Information Retrieval

Prof: Ehab Ezzat Hassanein

Introduction

Course Objectives

- How to do efficient (fast, compact) text indexing
- Retrieval models: Boolean, vector-space, probabilistic, and machine learning models
- Evaluation and IR interface issues
- Document clustering and classification
- Search on the web, including crawling, link-based algorithms, indirect feedback, metadata
- Trends: AI, chatGPT, Bard,….etc.

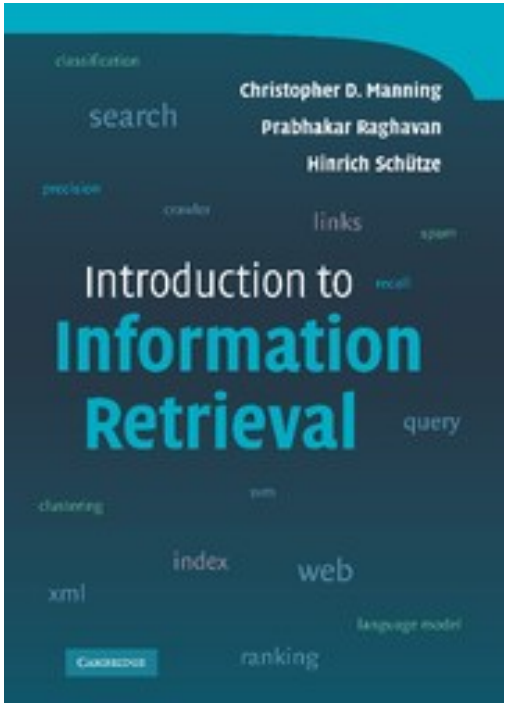
Course Plan

Week	Topic	Comments	Reference
1	Introduction to Information Retrieval		
2	Term Document Incidence Matrices		
3	The Inverted Index	HW1 - Announced	
4	Query Processing with the Inverted Index	HW1- Due	
5	The Boolean Retrieval Model		
6	Phrase Queries and Positional Indexes		
7	Introducing Ranked Retrieval	Project Announced	
8	Scoring with the Jaccard Coefficient	HW2 - Announced	
9	Term Frequency Weighting		
10	Inverse Document Frequency Weighting	HW2 Due	
11	TF IDF Weighting		
12	The Vector Space Model		
13	Calculating TF IDF Cosine Scores		
14	Evaluating Search Engines	Project Due	

Recommended Textbook

Introduction to Information Retrieval by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, 2008

ISBN-13: 9780521865715



Google Stock in 8-2-2023

Alphabet Cl A (GOOGL)

