UNSA
UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

INGENIERIA ISTEMAS

Universidad Nacional de San Agustín de Arequipa

Escuela Profesional de Ingeniería de Sistemas

Semestre 2025-A

TÓPICOS AVANZADOS EN BASES DE DATOS (E) – Grupo “B”

Tesis
“Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector”

Presentado por: Calcina Puma Esteven Antonio

Docente asesor : ING ANTONIO ARROYO PAZ

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

- **Realidad problemática** :

  La transformación digital en el sector financiero ha incrementado la diversidad de productos bancarios, pero los sistemas de búsqueda tradicionales (basados en palabras clave o formularios rígidos) no interpretan consultas en lenguaje natural. Esto limita la precisión de los resultados y dificulta que los usuarios encuentren productos adaptados a sus necesidades (ej.: cuentas de ahorro con condiciones específicas en determinada ubicación).

- **Problema principal:**

  Los métodos actuales de búsqueda carecen de capacidad semántica, lo que reduce su eficacia en dominios especializados como finanzas, afectando tanto a usuarios como a entidades financieras.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Objetivo principal :

Implementar un sistema de búsqueda semántica (all-MiniLM-L6-v2 + pgvector) para productos de ahorro de la SBS, procesando consultas en lenguaje natural y mejorando la precisión frente a métodos tradicionales.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Objetivos específicos:

- Automatizar la extracción de datos de tasas de la SBS mediante web scraping.
- Generar embeddings semánticos (all-MiniLM-L6-v2) de productos financieros.
- Almacenar vectores en PostgreSQL/pgvector para búsquedas por similitud.
- Implementar búsquedas en lenguaje natural (ej: "cuenta en soles sin comisiones").
- Desarrollar un frontend para interacción con usuarios.
- Validar el sistema con datos reales de entidades financieras peruanas.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Trabajos relacionados

Las referencias más relevantes son:

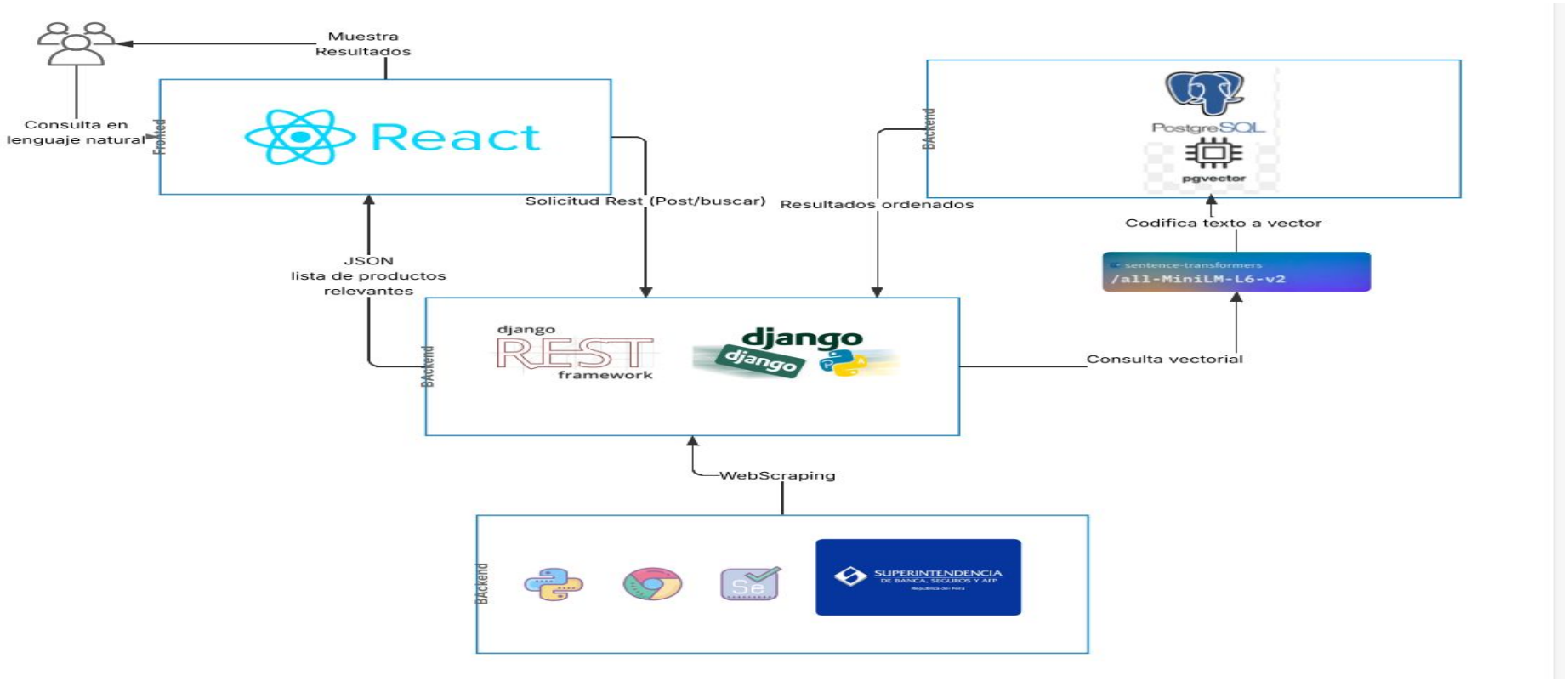
Nigam et al. [9] proponen un modelo de deep learning para búsqueda semántica en comercio electrónico, superando limitaciones de métodos léxicos. Usando embeddings compartidos y una función de pérdida "3-part hinge", su enfoque logra un Recall@100 del 79.4% en datos de Amazon, superando en 14.5% a modelos como DSSM. Destaca por su escalabilidad (entrenamiento distribuido en 8 GPUs) y aplicabilidad potencial en dominios como finanzas, aunque requiere adaptaciones para consultas coloquiales y datos estructurados.

