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

Yunhe Feng

Assistant Professor, Department of Computer Science and Engineering

7th November, 2024

*Teaching materials are reorganized and reformed based on Dr. Jure Leskovec & Dr. Mina Ghashami's Stanford CS246: Mining Massive Data Sets

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

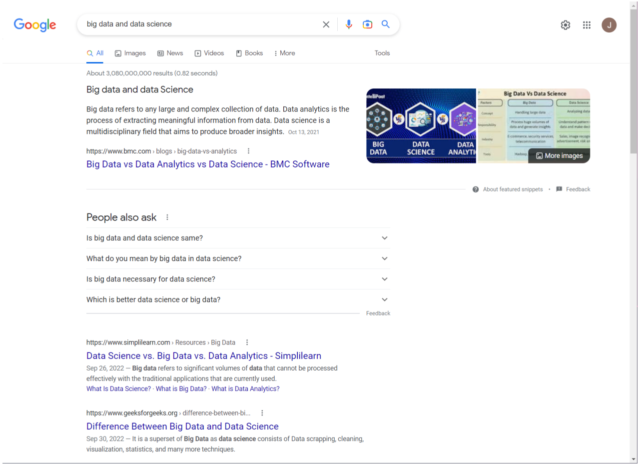
Topics Covered So Far

- 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

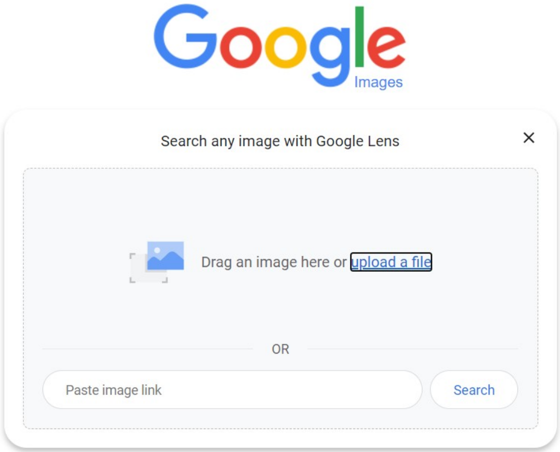
1. Topics Covered So Far
2. Special Topics: Finding Similar/Same Items in Big Data
3. Bloom Filter: A Probabilistic Data Structure for Membership Test
4. Assignment

