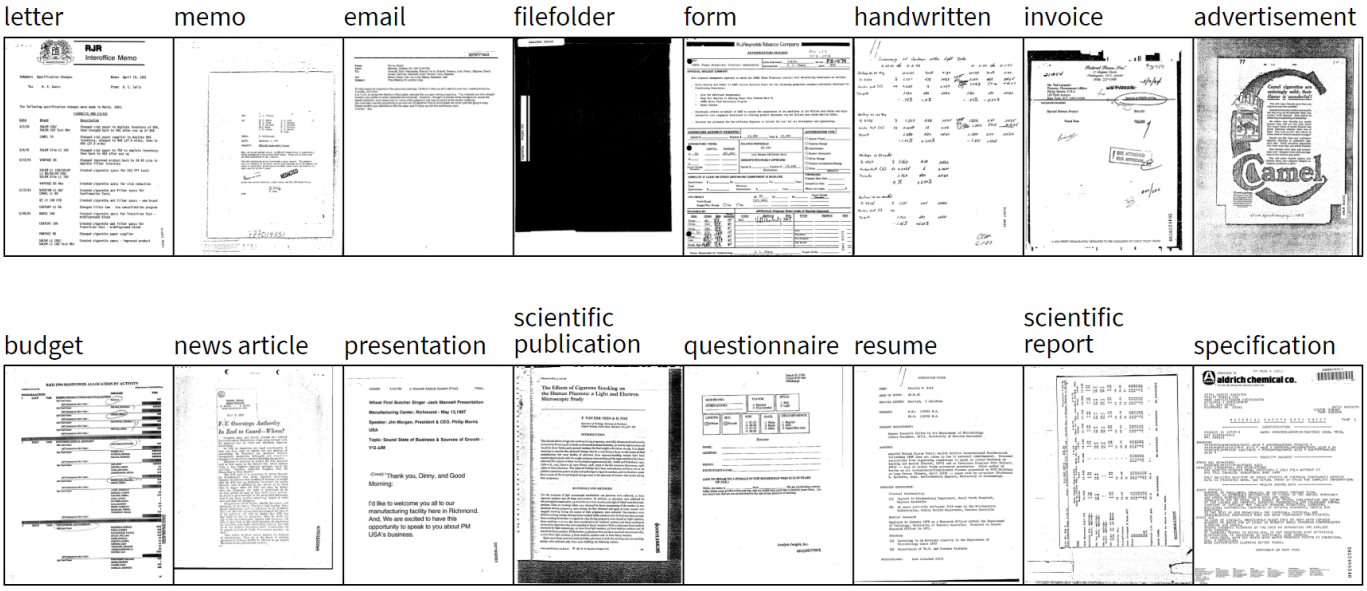
Document Image Classification

1. Business Problem

1.1 Description

The RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset consists of 400,000 grayscale images in 16 classes, with 25,000 images per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images. The images are sized so their largest dimension does not exceed 1000 pixels.

Here are the classes in the dataset, and an example from each:



Source : http://www.cs.cmu.edu/~aharley/rvl-cdip/ (http://www.cs.cmu.edu/~aharley/rvl-cdip/)

A. W. Harley, A. Ufkes, K. G. Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," in ICDAR, 2015

1.2 Problem Statemtent

Detection of different types of Document images and classify them in different classes like letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo. So It is a Multiclass Classification Problem or we can call it a Computer Vision task.

1.3 Data Description

The label files list the images and their categories in the following format:

path/to/the/image.tif category where the categories are numbered 0 to 15, in the following order:

```
0.letter                                1.form
2.email                             3.handwritten                          4.a
dvertisement                           5.scientific report              6.scie
ntific publication                   7.specification
 8.file folder
9.news article                          10.budget
11.invoice
12.presentation                         13.questionnaire
14.resume
15.memo
```

1.4 Usage

This dataset is a subset of the IIT-CDIP Test Collection 1.0 [1], which is publicly available here. The file structure of this dataset is the same as in the IIT collection, so it is possible to refer to that dataset for OCR and additional metadata. The IIT-CDIP dataset is itself a subset of the Legacy Tobacco Document Library [2].

1.5 Real-world/Business objectives and constraints

1. The cost of a mis-classification can be high.
2. No strict latency concerns.
3. Computationally Expensive

Importing Libraries

In [0]:

```
import os
import csv
import random
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import gc
import cv2
import time
import warnings
from scipy import sparse
import scipy.stats as ss
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
import seaborn as sns
from wordcloud import WordCloud ,STOPWORDS
from PIL import Image
import string
import re
import nltk
from nltk.corpus import stopwords
import spacy
from nltk import pos_tag
import warnings
warnings.filterwarnings("ignore")
from glob import glob
from tqdm import tqdm
from keras.layers.normalization import BatchNormalization
from keras.layers import Dropout
from keras.initializers import RandomNormal
import PIL
import cv2
from keras.callbacks import ReduceLROnPlateau
from keras.utils import np_utils
from keras.preprocessing import image
from sklearn.datasets import load_files
from keras.layers import Conv2D, MaxPooling2D, Dense, Flatten
from keras.models import Sequential
from keras.applications.vgg16 import VGG16
from keras.models import Model
from keras.layers import Conv2D, MaxPooling2D, Input, Dense, Flatten, concatenate
from keras.models import Model
from keras.applications.vgg16 import decode_predictions
from keras import optimizers
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.tokenize import TweetTokenizer
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer, HashingV
from sklearn.decomposition import TruncatedSVD
from sklearn.base import BaseEstimator, ClassifierMixin
from sklearn.utils.validation import check_X_y, check_is_fitted
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import log_loss
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import train_test_split
from wordcloud import WordCloud,STOPWORDS
import time
```

```
import warnings
warnings.filterwarnings("ignore")
from glob import glob
from tqdm import tqdm
import PIL
import cv2
from keras.utils import np_utils
from keras.preprocessing import image
from sklearn.datasets import load_files
from keras.layers import Conv2D, MaxPooling2D, Dense, Flatten
from keras.models import Sequential
from keras.applications.vgg16 import VGG16
from keras.models import Model
from keras.applications.vgg16 import decode_predictions
from keras import optimizers
from tensorboardcolab import *
import tensorflow as tf
from keras.layers import GlobalAveragePooling2D
import joblib
from keras.backend import set_session
from keras import backend as K
from keras.models import Sequential
from keras.layers import Input, Dropout, Flatten, Conv2D, MaxPooling2D, Dense, Activat
from keras.optimizers import RMSprop
from keras.callbacks import ModelCheckpoint, Callback, EarlyStopping
from keras.optimizers import Adam
from keras.utils import np_utils
from keras.models import Sequential
from keras.preprocessing.image import ImageDataGenerator
from keras.layers import Input, Dropout, Flatten, Conv2D, MaxPooling2D, Dense, Activat
from keras.optimizers import RMSprop
from keras.callbacks import ModelCheckpoint, Callback, EarlyStopping
from keras.utils import np_utils
from keras.preprocessing.image import ImageDataGenerator, array_to_img, img_to_array,
```

```
Using TensorFlow backend.
```

The default version of TensorFlow in Colab will soon switch to TensorFlow 2.x.
We recommend you upgrade (https://www.tensorflow.org/guide/migrate) now or ensure your notebook will continue to use TensorFlow 1.x via the `%tensorflow_version 1.x` magic:
more info (https://colab.research.google.com/notebooks/tensorflow_version.ipynb).

Mounting Google Drive

In [0]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?clien
t_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.co
m&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=email%20https%3A%2
F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.co
m%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.re
adonly%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fpeopleapi.readonly&respon
se_type=code (https://accounts.google.com/o/oauth2/auth?client_id=9473189898
03-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=
urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=email%20https%3A%2F%2Fwww.googleap
is.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive%
20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.readonly%20https%3
A%2F%2Fwww.googleapis.com%2Fauth%2Fpeopleapi.readonly&response_type=code)

Enter your authorization code:
··········
Mounted at /content/drive
```

Downloading the RVL-CDIP Dataset

In [0]:

```
!curl 'https://doc-10-b4-docs.googleusercontent.com/docs/securesc/phje98qeour5bfeeutelr
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Curr
ent
                                 Dload  Upload   Total   Spent    Left  Spee
d
100 36.1G    0 36.1G    0     0  60.8M      0 --:--:--  0:10:07 --:--:-- 54.
3M
```

Uncompressing the tar.gz file

In [0]:

```
import shutil
shutil.unpack_archive('/content/rvl-cdip.tar.gz')
```

In [0]:

```
labels = open('/content/labels/train.txt', 'r')
labels = labels.read()
```

In [0]:

```
import os
classes = {'0': 'letter', '1': 'form', '2': 'email', '3': 'handwritten', '4': 'adverti
           '6': 'scientific publication', '7': 'specification', '8': 'file folder', '9
           '11': 'invoice', '12': 'presentation', '13': 'questionnaire', '14': 'resume

root_path = '/content/images/images'

def make_doc_dirs(path):
    for label in classes.values():
        if not os.path.exists(path + str(label)):
            os.mkdir(path + str(label))

make_doc_dirs('/content/DocImages/training/')
make_doc_dirs('/content/DocImages/testing/')
make_doc_dirs('/content/DocImages/validation/')
```

In [0]:

