Напредна Тема: Векторска База на Податоци

Семантичко Пребарување на Производи со pgvector

1. Вовед

Класичното SQL пребарување со LIKE оператор бара точно совпаѓање на зборови — ако корисникот напише "топла јакна за зима", системот нема да најде производ со наслов "warm winter jacket". Векторската база го решава овој проблем преку семантичко пребарување.

Секој производ се претвора во нумерички вектор (embedding) кој го претставува неговото значење. Кога корисникот пребарува, неговиот текст исто се претвора во вектор и се бараат производи чии вектори се математички најблиски — без оглед на точните зборови.

За имплементација се користат:

- pgvector — PostgreSQL екстензија за векторски операции
- sentence-transformers — Python библиотека за генерирање embeddings
- Модел all-MiniLM-L6-v2 — компактен AI модел кој генерира вектори со 384 димензии

Практична примена во MarketNet:

- Семантичко пребарување — "евтин телефон со добра камера" наоѓа релевантни производи
- Слични производи — "корисниците кои го гледале ова, би гледале и..."

2. Инсталација на pgvector

pgvector е инсталиран на локална PostgreSQL 18 инсталација на Windows преку Visual Studio C++ Build Tools:

По успешна инсталација, екстензијата се активира во базата:

```
CREATE EXTENSION vector;
```

3. Креирање на табелата за embeddings

Се креира нова табела product_embeddings која чува текстуален опис и векторски embedding за секој производ:

```
CREATE TABLE product_embeddings (
    product_id INT PRIMARY KEY REFERENCES product(product_id),
    embedding_text TEXT NOT NULL,
    embedding VECTOR(384)
);
```

VECTOR(384) е колона која чува вектор со 384 бројни вредности — димензијата на моделот all-MiniLM-L6-v2.

4. Пополнување на embedding_text

4.1 Кои карактеристики се користат

За секој производ се комбинираат следните информации во еден текст:

- title — насловот на производот (главен идентификатор)
- description — описот на производот
- category name > parent category — хиерархија на категоријата
- attribute_name: attribute_value — сите атрибути (боја, бренд, материјал, состојба итн.)
- location — градот каде се наоѓа производот
- price + currency — цената

Колоните кои НЕ се вклучени (немаат семантичко значење): product_id, seller_id, created_at, is_active, quantity, category_id.

4.2 INSERT на embedding_text

```
INSERT INTO product_embeddings (product_id, embedding_text)
SELECT
    p.product_id,
    p.title || '. ' || p.description ||
    ' Category: ' || c.name ||
    COALESCE(' > ' || pc.name, '') || '. ' ||
    COALESCE(string_agg(ca.attribute_name || ': ' || pa.attribute_value, ', '), '') ||
    '. Location: ' || COALESCE(p.location, '') ||
    '. Price: ' || p.price::text || ' ' || p.currency
FROM product p
JOIN category c ON c.category_id = p.category_id
LEFT JOIN category pc ON pc.category_id = c.parent_category
LEFT JOIN productattributes pa ON pa.product_id = p.product_id
LEFT JOIN categoryattributes ca ON ca.category_attribute_id = pa.category_attribute_id
GROUP BY p.product_id, p.title, p.description, c.name, pc.name, p.location, p.price, p.currency;
```

Пример на генериран embedding_text:

	product_id [PK] integer	embedding_text text
1	2294	Jackets - 7b66b4fd. Quality Jackets product. Category: Jackets > Clothing. Brand: Nike, Condition: New, Size: XS, Color: Red, Gender: Male, Material: Silk, Type: Leather. Location: Bochum. Pri…

"Jackets - 7b66b4fd. Quality Jackets product. Category: Jackets > Clothing. Brand: Nike, Condition: New, Size: XS, Color: Red, Gender: Male, Material: Silk, Type: Leather. Location: Bochum. Price: 136.34 EUR"

5. Генерирање на векторски embeddings со Python

Python скриптата го вчитува AI моделот all-MiniLM-L6-v2 и за секој текст генерира вектор со 384 вредности кој го зачувува во базата:

```
from sentence_transformers import SentenceTransformer
import psycopg2
```

```
)
cur = conn.cursor()

# Вчитај модел
print("Вчитување модел...")
model = SentenceTransformer('all-MiniLM-L6-v2')

# Земи ги сите текстови
cur.execute("SELECT product_id, embedding_text FROM product_embeddings")
rows = cur.fetchall()
print(f"Генерирање embeddings за {len(rows)} производи...")

# Генерирај и зачувај embeddings
for i, (product_id, text) in enumerate(rows):
    embedding = model.encode(text).tolist()
    cur.execute(
        query: "UPDATE product_embeddings SET embedding = %s WHERE product_id = %s",
        vars: (embedding, product_id)
    )
    if i % 100 == 0:
        print(f"Обработено: {i}/{len(rows)}")
        conn.commit()

conn.commit()
cur.close()
conn.close()
print("Готово!")
```

