Derivative Security using AI

Zia Babar @ IEEE Toronto, February 2021

https://www.linkedin.com/in/zbabar/
https://twitter.com/ziababar

Abstract

Enterprises are increasingly storing large volumes of unstructured data. However, irrespective of the data format or type, unstructured data is difficult to secure and control its transfer. This is a major problem due to evolving compliance policies and the need to adhere to standards such as GDPR.

Through derivative data security practices, enterprises can utilize machine learning and deep learning techniques to determine and trace clones and derivatives of unstructured data across the enterprise. In this talk, Zia Babar will provide a background on data security approaches, and provide a demonstration on machine learning and deep learning techniques can be used for providing derivative data security.

About the Presenter...

Zia has over 20 years of professional industry experience. He has deep expertise in the design, development and deployment of enterprise applications, data engineering platforms and distributed systems, with a particular focus on incorporating machine learning practices and cognitive services into software applications.

Zia obtained his PhD from the Faculty of Information, University of Toronto.

Background

- Enterprise are increasingly expected to adhere to data governance requirements.
- These governance requirements may be mandated by industry regulators OR industry practices.

Regulations Governing Enterprise Data

- HIPAA (Health Insurance Portability and Accountability Act of 1996)
  - United States legislation that provides data privacy and security provisions for safeguarding medical information.
- Payment Card Industry Data Security Standard (PCI DSS)
  - Payment Card Industry Security Standards Council (PCI SSC) defined set of compliance requirements to safeguard credit card transactions and consumer personal and financial data
- Gramm-Leach-Bliley Act (GLBA)
  - U.S. federal law that focuses on ensuring financial institutions communicate, clearly, how they are protecting customer data.

Regulations Governing Enterprise Data

- Sarbanes-Oxley Act (SOX)
  - The key areas which impact data security are around the way that electronic records are stored, including optionally data encryption.
- Payment Services Directive (PSD2)
  - EU directive that places emphasis on the transfer of data during end-to-end payments.
- Basel III
  - Sets out measurements around financial risk and management, including securing critical infrastructure, managing operational security, and protecting confidential data access.

GDPR

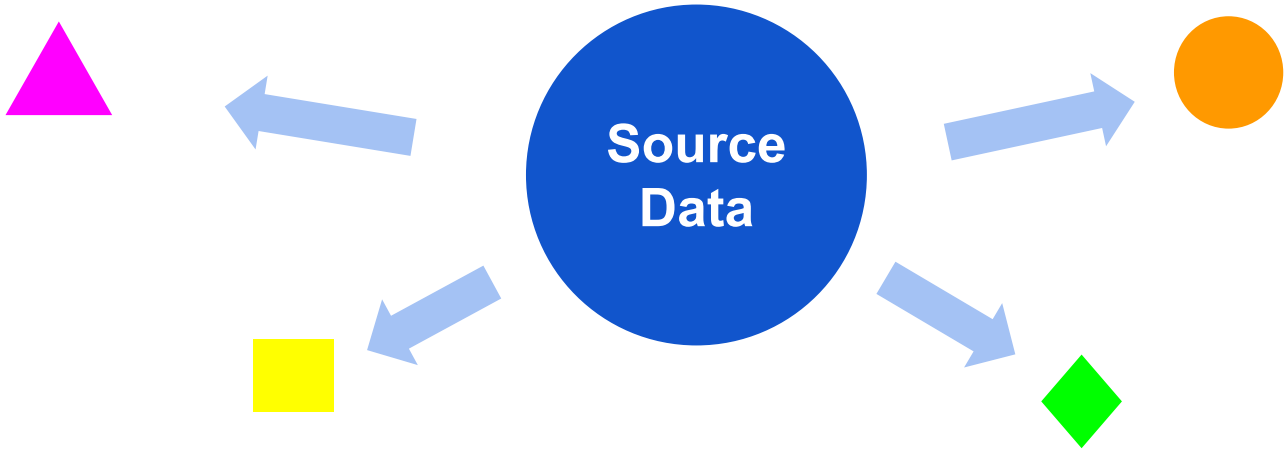
- General Data Protection Regulation (GDPR) is a set of rules designed to give EU citizens more control over their personal data.
- Aims is to simplify the regulatory environment for business so both citizens and businesses in the European Union can fully benefit from the digital economy.