Malmberg [10] compara embeddings neuronales (BERT, SBERT) con TF-IDF en tareas de similitud semántica. Aunque TF-IDF supera en precisión (77.2% vs. 72.9% de SBERT) en conjuntos como SQuAD 1.0, SBERT mejora la alineación semántica. El estudio resalta la vigencia de TF-IDF en contextos léxicos, pero sugiere evaluar arquitecturas modernas como ELECTRA para capturar relaciones complejas.

[9]S. Asmala, “Improving Semantic Search in Medical Literature with Sentence Transformers,” *HealthTech Proceedings*, vol. 3, no. 1, 2023.
[10]V. Mohan, Y. Song, P. Nigam, C. H. Teo, W. Ding, V. Lakshman, A. Shingavi, H. Gu, and B. Yin, “Semantic product search,” in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD), 2019, pp. 2876–2885.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Arquitectura



Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Arquitectura

1. **Frontend (React 19.1.0)**
   - Interfaz para consultas en lenguaje natural.
   - Renderiza resultados dinámicos ordenados por similitud.
2. **Backend (Django 4.2.23 + SentenceTransformers 3.2.1)**
   - Web scraping automatizado (Selenium 4.34.2) para recolectar datos financieros.
   - Procesamiento y limpieza de datos (ej: tipo de cuenta, moneda).
   - Generación de *embeddings* (all-MiniLM-L6-v2) y búsqueda por similitud coseno.
3. **Base de Datos (PostgreSQL 17.5 + pgvector 0.3.6)**
   - Almacena productos y vectores semánticos.
   - Búsquedas eficientes con índices HNSW.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Desarrollo

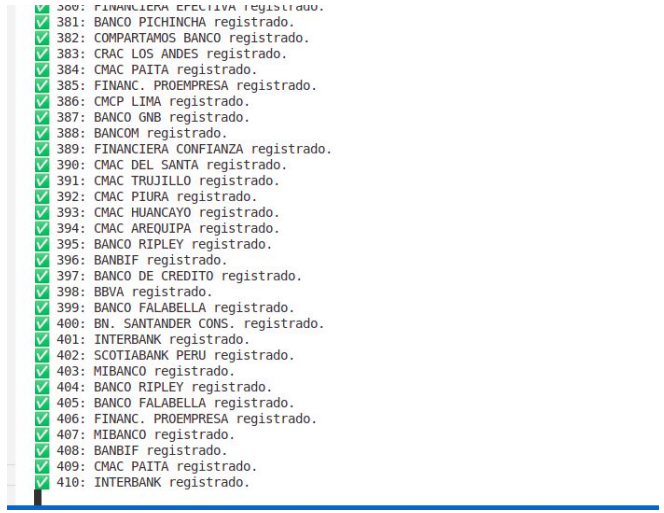
1. Extracción y Procesamiento de Datos
   - Se implementó *web scraping* con Selenium en Python para recolectar datos de productos financieros de la SBS, obteniendo 1,528 registros estructurados (Figura 2).
   - Los datos se normalizaron (ej: tipo de cuenta, moneda) y se almacenaron en PostgreSQL para su posterior análisis semántico.

