Duplicate Document Detection Solution

This plan outlines a comprehensive solution for detecting duplicate documents when users add more files to an existing batch, with options to Replace, Reject, or Keep Both.

Problem Statement:

When users upload multiple documents and then add more documents to the existing batch, they might accidentally add duplicates. The system needs to:

1. Detect duplicate documents intelligently
2. Show a user-friendly modal/dialog
3. Allow users to choose: Replace, Reject, or Keep Both
4. Handle multiple duplicates gracefully

User Review Required:

IMPORTANT

Duplicate Detection Strategy: The implementation uses multiple criteria for duplicate detection:

- Filename + Size (basic check)
- File hash (MD5/SHA-256 for accurate detection)
- Content comparison (optional, for backend validation)

Please confirm which strategy you prefer or if you want a combination approach.

IMPORTANT

User Experience Flow: When duplicates are detected:

1. Upload process pauses
2. Modal shows duplicate file details (name, size, preview)
3. User chooses action for each duplicate
4. Process continues with user's choices applied

For multiple duplicates, should we show them one-by-one or in a batch list?

Backend API

[NEW] Backend Duplicate Detection Endpoint

Endpoint: `POST /api/documents/check-duplicates`

Purpose: Server-side duplicate validation before upload

Request Body:

```
{
 "files": [
     {
       "name": "document.pdf",
       "size": 1024000,
       "hash": "abc123...",
       "type": "application/pdf"
     }
 ],
 "batchId": "batch-uuid"
}
```

Response:

```
{
 "duplicates": [
     {
       "newFile": {
         "name": "document.pdf",
         "size": 1024000,
         "hash": "abc123..."
       },
   "existingFile": {
         "id": "doc-uuid",
         "name": "document.pdf",
         "size": 1024000,
         "uploadedAt": "2026-01-21T10:00:00Z",
                "url": "https://..."
            },
        "matchType": "exact" // or "similar"
      }
    ]
  }
```

[NEW] Backend Duplicate Resolution Endpoint

Endpoint: `POST /api/documents/resolve-duplicates`
Purpose: Process user decisions on duplicates
Request Body:

```
{
    "batchId": "batch-uuid",
        "resolutions": [
            {
              "existingFileId": "doc-uuid",
              "newFile": File,
              "action": "replace" // or "reject" or "keep_both"
            }
    ]
}
```

Database Schema

Add fields for duplicate tracking:

```sql
ALTER TABLE documents ADD COLUMN file_hash VARCHAR(64);
ALTER TABLE documents ADD COLUMN original_name VARCHAR(255);
ALTER TABLE documents ADD COLUMN is_duplicate BOOLEAN DEFAULT FALSE;
ALTER TABLE documents ADD COLUMN duplicate_of UUID REFERENCES documents(id);
CREATE INDEX idx_documents_hash ON documents(file_hash);
CREATE INDEX idx_documents_batch ON documents(batch_id, file_hash);
```