Impact of GDPR on Organizations

- Organisations have to ensure that personal data is gathered legally and under strict conditions.
- Organisations that collect and manage data are obliged to protect it from misuse and exploitation
- Organisations have to respect the rights of data owners - or face penalties for not doing so.
- Organisations have to provide a '**right to be forgotten**' to individuals who no longer want their personal data processed, or wish for it to be deleted.

However!

The data is often not in a central data repository. In fact, it's replicated and transformed across many applications, endpoints and servers across the enterprise.



Derivative Data Security

*So by **Derivative Data Security** we mean the data security practices and mechanisms that have evolved to detect and protect derivatives from primary data sources.*

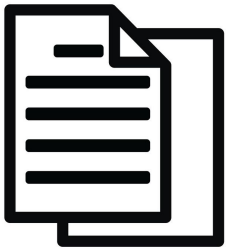
Derivative Data

- Derivative data is data that has been **derived** from primary data sources, particularly unstructured data
- For example,
  - An image that has been emailed to someone.
  - A text document that has been shared within the organization.
- It would also include unstructured data that has been modified from a source.
  - An image being cropped.
  - Portion of a text file being copied.
  - The format of a video file being changed.

Solving Derivative Data Security through ML

- Machine Learning and Deep Learning techniques are particularly helpful in processing large volumes of unstructured data.
- Particularly, they can be used to find similarities between unstructured data types belonging to the same category.
- Example,
    - Are two images similar?

Determining Derivatives of Text Documents



doc1.txt

doc2.txt

doc3.txt

All text was generated by an AI bot at https://deepai.org/machine-learning-model/text-generator

Determining Derivatives of Text Documents

- Step 1: Determine the corpus.
- Step 2: Convert the sentences in our corpus into tokens (individual words).
- Step 3: Create a dictionary of word frequency
- Step 4: Find the TF-IDF values for all the words in each sentence (across all documents)
- Step 5: TF-IDF values for each sentence are stored as vectors in the database.

TF-IDF

- TF-IDF stands for Term Frequency-Inverse Document Frequency.
- In TF-IDF, words that are more common in one sentence and less common in other sentences should be given higher heights.T

- "I like to play football"
- "Did you go outside to play tennis"
- "John and I play tennis"

Source: https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/

TF-IDF

- First we tokenize document

Sentence 1	Sentence 2	Sentence 3
I	Did	John
like	you	and
to	go	I
play	outside	play
football	to	tennis
	play	
	tennis	

Source: https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/

TF-IDF

- Then we calculate Term Frequency,
  - TF = Frequency of the word in a document / Total number of words in that document

Word	Frequency
play	3
tennis	2
to	2
I	2
football	1
Did	1
you	1

Source: https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/

TF-IDF

- Then we calculate Inverse Document Frequency,
  - IDF = Total number of documents / Number of documents containing the word

Word	Frequency	IDF
play	3	3/3 = 1
tennis	2	3/2 = 1.5
to	2	3/2 = 1.5
I	2	3/2 = 1.5
football	1	3/1 = 3

Source: https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/

TF-IDF

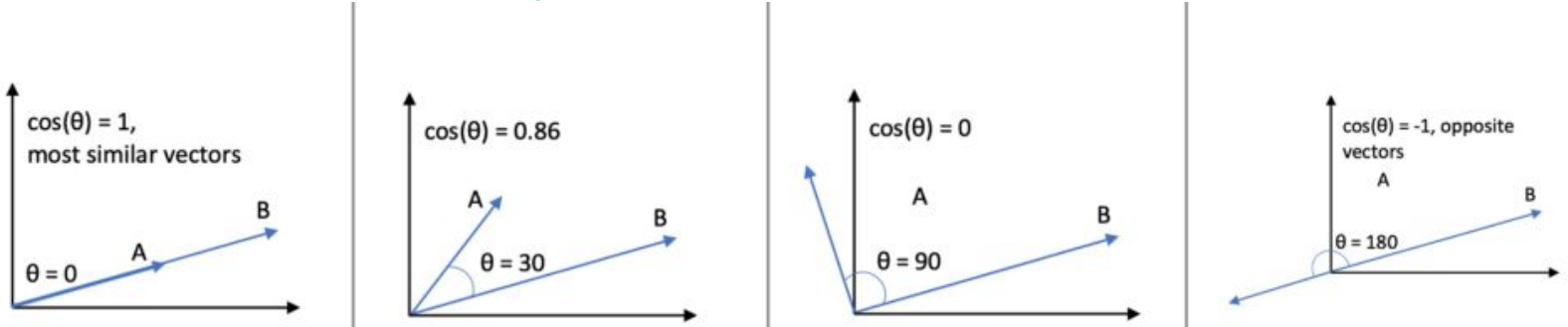
- TF-IDF
  - Multiply TF values with the corresponding IDF values