Проверка дека embeddings се генерирани успешно:

```
SELECT product_id, embedding_text,
       embedding IS NOT NULL AS has_embedding
FROM product_embeddings
LIMIT 5;
```

	product_id [PK] integer	embedding_text text	has_embe boolean
1	2024	Smart Lighting - 8232e119. Quality Smart Lighting product. Category: Smart Lighting > Smart Home. Condition: New, Brand: Generic, Power: 1000W, Connectivity: Wired. Location: Catania. Price: 1…	true
2	2025	Security Cameras - 7b1ce3d7. Quality Security Cameras product. Category: Security Cameras > Smart Home. Connectivity: Bluetooth, Brand: Arlo, Resolution: 2K, Condition: Used. Location: Konya. …	true
3	2026	Smart Sensors - e7e23670. Quality Smart Sensors product. Category: Smart Sensors > Smart Home. Brand: Generic, Connectivity: WiFi, Sensor Type: Door/Window, Condition: Used. Location: Paler…	true
4	2027	Home Automation Devices - 24f0d2c9. Quality Home Automation Devices product. Category: Home Automation Devices > Smart Home. Brand: Generic, Condition: New, Type: Hub, Connectivity: Bl…	true
5	2028	Cardio Equipment - c6335734. Quality Cardio Equipment product. Category: Cardio Equipment > Fitness. Brand: Life Fitness, Resistance Level: Light, Type: Jump Rope, Condition: Used. Location: A…	true

6. Преглед на embedding за конкретен производ

Преглед на embedding запис за производ со ID 2294 (Nike јакна):

```
select *
from product_embeddings
where product_id = 2294
```

	product_id [PK] integer	embedding_text text
1	2294	Jackets - 7b66b4fd. Quality Jackets product. Category: Jackets > Clothing. Brand: Nike, Condition: New, Size: XS, Color: Red, Gender: Male, Mat

embedding vector
[-0.050679285,0.06952143,0.0018344288,0.037880242,0.06871441,-0.015172314,0.13004196,0.0434043,-0.0005284489,0.035887465,0.029310467,-0.018464083,0.03244497,-0.062211964,0.026

ПРИМЕНА

1. Семантичко пребарување по текст

Python скриптата го претвора текстот за пребарување во вектор и ги наоѓа најсличните производи:

```
cur = conn.cursor()

model = SentenceTransformer('all-MiniLM-L6-v2')

query = "diesel car low mileage"

embedding = model.encode(query).tolist()

cur.execute( query: """
    SELECT pe.product_id, pe.embedding_text,
           1 - (pe.embedding <=> %s::vector) AS similarity
    FROM product_embeddings pe
    ORDER BY pe.embedding <=> %s::vector
    LIMIT 5
""", vars: (embedding, embedding))

results = cur.fetchall()
for product_id, text, similarity in results:
    print(f"\nProduct ID: {product_id}")
    print(f"Similarity: {similarity:.4f}")
    print(f"Text: {text[:150]}")

cur.close()
conn.close()
```

Операторот <=> пресметува cosine растојание. Similarity = 1 - растојание, каде 1.0 е совршено совпаѓање.

Резултати

Влез:

```
query = "diesel car low mileage"
```

Излез:

```
Product ID: 3224
Similarity: 0.4188
Text: Cars - 1a04f965. Quality Cars product.  Category: Cars > Automotive. Mileage: 0-10,000 km, Condition: New, Brand: Volkswagen, Model: A6, Fuel Type: Pe

Product ID: 4725
Similarity: 0.3853
Text: Cars - 522e1ea4. Quality Cars product.  Category: Cars > Automotive. Fuel Type: Diesel, Year: 2007, Transmission: Automatic, Mileage: 0-10,000 km, Con

Product ID: 2118
Similarity: 0.3821
Text: Cars - 1145a30f. Quality Cars product.  Category: Cars > Automotive. Fuel Type: Diesel, Condition: Used, Brand: Mercedes-Benz, Model: A4, Mileage: 10,

Product ID: 4804
Similarity: 0.3733
Text: Cars - 13bf4a96. Quality Cars product.  Category: Cars > Automotive. Brand: Volkswagen, Transmission: Automatic, Year: 1994, Condition: New, Model: Ci

Product ID: 3935
Similarity: 0.3712
Text: Cars - fc761507. Quality Cars product.  Category: Cars > Automotive. Brand: Volkswagen, Year: 2024, Transmission: Manual, Mileage: 50,000-100,000 km,
```