```
labels_split = labels.split('images')
full_paths = []
for image_label in labels_split:
    label = root_path + image_label
    full_paths.append(label)
```

In [0]:

```
print(len(full_paths))
print(full_paths[0:10])
```

```
320001
['/content/images/images', '/content/images/imagesq/q/o/c/qoc54c00/80035521.
tif 15\n', '/content/images/imagese/e/w/c/ewc23d00/513280028.tif 1\n', '/con
tent/images/imagesw/w/b/t/wbt26e00/2053453161.tif 7\n', '/content/images/ima
gesm/m/k/m/mkm05e00/2040792992_2040792994.tif 10\n', '/content/images/images
o/o/e/x/oex80d00/522787731+-7732.tif 3\n', '/content/images/imagesq/q/v/t/qv
t06d00/50525666-5667.tif 14\n', '/content/images/imagesr/r/l/z/rlz20d00/5211
07137+-7140.tif 11\n', '/content/images/imagesk/k/n/i/kni98c00/87538356_835
9.tif 13\n', '/content/images/imagesm/m/q/v/mqv03f00/0011846871.tif 0\n']
```

In [0]:

```
full_paths = full_paths[1:]
paths = []
for path in full_paths:
    path = path.split('\n')
    for p in path:
        p = p.rsplit(' ',1)
        paths.append(p)

paths = [x for x in paths if x != ['']]
print(len(paths))
print(paths[:5])
```

```
320000
[['/content/images/imagesq/q/o/c/qoc54c00/80035521.tif', '15'], ['/content/i
mages/imagese/e/w/c/ewc23d00/513280028.tif', '1'], ['/content/images/images
w/w/b/t/wbt26e00/2053453161.tif', '7'], ['/content/images/imagesm/m/k/m/mkm0
5e00/2040792992_2040792994.tif', '10'], ['/content/images/imageso/o/e/x/oex8
0d00/522787731+-7732.tif', '3']]
```

In [0]:

```
i = 0
for p in paths[:320000]:
    old_path = p[0]
    full_path = p[0].split('images/')
    doctype = p[1]
    for key, value in classes.items():
        if doctype == key:
            doctype = value
    root_path = full_path[0]
    image_filename = full_path[1].rsplit('/', 1)[1]
    new_path = root_path +'DocImages/training/'+str(doctype) + '/' + image_filename

    if not os.path.exists(new_path):
        os.rename(old_path, new_path)
        i += 1
        if i % 1000 == 0:
            print(str(i) + "/320000")
            print(old_path)
            print(new_path)
```

```
1000/320000
/content/images/imagesk/k/n/i/kni87c00/506818313+-8313.tif
/content/DocImages/training/handwritten/506818313+-8313.tif
2000/320000
/content/images/imagesm/m/a/v/mav71a00/2057467808_2057467811.tif
/content/DocImages/training/specification/2057467808_2057467811.tif
3000/320000
/content/images/imagesh/h/w/r/hwr02e00/2028728622.tif
/content/DocImages/training/invoice/2028728622.tif
4000/320000
/content/images/imagesw/w/h/k/whk16d00/50590758-0759.tif
/content/DocImages/training/resume/50590758-0759.tif
5000/320000
/content/images/imagesk/k/h/b/khb84e00/2061193717.tif
/content/DocImages/training/advertisement/2061193717.tif
6000/320000
/content/images/imagesn/n/k/q/nkq13d00/513614517_513614518.tif
/content/DocImages/training/handwritten/513614517_513614518.tif
7000/320000
/content/images/imagesq/q/t/v/qtv00e00/03011899.tif
/content/DocImages/training/news article/03011899.tif
8000/320000
/content/images/imagesy/y/t/l/ytl65c00/2082867907.tif
/content/DocImages/training/form/2082867907.tif
9000/320000
/content/images/imagesz/z/n/f/znf51e00/01616306.tif
/content/DocImages/training/questionnaire/01616306.tif
10000/320000
/content/images/imagesw/w/w/i/wwi91e00/00921642.tif
/content/DocImages/training/invoice/00921642.tif
11000/320000
/content/images/imageso/o/e/w/oew36c00/2084412970_2971.tif
/content/DocImages/training/advertisement/2084412970_2971.tif
12000/320000
/content/images/imagesc/c/n/f/cnf02d00/71513746.tif
/content/DocImages/training/advertisement/71513746.tif
13000/320000
/content/images/imagesr/r/z/r/rzr58c00/88198229_8234.tif
```

```
/content/DocImages/training/scientific report/88198229_8234.tif
14000/320000
/content/images/imagesv/v/f/y/vfy44a00/93432593_2596.tif
/content/DocImages/training/memo/93432593_2596.tif
15000/320000
/content/images/imagesx/x/u/l/xul14c00/2069723544.tif
/content/DocImages/training/specification/2069723544.tif
16000/320000
/content/images/imagesn/n/v/p/nvp03e00/2044163463.tif
/content/DocImages/training/memo/2044163463.tif
17000/320000
/content/images/imagest/t/o/l/tol31d00/522731336+-1336.tif
/content/DocImages/training/handwritten/522731336+-1336.tif
18000/320000
/content/images/imagesi/i/n/e/ine45e00/2040768943_2040768945.tif
/content/DocImages/training/budget/2040768943_2040768945.tif
19000/320000
/content/images/imagesp/p/r/q/prq25a00/92872435.tif
/content/DocImages/training/invoice/92872435.tif
20000/320000
/content/images/imagesk/k/c/w/kcw43a00/2065401350.tif
/content/DocImages/training/form/2065401350.tif
21000/320000
/content/images/imagesc/c/p/q/cpq40e00/04235083.tif
/content/DocImages/training/news article/04235083.tif
22000/320000
/content/images/imagesp/p/m/r/pmr52d00/2501722082_2097.tif
/content/DocImages/training/scientific publication/2501722082_2097.tif
23000/320000
/content/images/imagesy/y/b/p/ybp77d00/2076276140.tif
/content/DocImages/training/specification/2076276140.tif
24000/320000
/content/images/imagesk/k/y/h/kyh90f00/0011845031.tif
/content/DocImages/training/questionnaire/0011845031.tif
25000/320000
/content/images/imagesv/v/f/l/vfl57c00/2072957905.tif
/content/DocImages/training/invoice/2072957905.tif
26000/320000
/content/images/imagesp/p/k/h/pkh45c00/2074251564.tif
/content/DocImages/training/advertisement/2074251564.tif
27000/320000
/content/images/imagesl/l/l/v/llv20c00/2080927981.tif
/content/DocImages/training/questionnaire/2080927981.tif
28000/320000
/content/images/imagesi/i/q/y/iqy35c00/2083608693.tif
/content/DocImages/training/file folder/2083608693.tif
29000/320000
/content/images/imagesq/q/s/s/qss39e00/2501237973.tif
/content/DocImages/training/letter/2501237973.tif
30000/320000
/content/images/imagesw/w/r/a/wra99c00/40008854-8854.tif
/content/DocImages/training/scientific publication/40008854-8854.tif
31000/320000
/content/images/imagesz/z/y/s/zys14e00/2023205437.tif
/content/DocImages/training/budget/2023205437.tif
32000/320000
/content/images/imagest/t/v/k/tvk41f00/0000969506.tif
/content/DocImages/training/scientific report/0000969506.tif
33000/320000
/content/images/imagesb/b/l/u/blu89c00/50286997-6997.tif
/content/DocImages/training/memo/50286997-6997.tif
```

```
34000/320000
/content/images/imagesk/k/i/d/kid10f00/0000544661.tif
/content/DocImages/training/advertisement/0000544661.tif
35000/320000
/content/images/imagesa/a/i/p/aip90d00/518294948+-4949.tif
/content/DocImages/training/handwritten/518294948+-4949.tif
36000/320000
/content/images/imagesr/r/m/v/rmv80f00/0011907157.tif
/content/DocImages/training/letter/0011907157.tif
37000/320000
/content/images/imagesj/j/b/x/jbx60c00/2077315599.tif
/content/DocImages/training/news article/2077315599.tif
38000/320000
/content/images/imagesq/q/m/y/qmy42e00/2031298788.tif
/content/DocImages/training/specification/2031298788.tif
39000/320000
/content/images/imagesj/j/s/d/jsd30e00/89663484_89663487.tif
/content/DocImages/training/memo/89663484_89663487.tif
40000/320000
/content/images/imagese/e/o/j/eoj24c00/2069712219.tif
/content/DocImages/training/specification/2069712219.tif
41000/320000
/content/images/imagesr/r/z/g/rzg41a00/0012946875.tif
/content/DocImages/training/letter/0012946875.tif
42000/320000
/content/images/imagesd/d/m/v/dmv84c00/93421699_1700.tif
/content/DocImages/training/advertisement/93421699_1700.tif
43000/320000
/content/images/imagesq/q/u/h/quh01e00/89280868.tif
/content/DocImages/training/form/89280868.tif
44000/320000
/content/images/imagesd/d/c/m/dcm88e00/2020137465.tif
/content/DocImages/training/form/2020137465.tif
45000/320000
/content/images/imagesy/y/x/i/yxi36e00/2056698419.tif
/content/DocImages/training/scientific report/2056698419.tif
46000/320000
/content/images/imagesq/q/l/v/qlv26d00/50681016-1016.tif
/content/DocImages/training/resume/50681016-1016.tif
47000/320000
/content/images/imagesz/z/p/w/zpw91a00/2059095268.tif
/content/DocImages/training/file folder/2059095268.tif
48000/320000
/content/images/imagesi/i/y/p/iyp30d00/524359675+-9676.tif
/content/DocImages/training/handwritten/524359675+-9676.tif
49000/320000
/content/images/imagesw/w/o/u/wou77c00/504572138+-2138.tif
/content/DocImages/training/specification/504572138+-2138.tif
50000/320000
/content/images/imagesd/d/g/d/dgd86d00/ti16351637.tif
/content/DocImages/training/budget/ti16351637.tif
51000/320000
/content/images/imagesd/d/a/g/dag93c00/86459025.tif
/content/DocImages/training/memo/86459025.tif
52000/320000
/content/images/imagesf/f/b/e/fbe00d00/50315434-5434.tif
/content/DocImages/training/resume/50315434-5434.tif
53000/320000
/content/images/imagese/e/e/w/eew94e00/2040771534_2040771550.tif
/content/DocImages/training/budget/2040771534_2040771550.tif
54000/320000
```

