

Introduction to Big Data and Data Science (CSCE 5300)*

Yunhe Feng

Assistant Professor, Department of Computer Science and Engineering

7th November, 2024



1 Topics Covered So Far

2 Special Topics: Finding Similar/Same Items in Big Data

3 Bloom Filter: A Probabilistic Data Structure for Membership Test

4 Assignment

Topics Covered So Far

- Introduction to Python Programming
- Data Visualization Techniques
- Working with DataFrames, Pandas, and PySpark
- An Introduction to Machine Learning
- Logistic Regression and Confusion Matrix Analysis
- Deep Learning Using PyTorch
- Fundamentals of Image Processing
- K-Nearest Neighbors (KNN) Classification
- Hadoop for Distributed Computing
- Introduction to Parallel Computing

- 1 Topics Covered So Far
- 2 Special Topics: Finding Similar/Same Items in Big Data
- 3 Bloom Filter: A Probabilistic Data Structure for Membership Test
- 4 Assignment

Finding Similar/Same Items Matters in Search Engines



DuckDuckGo



Search by Text on Google

The screenshot shows a Google search results page for the query "big data and data science". The search bar at the top contains the text "big data and data science" with a search icon. Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Books", and "More". The "All" tab is selected. The search results show "About 3,080,000,000 results (0.82 seconds)". The first result is titled "Big data and data Science" and includes a snippet: "Big data refers to any large and complex collection of data. Data analytics is the process of extracting meaningful information from data. Data science is a multidisciplinary field that aims to produce broader insights." Below this snippet is a link to "https://www.bmc.com › blogs › big-data-vs-analytics" and a date "Oct 13, 2021". To the right of the text is a thumbnail image titled "Big Data Vs Data Science" which is a comparison chart. Below the first result is a "People also ask" section with four questions: "Is big data and data science same?", "What do you mean by big data in data science?", "Is big data necessary for data science?", and "Which is better data science or big data?". Each question has a dropdown arrow. Below this section is a link to "https://www.simplilearn.com › Resources › Big Data" and a title "Data Science vs. Big Data vs. Data Analytics - Simplilearn". The snippet for this result says: "Sep 26, 2022 – Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. What is Data Science? - What is Big Data? - What is Data Analytics?". Below this is another link to "https://www.geeksforgeeks.org › difference-between-bi..." and a title "Difference Between Big Data and Data Science". The snippet for this result says: "Sep 30, 2022 – It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques."

Google

big data and data science

Q All Images News Videos Books More Tools

About 3,080,000,000 results (0.82 seconds)

Big data and data Science

Big data refers to any large and complex collection of data. Data analytics is the process of extracting meaningful information from data. Data science is a multidisciplinary field that aims to produce broader insights. Oct 13, 2021

<https://www.bmc.com › blogs › big-data-vs-analytics>

Big Data vs Data Analytics vs Data Science - BMC Software

Big Data Vs Data Science

People also ask

Is big data and data science same?

What do you mean by big data in data science?

Is big data necessary for data science?

Which is better data science or big data?

<https://www.simplilearn.com › Resources › Big Data>

Data Science vs. Big Data vs. Data Analytics - Simplilearn

Sep 26, 2022 – Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. What is Data Science? - What is Big Data? - What is Data Analytics?

<https://www.geeksforgeeks.org › difference-between-bi...>


Difference Between Big Data and Data Science

Sep 30, 2022 – It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques.

Search by an Image on Google <https://images.google.com/>



Search any image with Google Lens ×

 Drag an image here or [upload a file](#)

OR

Finding Similar Items Matters in Recommender System



Recommendations on Amazon

Explore more from across the store



Deals on frequently repurchased items



Recommendations on Google Scholar <https://scholar.google.com/>

Google Scholar



☒ Articles ☐ Case law

Recommended articles



Fairness task assignment strategy with distance constraint in Mobile CrowdSensing



X Song, E Wang, W Liu, Y Liu, Y Dong

CCF Transactions on Pervasive Computing and Intera... - 2 days ago



Stable Worker-Task Assignment in Mobile Crowdsensing Applications



F Yucel, M Yuksel, E Bulut

people.vcu.edu - 4 days ago

PDF



Similarity Metrics

- Euclidean distance

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

- Cosine similarity

$$\text{cos_sim}(p, q) = \frac{\sum_{i=1}^n p_i \cdot q_i}{\sqrt{\sum_{i=1}^n p_i^2} \cdot \sqrt{\sum_{i=1}^n q_i^2}}$$