GOOGL

Enter Symbol/Company GET QUOTE

After Hours 07:59 PM ET 02/10/2023

$94.74 0.17 ↑ 0.18%

Previous Close

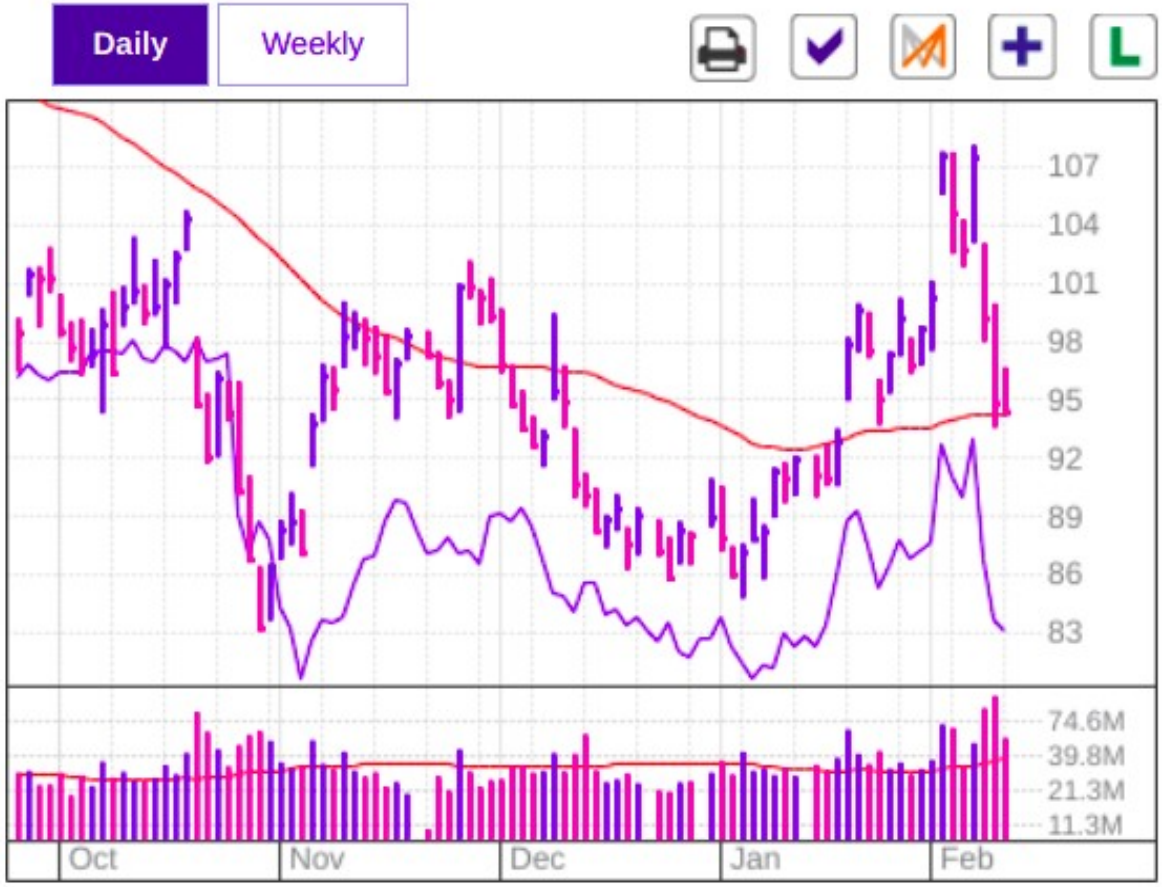
$94.57

0.44 ↓ **0.46%**

Volume: 55 Mil

Volume % Chg: ↑ 49%

Get a Leaderboard Chart for **GOOGL**



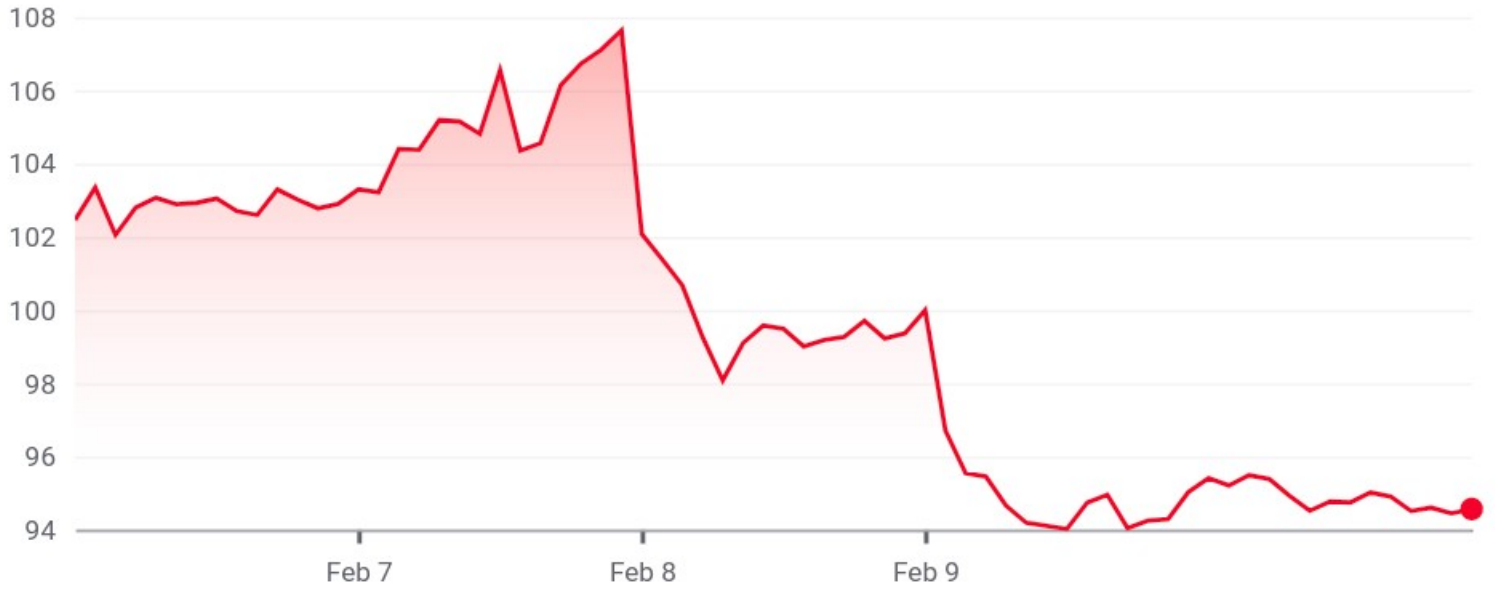
Alphabet Inc Class A

$94.57 ↓7.70% -7.89 5D

Feb 10, 8:00:00 PM UTC-5 · USD · NASDAQ · Disclaimer

1D 5D 1M 6M YTD 1Y 5Y MAX

Key events >



Google Stock Keeps Falling After Bard Ad Shows Inaccurate Answer, AI Race Heats Up

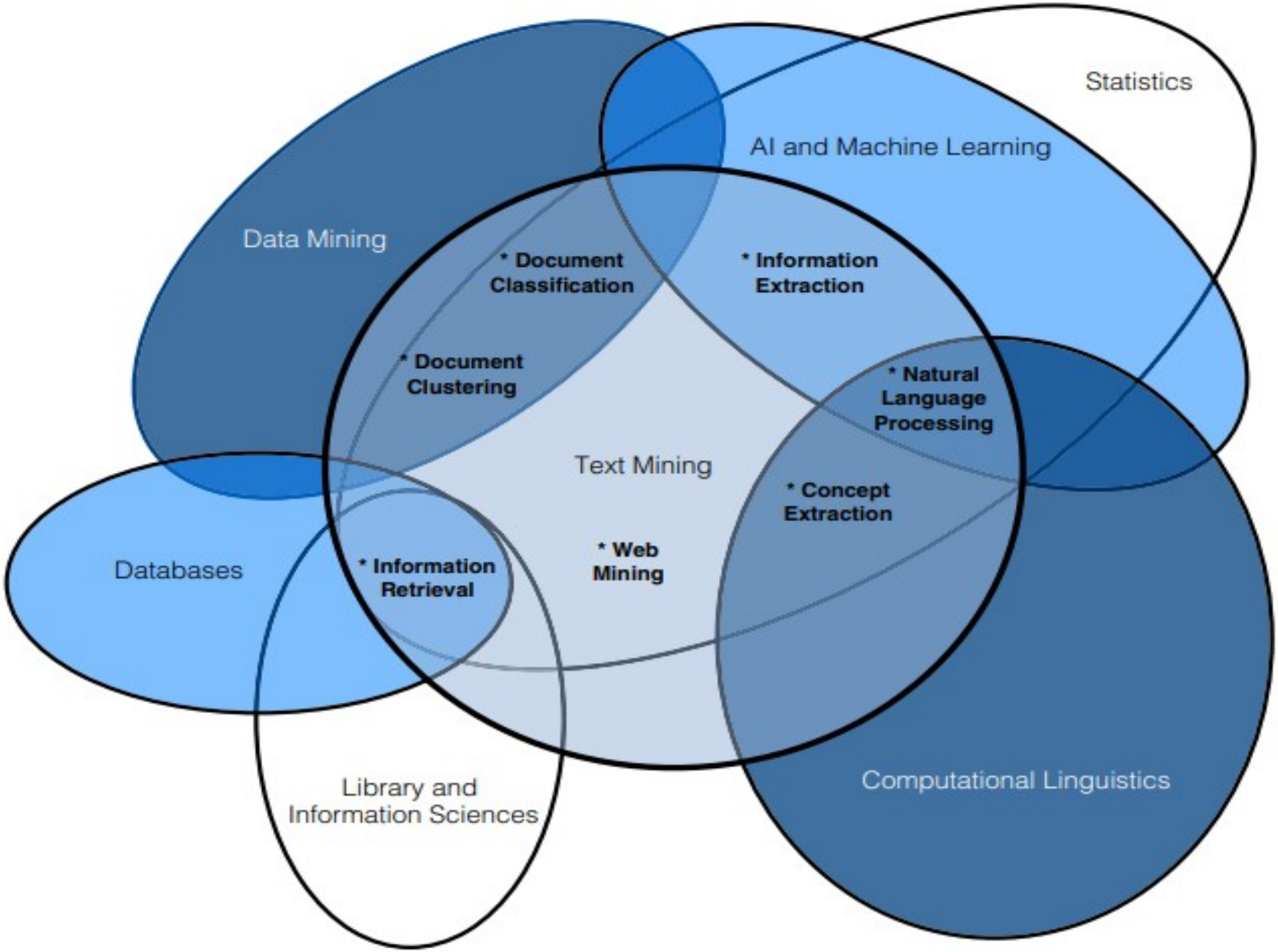
Google posted a video on Twitter demonstrating the "experimental conversational AI service powered by LaMDA," the company wrote. LaMDA is Google's Language Model for Dialogue Applications, which applies machine learning to chatbots and allows them to engage in "free-flowing" conversations, the company says.

