[3] Вологдин Э.И. Методы и алгоритмы обработки звуковых сигналов / – СПб.: СПбГУТ, 2009. – 96 с.

[4] Петровский А.А. Анализаторы речевых и звуковых сигналов: Методы, алгоритмы и практика / – Минск: Бестпринт, 2009. – 456 с.

Roman Saveliev, Mykhailo Dendiuk*

Ph.D. student, Ukrainian National Forestry University, Lviv, Ukraine.
romansavelyev6@gmail.com.

*senior teacher of CS Department, Ukrainian National Forestry University, Lviv, Ukraine. https://orcid.org/0000-0002-7631-022X. dendyuk@nltu.edy.ua.

Application of synthetic data for analytics with generative ai and vector search capabilities

Abstract – This research investigates the transformative potential of Generative Artificial Intelligence (Generative AI) in the realm of data analytics within distributed systems. Specifically, we explore the utilization of Generative AI models, such as Generative Adversarial Networks, to generate synthetic datasets that closely mimic real-world data. Our focus extends to the innovative application of vector search within these synthetic datasets, examining its implications for data analytics and exploring the ethical considerations surrounding the use of synthetic data.

Keywords – Generative AI, Data Analytics, Distributed Systems, Synthetic Data, Vector Search

Generative AI, a groundbreaking technology, introduces a new paradigm in data analytics within distributed systems. The research centers around the application of Generative AI models, for instance Generative Adversarial Networks, to create synthetic datasets that closely mirror real-world data.

Generative AI addresses the constraint of limited access to diverse datasets by enabling the creation of synthetic data. The study delves into the technical intricacies of synthetic data generation, emphasizing the preservation of statistical characteristics from the original datasets.

* науковий керівник

Let G(z) be the generator function of the Generative Adversarial Network (GAN), mapping a latent space vector z to synthetic data. The generator aims to approximate the true data distribution p_{data} . The training process involves minimizing the Jensen-Shannon divergence, formulated as:

$$min_G max_D V(D, G) = E_{data}[\log D(X)] + E_z[\log(1 - D(G(z)))]$$

Here, D(x) represents the discriminator's output on real data, and $D(G(z))$ is the output on synthetic data. This adversarial process converges to a point where the synthetic data distribution is indistinguishable from the real data distribution.

Our research introduces vector search as a novel application within synthetic datasets. Assuming a synthetic dataset of customer profiles generated by a Generative AI model, a vector search query aims to find customers with similar preferences to a given profile:

```
DECLARE @GivenProfileVector NVARCHAR(MAX);
SET @GivenProfileVector = '[0.2, 0.8, -0.5, 0.1]';

CREATE TABLE #SimilarCustomers (
    CustomerID INT,
    SimilarityScore FLOAT
);

INSERT INTO #SimilarCustomers (CustomerID, SimilarityScore)
SELECT
    CustomerID,
    dbo.VECTOR_SEARCH(ProfileVector, @GivenProfileVector) AS SimilarityScore
FROM
    SyntheticCustomers
ORDER BY
    SimilarityScore DESC;

SELECT TOP 10 *
FROM
    #SimilarCustomers
ORDER BY
    SimilarityScore DESC;

DROP TABLE #SimilarCustomers;
```