- Jaccard index

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Question - Removing Duplicated Text in Big Data

Any cell values that repeat
are highlighted

	A	B	C	D
1	Date	Sales Rep	Region	Amount
2	22-07-2015	John	China	\$ 16,543
3	22-07-2015	Jack	US	\$ 32,434
4	23-07-2015	Jill	Canada	\$ 534
5	22-07-2015	Joe	Brazil	\$ 5,243
6	22-07-2015	Jinie	US	\$ 34,536
7	22-07-2015	Jasmine	Canada	\$ 23,424
8	22-07-2015	John	Brazil	\$ 2,342
9	23-07-2015	Jack	China	\$ 6,547
10	23-07-2015	Jill	US	\$ 5,000
11	23-07-2015	Joe	Canada	\$ 31,235
12	23-07-2015	Jinie	Brazil	\$ 6,465
13	23-07-2015	Jill	US	\$ 5,000
14	23-07-2015	Joe	Canada	\$ 4,325
15	22-07-2015	Jinie	Brazil	\$ 2,346
16	22-07-2015	John	China	\$ 16,543

Only duplicate rows are
highlighted

	A	B	C	D
1	Date	Sales Rep	Region	Amount
2	22-07-2015	John	China	\$ 16,543
3	22-07-2015	Jack	US	\$ 32,434
4	23-07-2015	Jill	Canada	\$ 534
5	22-07-2015	Joe	Brazil	\$ 5,243
6	22-07-2015	Jinie	US	\$ 34,536
7	22-07-2015	Jasmine	Canada	\$ 23,424
8	22-07-2015	John	Brazil	\$ 2,342
9	23-07-2015	Jack	China	\$ 6,547
10	24-07-2015	Jill	US	\$ 5,000
11	23-07-2015	Joe	Canada	\$ 31,235
12	23-07-2015	Jinie	Brazil	\$ 6,465
13	24-07-2015	Jill	US	\$ 5,000
14	23-07-2015	Joe	Canada	\$ 4,325
15	22-07-2015	Jinie	Brazil	\$ 2,346
16	22-07-2015	John	China	\$ 16,543

Question - Removing Duplicated Text in Big Data

- ① Preprocessing and normalization (e.g., case normalization, and whitespace normalization)
- ② Hashing for quick comparison (e.g., SimHash or MinHash for near-duplicate detections)
- ③ Scalable data structures and algorithms (e.g., Bloom Filters, Hadoop or Spark for distributed data processing)

Question - Removing Duplicated Images in Big Data



Question - Removing Duplicated Images in Big Data

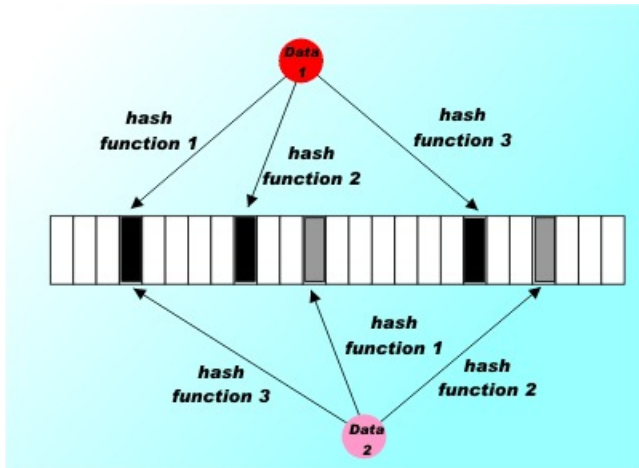
- ① Preliminary sorting based on metadata (e.g., file size, creation date, and dimensions)
- ② Hashing for quick comparison (e.g., Average Hash, Perceptual Hash, Difference Hash for near-duplicate detections)
- ③ Detailed comparison for near-duplicates (e.g., structural similarity indexes (SSIM))