In the advertisement, Bard is prompted with the question, "What new discoveries from the James Webb Space Telescope can I tell my 9-year old about?"

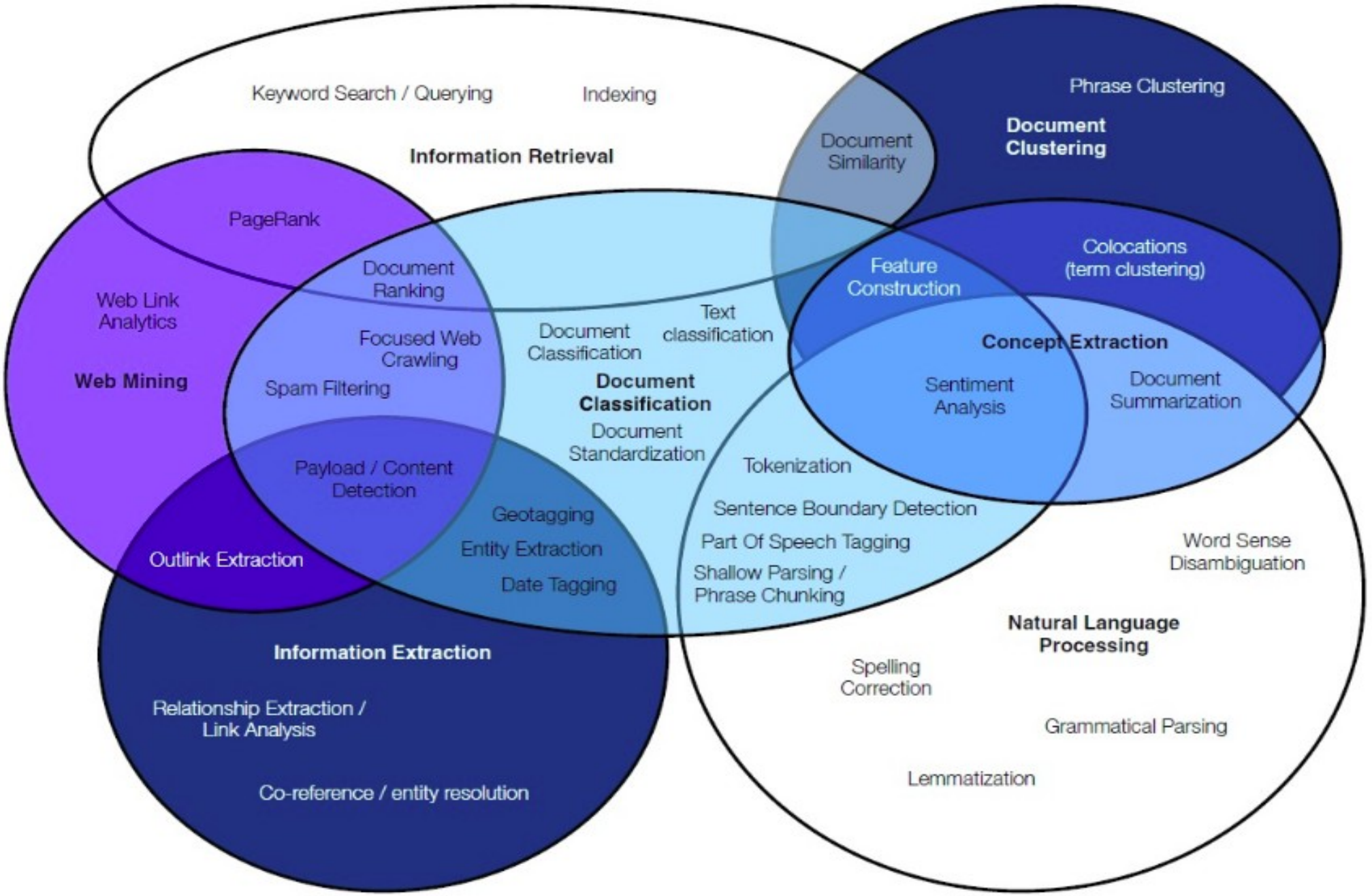
Bard quickly rattles off two correct answers. But its final response was inaccurate. Bard wrote that the telescope took the very first pictures of a planet outside our solar system. In fact, the first pictures of these "exoplanets" were taken by the European Southern Observatory's Very Large Telescope, according to NASA records.

Google touted the new tech at an event in Paris on Wednesday, where it announced plans to roll out AI-powered search results and maps, the Wall Street Journal reported. The new feature will generate lengthy text responses to complex questions, similar to ChatGPT. And Google will launch the feature when it is confident in the quality of answers, the company says.

Text mining interaction with other fields



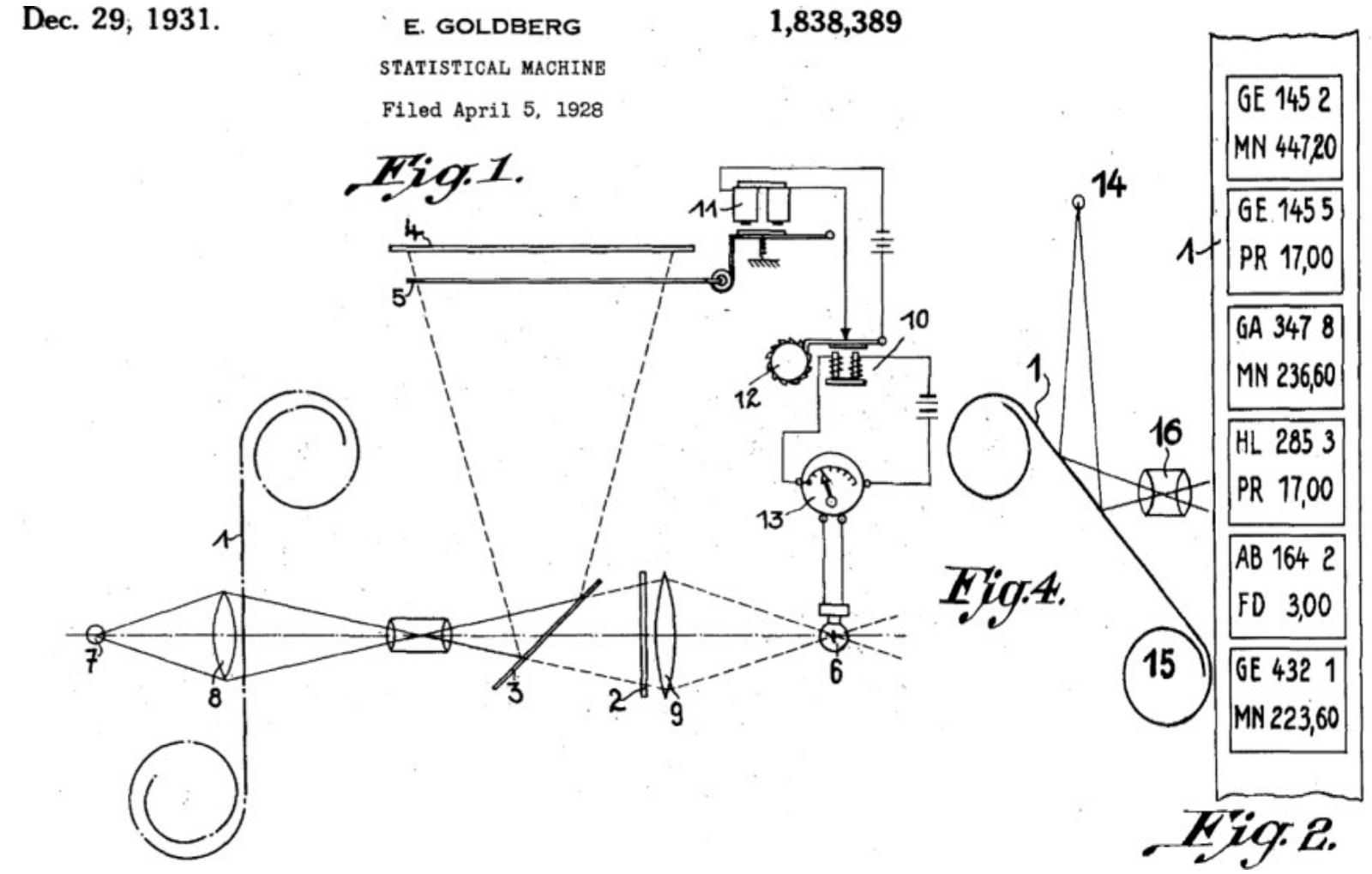
Inter-relationship among different text mining techniques and their core functionalities