```
/content/images/imagesx/x/k/k/xkk49c00/2082345930.tif
/content/DocImages/training/scientific report/2082345930.tif
55000/320000
/content/images/imagesp/p/p/v/ppv48c00/2083789982.tif
/content/DocImages/training/news article/2083789982.tif
56000/320000
/content/images/imagesk/k/g/g/kgg21d00/515627440+-7441.tif
/content/DocImages/training/handwritten/515627440+-7441.tif
57000/320000
/content/images/imagesz/z/r/f/zrf39e00/2501593586.tif
/content/DocImages/training/scientific report/2501593586.tif
58000/320000
/content/images/imagesz/z/s/l/zsl96e00/2057809686.tif
/content/DocImages/training/form/2057809686.tif
59000/320000
/content/images/imagesd/d/m/w/dmw15d00/505438061.tif
/content/DocImages/training/handwritten/505438061.tif
60000/320000
/content/images/imagesg/g/m/r/gmr71f00/2022201771.tif
/content/DocImages/training/scientific report/2022201771.tif
61000/320000
/content/images/imagesv/v/u/q/vuq47c00/2071537221.tif
/content/DocImages/training/news article/2071537221.tif
62000/320000
/content/images/imagesq/q/c/w/qcw02e00/2028719704.tif
/content/DocImages/training/invoice/2028719704.tif
63000/320000
/content/images/imagesx/x/m/q/xmq3aa00/11238535.tif
/content/DocImages/training/file folder/11238535.tif
64000/320000
/content/images/imagesz/z/p/h/zph92d00/515816457_515816459.tif
/content/DocImages/training/handwritten/515816457_515816459.tif
65000/320000
/content/images/imagesn/n/z/b/nzb31d00/514879390+-9392.tif
/content/DocImages/training/handwritten/514879390+-9392.tif
66000/320000
/content/images/imagesz/z/g/l/zgl86c00/2081190836.tif
/content/DocImages/training/email/2081190836.tif
67000/320000
/content/images/imagesy/y/l/u/ylu05e00/2041719684.tif
/content/DocImages/training/budget/2041719684.tif
68000/320000
/content/images/imagesu/u/g/o/ugo62e00/2060003478.tif
/content/DocImages/training/questionnaire/2060003478.tif
69000/320000
/content/images/imagesm/m/p/p/mpp96c00/50736931-6949.tif
/content/DocImages/training/resume/50736931-6949.tif
70000/320000
/content/images/imagesa/a/g/g/agg05e00/2041076518.tif
/content/DocImages/training/questionnaire/2041076518.tif
71000/320000
/content/images/imagesr/r/p/p/rpp71a00/2057450354_2057450355.tif
/content/DocImages/training/specification/2057450354_2057450355.tif
72000/320000
/content/images/imagest/t/v/p/tvp81a00/1000385350.tif
/content/DocImages/training/questionnaire/1000385350.tif
73000/320000
/content/images/imagesn/n/z/i/nzi97d00/2063676513_6515.tif
/content/DocImages/training/presentation/2063676513_6515.tif
74000/320000
/content/images/imagesn/n/t/e/nte81d00/602242.tif
```

```
/content/DocImages/training/letter/602242.tif
75000/320000
/content/images/imagesa/a/n/e/ane51e00/01187623.tif
/content/DocImages/training/scientific report/01187623.tif
76000/320000
/content/images/imageso/o/r/b/orb02d00/71388960.tif
/content/DocImages/training/advertisement/71388960.tif
77000/320000
/content/images/imagesy/y/t/v/ytv43c00/89310988.tif
/content/DocImages/training/letter/89310988.tif
78000/320000
/content/images/imagesg/g/k/v/gkv09e00/2050663636.tif
/content/DocImages/training/form/2050663636.tif
79000/320000
/content/images/imagesd/d/d/g/ddg60c00/2073264245.tif
/content/DocImages/training/email/2073264245.tif
80000/320000
/content/images/imagesi/i/z/q/izq72c00/2078373784.tif
/content/DocImages/training/email/2078373784.tif
81000/320000
/content/images/imagesn/n/n/c/nnc80d00/517682383+-2384.tif
/content/DocImages/training/handwritten/517682383+-2384.tif
82000/320000
/content/images/imagesr/r/g/s/rgs09c00/95510051.tif
/content/DocImages/training/memo/95510051.tif
83000/320000
/content/images/imagesq/q/w/q/qwq1aa00/10132891_10132930.tif
/content/DocImages/training/scientific report/10132891_10132930.tif
84000/320000
/content/images/imagesi/i/g/w/igw53e00/2030179651.tif
/content/DocImages/training/form/2030179651.tif
85000/320000
/content/images/imagesh/h/k/b/hkb23d00/513320271_513320272.tif
/content/DocImages/training/letter/513320271_513320272.tif
86000/320000
/content/images/imagest/t/y/b/tyb74f00/0060277309.tif
/content/DocImages/training/letter/0060277309.tif
87000/320000
/content/images/imagesq/q/i/s/qis08e00/1001604170_1001604176.tif
/content/DocImages/training/scientific report/1001604170_1001604176.tif
88000/320000
/content/images/imagesp/p/v/h/pvh60c00/2073857895.tif
/content/DocImages/training/news article/2073857895.tif
89000/320000
/content/images/imagesq/q/m/k/qmk27d00/2072165979_5986.tif
/content/DocImages/training/presentation/2072165979_5986.tif
90000/320000
/content/images/imagesa/a/n/e/ane23e00/2058502554.tif
/content/DocImages/training/advertisement/2058502554.tif
91000/320000
/content/images/imageso/o/e/s/oes30e00/87427684.tif
/content/DocImages/training/scientific report/87427684.tif
92000/320000
/content/images/imagesz/z/r/k/zrk53e00/2022262612.tif
/content/DocImages/training/memo/2022262612.tif
93000/320000
/content/images/imagesm/m/a/p/map37e00/2029023698.tif
/content/DocImages/training/scientific report/2029023698.tif
94000/320000
/content/images/imagesa/a/l/z/alz79e00/0000176566.tif
/content/DocImages/training/questionnaire/0000176566.tif
```

```
95000/320000
/content/images/imagesk/k/j/m/kjm09d00/50446270-6271.tif
/content/DocImages/training/resume/50446270-6271.tif
96000/320000
/content/images/imagesd/d/y/e/dye50c00/ti17311427.tif
/content/DocImages/training/invoice/ti17311427.tif
97000/320000
/content/images/imagesu/u/b/x/ubx22f00/tim00832.50_tim00832.51.tif
/content/DocImages/training/budget/tim00832.50_tim00832.51.tif
98000/320000
/content/images/imagesn/n/l/m/nlm22c00/2069730621.tif
/content/DocImages/training/specification/2069730621.tif
99000/320000
/content/images/imagesb/b/u/p/bup48e00/2023117124_2023117137.tif
/content/DocImages/training/presentation/2023117124_2023117137.tif
100000/320000
/content/images/imagesa/a/l/d/ald92f00/tob06107.98.tif
/content/DocImages/training/letter/tob06107.98.tif
101000/320000
/content/images/imagesk/k/c/p/kcp11e00/87092526.tif
/content/DocImages/training/file folder/87092526.tif
102000/320000
/content/images/imagesr/r/a/b/rab98e00/2024114225_2024114226.tif
/content/DocImages/training/presentation/2024114225_2024114226.tif
103000/320000
/content/images/imagesj/j/v/a/jva08d00/2071791407_1408.tif
/content/DocImages/training/presentation/2071791407_1408.tif
104000/320000
/content/images/imagest/t/r/h/trh30c00/corti0005406.tif
/content/DocImages/training/budget/corti0005406.tif
105000/320000
/content/images/imagesa/a/x/h/axh93c00/2060927273.tif
/content/DocImages/training/form/2060927273.tif
106000/320000
/content/images/imagesp/p/r/o/pro75d00/503950514.tif
/content/DocImages/training/advertisement/503950514.tif
107000/320000
/content/images/imagesy/y/y/l/yyl75f00/0060057552.tif
/content/DocImages/training/invoice/0060057552.tif
108000/320000
/content/images/imagesi/i/h/a/iha15c00/2073737168.tif
/content/DocImages/training/file folder/2073737168.tif
109000/320000
/content/images/imagesh/h/w/n/hwn57e00/2031022772.tif
/content/DocImages/training/file folder/2031022772.tif
110000/320000
/content/images/imagesk/k/i/w/kiw22c00/2069737666.tif
/content/DocImages/training/specification/2069737666.tif
111000/320000
/content/images/imagesm/m/y/b/myb88e00/2020021643_2020021649.tif
/content/DocImages/training/presentation/2020021643_2020021649.tif
112000/320000
/content/images/imagesl/l/x/w/lxw35a00/86377464_7490.tif
/content/DocImages/training/presentation/86377464_7490.tif
113000/320000
/content/images/imagesy/y/u/k/yuk55e00/2045467669.tif
/content/DocImages/training/budget/2045467669.tif
114000/320000
/content/images/imagesf/f/x/o/fxo78e00/2015057716.tif
/content/DocImages/training/budget/2015057716.tif
115000/320000
```