Word	Sentence 1	Sentence 2	Sentence 3
play	0.20 x 1 = 0.20	0.14 x 1 = 0.14	0.20 x 1 = 0.20
tennis	0 x 1.5 = 0	0.14 x 1.5 = 0.21	0.20 x 1.5 = 0.30
to	0.20 x 1.5 = 0.30	0.14 x 1.5 = 0.21	0 x 1.5 = 0
I	0.20 x 1.5 = 0.30	0 x 1.5 = 0	0.20 x 1.5 = 0.30
football	0.20 x 3 = 0.6	0 x 3 = 0	0 x 3 = 0

Source: https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/

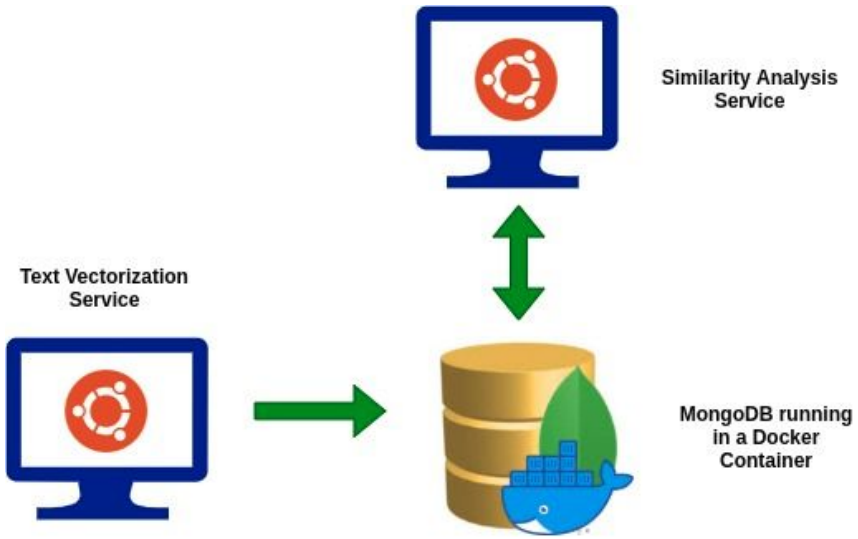
Similarity Analysis

- Step 1: Similarity between two vectors can be determined by calculating the cosine of the angle between the two vectors.
- Step 2: For each sentence vector, cycle through all the other sentence vectors and calculate cosine similarity.

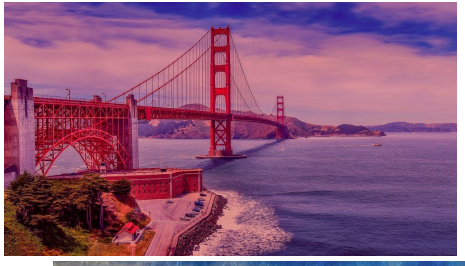


Source: https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db

Demo....


Similarity Analysis Service
Text Vectorization Service
MongoDB running in a Docker Container

Determining Derivatives of Image Documents



SIFT

- SIFT (Scale Invariant Feature Transform) is a feature detection algorithm in Computer Vision.
- SIFT helps locate the local features (keypoints) in an image. Keypoints are scale & rotation invariant that can be used for image matching.
- Keypoints are not affected by the size or orientation of the image (unlike over edge or hog features).

SIFT

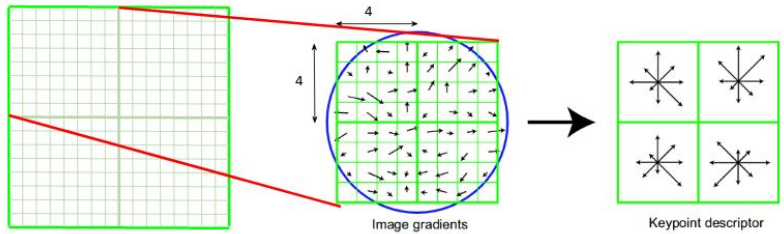
- Step 1: Constructing a Scale Space to make sure that features are scale-independent.