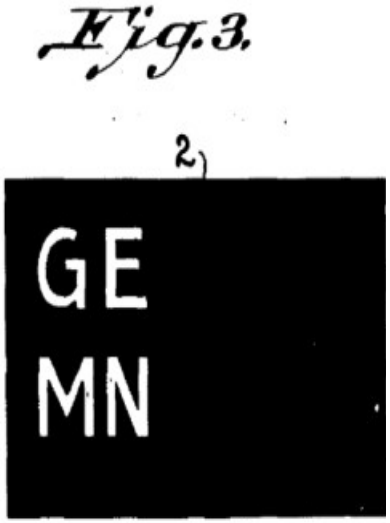


History

Goldberg machine

Goldberg machine is a mechanical machine that searched for a pattern of dots or letters across catalog entries stored on a roll of microfilm.





Goldberg machine cont.

- Here it can be seen that catalog entries were stored on a roll of film (No. 1 of the figure).
- A query (2) was also on film showing a negative image of the part of the catalog being searched for; in this case the 1 st and 6 th entries on the roll.
- A light source (7) was shone through the catalog roll and query film, focused onto a photocell (6).
- If an exact match was found, all light was blocked to the cell causing a relay to move a counter forward (12) and for an image of the match to be shown via a half silvered mirror (3), reflecting the match onto a screen or photographic plate (4 & 5).

The number of websites

- While the exact number of websites keeps changing every second, there are well over 1 billion sites on the world wide web (1,197,982,359 according to Netcraft’s January 2021 Web Server Survey

  January 2020 1 295 973 827 (189 000 000)
- January 2018 1 805 260 010 (171 648 771)
- January 2016 906 616 188 (170 258 872)
- January 2014 861 379 152 (180 067 270)
- January 2012 582 716 657 (182 441 983)
- January 2010 206 741 990 (83 456 669)
- January 2008 155 583 825 (68 274 154)

Basic Definitions

Information retrieval (IR)

Information retrieval (IR) is

finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

That include not only **Web Search** but also **:**

- **Email Search**
- **Searching your laptop**
- **Corporate knowledge bases**
- **Legal Information retrieval**

Data extraction & Information extraction

Data extraction is a process that involves retrieval of data from various sources. Frequently, companies extract data in order to process it further, migrate the data to a data repository or to further analyze it.

Information extraction (IE) is the automated retrieval of specific information related to a selected topic from a body or bodies of text. Information extraction tools make it possible to pull information from text documents, databases, websites or multiple sources.

Data mining & Web mining

Data mining Data mining is the process of analyzing dense volumes of data to find patterns, discover trends, and gain insight into how that data can be used. Data miners can then use those findings to make decisions or predict an outcome. Data mining is an interconnected discipline, blending the fields of statistics, machine learning, and artificial intelligence.

Web Mining is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents and services, Web content, hyperlinks and server logs.

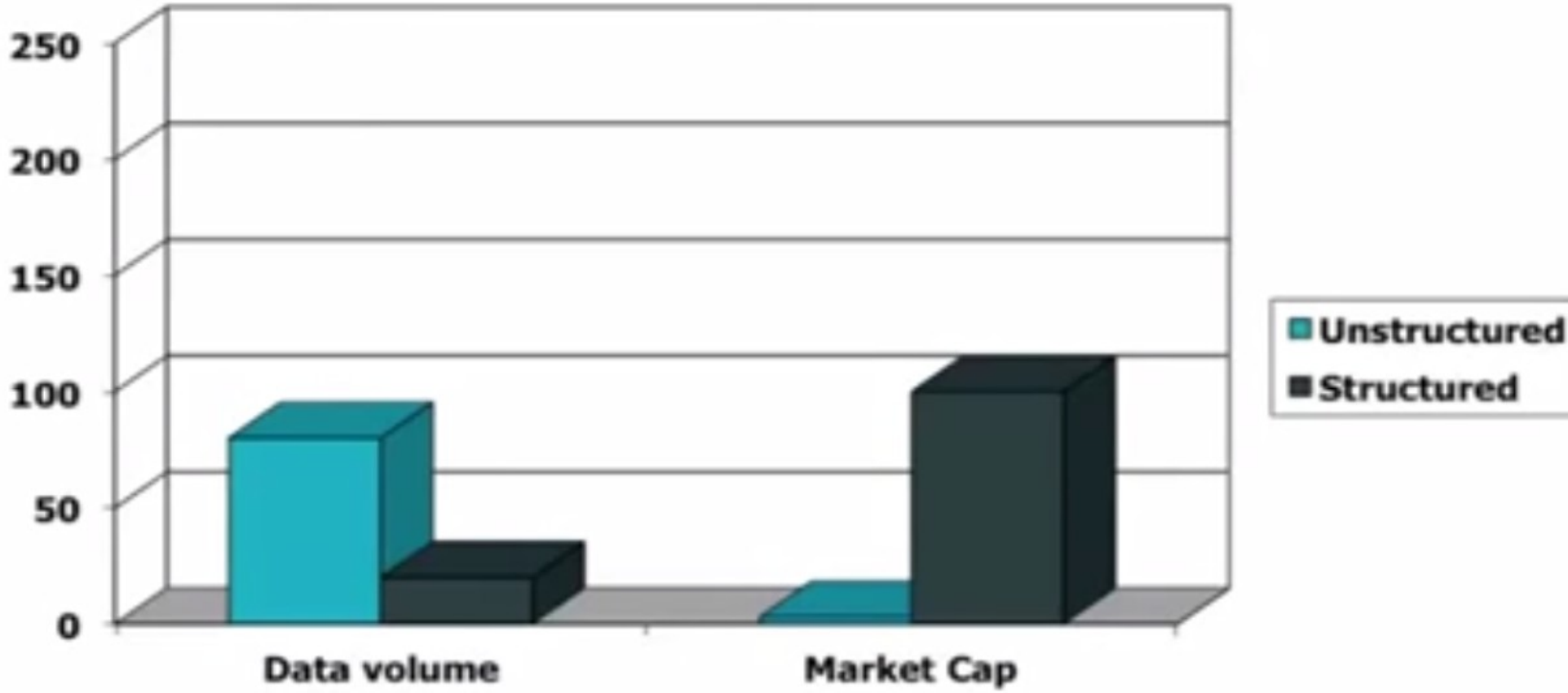
The goal of Web mining is to look for patterns in Web data by collecting and analyzing information in order to gain insight into trends, the industry and users in general.