Влез:

```
query = "mountain bike for outdoor cycling"
```

Излез:

```
Product ID: 136 | Similarity: 0.5558
Cycling - a8baa565. Quality Cycling product.  Category: Cycling > Outdoor. Type: Mountain Bike, Brand: The North Face, Gear Count: 24, Frame Material: Carbon Fiber, Condition: New. Location: Oradea. P

Product ID: 2427 | Similarity: 0.5518
Cycling - 070dbb60. Quality Cycling product.  Category: Cycling > Outdoor. Condition: Used, Brand: Columbia, Type: Mountain Bike, Frame Material: Steel, Gear Count: 3. Location: Verona. Price: 2364.32

Product ID: 4797 | Similarity: 0.5336
Cycling - 4ffbd5c8. Quality Cycling product.  Category: Cycling > Outdoor. Gear Count: 27, Brand: Salomon, Frame Material: Aluminum, Type: Road Bike, Condition: Used. Location: Olomouc. Price: 2973.88
```

Влез:

```
query = "comfortable sofa for living room"
```

Излез:

```
Product ID: 2246 | Similarity: 0.5700
Furniture & Seating - f7e0b956. Quality Furniture & Seating product.  Category: Furniture & Seating > Home Living. Type: Sofa, Color: Black, Brand: Vox, Condition: New, Material: Glass. Location: Reim

Product ID: 2088 | Similarity: 0.5329
Furniture & Seating - bd0cc810. Quality Furniture & Seating product.  Category: Furniture & Seating > Home Living. Condition: Used, Brand: Tempur, Color: Black, Material: Leather, Type: Sofa. Location

Product ID: 2957 | Similarity: 0.5107
Furniture & Seating - 6917ff2a. Quality Furniture & Seating product.  Category: Furniture & Seating > Home Living. Type: Sofa, Brand: Kika, Color: Gold, Material: Leather, Condition: Used. Location: R
```

2. Пребарување на слични производи

Наместо текст, се зема embedding на постоечки производ и се бараат семантички слични:

```
product_id = 2294

# Земи го неговиот embedding
cur.execute( query: "SELECT embedding_text, embedding FROM product_embeddings WHERE product_id = %s", vars: (product_id,))
row = cur.fetchone()
print(f"Базен производ: {row[0][:200]}")

# Барај слични - исклучи го самиот производ
cur.execute( query: """
    SELECT pe.product_id, pe.embedding_text,
           1 - (pe.embedding <=> %s::vector) AS similarity
    FROM product_embeddings pe
    WHERE pe.product_id != %s
    ORDER BY pe.embedding <=> %s::vector
    LIMIT 5
""", vars: (row[1], product_id, row[1]))

results = cur.fetchall()
print(f"\nСлични производи:")
for pid, text, similarity in results:
    print(f"\n  Product ID: {pid} | Similarity: {similarity:.4f}")
    print(f"  {text[:200]}")

cur.close()
conn.close()
```

Резултати:

Влез:

```
product_id = 2294
```

Излез:

```
Product ID: 793 | Similarity: 0.8228
Jackets - 632cee94. Quality Jackets product.  Category: Jackets > Clothing. Material: Wool, Condition: Used, Brand: Nike, Type: Windbreaker, Size: XS, Color: White, Gender: Unisex. Location: Esbjerg.

Product ID: 4032 | Similarity: 0.8029
Jackets - b207f5c5. Quality Jackets product.  Category: Jackets > Clothing. Type: Denim, Size: M, Color: Gold, Material: Wool, Brand: Nike, Condition: Used, Gender: Female. Location: Zadar. Price: 36.

Product ID: 2531 | Similarity: 0.8025
Jackets - 4de81d91. Quality Jackets product.  Category: Jackets > Clothing. Condition: Used, Brand: Nike, Type: Leather, Size: XXXL, Color: Black, Gender: Female, Material: Denim. Location: Rijeka. Pr

Product ID: 4111 | Similarity: 0.7910
Jackets - fb8e51c5. Quality Jackets product.  Category: Jackets > Clothing. Condition: Used, Size: L, Color: Red, Gender: Unisex, Material: Denim, Type: Denim, Brand: Nike. Location: Granada. Price: 8

Product ID: 2926 | Similarity: 0.7804
Jackets - 10ce03a1. Quality Jackets product.  Category: Jackets > Clothing. Color: Brown, Gender: Female, Material: Silk, Type: Trench, Brand: Mango, Condition: New, Size: XXL. Location: Bonn. Price:
```