```
/content/images/imagesf/f/v/q/fvq94c00/2085799496.tif
/content/DocImages/training/email/2085799496.tif
116000/320000
/content/images/imagesu/u/j/h/ujh70f00/0011985096.tif
/content/DocImages/training/presentation/0011985096.tif
117000/320000
/content/images/imagesu/u/n/c/unc45a00/82341281_1285.tif
/content/DocImages/training/scientific report/82341281_1285.tif
118000/320000
/content/images/imagesl/l/q/v/lqv05f00/0060189215.tif
/content/DocImages/training/questionnaire/0060189215.tif
119000/320000
/content/images/imagesq/q/w/d/qwd42e00/2500100014_2500100032.tif
/content/DocImages/training/presentation/2500100014_2500100032.tif
120000/320000
/content/images/imagesn/n/x/s/nxs76c00/2084399692_9693.tif
/content/DocImages/training/form/2084399692_9693.tif
121000/320000
/content/images/imagesb/b/w/u/bwu57e00/2031578069_2031578070.tif
/content/DocImages/training/news article/2031578069_2031578070.tif
122000/320000
/content/images/imagesg/g/q/g/gqg64e00/1001762169_1001762171.tif
/content/DocImages/training/scientific report/1001762169_1001762171.tif
123000/320000
/content/images/imagesy/y/x/v/yxv99c00/50295188-5189.tif
/content/DocImages/training/resume/50295188-5189.tif
124000/320000
/content/images/imagesa/a/s/g/asg64a00/89228432.tif
/content/DocImages/training/news article/89228432.tif
125000/320000
/content/images/imagesv/v/o/f/vof34f00/0060258940.tif
/content/DocImages/training/file folder/0060258940.tif
126000/320000
/content/images/imagesx/x/w/a/xwa35f00/0060035687.tif
/content/DocImages/training/memo/0060035687.tif
127000/320000
/content/images/imagesh/h/g/g/hgg96d00/tcal0469676.tif
/content/DocImages/training/budget/tcal0469676.tif
128000/320000
/content/images/imagesa/a/t/m/atm53a00/2047348388.tif
/content/DocImages/training/news article/2047348388.tif
129000/320000
/content/images/imagesc/c/m/e/cme41c00/2085756375a.tif
/content/DocImages/training/email/2085756375a.tif
130000/320000
/content/images/imageso/o/r/m/orm19e00/2500138911_2500138916.tif
/content/DocImages/training/budget/2500138911_2500138916.tif
131000/320000
/content/images/imagesi/i/s/l/isl6aa00/11320376_11320377.tif
/content/DocImages/training/scientific publication/11320376_11320377.tif
132000/320000
/content/images/imagest/t/g/y/tgy08d00/2070114895_4910.tif
/content/DocImages/training/presentation/2070114895_4910.tif
133000/320000
/content/images/imagesl/l/k/t/lkt46c00/2505106787b.tif
/content/DocImages/training/email/2505106787b.tif
134000/320000
/content/images/imagesl/l/p/c/lpc74a00/2051825981_5989.tif
/content/DocImages/training/presentation/2051825981_5989.tif
135000/320000
/content/images/imagesu/u/d/s/uds02c00/2085103708c.tif
```

```
/content/DocImages/training/email/2085103708c.tif
136000/320000
/content/images/imagesu/u/b/x/ubx66c00/2073405731.tif
/content/DocImages/training/advertisement/2073405731.tif
137000/320000
/content/images/imagesn/n/u/r/nur77e00/2061011479.tif
/content/DocImages/training/advertisement/2061011479.tif
138000/320000
/content/images/imagesl/l/u/a/lua88e00/2026189403.tif
/content/DocImages/training/file folder/2026189403.tif
139000/320000
/content/images/imagesu/u/r/m/urm11d00/522812099+-2103.tif
/content/DocImages/training/handwritten/522812099+-2103.tif
140000/320000
/content/images/imagesb/b/q/n/bqn72a00/2021336009.tif
/content/DocImages/training/news article/2021336009.tif
141000/320000
/content/images/imagesy/y/p/t/ypt25a00/81887824_7826.tif
/content/DocImages/training/email/81887824_7826.tif
142000/320000
/content/images/imagesa/a/l/w/alw14f00/0000099355.tif
/content/DocImages/training/presentation/0000099355.tif
143000/320000
/content/images/imagesh/h/s/p/hsp19e00/2062556170_2062556177.tif
/content/DocImages/training/budget/2062556170_2062556177.tif
144000/320000
/content/images/imagesr/r/s/t/rst14e00/2023075430_2023075461.tif
/content/DocImages/training/presentation/2023075430_2023075461.tif
145000/320000
/content/images/imagese/e/j/j/ejj23e00/1002608961-a.tif
/content/DocImages/training/news article/1002608961-a.tif
146000/320000
/content/images/imagesk/k/w/h/kwh11e00/87846000.tif
/content/DocImages/training/file folder/87846000.tif
147000/320000
/content/images/imagesc/c/p/f/cpf8aa00/11009646.tif
/content/DocImages/training/invoice/11009646.tif
148000/320000
/content/images/imagesc/c/z/b/czb4aa00/10404340.tif
/content/DocImages/training/scientific publication/10404340.tif
149000/320000
/content/images/imagesy/y/d/s/yds57d00/2045863295_3300.tif
/content/DocImages/training/presentation/2045863295_3300.tif
150000/320000
/content/images/imagesg/g/l/n/gln43e00/2025371830.tif
/content/DocImages/training/file folder/2025371830.tif
151000/320000
/content/images/imagest/t/d/x/tdx27d00/2072907978.tif
/content/DocImages/training/file folder/2072907978.tif
152000/320000
/content/images/imagesl/l/g/x/lgx93c00/2064018373.tif
/content/DocImages/training/letter/2064018373.tif
153000/320000
/content/images/imagesy/y/p/c/ypc63f00/0001213593.tif
/content/DocImages/training/memo/0001213593.tif
154000/320000
/content/images/imagesk/k/g/q/kgq05c00/2505348500.tif
/content/DocImages/training/file folder/2505348500.tif
155000/320000
/content/images/imageso/o/o/m/oom60c00/2074905265_5271.tif
/content/DocImages/training/scientific publication/2074905265_5271.tif
```

```
156000/320000
/content/images/imagesz/z/l/c/zlc02e00/2028848779.tif
/content/DocImages/training/file folder/2028848779.tif
157000/320000
/content/images/imagesy/y/a/x/yax65a00/528588958+-8958.tif
/content/DocImages/training/email/528588958+-8958.tif
158000/320000
/content/images/imagesb/b/e/d/bed10f00/0000545039.tif
/content/DocImages/training/advertisement/0000545039.tif
159000/320000
/content/images/imagesf/f/c/e/fce30d00/522897718+-7733.tif
/content/DocImages/training/invoice/522897718+-7733.tif
160000/320000
/content/images/imagesk/k/c/n/kcn65d00/504528858.tif
/content/DocImages/training/specification/504528858.tif
161000/320000
/content/images/imagesl/l/k/n/lkn16d00/50598921-8922.tif
/content/DocImages/training/resume/50598921-8922.tif
162000/320000
/content/images/imagese/e/h/q/ehq31c00/2085724202b.tif
/content/DocImages/training/email/2085724202b.tif
163000/320000
/content/images/imagesq/q/p/i/qpi09d00/50439402-9405.tif
/content/DocImages/training/scientific publication/50439402-9405.tif
164000/320000
/content/images/imagesa/a/s/j/asj14c00/2069724632.tif
/content/DocImages/training/specification/2069724632.tif
165000/320000
/content/images/imagesw/w/c/n/wcn88e00/2020175913.tif
/content/DocImages/training/memo/2020175913.tif
166000/320000
/content/images/imagesz/z/z/c/zzc50c00/ti17180287.tif
/content/DocImages/training/news article/ti17180287.tif
167000/320000
/content/images/imagesr/r/d/l/rdl30e00/88027193_88027198.tif
/content/DocImages/training/scientific publication/88027193_88027198.tif
168000/320000
/content/images/imagesr/r/k/b/rkb86d00/ti16371218.tif
/content/DocImages/training/budget/ti16371218.tif
169000/320000
/content/images/imagesn/n/e/k/nek94e00/1003044079-b.tif
/content/DocImages/training/news article/1003044079-b.tif
170000/320000
/content/images/imagesr/r/u/g/rug07e00/2055113518.tif
/content/DocImages/training/file folder/2055113518.tif
171000/320000
/content/images/imagesy/y/o/h/yoh27e00/2028717014.tif
/content/DocImages/training/invoice/2028717014.tif
172000/320000
/content/images/imagest/t/k/w/tkw55c00/2082560976.tif
/content/DocImages/training/email/2082560976.tif
173000/320000
/content/images/imagesk/k/y/m/kym80d00/515958471+-8472.tif
/content/DocImages/training/handwritten/515958471+-8472.tif
174000/320000
/content/images/imagesb/b/x/x/bxx24a00/2067686276_2067686277.tif
/content/DocImages/training/email/2067686276_2067686277.tif
175000/320000
/content/images/imagesk/k/z/u/kzu82a00/524555308+-5314.tif
/content/DocImages/training/handwritten/524555308+-5314.tif
176000/320000
```