web crawler & web scraper

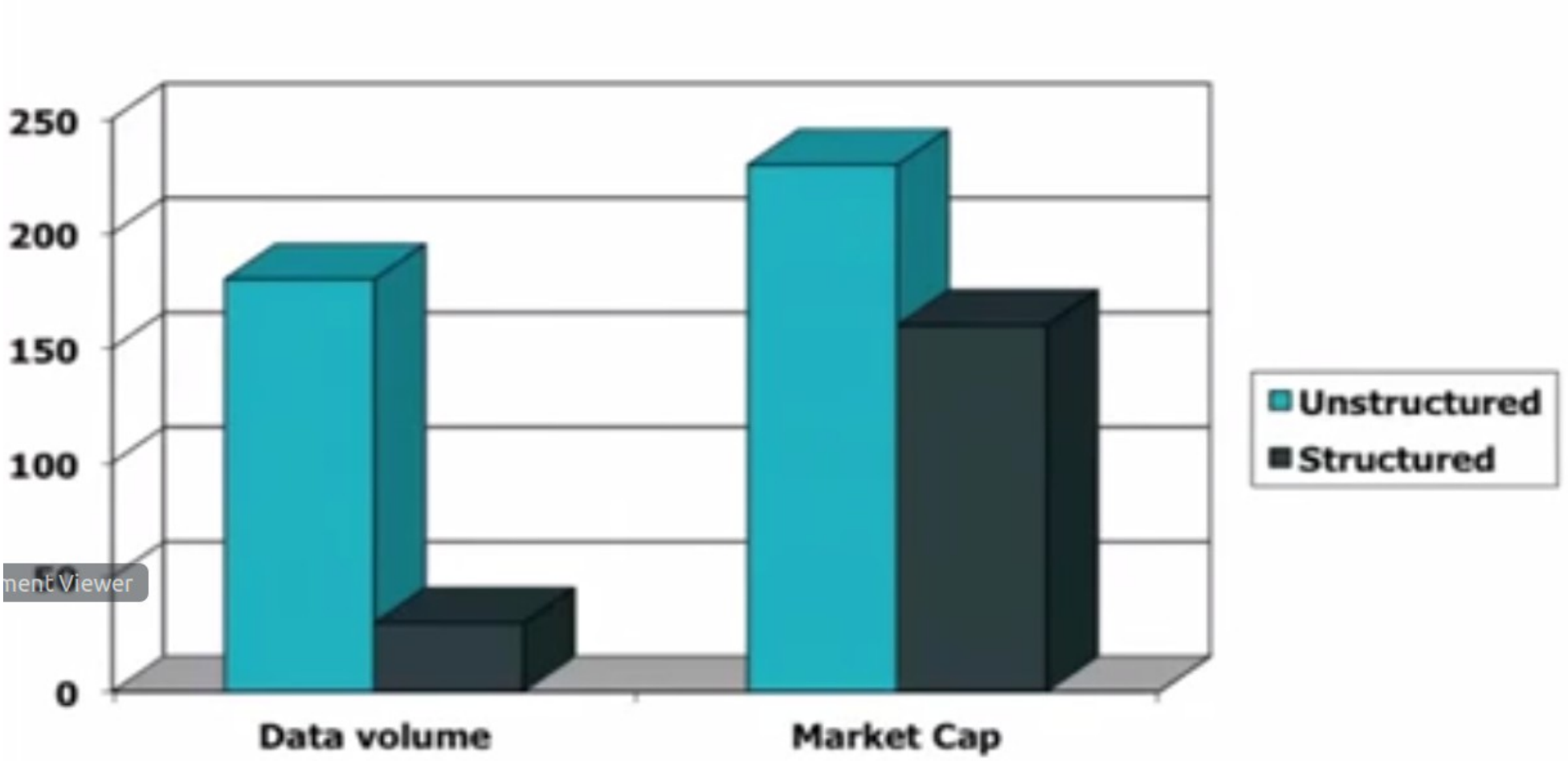
A web crawler sometimes called a “spider,” is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.

A web scraper is a process of extracting specific data. Unlike web crawling, a web scraper searches for specific information on specific websites or pages.