SIFT

- Step 2: Identifying the suitable features or keypoints.

SIFT

- Step 3: Ensure the keypoints are rotation invariant
- Step 4: Assign a unique fingerprint to each keypoint to come up with descriptors

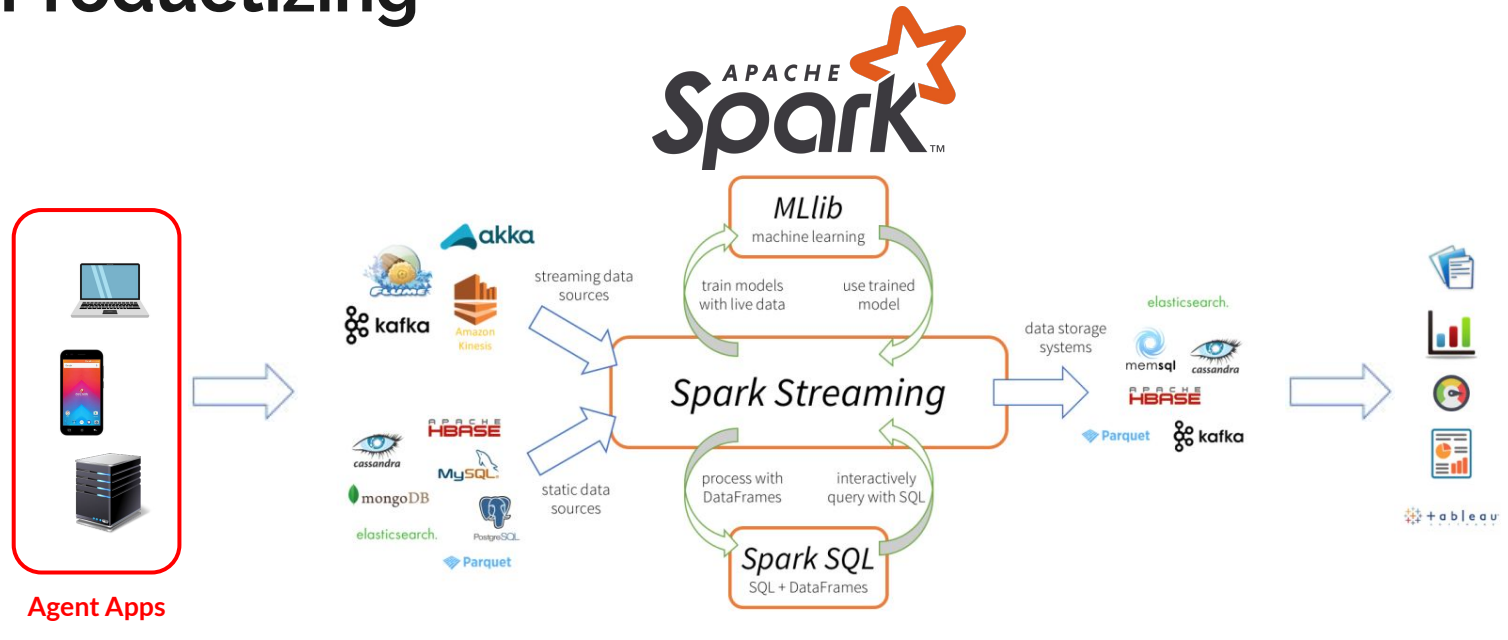


Source:
https://www.analyticsvidhya.com/blog/2019/10/detailed-guide-powerful-sift-technique-image-matching-python/

Demo...

- Python Programs
- MongoDB
-

Productizing



Agent Applications

- Continuous monitor and scan all the unstructured data on the endpoint.
  - Laptops
  - PCs
  - Mobiles
  - Servers
- The convert the unstructured document file into a numeric form.
- This is then sent to the central data platform for processing.
- This data is compared against data in the central data stores.

Conclusions

- Solving the problem of securing derivative data is an important for enterprises.
- Organizations can leverage the power of AI to detect derived data by actively scanning data storage areas.
- These can then be centrally present through some solution offering.

Thank You!

Zia Babar

https://www.linkedin.com/in/zbabar/
https://twitter.com/ziababar