```
/content/images/imagesm/m/f/t/mft43d00/511522416.tif
/content/DocImages/training/specification/511522416.tif
177000/320000
/content/images/imagesc/c/a/v/cav18d00/2071235600.tif
/content/DocImages/training/form/2071235600.tif
178000/320000
/content/images/imagesv/v/o/z/voz24d00/507383078.tif
/content/DocImages/training/letter/507383078.tif
179000/320000
/content/images/imagesp/p/l/o/plo13d00/513851715_513851716.tif
/content/DocImages/training/letter/513851715_513851716.tif
180000/320000
/content/images/imagesv/v/r/s/vrs90e00/91514855.tif
/content/DocImages/training/invoice/91514855.tif
181000/320000
/content/images/imagesl/l/f/n/lfn53f00/0001224757.tif
/content/DocImages/training/memo/0001224757.tif
182000/320000
/content/images/imagese/e/q/o/eqo55e00/2044191009.tif
/content/DocImages/training/budget/2044191009.tif
183000/320000
/content/images/imagesi/i/q/h/iqh51a00/0071087714.tif
/content/DocImages/training/advertisement/0071087714.tif
184000/320000
/content/images/imagesk/k/n/c/knc74a00/2051827371.tif
/content/DocImages/training/news article/2051827371.tif
185000/320000
/content/images/imagesy/y/r/d/yrd16e00/2044400306.tif
/content/DocImages/training/advertisement/2044400306.tif
186000/320000
/content/images/imagesl/l/j/e/lje95a00/500028605+-8608.tif
/content/DocImages/training/questionnaire/500028605+-8608.tif
187000/320000
/content/images/imagesu/u/w/p/uwp64a00/82262519.tif
/content/DocImages/training/advertisement/82262519.tif
188000/320000
/content/images/imagesw/w/i/h/wih01d00/514875815+-5816.tif
/content/DocImages/training/handwritten/514875815+-5816.tif
189000/320000
/content/images/imagesp/p/z/i/pzi29e00/2504079068.tif
/content/DocImages/training/news article/2504079068.tif
190000/320000
/content/images/imagesj/j/m/o/jmo59e00/2024482986.tif
/content/DocImages/training/file folder/2024482986.tif
191000/320000
/content/images/imagesx/x/i/j/xij22e00/2501100601_2501100603.tif
/content/DocImages/training/presentation/2501100601_2501100603.tif
192000/320000
/content/images/imagesl/l/p/l/lpl53a00/2048374799_4800.tif
/content/DocImages/training/letter/2048374799_4800.tif
193000/320000
/content/images/imagesj/j/a/s/jas81f00/2046936212.tif
/content/DocImages/training/budget/2046936212.tif
194000/320000
/content/images/imagesm/m/i/n/min53c00/98393391_3393.tif
/content/DocImages/training/budget/98393391_3393.tif
195000/320000
/content/images/imagesr/r/m/c/rmc33f00/0000946523.tif
/content/DocImages/training/memo/0000946523.tif
196000/320000
/content/images/imagesb/b/w/v/bwv42f00/tob16526.05_tob16526.11.tif
```

```
/content/DocImages/training/presentation/tob16526.05_tob16526.11.tif
197000/320000
/content/images/imagesv/v/e/u/veu62f00/tob11423.94_tob11424.04.tif
/content/DocImages/training/news article/tob11423.94_tob11424.04.tif
198000/320000
/content/images/imagesk/k/v/q/kvq50d00/524359510+-9511.tif
/content/DocImages/training/questionnaire/524359510+-9511.tif
199000/320000
/content/images/imagesv/v/r/l/vrl67d00/2063531730.tif
/content/DocImages/training/file folder/2063531730.tif
200000/320000
/content/images/imagesz/z/a/x/zax17e00/2024485516_2024485517.tif
/content/DocImages/training/form/2024485516_2024485517.tif
201000/320000
/content/images/imagesz/z/w/d/zwd64e00/1001867678.tif
/content/DocImages/training/memo/1001867678.tif
202000/320000
/content/images/imagesb/b/k/t/bkt79d00/500394405_500394410.tif
/content/DocImages/training/letter/500394405_500394410.tif
203000/320000
/content/images/imagesk/k/h/r/khr18d00/2044046077_6079.tif
/content/DocImages/training/presentation/2044046077_6079.tif
204000/320000
/content/images/imagesb/b/s/w/bsw07d00/tnwl0015552_5555.tif
/content/DocImages/training/budget/tnwl0015552_5555.tif
205000/320000
/content/images/imagesh/h/v/z/hvz92c00/2084391846a.tif
/content/DocImages/training/email/2084391846a.tif
206000/320000
/content/images/imagesx/x/f/h/xfh05f00/0060178462.tif
/content/DocImages/training/memo/0060178462.tif
207000/320000
/content/images/imagesb/b/p/m/bpm67c00/2076805982.tif
/content/DocImages/training/email/2076805982.tif
208000/320000
/content/images/imagesv/v/r/b/vrb64e00/1003369563.tif
/content/DocImages/training/scientific report/1003369563.tif
209000/320000
/content/images/imageso/o/y/z/oyz94d00/506089124_506089126.tif
/content/DocImages/training/handwritten/506089124_506089126.tif
210000/320000
/content/images/imagesa/a/i/e/aie07d00/tnwl0043816_3817.tif
/content/DocImages/training/budget/tnwl0043816_3817.tif
211000/320000
/content/images/imagesp/p/i/t/pit55f00/0060135823.tif
/content/DocImages/training/file folder/0060135823.tif
212000/320000
/content/images/imagesg/g/r/e/gre79d00/500539689.tif
/content/DocImages/training/letter/500539689.tif
213000/320000
/content/images/imagesm/m/z/m/mzm13f00/0000395118.tif
/content/DocImages/training/memo/0000395118.tif
214000/320000
/content/images/imagesd/d/n/o/dno70c00/2078583974_3988.tif
/content/DocImages/training/presentation/2078583974_3988.tif
215000/320000
/content/images/imagesj/j/t/c/jtc18d00/2070316981.tif
/content/DocImages/training/presentation/2070316981.tif
216000/320000
/content/images/imagesr/r/q/r/rqr11a00/0071000452.tif
/content/DocImages/training/file folder/0071000452.tif
```

```
217000/320000
/content/images/imagesa/a/f/j/afj14c00/2069724829.tif
/content/DocImages/training/specification/2069724829.tif
218000/320000
/content/images/imageso/o/a/r/oar95e00/2025835283.tif
/content/DocImages/training/memo/2025835283.tif
219000/320000
/content/images/imagesq/q/j/s/qjs23e00/2061844666.tif
/content/DocImages/training/news article/2061844666.tif
220000/320000
/content/images/imagesf/f/u/d/fud25c00/2505333366_3388.tif
/content/DocImages/training/scientific report/2505333366_3388.tif
221000/320000
/content/images/imagest/t/x/t/txt26e00/2053462360.tif
/content/DocImages/training/specification/2053462360.tif
222000/320000
/content/images/imagesb/b/r/o/bro94c00/2065441024.tif
/content/DocImages/training/file folder/2065441024.tif
223000/320000
/content/images/imagese/e/h/b/ehb93e00/2043123087_2043123088.tif
/content/DocImages/training/questionnaire/2043123087_2043123088.tif
224000/320000
/content/images/imagesc/c/g/l/cgl05e00/2040787215_2040787216.tif
/content/DocImages/training/budget/2040787215_2040787216.tif
225000/320000
/content/images/imagesq/q/j/t/qjt50f00/0000441424.tif
/content/DocImages/training/questionnaire/0000441424.tif
226000/320000
/content/images/imagese/e/k/k/ekk57d00/2047320619_0620.tif
/content/DocImages/training/resume/2047320619_0620.tif
227000/320000
/content/images/imagesr/r/i/u/riu21e00/87680976_87680979.tif
/content/DocImages/training/letter/87680976_87680979.tif
228000/320000
/content/images/imagest/t/b/p/tbp22d00/2062199219.tif
/content/DocImages/training/presentation/2062199219.tif
229000/320000
/content/images/imagesb/b/o/c/boc83c00/2040985330a_5331a.tif
/content/DocImages/training/news article/2040985330a_5331a.tif
230000/320000
/content/images/imagesh/h/h/e/hhe41a00/0012944918.tif
/content/DocImages/training/file folder/0012944918.tif
231000/320000
/content/images/imagesh/h/o/k/hok08c00/527851937+-1940.tif
/content/DocImages/training/email/527851937+-1940.tif
232000/320000
/content/images/imagesy/y/b/x/ybx65d00/504437542_504437556.tif
/content/DocImages/training/handwritten/504437542_504437556.tif
233000/320000
/content/images/imagesx/x/s/t/xst22d00/2028371770.tif
/content/DocImages/training/invoice/2028371770.tif
234000/320000
/content/images/imagesc/c/q/k/cqk25c00/2078636921.tif
/content/DocImages/training/email/2078636921.tif
235000/320000
/content/images/imagesa/a/h/c/ahc92d00/515976187.tif
/content/DocImages/training/letter/515976187.tif
236000/320000
/content/images/imagesb/b/z/h/bzh14a00/95732949.tif
/content/DocImages/training/budget/95732949.tif
237000/320000
```