Unstructured (text) vs. Structurer (database) data In the mid nineties



Unstructured (text) vs. Structurer (database) data Today



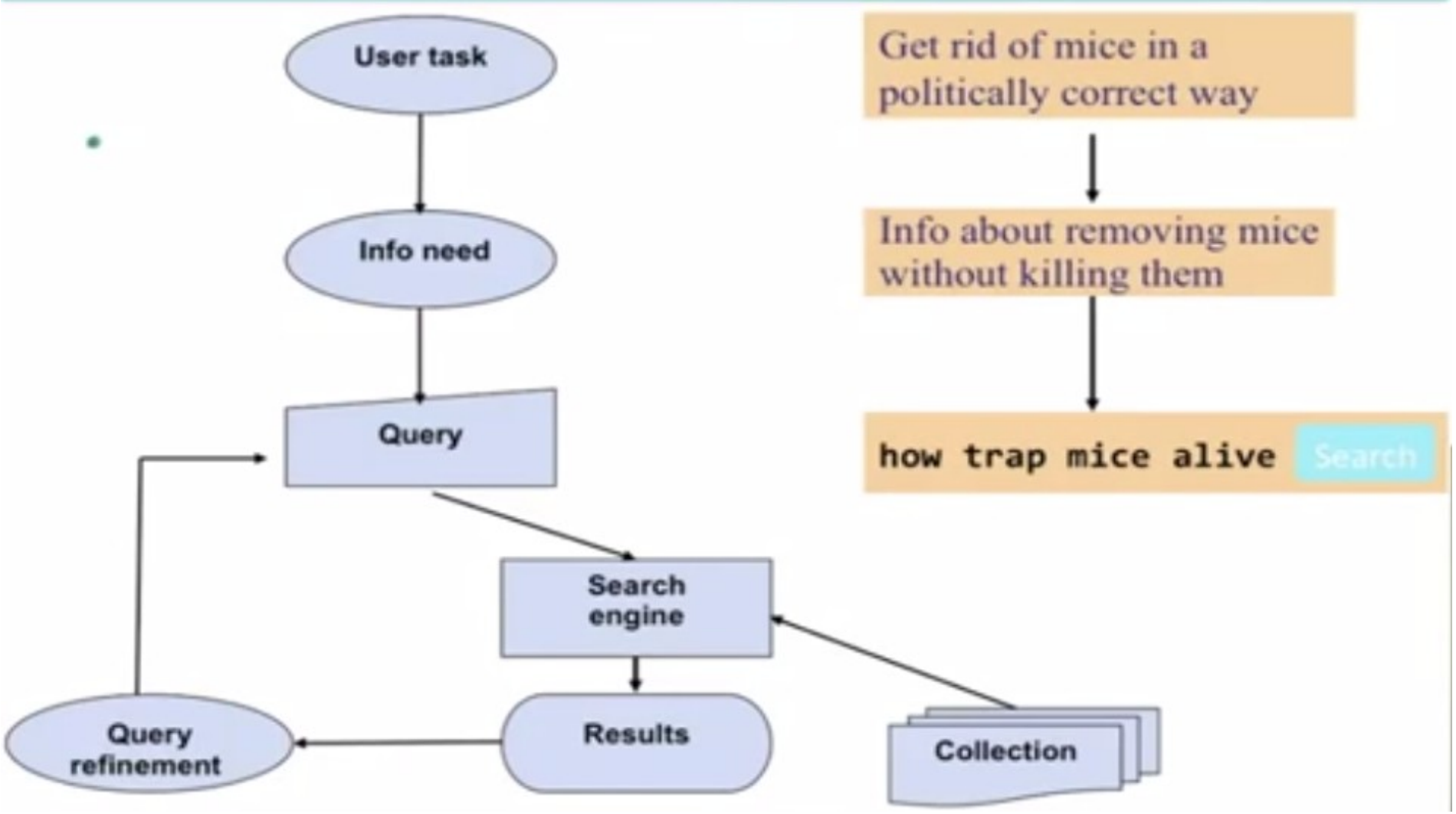
Basic Assumptions of Information Retrieval

- Collection: a set of documents

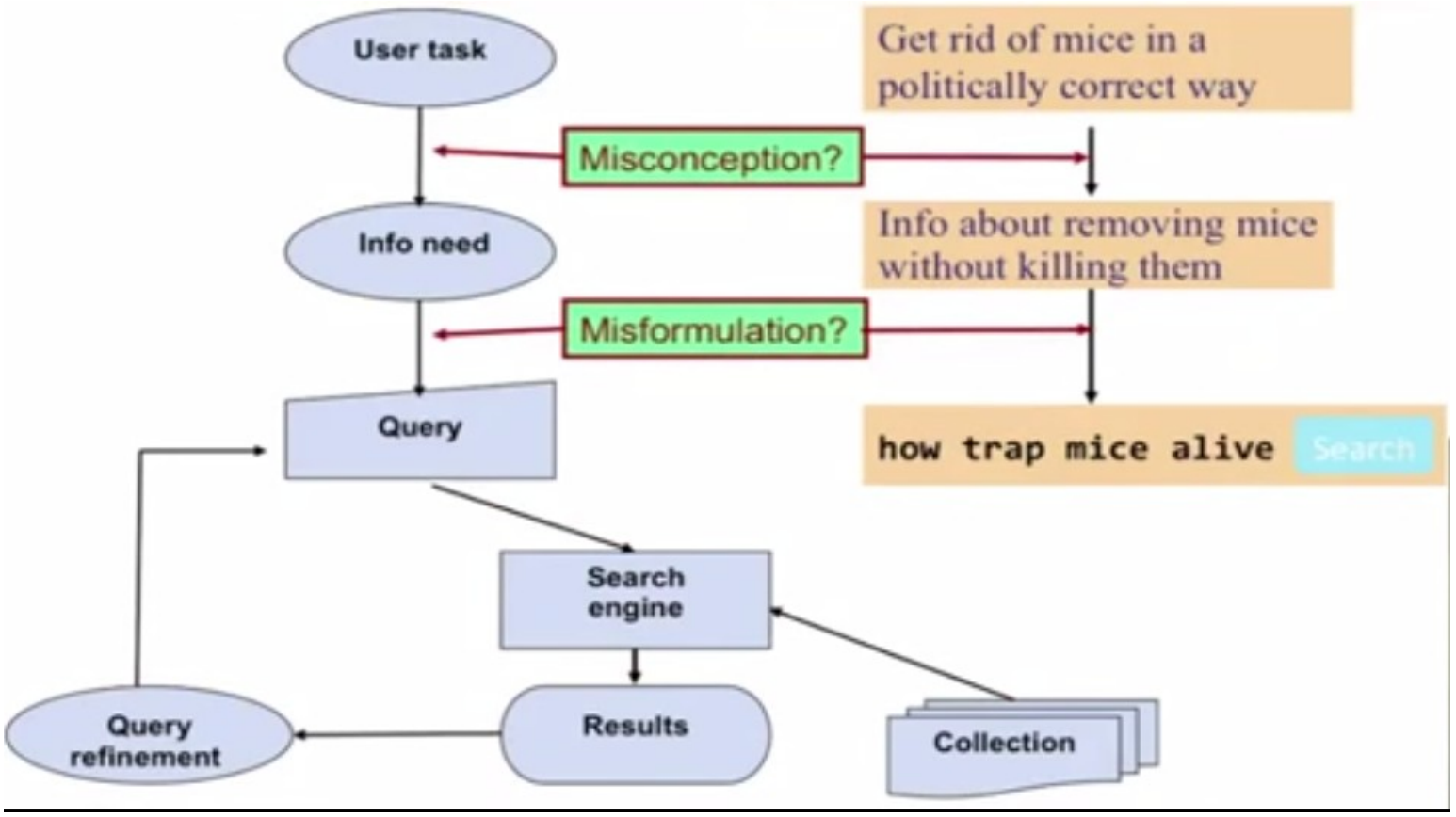
  Assume it is a static collection for now..

- Goal: retrieve documents with Information that is relevant to the user's information need and help the user to complete a task

The Classic Search Model



The Classic Search Model what can go wrong..



Information need

- An information need is the topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need.

Relevance

- **Relevant** if it is one that the user perceives as containing information of value with respect to their personal information need.

The Effectiveness

To assess the effectiveness of an IR system (i.e., the quality of its search results), a user will usually want to know two key statistics about the system’s returned results for a query: Percesion and Recall

How good are the retrieved documents

- **PRECISION**

  Precision: What fraction of the returned results are relevant to the information need?