Verification Plan

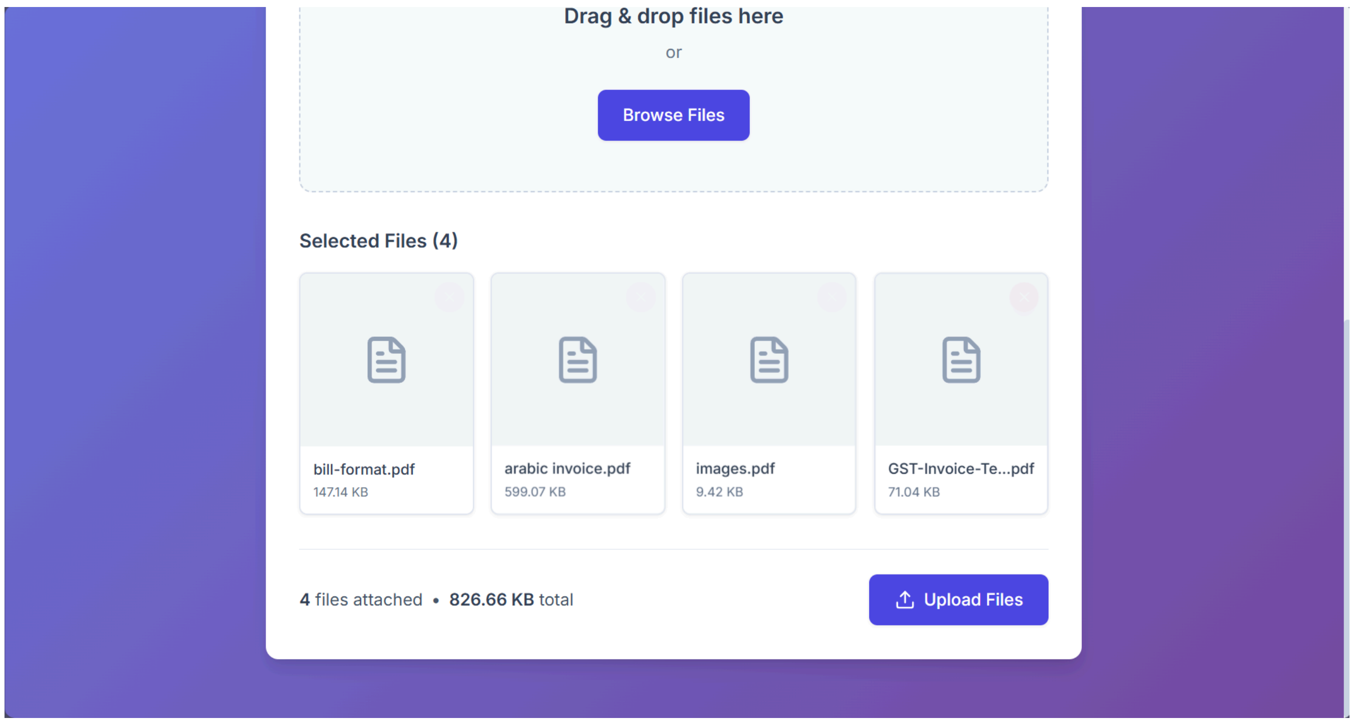
Automated Tests

1. **Frontend Unit Tests:**
   - **Test duplicate detection logic**
   - **Test modal interactions**
   - **Test file replacement logic**
   - **Test "keep both" renaming logic**
2. **Backend API Tests:**
   - **Test duplicate detection endpoint**
   - **Test hash generation accuracy**
   - **Test resolution endpoint**
   - **Test database updates**

Manual Verification

1. **Single Duplicate Scenario:**
   - **Upload 3 documents**
   - **Add 1 duplicate document**
   - **Verify modal appears**
   - **Test each action (replace, reject, keep)**
2. **Multiple Duplicates Scenario:**
   - **Upload 5 documents**
   - **Add 3 duplicates**
   - **Verify queue handling**
   - **Test mixed actions**
3. **Edge Cases:**
   - **Same name, different content**
   - **Different name, same content**
   - **Partial duplicates**
   - **Large file duplicates**
4. **User Experience:**
   - **Verify smooth animations**
   - **Check responsive design**
   - **Test keyboard navigation**
   - **Verify accessibility**

Results:


Drag & drop files here

or

Browse Files

Selected Files (4)