```
/content/images/imagesm/m/s/b/msb96d00/tcal0414376.tif
/content/DocImages/training/budget/tcal0414376.tif
238000/320000
/content/images/imagesm/m/w/y/mwy74e00/2044266691_2044266693.tif
/content/DocImages/training/news article/2044266691_2044266693.tif
239000/320000
/content/images/imagesh/h/a/x/hax33e00/2045042786.tif
/content/DocImages/training/file folder/2045042786.tif
240000/320000
/content/images/imagesa/a/n/x/anx33d00/512279814_512279815.tif
/content/DocImages/training/letter/512279814_512279815.tif
241000/320000
/content/images/imagesj/j/u/m/jum12d00/71213115.tif
/content/DocImages/training/advertisement/71213115.tif
242000/320000
/content/images/imagesc/c/s/b/csb94e00/1000327912.tif
/content/DocImages/training/memo/1000327912.tif
243000/320000
/content/images/imagesz/z/a/f/zaf80d00/517610991+-0992.tif
/content/DocImages/training/handwritten/517610991+-0992.tif
244000/320000
/content/images/imagesn/n/y/c/nyc12a00/2504011526.tif
/content/DocImages/training/advertisement/2504011526.tif
245000/320000
/content/images/imagesy/y/q/k/yqk65e00/2049420696.tif
/content/DocImages/training/budget/2049420696.tif
246000/320000
/content/images/imagesm/m/b/w/mbw03e00/2043460331_2043460333.tif
/content/DocImages/training/form/2043460331_2043460333.tif
247000/320000
/content/images/imagesy/y/h/i/yhi24c00/2069712746.tif
/content/DocImages/training/specification/2069712746.tif
248000/320000
/content/images/imageso/o/z/h/ozh95a00/3000155899.tif
/content/DocImages/training/file folder/3000155899.tif
249000/320000
/content/images/imagesf/f/m/a/fma99c00/40007001-7002.tif
/content/DocImages/training/resume/40007001-7002.tif
250000/320000
/content/images/imagesh/h/s/n/hsn21d00/515957488+-7490.tif
/content/DocImages/training/handwritten/515957488+-7490.tif
251000/320000
/content/images/imagesw/w/i/l/wil03f00/tob00704.12_tob00704.18.tif
/content/DocImages/training/presentation/tob00704.12_tob00704.18.tif
252000/320000
/content/images/imagesv/v/k/h/vkh58c00/92742251_2253.tif
/content/DocImages/training/news article/92742251_2253.tif
253000/320000
/content/images/imagesh/h/x/s/hxs90e00/91514651.tif
/content/DocImages/training/invoice/91514651.tif
254000/320000
/content/images/imagest/t/x/l/txl67e00/2063610133.tif
/content/DocImages/training/invoice/2063610133.tif
255000/320000
/content/images/imagesz/z/d/f/zdf55f00/0060128826.tif
/content/DocImages/training/questionnaire/0060128826.tif
256000/320000
/content/images/imagesc/c/l/w/clw62c00/2077820213_0214.tif
/content/DocImages/training/memo/2077820213_0214.tif
257000/320000
/content/images/imagesn/n/v/v/nvv32f00/titx1318.15.tif
```

```
/content/DocImages/training/memo/titx1318.15.tif
258000/320000
/content/images/imagesj/j/j/i/jji03d00/514932296_514932297.tif
/content/DocImages/training/letter/514932296_514932297.tif
259000/320000
/content/images/imagesb/b/q/v/bqv38d00/2000599180_9181.tif
/content/DocImages/training/budget/2000599180_9181.tif
260000/320000
/content/images/imagesb/b/b/i/bbi38e00/1001823159_1001823171.tif
/content/DocImages/training/scientific report/1001823159_1001823171.tif
261000/320000
/content/images/imagesn/n/c/e/nce24c00/2048295292_5294.tif
/content/DocImages/training/news article/2048295292_5294.tif
262000/320000
/content/images/imagesy/y/y/x/yyx35a00/86184339_4377.tif
/content/DocImages/training/presentation/86184339_4377.tif
263000/320000
/content/images/imagesg/g/m/q/gmq06d00/50519934-9934.tif
/content/DocImages/training/file folder/50519934-9934.tif
264000/320000
/content/images/imagesw/w/l/d/wld02f00/okl00215.94.tif
/content/DocImages/training/memo/okl00215.94.tif
265000/320000
/content/images/imagest/t/k/u/tku07c00/50639239-9240.tif
/content/DocImages/training/resume/50639239-9240.tif
266000/320000
/content/images/imagesz/z/k/q/zkq82d00/517125620_517125621.tif
/content/DocImages/training/letter/517125620_517125621.tif
267000/320000
/content/images/imagesg/g/b/f/gbf71f00/2000459645.tif
/content/DocImages/training/scientific report/2000459645.tif
268000/320000
/content/images/imagesq/q/q/y/qqy24a00/2067679638_2067679639.tif
/content/DocImages/training/email/2067679638_2067679639.tif
269000/320000
/content/images/imagesm/m/p/x/mpx45f00/0060035652.tif
/content/DocImages/training/invoice/0060035652.tif
270000/320000
/content/images/imagesu/u/o/l/uol15f00/0060197831.tif
/content/DocImages/training/questionnaire/0060197831.tif
271000/320000
/content/images/imagesr/r/m/q/rmq33e00/2050681995.tif
/content/DocImages/training/file folder/2050681995.tif
272000/320000
/content/images/imagesp/p/y/u/pyu39c00/2505226832a_6833.tif
/content/DocImages/training/email/2505226832a_6833.tif
273000/320000
/content/images/imagesq/q/h/n/qhn94c00/2505511294_1297.tif
/content/DocImages/training/scientific publication/2505511294_1297.tif
274000/320000
/content/images/imagesq/q/p/t/qpt30d00/523754122+-4122.tif
/content/DocImages/training/advertisement/523754122+-4122.tif
275000/320000
/content/images/imagesw/w/e/q/weq30c00/ti01411013.tif
/content/DocImages/training/invoice/ti01411013.tif
276000/320000
/content/images/imagesu/u/j/j/ujj89c00/50267224-7227.tif
/content/DocImages/training/resume/50267224-7227.tif
277000/320000
/content/images/imagesw/w/h/d/whd10d00/50367660-7661.tif
/content/DocImages/training/resume/50367660-7661.tif
```

```
278000/320000
/content/images/imagesa/a/r/t/art41c00/2085773934c.tif
/content/DocImages/training/email/2085773934c.tif
279000/320000
/content/images/imagesp/p/h/t/pht06d00/50524934-4935.tif
/content/DocImages/training/resume/50524934-4935.tif
280000/320000
/content/images/imagesv/v/s/r/vsr3aa00/11241352.tif
/content/DocImages/training/scientific report/11241352.tif
281000/320000
/content/images/imagesf/f/f/g/ffg14f00/0000184351.tif
/content/DocImages/training/budget/0000184351.tif
282000/320000
/content/images/imagesx/x/g/y/xgy36c00/2070494271_4274.tif
/content/DocImages/training/advertisement/2070494271_4274.tif
283000/320000
/content/images/imagesf/f/c/r/fcr85e00/2025857564.tif
/content/DocImages/training/memo/2025857564.tif
284000/320000
/content/images/imagesf/f/r/l/frl95c00/2073184659.tif
/content/DocImages/training/memo/2073184659.tif
285000/320000
/content/images/imagesc/c/l/i/cli78e00/2023701738_2023701739.tif
/content/DocImages/training/budget/2023701738_2023701739.tif
286000/320000
/content/images/imagesc/c/i/f/cif12e00/2028707103.tif
/content/DocImages/training/invoice/2028707103.tif
287000/320000
/content/images/imagesw/w/s/r/wsr35d00/504956341_504956342.tif
/content/DocImages/training/questionnaire/504956341_504956342.tif
288000/320000
/content/images/imagesl/l/b/a/lba65f00/0060219777.tif
/content/DocImages/training/memo/0060219777.tif
289000/320000
/content/images/imagesg/g/l/a/gla51f00/0000956232.tif
/content/DocImages/training/scientific report/0000956232.tif
290000/320000
/content/images/imagesr/r/i/j/rij82f00/tob07801.64_tob07801.65.tif
/content/DocImages/training/budget/tob07801.64_tob07801.65.tif
291000/320000
/content/images/imagesm/m/y/r/myr30e00/87430331.tif
/content/DocImages/training/scientific report/87430331.tif
292000/320000
/content/images/imagesg/g/y/a/gya92a00/522934200+-4200.tif
/content/DocImages/training/advertisement/522934200+-4200.tif
293000/320000
/content/images/imagesn/n/z/c/nzc33e00/2021350970_2021350971.tif
/content/DocImages/training/presentation/2021350970_2021350971.tif
294000/320000
/content/images/imagesh/h/u/f/huf37c00/2064983487.tif
/content/DocImages/training/form/2064983487.tif
295000/320000
/content/images/imagesp/p/k/j/pkj07e00/2058096006.tif
/content/DocImages/training/specification/2058096006.tif
296000/320000
/content/images/imagesy/y/x/y/yxy80d00/522738019+-8020.tif
/content/DocImages/training/questionnaire/522738019+-8020.tif
297000/320000
/content/images/imageso/o/u/y/ouy47d00/2057914164_4165.tif
/content/DocImages/training/presentation/2057914164_4165.tif
298000/320000
```

