AWS Textract

- Textract es un servicio que nos permite realizar reconocimiento de caracteres sobre imagenes o pdf.
- Podemos invocarlo desde python.
- Nos devuelve una serie de jsons con la información del contenido del documento.
- FreeTier: 1000 paginas al mes durante 3 meses.
- Documentación:
  - https://docs.aws.amazon.com/textract/latest/dg/what-is.html
  - https://docs.aws.amazon.com/textract/latest/dg/analyzing-document-text.html

```
import boto3
aws_region = "eu-central-1"

textract = boto3.client("textract", region_name=aws_region)

textract.start_document_analysis(
     DocumentLocation={
         'S3Object': {'Bucket': bucket, 'Name': key}
     },
     FeatureTypes=["TABLES", "FORMS"],
     OutputConfig={'S3Bucket': 'outbucketeuappoc', 'S3Prefix': key}
)
```

```
import boto3
aws_region = "eu-central-1"

textract = boto3.client("textract", region_name=aws_region)

textract.start_document_text_detection(
    DocumentLocation={
        "S3Object": {
            "Bucket": in_bucket,
            "Name": in_key
        }
    },
    #NotificationChannel={"RoleArn": rolearn, "SNSTopicArn": snstarn},
    OutputConfig={"S3Bucket": out_bucket, "S3Prefix": out_prefix},
)
```

- Formato de salida: https://docs.aws.amazon.com/textract/latest/dg/API_Block.html
- Ejemplo en la carpeta: output_format.

Ejercicio

- Pruebalo usando algun pdf o imagen.