Analytics Vidhya

Data Preprocessing

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Introduction

High quality training data leads to powerful models

Datasets needs to be cleaned

Data Processing

Filtering out raw data to create a high-quality training dataset

Data Sampling

Sampling datasets to handle the training data distribution

Data Sampling in GPT-3

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 Billion	60%	0.44
WebText2	19 Billion	22%	2.9
Books1	12 Billion	8%	1.9
Books2	55 Billion	8%	0.43
Wikipedia	3 Billion	3%	3.4

Data Deduplication

The process of removing duplicate text across training data

Why Data Deduplication?

- Efficient model training
- Accelerate training process
- Accurate evaluation

Methods of Deduplication

Based on the Jaccard similarity b/w document pair-wise

Document No.	Text
Document 1	Generative AI is fun
Document 2	Generative AI is fun and easy

Methods of Deduplication

Based on the Jaccard similarity

Document No.	Text	Bigrams
Document 1	Generative AI is fun	Generative AI, AI is, is fun
Document 2	Generative AI is fun and easy	Generative AI, AI is, is fun, fun and, and easy

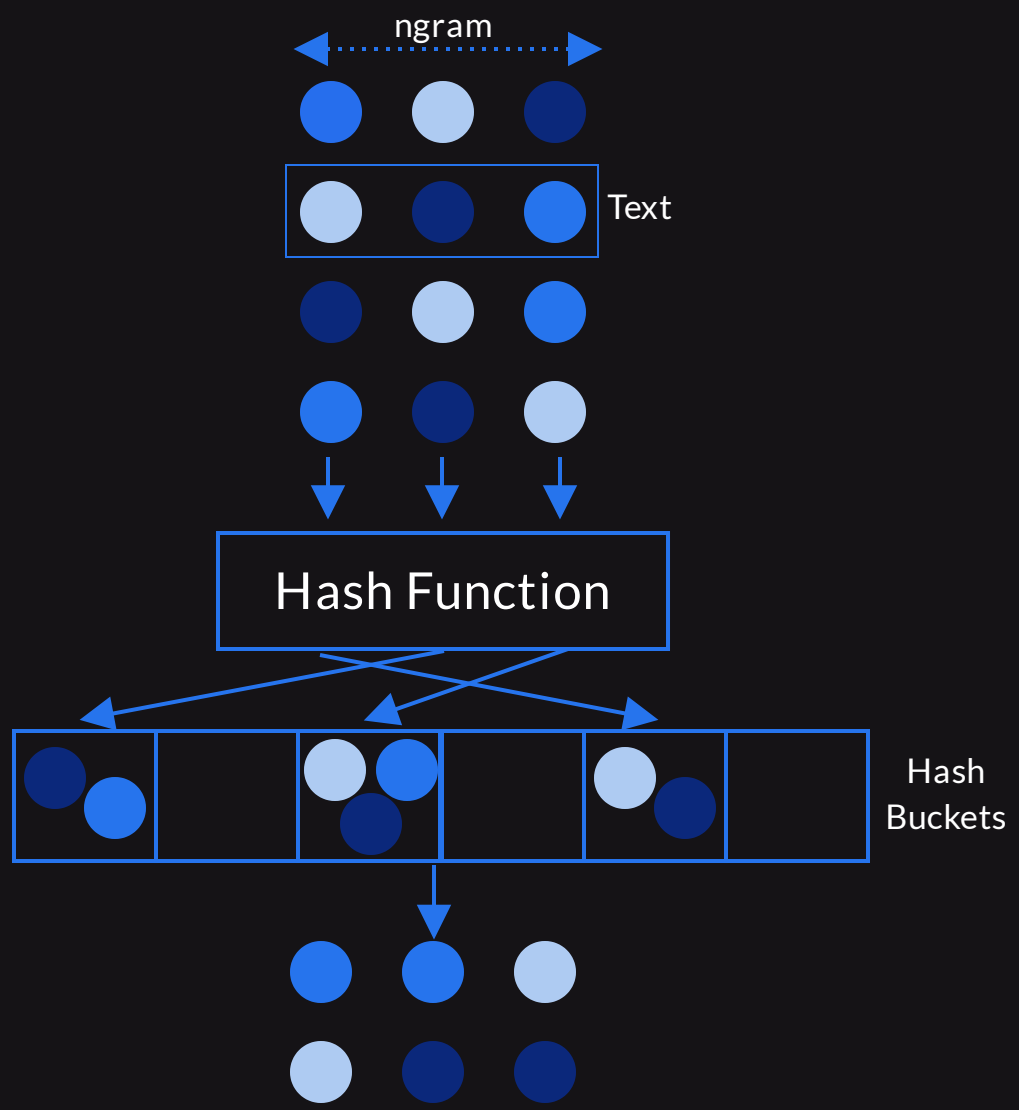
Limitation

Not Scalable

MinHash

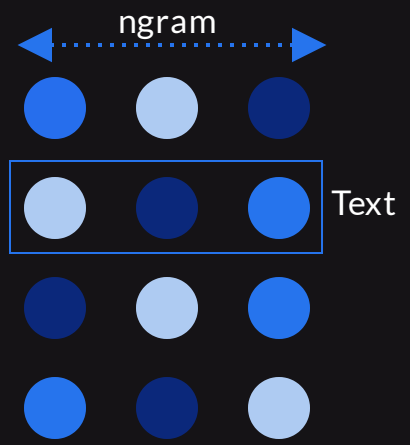
MinHash

Novel technique to compute the jaccard similarity based on Hashing



