Database Table Definitions

# 1. Table: historical\_news\_data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Name | Data Type | Nullable | Description | Notes |
| article\_id | UUID | No | Unique identifier for the article | Primary key; auto-generated using gen\_random\_uuid() |
| company\_name | TEXT | No | Name of the company the article is about | Required field |
| industry\_name | TEXT | Yes | Name of the industry the company belongs to | Optional |
| type\_of\_article | TEXT | Yes | Category or classification of the article | e.g., News, Analysis, Interview |
| article\_url | TEXT | Yes | URL from which the article was extracted | Should be a valid web URL |
| article\_search\_term | TEXT | Yes | Search term used to find the article | Useful for tracing back the query context |
| article\_title | TEXT | Yes | Title of the article | Extracted from the article page |
| article\_date | DATE | Yes | Date the article was published | Format: YYYY-MM-DD |
| article\_text | TEXT | Yes | Full textual content of the article | Can be lengthy |
| article\_summary | TEXT | Yes | Summary or extract of the article | Optional, may be generated by LLM or summarization model |
| date\_of\_extraction | DATE | Yes (default) | Date the article was extracted into the system | Defaults to CURRENT\_DATE |
| relevance\_score | REAL | Yes | Numerical score indicating article's relevance | Range typically between 0 and 1 |
| article\_embedding | VECTOR(1536) | Yes | Embedding vector representing the full article content | Requires pgvector extension; used for semantic search |
| article\_summary\_embedding | VECTOR(1536) | Yes | Embedding vector representing the summary | Similar to article\_embedding, but for summary |
| inserted\_at | TIMESTAMP | Yes (default) | Timestamp when the row was inserted | Defaults to CURRENT\_TIMESTAMP |

# 2. Table: company\_websites\_data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Name | Data Type | Nullable | Description | Notes |
| id | SERIAL | No | Auto-incremented unique identifier for each record | Primary key; automatically generated by PostgreSQL |
| client\_name | VARCHAR(100) | Yes | Name of the client whose website was extracted | Can be null if client is not identified |
| date\_of\_extraction | DATE | Yes | Date when the website data was extracted | Format: YYYY-MM-DD |
| url | VARCHAR(250) | Yes | URL of the extracted web page | Should be a valid URL |
| page\_content | TEXT | Yes | Raw textual content scraped from the web page | Could be large depending on webpage length |
| page\_embedding | VECTOR(1536) | Yes | Embedding vector representing the page content | Requires pgvector; used for semantic search and similarity matching |

# 3. Table: external\_raw\_data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Name | Data Type | Nullable | Description | Notes |
| memo\_id | UUID | No | Unique identifier for the memo or record | Primary key |
| client\_name | TEXT | No | Name of the client associated with the data | Required for identifying the data owner |
| article\_id | TEXT | No | Identifier of the source article | Could be external or internal ID |
| chunk\_id | TEXT | No | Unique identifier for a text chunk within an article | Helps in breaking down large documents into smaller manageable pieces |
| chunk\_text | TEXT | No | Actual textual content of the chunk | Critical for downstream NLP tasks |
| embedding | VECTOR(1536) | Yes | Embedding vector representing the chunk text | Generated by an embedding model (e.g., Azure OpenAI) |
| text\_type | TEXT | Yes | Label/type of the text chunk (e.g., summary, body, intro) | Optional classification tag |
| url | TEXT | Yes | Source URL from where the text was extracted | Useful for tracing back to the original content |
| inserted\_at | TIMESTAMP | Yes (default) | Timestamp when the record was inserted into the system | Defaults to CURRENT\_TIMESTAMP |