Finding Similar/Same Items Matters in Search Engines

Search by Text on Google

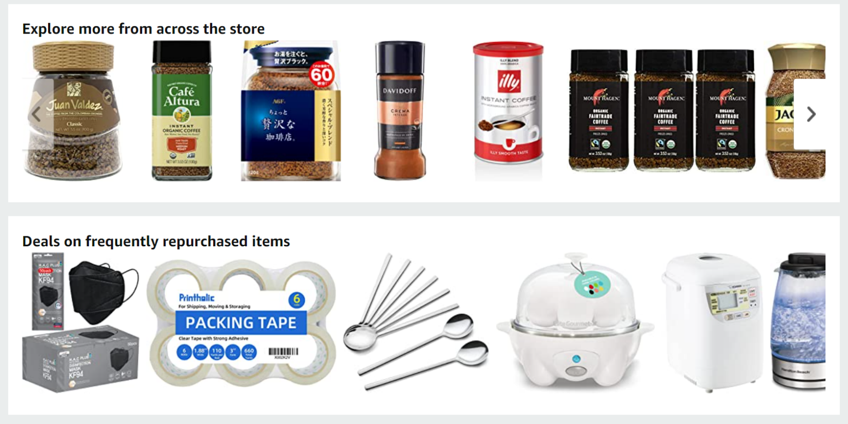


Search by an Image on Google `https://images.google.com/`

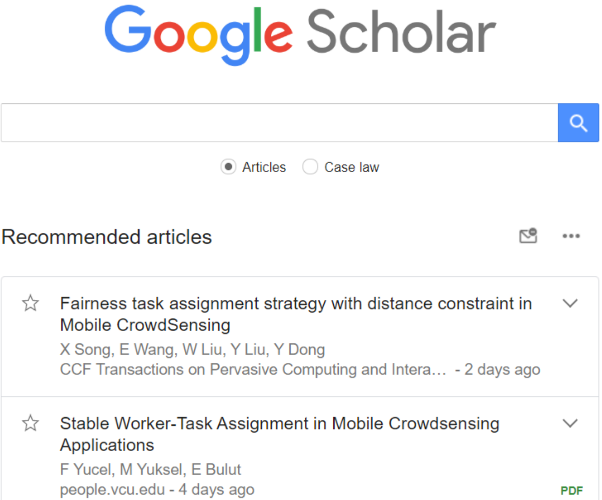


Finding Similar Items Matters in Recommender System

Recommendations on Amazon



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



Similarity Metrics

- Euclidean distance
  $d(\mathrm{p}, \mathrm{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$
- Cosine similarity
  $\mathrm{cos\_sim}(\mathrm{p}, \mathrm{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$
- 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

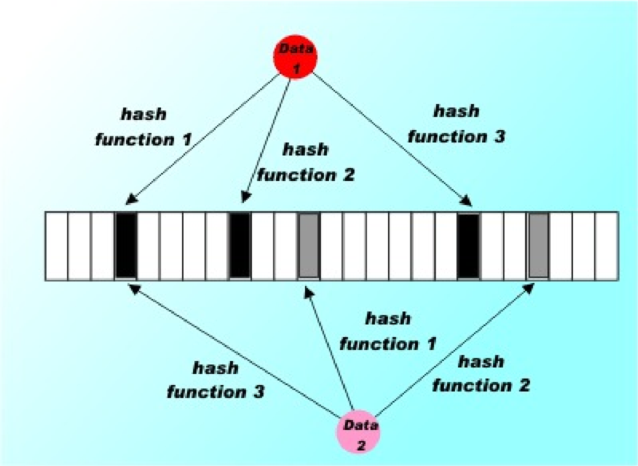
Question - Removing Duplicated Images in Big Data

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))

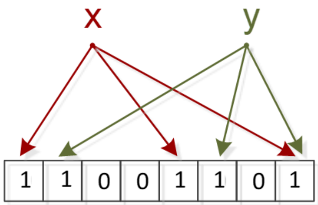
Question - Removing Duplicated Videos in Big Data

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

What is Bloom Filter



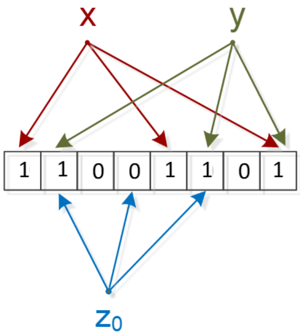
Bloom Filter



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

- 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.

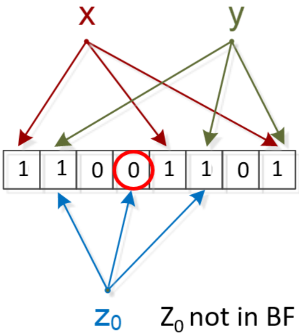
Bloom Filter



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.

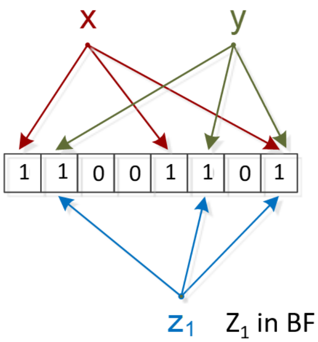
Bloom Filter



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.

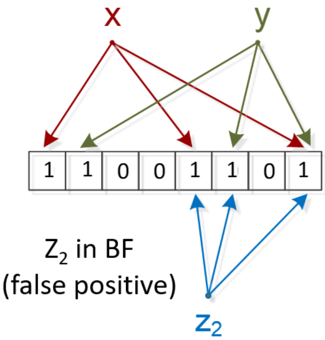
Bloom Filter



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.)