bill-format.pdf
147.14 KB

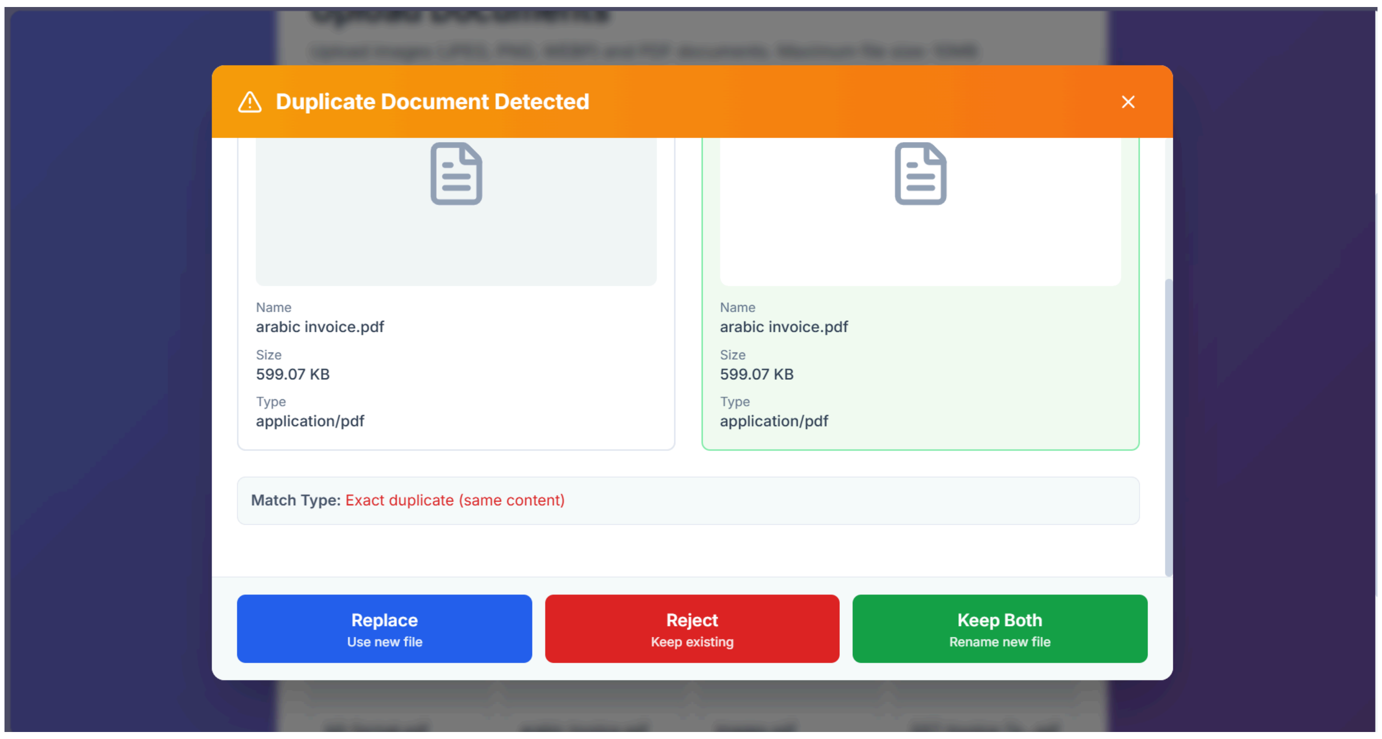
arabic invoice.pdf
599.07 KB

images.pdf
9.42 KB

GST-Invoice-Te...pdf
71.04 KB

4 files attached • 826.66 KB total

Upload Files


Duplicate Document Detected

Name
arabic invoice.pdf
Size
599.07 KB
Type
application/pdf

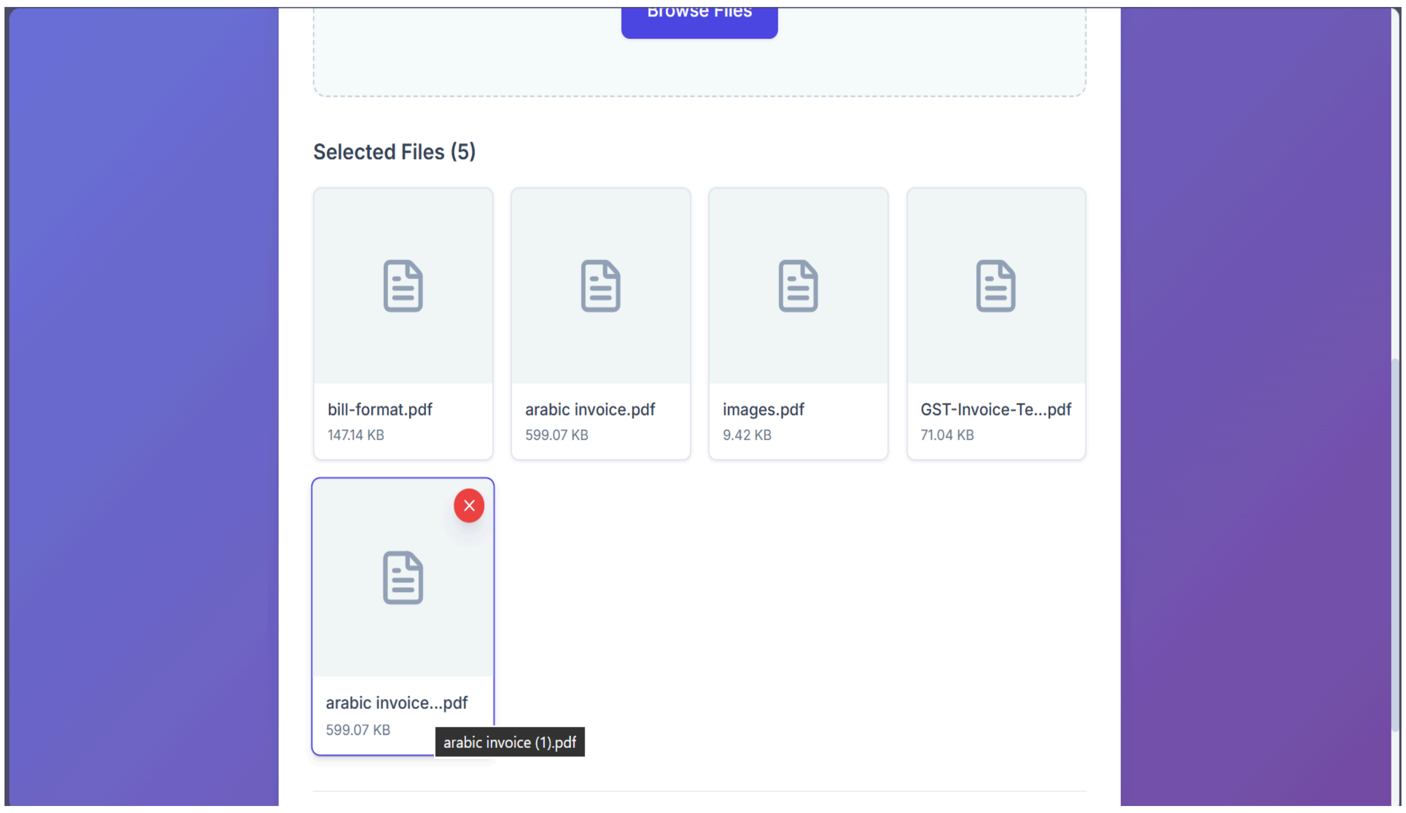
Name
arabic invoice.pdf
Size
599.07 KB
Type
application/pdf

Match Type: Exact duplicate (same content)

Replace
Use new file

Reject
Keep existing

Keep Both
Rename new file


Browse Files

Selected Files (5)

bill-format.pdf
147.14 KB

arabic invoice.pdf
599.07 KB

images.pdf
9.42 KB

GST-Invoice-Te...pdf
71.04 KB

arabic invoice...pdf
599.07 KB

arabic invoice (1).pdf