Question - Removing Duplicated Videos in Big Data

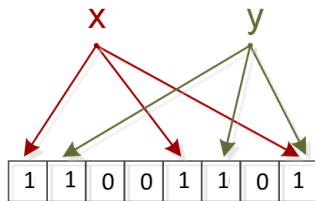
- 1 Video fingerprinting (e.g., capturing key features such as color distribution, shape, motion patterns, and scene changes across frames)
- 2 Keyframe extraction
- 3 Process duplicated images

- 1 Topics Covered So Far
- 2 Special Topics: Finding Similar/Same Items in Big Data
- 3 Bloom Filter: A Probabilistic Data Structure for Membership Test**
- 4 Assignment

What is Bloom Filter



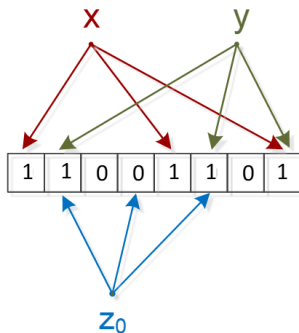
Bloom Filter



- $m = 8$
- $n = 2$
- $k = 3$
- $t = 5$

- m : the size (total number of bits) of the bloom filter;
- k : the number of hash functions;
- n : the number of elements inserted in the bloom filter;
- t : the number of bits flipped to one;
- p : the false positive probability of the bloom filter.

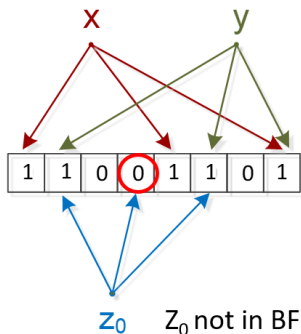
Bloom Filter



To query for z_0

- m : the size (total number of bits) of the bloom filter;
- k : the number of hash functions;
- n : the number of elements inserted in the bloom filter;
- t : the number of bits flipped to one;
- p : the false positive probability of the bloom filter.

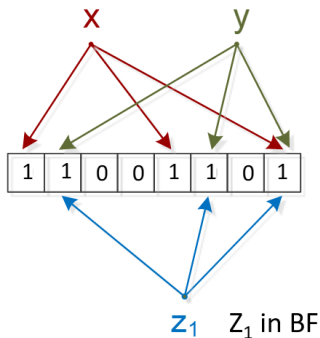
Bloom Filter



To query for z_0

- m : the size (total number of bits) of the bloom filter;
- k : the number of hash functions;
- n : the number of elements inserted in the bloom filter;
- t : the number of bits flipped to one;
- p : the false positive probability of the bloom filter.

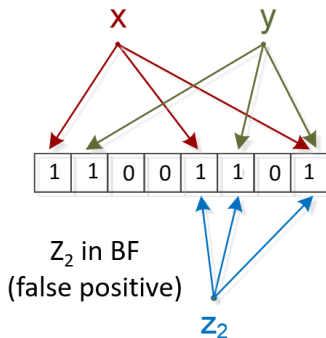
Bloom Filter



To query for z_1

- m : the size (total number of bits) of the bloom filter;
- k : the number of hash functions;
- n : the number of elements inserted in the bloom filter;
- t : the number of bits flipped to one;
- p : the false positive probability of the bloom filter.

Bloom Filter



To query for z_2

- m : the size (total number of bits) of the bloom filter;
- k : the number of hash functions;
- n : the number of elements inserted in the bloom filter;
- t : the number of bits flipped to one;
- p : the false positive probability of the bloom filter.

Bloom Filter

- The probability of **false positives** p given a parameter setting $(m; k; n)$ can be calculated as:

$$p \approx (1 - e^{-\frac{kn}{m}})^k$$

- For a given bloom filter size m and the number of inserted elements n , the **number of hash functions** k that minimizes the false positive is:

$$k = \frac{m}{n} \ln 2$$

- For a given number of inserted elements n and the desired false positive p , the required **bloom filter size** m is:

$$m = -\frac{n \ln p}{(\ln 2)^2}$$

- 1 Topics Covered So Far
- 2 Special Topics: Finding Similar/Same Items in Big Data
- 3 Bloom Filter: A Probabilistic Data Structure for Membership Test
- 4 Assignment

Assignment-9 (2.0 pts.)

- Implement Bloom Filter (2.0 pts.)