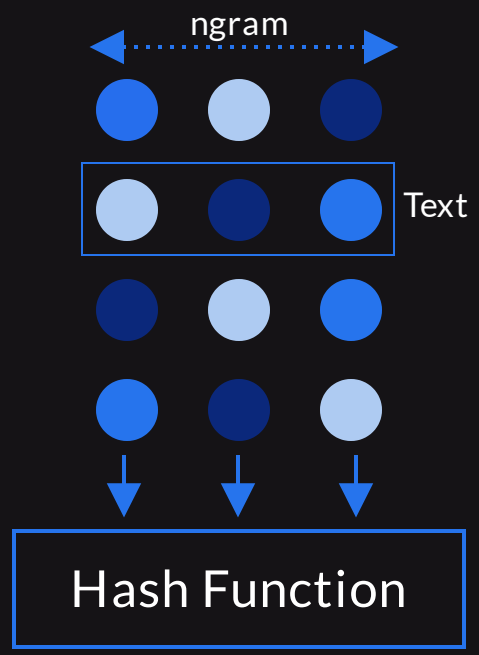
MinHash

- Tokenization



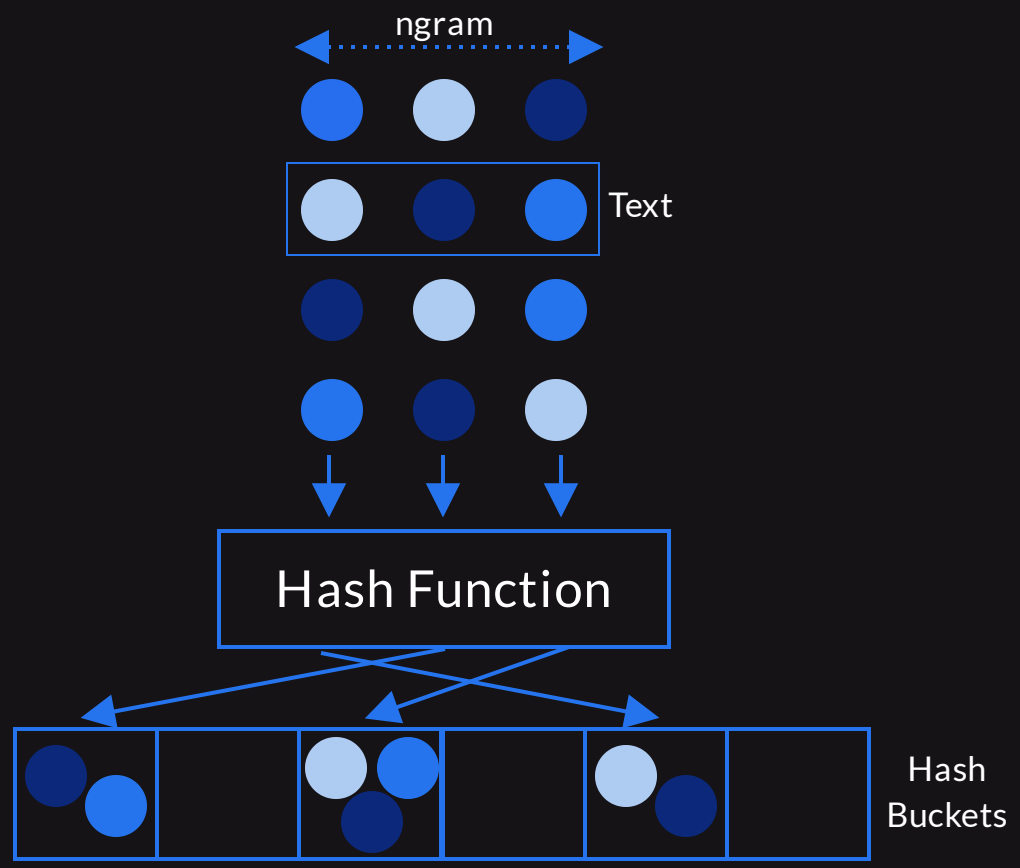
MinHash

- Tokenization
- Fingerprinting



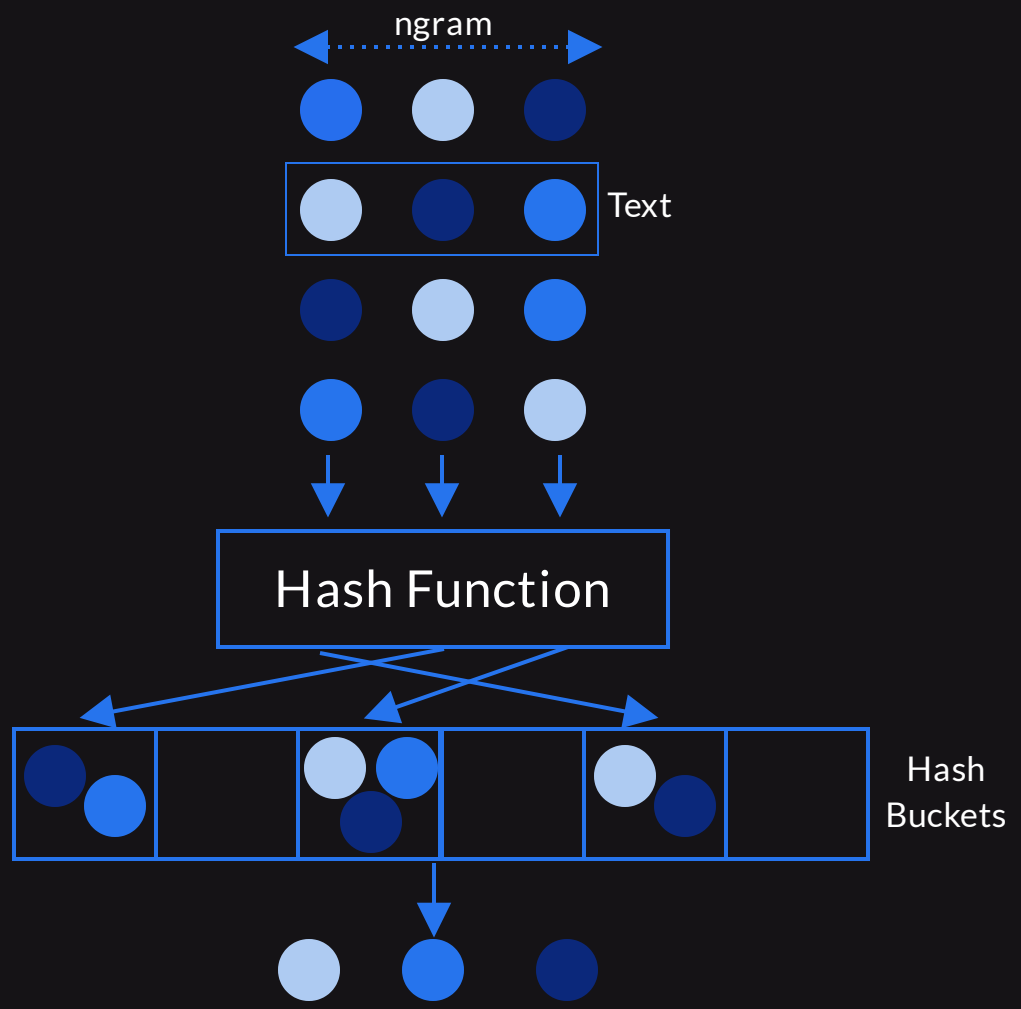
MinHash

- Tokenization
- Fingerprinting
- Locality-sensitive hashing (LSH)



MinHash

- Tokenization
- Fingerprinting
- Locality-sensitive hashing (LSH)
- Duplicate removal



Common Data Preprocessing Steps

- Removing Boilerplate text
- Eliminating HTML code
- Removing bias/harmful content

What is it for you?

1. Data Sampling
2. Data Deduplication
3. Specific data preprocessing steps as per your data

Thank You

Analytics Vidhya