SnapPDF API

Notes

Trước mắt mình sẽ dựng 4 services mà a đã làm trong demo. Bước đầu luôn là upload files lên S3 ngay khi user pick files. Sau đó user sẽ có các lựa chọn tương ứng cho mỗi services (eg. số trang…) rồi mình bắn API cho mỗi services và trả về link S3, rồi cho browser tải về khi user click tải. Nguyên tắc theo chuẩn thị trường:

- Nếu chỉ 1 file, sẽ trả về đuôi đúng. Ví dụ tải pdf sẽ là đuôi pdf, tải ảnh jpg sẽ là đuôi jpg.
- Nếu nhiều file thì sẽ trả về đuôi zip, mở file zip ra sẽ đc các file tương ứng. Lượng cứ trả nguyên file zip cho user

Tên file tải về sẽ là 1 random string, để tránh conflict tên file trên S3. Lượng tự đổi tên lại cho phù hợp.

(gọi là s3 nhưng BE chỗ nào rẻ thì đẩy lên thôi)

upload files gốc

Dùng multipart/form-data. Upload tất cả các file và field “service” (là cái service mà user định dùng). Dựa vào cái field “service” mà BE biết để preprocess nếu có thể.

BE sẽ upload file lên s3, rồi trả về 1 session id và 1 array các string là id của mỗi file (nếu có k files, thì array có length k), theo đúng thứ tự file.

```
path: /v1/first_upload_files

input = {
  "service": str // eg. split_pdf, merge_pdf, img2pdf...

  "file_1": b"....",
  "file_2": b"....",
...
  "file_**":...
}

response = {
```

```
  "session_id": str
  "input_file_ids": [str, str, ...] // array
}
```

Ngoài ra, sau khi đã upload, user còn có ý định upload more files lúc đã ở trong màn hình chọn option.

```
path: /v1/add_more_files // Thực ra path này và path trên có thể nhập làm 1. để tính xem

input = {
  "service": str // eg. split_pdf, merge_pdf, img2pdf...
  "session_id": str

  "file_1": b"....",
  "file_2": b"....",
...
  "file_**":...
}

response = {
  "session_id": str
  "input_file_ids": [str, str, ...] // array
}
```

Split

```
path: /v1/split_pdf // Thực ra path này và path trên có thể nhập làm 1. để tính xem

input = {
  "session_id": str
  "input_file_ids": [str, str, ...]
  "split_method": str // extract, fixed, custom
  "split_rule": [str, str,...]
}

response = {
  "output_file_id": str // nếu nhiều hơn 1 file thì đã đc zip lại, nên luôn chỉ có 1 file
}
```

Gửi cả sesion_id và input_file_ids, vì user có thể bỏ bớt files ra khỏi list.

“split_method” sẽ cover đc hết mọi cases:

- “extract”: Nếu user muốn split mỗi cái ra 1 trang duy nhất (n pages → n file pdfs)

- “fixed”: to ceil(n/k) files, mỗi file có k trang: 1 đến k, k+1 đến 2k…
- “custom”: phân ra vài files pdf (có thể 1) và mỗi file theo rule nhất định

“split_rule”:

- nếu method là “fixed”, thì rule chỉ cần ví dụ "split_rule": [6], nghĩa là mỗi pdf có 6 trang: 1-6, 7-12,…
- Nếu method là “custom” thì split_rule là 1 array gồm các string, ví dụ [”1-3, 7, 9”, “12-15”]. Nghĩa là sẽ có 2 file pdf (vì array length 2), mỗi file pdf sẽ đc tạo ra từ cái string tương ứng ví dụ file đầu tiên sẽ bao gồm trang 1,2,3,7,9, file thứ 2 gồm trang 12, 13, 14, 15. BE sẽ parse string nếu a-b nghĩa là trang a đến trang b inclusive, số c đứng 1 mình nghĩa là mỗi trang c.

Merge

```
path: /v1/merge_pdf // Thực ra path này và path trên có thể nhập làm 1. để tính xem

input = {
  "session_id": str
  "input_file_ids": [str, str, ...]
}

response = {
  "output_file_id": str // nếu nhiều hơn 1 file thì đã đc zip lại, nên luôn chỉ có 1 file
}
```

jpg2pdf

```
path: /v1/jpg2pdf // Thực ra path này và path trên có thể nhập làm 1. để tính xem

input = {
  "session_id": str
  "input_file_ids": [str, str, ...]
  "page_size": str // eg. A4
  "orientation": str // portrait, landscape
  "margin": float // chắc đo bằng mm, hoặc sẽ phân ra làm to, vừa, nhỏ
  "single_output": true/false // true thì ra 1 file, false thì ra n files
}
```

```
response = {
  "output_file_id": str // nếu nhiều hơn 1 file thì đã đc zip lại, nên luôn chỉ có 1 file
}
```

pdf2jpg

```
path: /v1/pdf2jpg // Thực ra path này và path trên có thể nhập làm 1. để tính xem

input = {
  "session_id": str
  "input_file_ids": [str, str, ...]
  "extract_type": str // page hoặc image
  "quality": int // chất lượng DPI
  "extracted_pages": str // các page để extract, default là all
}

response = {
  "output_file_id": str // nếu nhiều hơn 1 file thì đã đc zip lại, nên luôn chỉ có 1 file
}
```

- Nếu extract_type là page thì extract cả page, nếu image thì lọc hết các images trong file pdf
- Nếu extracted_pages là null hoặc “all” thì extract tất cả. nếu extract vài page thì dùng string ví dụ ”1-3, 7, 9”