- **RECALL**

  Recall: What fraction of the relevant documents in the collection were returned by the system?

Term-document Incidence Matrix And Inverted Index

Information Need

- An information need is the topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need.

AD HOC RETRIEVAL

- Our goal is to develop a system to address the ad hoc retrieval task.
- This is the most standard IR task. In it, a system aims to provide documents from within the collection that are relevant to an arbitrary user information need, communicated to the system by means of a one-off, user-initiated query

Relevance

- **Relevant** if it is one that the user perceives as containing information of value with respect to their personal information need.

The Effectiveness

To assess the effectiveness of an IR system (i.e., the quality of its search results), a user will usually want to know two key statistics about the system’s returned results for a query: Percesion and Recall

How good are the retrieved documents

- **PRECISION**

  Precision: What fraction of the returned results are relevant to the information need?

- **RECALL**

  Recall: What fraction of the relevant documents in the collection were returned by the system?

Grepping

- This process is commonly referred to as grepping through text, after the Unix command grep, which performs this process.
- Grepping through text can be a very effective process, especially given the speed of modern computers, and often allows useful possibilities for wildcard pattern matching through the use of regular expressions.
- for simple querying of modest collections (the size of Shakespeare’s Collected Works is a bit under one million words of text in total), you really need nothing more

Unstructured data in 1620

- Which plays of Shakespeare contain the words ***Brutus*** *AND* ***Caesar*** but *NOT* ***Calpurnia***?
- One could `grep` all of Shakespeare's plays for ***Brutus*** and ***Caesar,*** then strip out lines containing ***Calpurnia***?
- Why is that not the answer?
  - Slow (for large corpora)
  - *NOT* ***Calpurnia*** is non-trivial
  - Other operations (e.g., find the word ***Romans*** near ***countrymen***) not feasible
  - Ranked retrieval (best documents to return)
    - Later lectures

10

Shortfalls of Grepping

1. To process large document collections quickly. The amount of online data has grown at least as quickly as the speed of computers, and we would now like to be able to search collections that total in the order of billions to trillions of words.

2. To allow more flexible matching operations. For example, it is impractical to perform the query Romans NEAR countrymen with grep, where NEAR might be defined as “within 5 words” or “within the same sentence”.

3. To allow ranked retrieval: in many cases you want the **best** answer to an information need among many documents that contain certain words.

term-document incidence matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

the query: Brutus AND Caesar AND NOT Calpurnia

we take the vectors for Brutus , Caesar and Calpurnia, complement the last, and then do a bitwise AND :

110100 AND 110111 AND 101111 = 100100

Answer to the Query:

- Antony and Cleopatra, Act III, Scene ii

Agrippa [Aside to DOMITIUS ENOBARBUS]: Why, Enobarbus,
When Antony found Julius ***Caesar*** dead,
He cried almost to roaring; and he wept
When at Philippi he found ***Brutus*** slain.

- Hamlet, Act III, Scene ii

Lord Polonius: I did enact Julius ***Caesar*** I was killed i' the Capitol; ***Brutus*** killed me.

BOOLEAN RETRIEVAL MODEL

- The Boolean retrieval model is a model for information retrieval in which we can pose any query which is in the form of a Boolean expression of terms, that is, in which terms are combined with the operators AND , OR , and NOT .
- The model views each document as just a set of words.

Bigger Collection

- Suppose we have **N = 1** million documents.
- Suppose each document is about 1000 words long (2–3 book pages)
- assume an average of 6 bytes per word including spaces and punctuation,
- This is a document collection about 6 GB in size
- Typically, there might be about **M =** 500,000 distinct terms in these documents (corresponds to the number of rows in the matrix)

Can’t build the Matrix!

- 500K x 1M matrix => half a trillion 0’s and 1’s

  BUT

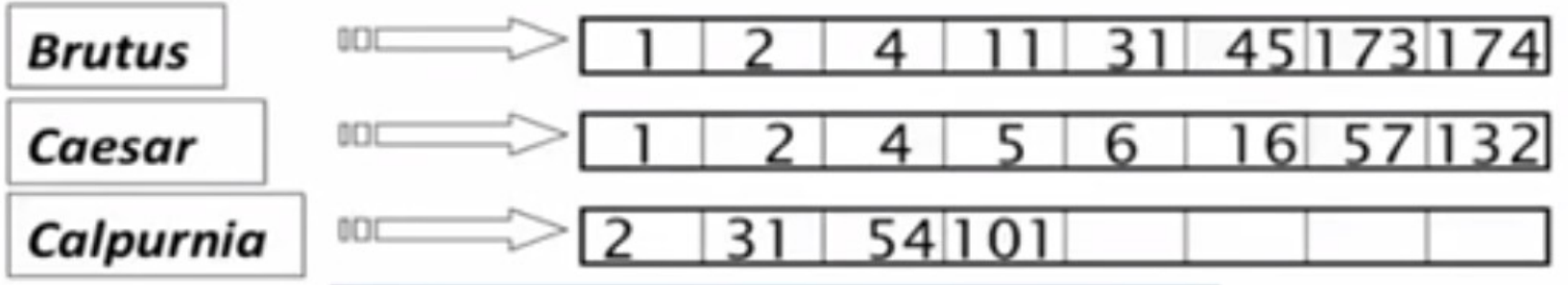
- **Almost all of the entries are 0’s**
- **The documents at most has 1 billion 1’s**
  - Since we assume that we have 1 M document each with 1000 words then even if w have distinct terms for each documents we at most have 1000M 1’s
- Such a matrix is extremely sparse. Almost all entries are 0’. We need better representation. A representation that records only the 1’s

Inverted Index.

- The **key data structure** that underlay all modern IR systems
- It is a data structure that exploits the sparsity of the term document matrix and allow for very efficient retrieval
- The name is actually redundant: an index always maps back from terms to the parts of a document where they occur.
- Nevertheless, inverted index, or sometimes inverted file, has become the standard term in information retrieval.

Inverted Index.