```
/content/images/imagesa/a/v/i/avi71a00/2057434720.tif
/content/DocImages/training/specification/2057434720.tif
299000/320000
/content/images/imagesk/k/v/j/kvj75f00/0060207922.tif
/content/DocImages/training/file folder/0060207922.tif
300000/320000
/content/images/imagesf/f/c/r/fcr84c00/98890531_0532.tif
/content/DocImages/training/form/98890531_0532.tif
301000/320000
/content/images/imagesn/n/q/v/nqv00a00/10321319_10321323.tif
/content/DocImages/training/scientific publication/10321319_10321323.tif
302000/320000
/content/images/imagesz/z/a/a/zaa43a00/71895677.tif
/content/DocImages/training/advertisement/71895677.tif
303000/320000
/content/images/imagesa/a/h/w/ahw41e00/80412046.tif
/content/DocImages/training/letter/80412046.tif
304000/320000
/content/images/imagesh/h/v/s/hvs62c00/2077747041.tif
/content/DocImages/training/email/2077747041.tif
305000/320000
/content/images/imagesf/f/b/a/fba81c00/2501062679_2684.tif
/content/DocImages/training/presentation/2501062679_2684.tif
306000/320000
/content/images/imagesx/x/u/e/xue20e00/01754062_01754063.tif
/content/DocImages/training/invoice/01754062_01754063.tif
307000/320000
/content/images/imagesv/v/a/a/vaa42e00/2504026229.tif
/content/DocImages/training/presentation/2504026229.tif
308000/320000
/content/images/imagesr/r/g/d/rgd84c00/92237670_7671.tif
/content/DocImages/training/advertisement/92237670_7671.tif
309000/320000
/content/images/imagesr/r/v/e/rve47e00/2030239239_2030239240.tif
/content/DocImages/training/form/2030239239_2030239240.tif
310000/320000
/content/images/imagesz/z/w/j/zwj13d00/514020149_514020176.tif
/content/DocImages/training/specification/514020149_514020176.tif
311000/320000
/content/images/imagesj/j/k/l/jkl25c00/2078618140_8144.tif
/content/DocImages/training/presentation/2078618140_8144.tif
312000/320000
/content/images/imagest/t/y/k/tyk54a00/92206388_6390.tif
/content/DocImages/training/invoice/92206388_6390.tif
313000/320000
/content/images/imagesz/z/x/x/zxx1aa00/10041635_10041639.tif
/content/DocImages/training/scientific publication/10041635_10041639.tif
314000/320000
/content/images/imagesh/h/f/x/hfx45c00/2076287354_7373.tif
/content/DocImages/training/scientific publication/2076287354_7373.tif
315000/320000
/content/images/imageso/o/d/j/odj84c00/93127466_7471.tif
/content/DocImages/training/budget/93127466_7471.tif
316000/320000
/content/images/imagesi/i/d/q/idq64e00/1000359230_1000359236.tif
/content/DocImages/training/scientific report/1000359230_1000359236.tif
317000/320000
/content/images/imagesm/m/j/u/mju33a00/504001920+-1922.tif
/content/DocImages/training/specification/504001920+-1922.tif
318000/320000
/content/images/imagesj/j/v/f/jvf87d00/2076881857.tif
```

```
/content/DocImages/training/scientific report/2076881857.tif
319000/320000
/content/images/imagesw/w/s/c/wsc07c00/50520540-0540.tif
/content/DocImages/training/memo/50520540-0540.tif
```

In [0]:

```
labels = open('/content/labels/test.txt', 'r')
labels = labels.read()
```

In [0]:

```
import os
classes = {'0': 'letter', '1': 'form', '2': 'email', '3': 'handwritten', '4': 'adverti
           '6': 'scientific publication', '7': 'specification', '8': 'file folder', '9
           '11': 'invoice', '12': 'presentation', '13': 'questionnaire', '14': 'resume

root_path = '/content/images/images'

def make_doc_dirs(path):
    for label in classes.values():
        if not os.path.exists(path + str(label)):
            os.mkdir(path + str(label))

make_doc_dirs('/content/DocImages/training/')
make_doc_dirs('/content/DocImages/testing/')
make_doc_dirs('/content/DocImages/validation/')
```

In [0]:

```
labels_split = labels.split('images')
full_paths = []
for image_label in labels_split:
    label = root_path + image_label
    full_paths.append(label)
```

In [0]:

```
print(len(full_paths))
print(full_paths[0:10])
```

```
40001
['/content/images/images', '/content/images/imagesr/r/g/e/rge31d00/503210033
+-0034.tif 3\n', '/content/images/imagesc/c/e/j/cej80d00/517306722+-6724.tif
3\n', '/content/images/imagesm/m/r/r/mrr36d00/50603620-3621.tif 14\n', '/con
tent/images/imagesg/g/t/u/gtu29c00/2084573574a.tif 2\n', '/content/images/im
agesh/h/o/f/hof08d00/2071783492.tif 9\n', '/content/images/imagesx/x/a/b/xab
71f00/1002977593_1002977622.tif 6\n', '/content/images/imageso/o/k/s/oks31f0
0/0001437969.tif 13\n', '/content/images/imagesj/j/t/o/jto61f00/2050283643.t
if 8\n', '/content/images/imagesh/h/j/g/hjg89e00/0000049717.tif 0\n']
```

In [0]:

```
full_paths = full_paths[1:]
paths = []
for path in full_paths:
    path = path.split('\n')
    for p in path:
        p = p.rsplit(' ',1)
        paths.append(p)

paths = [x for x in paths if x != ['']]
print(len(paths))
print(paths[:5])
```

```
40000
[['/content/images/imagesr/r/g/e/rge31d00/503210033+-0034.tif', '3'], ['/con
tent/images/imagesc/c/e/j/cej80d00/517306722+-6724.tif', '3'], ['/content/im
ages/imagesm/m/r/r/mrr36d00/50603620-3621.tif', '14'], ['/content/images/ima
gesg/g/t/u/gtu29c00/2084573574a.tif', '2'], ['/content/images/imagesh/h/o/f/
hof08d00/2071783492.tif', '9']]
```

In [0]:

```
i = 0
for p in paths[:39000]:
    old_path = p[0]
    full_path = p[0].split('images/')
    doctype = p[1]
    for key, value in classes.items():
        if doctype == key:
            doctype = value
    root_path = full_path[0]
    image_filename = full_path[1].rsplit('/', 1)[1]
    new_path = root_path +'DocImages/testing/'+str(doctype) + '/' + image_filename

    if not os.path.exists(new_path):
        os.rename(old_path, new_path)
        i += 1
        if i % 1000 == 0:
            print(str(i) + "/39000")
            print(old_path)
            print(new_path)
```

```
1000/39000
/content/images/imagesm/m/d/o/mdo43e00/2024216699.tif
/content/DocImages/testing/file folder/2024216699.tif
2000/39000
/content/images/imagesa/a/b/l/abl24e00/2024961531.tif
/content/DocImages/testing/file folder/2024961531.tif
3000/39000
/content/images/imagesu/u/p/z/upz84c00/93503341.tif
/content/DocImages/testing/budget/93503341.tif
4000/39000
/content/images/imagesm/m/m/e/mme21d00/515944614+-4616.tif
/content/DocImages/testing/handwritten/515944614+-4616.tif
5000/39000
/content/images/imagesp/p/i/c/pic03a00/518599493+-9497.tif
/content/DocImages/testing/form/518599493+-9497.tif
6000/39000
/content/images/imagese/e/y/d/eyd73c00/2065518028_8030.tif
/content/DocImages/testing/news article/2065518028_8030.tif
7000/39000
/content/images/imagesl/l/n/x/lnx01f00/0001254743.tif
/content/DocImages/testing/memo/0001254743.tif
8000/39000
/content/images/imagesn/n/t/m/ntm40f00/0000525355.tif
/content/DocImages/testing/memo/0000525355.tif
9000/39000
/content/images/imagesc/c/p/t/cpt21e00/87596052_87596054.tif
/content/DocImages/testing/scientific report/87596052_87596054.tif
10000/39000
/content/images/imagesp/p/a/j/paj13e00/2060540708_2060540715.tif
/content/DocImages/testing/presentation/2060540708_2060540715.tif
11000/39000
/content/images/imagesm/m/e/k/mek30c00/corti0013538b.tif
/content/DocImages/testing/news article/corti0013538b.tif
12000/39000
/content/images/imagesg/g/w/g/gwg85e00/2024002176_2024002177.tif
/content/DocImages/testing/questionnaire/2024002176_2024002177.tif
13000/39000
/content/images/imagesq/q/q/t/qqt01d00/518220097+-0098.tif
```

