured data that keeps track of where certain words, phrases, or concepts are mentioned.
- As you progress through this course, you'll see that modern search engines are based on ideas that are similar to the catalogs, indexes, and concordances in this exploration.


Any
Questions