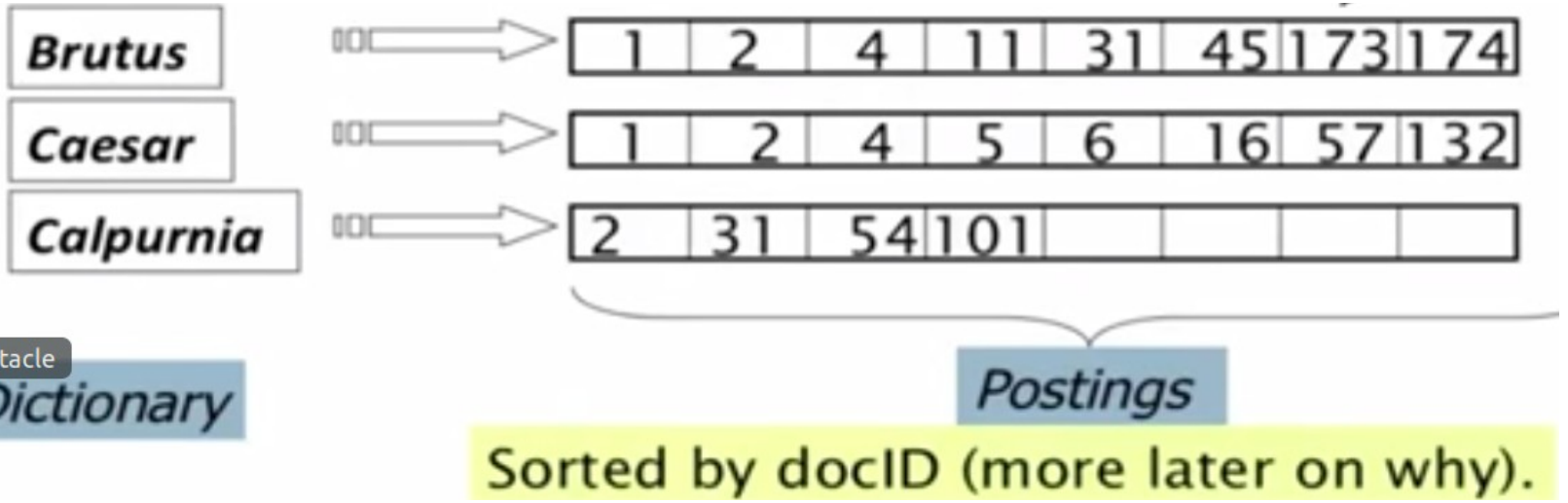
- For each term t, we must store all the documents that contain t.
  - Identify each document by docID, a document serial number
  - Can we us Fixed-size arrays for this?
    - Verv inefficient



What happens if the word ***Caesar*** is added to document 14?

Inverted Index.

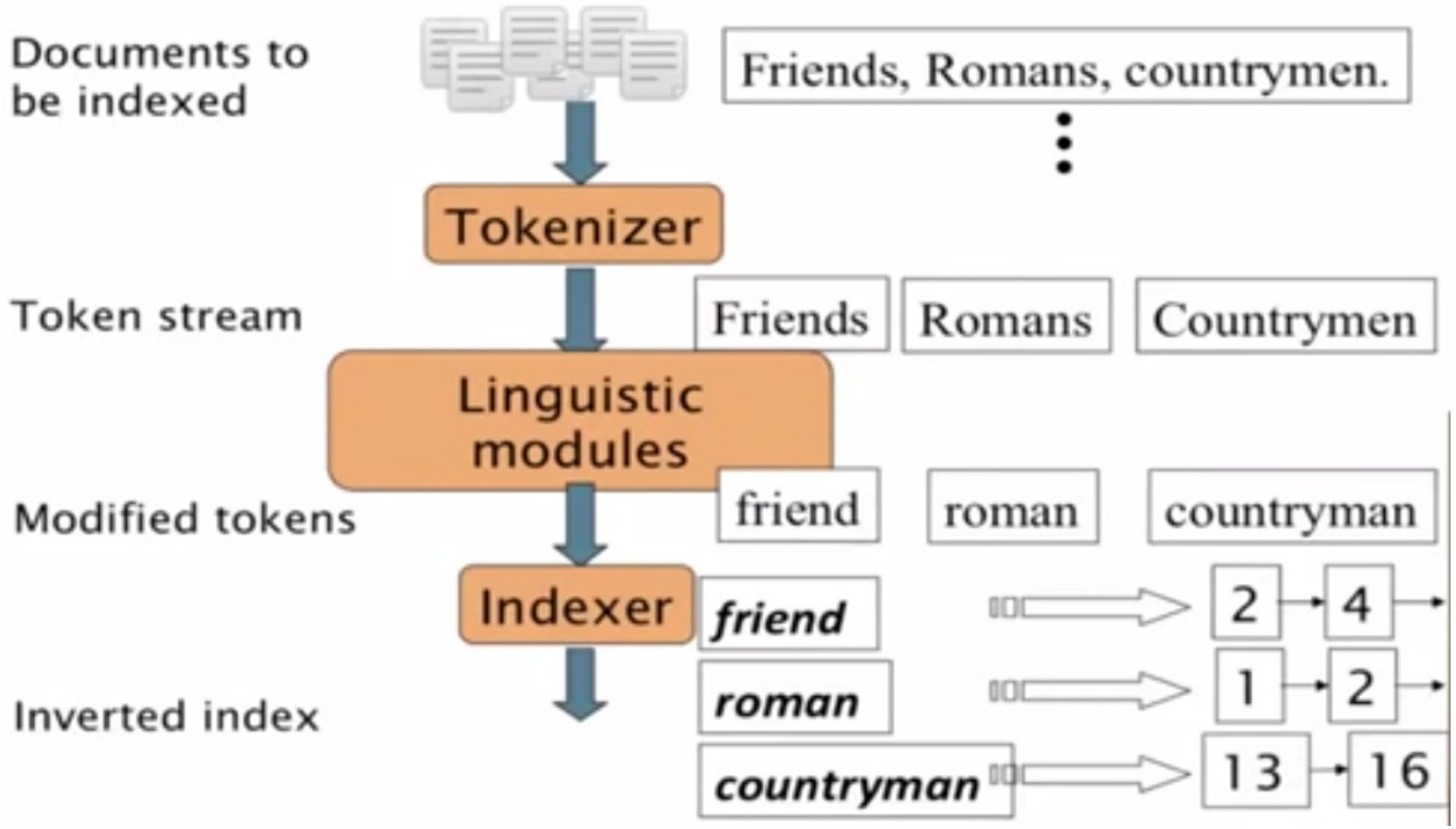
- We need variable-size posting lists
  - In disk a continuous run of postings is normal and best.
  - In memory, can use linked lists or variable length arrays.
  - Dictionary is small so it can be stored in memory; whereas, postings are large and may be stored in disks.



Inverted index vs. Forward Index

- In a search engine you have a list of documents (pages on web sites), where you enter some keywords and get results back.
- A **forward index** (or just index) is the list of documents, and which words appear in them. In the web search example, Google crawls the web, building the list of documents, figuring out which words appear in each page.
- The **inverted index** is the list of words, and the documents in which they appear. In the web search example, you provide the list of words (your search query), and Google produces the documents (search result links).

Inverted Index construction



Initial stages of text processing

- **Tokenization**
  - Cut character sequence into words tokens
    - Deal with “John’s”, a state-of-the-art solution
- **Normalization**
  - Map text and query term to the same form
    - USA and U.S.A to match
- **Stemming**
  - We may wish different forms of a root to match
    - authorize and authorization
- **Stop words**
  - We may omit very common words (or not!)
    - The, a, to, of
      - Query the song to be or not to be!!