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Desarrollo

Generación de *Embeddings*

- Se utilizó el modelo all-MiniLM-L6-v2 para convertir atributos clave (ubicación, entidad, tasa) en vectores de 384 dimensiones.
- La codificación se automatizó en Django, generando *embeddings* al crear/actualizar productos



Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Desarroll

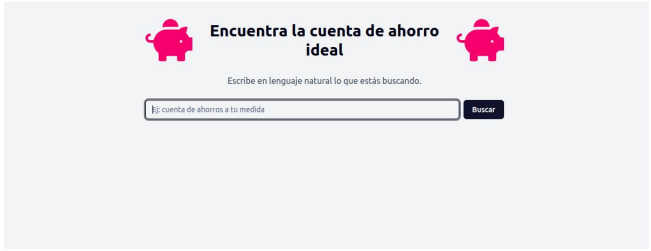
1. Almacenamiento y Búsqueda Semántica
   - PostgreSQL con pgvector almacena los vectores y ejecuta búsquedas por similitud coseno (`<=>`).
   - Un endpoint API en Django procesa consultas en lenguaje natural, devolviendo los 5 productos más relevantes .

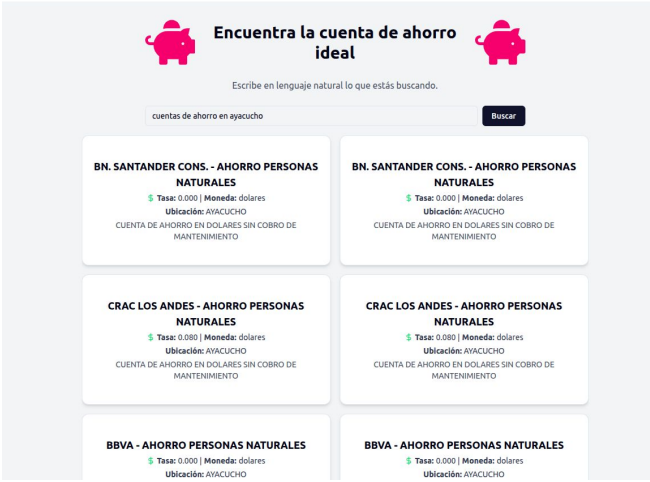
```
print(query_embedding)

# Buscar los más similares usando pgvector (asumiendo que el campo se llama 'embedding')
# Esto usa la distancia coseno (o el operador <-> depende de tu configuración)
with connection.cursor() as cursor:
    cursor.execute("""
        SELECT id, ubicacion, entidad, tasa, tipo_cuenta, condiciones, moneda, embedding
        FROM productos_productoahorrosbs
        ORDER BY embedding <#> %s::vector
        LIMIT 10;
    """, [query_embedding])
    resultados = cursor.fetchall()

# Mapear resultados a diccionarios
campos = ["id", "ubicacion", "entidad", "tasa", "tipo_cuenta", "condiciones", "moneda", "embeddinc
```

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Implementación del sistema -Desarrollo

1. Frontend (React + Tailwind CSS)
   - Interfaz responsive con búsqueda en lenguaje natural (ej: *"cuenta en soles sin mantenimiento"*).
   - Muestra resultados en tarjetas con tasas, condiciones y entidades bancarias (Figuras 5 y 6).