Fig. 1 – SQL script that calculates the Similarity score

* науковий керівник

VECTOR_SEARCH Function can be implemented by incorporating cosine similarity as a measure of vector similarity. This implementation considers the vectors as arrays for better representation:

```sql
-- cosine similarity
CREATE FUNCTION dbo.VECTOR_SEARCH(@Vector1 NVARCHAR(MAX), @Vector2 NVARCHAR(MAX))
RETURNS FLOAT
AS
BEGIN
    DECLARE @Result FLOAT;
    DECLARE @DotProduct FLOAT;
    DECLARE @Magnitude1 FLOAT;
    DECLARE @Magnitude2 FLOAT;

    -- Convert the input vectors from JSON arrays to tables
    DECLARE @Table1 TABLE (Value FLOAT);
    DECLARE @Table2 TABLE (Value FLOAT);

    INSERT INTO @Table1
    SELECT Value
    FROM OPENJSON(@Vector1) WITH (Value FLOAT);

    INSERT INTO @Table2
    SELECT Value
    FROM OPENJSON(@Vector2) WITH (Value FLOAT);

    -- Calculate dot product
    SELECT @DotProduct = SUM(t1.Value * t2.Value)
    FROM @Table1 t1
    INNER JOIN @Table2 t2 ON t1.[key] = t2.[key];

    -- Calculate magnitudes
    SELECT @Magnitude1 = SQRT(SUM(Value * Value))
    FROM @Table1;

    SELECT @Magnitude2 = SQRT(SUM(Value * Value))
    FROM @Table2;

    -- Calculate cosine similarity
    IF (@Magnitude1 > 0 AND @Magnitude2 > 0)
        SET @Result = @DotProduct / (@Magnitude1 * @Magnitude2);
    ELSE
        SET @Result = 0.0; -- Default to 0 if magnitudes are 0

    RETURN @Result;
END;
```

Fig. 2 – SQL function that shows the implementation of Vector Search function

So the output of the VECTOR_SEARCH Function would be the following:

* науковий керівник

CustomerID	SimilarityScore
101	0.92
204	0.89
307	0.85
512	0.82
609	0.78
718	0.75
815	0.72
923	0.68
104	0.65
206	0.62

Fig. 3 – Output of the Similarity search function

Privacy and Security Implications. Our research further delves into the privacy and security implications of using synthetic data. By generating datasets that maintain statistical nuances without revealing sensitive information, organizations can conduct robust analyses while adhering to privacy regulations.

Augmentation of Labeled Data**.** An exploration into the scarcity of labeled data in machine learning tasks reveals the potential of Generative AI in generating synthetic labeled datasets. This augmentation enhances the training of models, leading to improved performance and generalization on real-world data.

Conclusion. The research reveals the transformative impact of Generative Artificial Intelligence (Generative AI) in distributed systems' data analytics. From crafting realistic synthetic datasets with Generative Adversarial Networks (GANs) to innovative vector search applications, the study unveils significant technical advancements.

Generative AI, particularly GANs, proves highly effective in creating synthetic datasets that closely resemble real-world data, maintaining statistical fidelity. The introduction of vector search as an application within synthetic datasets represents a pioneering approach to data analytics. Ethical considerations encompass privacy and security implications, emphasizing responsible and compliant use.

Looking ahead, the research advocates ongoing exploration and collaborative efforts to harness Generative AI's full potential in data analytics. Future studies may

* науковий керівник

delve into refining synthetic data generation techniques, advancing vector search methodologies, and addressing emerging ethical challenges. In essence, this research shapes a dynamic interplay between technological advancements and ethical considerations, ensuring Generative AI's positive contribution to societal progress and knowledge generation.

References

[1] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, A. (2016). Deep Learning. MIT Press.

[2] Dhillon, G. S. (2020). Generative AI Algorithms. O'Reilly Media.

[3] Brock, A., Donahue, J., & Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis.

Олег Яворський, Олександр Сторожук*, Наталія Гіссовська*

магістрант каф. КН НЛТУ України, Львів, Україна.
yavorskyy.o@nltu.lviv.ua.
*доцент каф. ІСКМ НЛТУ України, Львів, Україна. https://orcid.org/0000-0001-6566-5271. storozhuk@nltu.lviv.ua.
*асистент каф. ІСКМ НЛТУ України, Львів, Україна.
gissovska.n@nltu.edu.ua.

Розроблення інтелектуальної системи створення тестових запитань для Moodle 4

Анотація - Робота присвячена розробці десктопного застосунку для автоматизації створення великих об’ємів тестових запитань у Moodle 4 з можливістю легкого експорту та імпорту XML структури тестових запитань у віртуальне навчальне середовище Moodle. Застосунок написаний об’єктно орієнтованою мовою C#. Під час його розробки було використано інтегроване середовище Microsoft Visual Studio та його платформу користувацького інтерфейсу для створення класичних застосунків Windows – .NET Windows Forms, що і було використано, як графічну основу для інтерфейсу програми.

Ключові слова - Moodle 4, Moodle XML, Microsoft Visual Studio, .NET Windows Forms, C#.

* науковий керівник