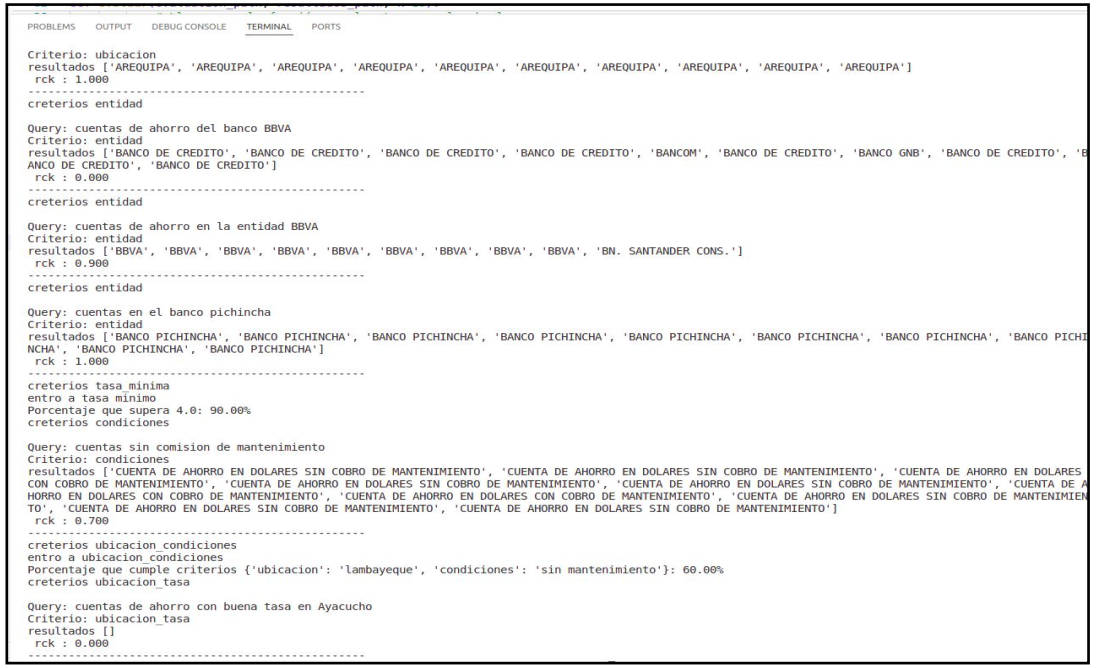


Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Resultados

Métricas Personalizadas

- Recall@k: Evalúa la capacidad de recuperar productos relevantes entre los *k* primeros resultados.
  - Consultas simples (ej: *"Arequipa"*, *"Banco Pichincha"*) lograron Recall@k = 1.0.
  - Consultas más complejas (ej: *"Ayacucho"*) obtuvieron Recall@k = 0.8 debido a resultados irrelevantes.
- %_superan_minimo: Mide cuántos resultados superan una tasa mínima deseada.
  - 90% de los productos cumplieron con una tasa ≥ 4.0%.
- %_cumplen_criterios: Evalúa coincidencia en múltiples atributos (ubicación + condiciones).
  - 60% de éxito en consultas combinadas (ej: *"sin mantenimiento en Lima"*).
  - 0% en consultas muy exigentes (ej: *"tasa alta + ubicación específica"*).



Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Resultados

Métricas Personalizadas

Hallazgos Clave

- Mayor precisión en consultas simples (ej: ubicación o entidad bancaria).
- Dificultad en consultas con múltiples restricciones (ej: *"tasa alta + condiciones específicas"*).


```
[
  {
    "query": "cuentas de ahorro en ayacucho",
    "results": [
      {
        "id": 1848,
        "ubicacion": "AYACUCHO",
        "entidad": "BN. SANTANDER CONS.",
        "tasa": "0.000",
        "tipo_cuenta": "AHORRO PERSONAS NATURALES",
        "condiciones": "CUENTA DE AHORRO EN DOLARES SIN COBRO DE MANTENIMIENTO",
        "moneda": "dolares"
      },
      {
        "id": 1843,
        "ubicacion": "AYACUCHO",
        "entidad": "CRAC LOS ANDES",
        "tasa": "0.080",
        "tipo_cuenta": "AHORRO PERSONAS NATURALES",
        "condiciones": "CUENTA DE AHORRO EN DOLARES SIN COBRO DE MANTENIMIENTO",
        "moneda": "dolares"
      },
      {
        "id": 1847,
        "ubicacion": "AYACUCHO",
        "entidad": "BBVA",
        "tasa": "0.000",
        "tipo_cuenta": "AHORRO PERSONAS NATURALES",
        "condiciones": "CUENTA DE AHORRO EN DOLARES SIN COBRO DE MANTENIMIENTO",
        "moneda": "dolares"
```

Búsqueda Semántica de Productos Financieros Usando Embeddings y pgvector

Conclusiones

Automatización efectiva del web scraping
Se logró automatizar con éxito la recolección de datos de productos financieros, eliminando la necesidad de intervención manual y garantizando actualizaciones constantes y precisas.

Procesamiento y limpieza de datos
Los datos obtenidos fueron tratados mediante técnicas de limpieza y estandarización, asegurando calidad, uniformidad y una estructura óptima para su análisis posterior.

Representación semántica mediante embeddings
Se utilizaron modelos preentrenados (all-MiniLM-L6-v2) para transformar textos financieros en vectores, permitiendo capturar similitudes contextuales entre productos.

Almacenamiento vectorial eficiente con pgvector
La integración de embeddings en PostgreSQL con pgvector permitió consultas rápidas y escalables basadas en similitud de vectores, optimizando el rendimiento del sistema.

Interpretación de lenguaje natural compleja
El sistema demostró capacidad para comprender consultas en lenguaje natural, incluso aquellas con términos implícitos o expresiones ambiguas, entregando resultados más precisos y relevantes.

Validación con métricas de búsqueda realistas
Las pruebas con datos reales evidenciaron altos valores de recall@10 y precisión, confirmando que el sistema es funcional y eficaz en contextos reales del mercado financiero peruano.

Gracias…

- "Los datos financieros no valen por lo que son, sino por lo que revelan."
  Adaptación de John Maynard Keynes