```
/content/DocImages/testing/handwritten/518220097+-0098.tif
14000/39000
/content/images/imagesb/b/q/x/bqx80a00/0060006641.tif
/content/DocImages/testing/memo/0060006641.tif
15000/39000
/content/images/imagesp/p/h/v/phv62f00/tob11409.31_tob11409.42.tif
/content/DocImages/testing/news article/tob11409.31_tob11409.42.tif
16000/39000
/content/images/imagesn/n/x/d/nxd26d00/50554833-4833.tif
/content/DocImages/testing/scientific publication/50554833-4833.tif
17000/39000
/content/images/imagesf/f/b/n/fbn22c00/2069730841.tif
/content/DocImages/testing/specification/2069730841.tif
18000/39000
/content/images/imagesc/c/h/d/chd43a00/96652988.tif
/content/DocImages/testing/budget/96652988.tif
19000/39000
/content/images/imagesu/u/p/b/upb16c00/2028997809_7833.tif
/content/DocImages/testing/scientific report/2028997809_7833.tif
20000/39000
/content/images/imagesz/z/v/y/zvy29e00/2501560754.tif
/content/DocImages/testing/specification/2501560754.tif
21000/39000
/content/images/imagese/e/l/r/elr16c00/2066004247_4251.tif
/content/DocImages/testing/presentation/2066004247_4251.tif
22000/39000
/content/images/imagesp/p/a/g/pag23e00/2058501653.tif
/content/DocImages/testing/advertisement/2058501653.tif
23000/39000
/content/images/imagest/t/h/g/thg32d00/2064266165.tif
/content/DocImages/testing/news article/2064266165.tif
24000/39000
/content/images/imagesb/b/z/h/bzh42e00/2500013473.tif
/content/DocImages/testing/scientific report/2500013473.tif
25000/39000
/content/images/imagesu/u/v/j/uvj75e00/2046399662_2046399669.tif
/content/DocImages/testing/scientific publication/2046399662_2046399669.ti
f
26000/39000
/content/images/imagesj/j/r/h/jrh75a00/2077583999_4000.tif
/content/DocImages/testing/letter/2077583999_4000.tif
27000/39000
/content/images/imagesg/g/a/w/gaw05f00/0060166458.tif
/content/DocImages/testing/file folder/0060166458.tif
28000/39000
/content/images/imagesq/q/f/h/qfh23e00/2058500853.tif
/content/DocImages/testing/advertisement/2058500853.tif
29000/39000
/content/images/imagesr/r/s/g/rsg44d00/506789172_506789177.tif
/content/DocImages/testing/handwritten/506789172_506789177.tif
30000/39000
/content/images/imageso/o/d/f/odf70e00/93774021_93774022.tif
/content/DocImages/testing/news article/93774021_93774022.tif
31000/39000
/content/images/imagesq/q/t/n/qtn40f00/0000511830.tif
/content/DocImages/testing/budget/0000511830.tif
32000/39000
/content/images/imagesd/d/k/m/dkm26d00/50617340-7341.tif
/content/DocImages/testing/resume/50617340-7341.tif
33000/39000
/content/images/imagese/e/a/u/eau67c00/2078250300.tif
```

```
/content/DocImages/testing/memo/2078250300.tif
34000/39000
/content/images/imagesj/j/d/a/jda15c00/2073932077.tif
/content/DocImages/testing/file folder/2073932077.tif
35000/39000
/content/images/imagesv/v/v/y/vvy06d00/50535944-5945.tif
/content/DocImages/testing/resume/50535944-5945.tif
36000/39000
/content/images/imagesz/z/h/n/zhn33c00/92203957_3964.tif
/content/DocImages/testing/advertisement/92203957_3964.tif
37000/39000
/content/images/imagesv/v/s/b/vsb66e00/2041953025.tif
/content/DocImages/testing/budget/2041953025.tif
38000/39000
/content/images/imagesh/h/u/y/huy36d00/50651670-1671.tif
/content/DocImages/testing/resume/50651670-1671.tif
```

In [0]:

```
labels = open('/content/labels/val.txt', 'r')
labels = labels.read()
```

In [0]:

```
import os
classes = {'0': 'letter', '1': 'form', '2': 'email', '3': 'handwritten', '4': 'adverti
           '6': 'scientific publication', '7': 'specification', '8': 'file folder', '9
           '11': 'invoice', '12': 'presentation', '13': 'questionnaire', '14': 'resume

root_path = '/content/images/images'

def make_doc_dirs(path):
    for label in classes.values():
        if not os.path.exists(path + str(label)):
            os.mkdir(path + str(label))

make_doc_dirs('/content/DocImages/training/')
make_doc_dirs('/content/DocImages/testing/')
make_doc_dirs('/content/DocImages/validation/')
```

In [0]:

```
labels_split = labels.split('images')
full_paths = []
for image_label in labels_split:
    label = root_path + image_label
    full_paths.append(label)
```

In [0]:

```
print(len(full_paths))
print(full_paths[0:10])
```

```
40001
['/content/images/images', '/content/images/imagesg/g/t/h/gth35e00/202452566
1.tif 11\n', '/content/images/imagesi/i/y/k/iyk38c00/512015827+-5827.tif 0
\n', '/content/images/imagesr/r/r/e/rre21e00/87103403.tif 0\n', '/content/im
ages/imagesk/k/s/u/ksu44c00/03636607.tif 4\n', '/content/images/imagesr/r/a/
i/rai09d00/50437856-7857.tif 14\n', '/content/images/imagesd/d/q/j/dqj13f00/
0000457436.tif 12\n', '/content/images/imagesx/x/o/g/xog20a00/10121367.tif 6
\n', '/content/images/imagesh/h/p/z/hpz84c00/93503327.tif 10\n', '/content/i
mages/imagesa/a/h/t/aht78d00/502596897.tif 4\n']
```

In [0]:

```
full_paths = full_paths[1:]
paths = []
for path in full_paths:
    path = path.split('\n')
    for p in path:
        p = p.rsplit(' ',1)
        paths.append(p)

paths = [x for x in paths if x != ['']]
print(len(paths))
print(paths[:5])
```

```
40000
[['/content/images/imagesg/g/t/h/gth35e00/2024525661.tif', '11'], ['/conten
t/images/imagesi/i/y/k/iyk38c00/512015827+-5827.tif', '0'], ['/content/image
s/imagesr/r/r/e/rre21e00/87103403.tif', '0'], ['/content/images/imagesk/k/s/
u/ksu44c00/03636607.tif', '4'], ['/content/images/imagesr/r/a/i/rai09d00/504
37856-7857.tif', '14']]
```

In [0]:

```
i = 0
for p in paths[:39000]:
    old_path = p[0]
    full_path = p[0].split('images/')
    doctype = p[1]
    for key, value in classes.items():
        if doctype == key:
            doctype = value
    root_path = full_path[0]
    image_filename = full_path[1].rsplit('/', 1)[1]
    new_path = root_path +'DocImages/validation/'+str(doctype) + '/' + image_filename

    if not os.path.exists(new_path):
        os.rename(old_path, new_path)
        i += 1
        if i % 1000 == 0:
            print(str(i) + "/39000")
            print(old_path)
            print(new_path)
```

```
1000/39000
/content/images/imagesu/u/o/v/uov22f00/tim01175.99.tif
/content/DocImages/validation/budget/tim01175.99.tif
2000/39000
/content/images/imagese/e/r/c/erc10d00/50365965-5965.tif
/content/DocImages/validation/file folder/50365965-5965.tif
3000/39000
/content/images/imagesx/x/g/o/xgo33e00/2053630098.tif
/content/DocImages/validation/file folder/2053630098.tif
4000/39000
/content/images/imagesa/a/j/v/ajv27e00/2028871490.tif
/content/DocImages/validation/questionnaire/2028871490.tif
5000/39000
/content/images/imagesj/j/a/h/jah72f00/tob09830.73.tif
/content/DocImages/validation/news article/tob09830.73.tif
6000/39000
/content/images/imagesb/b/d/l/bdl37e00/2028953543.tif
/content/DocImages/validation/letter/2028953543.tif
7000/39000
/content/images/imagesc/c/h/d/chd23e00/2058503762.tif
/content/DocImages/validation/advertisement/2058503762.tif
8000/39000
/content/images/imagesl/l/f/r/lfr99c00/40045130-5131.tif
/content/DocImages/validation/resume/40045130-5131.tif
9000/39000
/content/images/imagesw/w/h/g/whg39d00/501762244.tif
/content/DocImages/validation/letter/501762244.tif
10000/39000
/content/images/imagesp/p/i/k/pik06d00/50507336-7336.tif
/content/DocImages/validation/scientific publication/50507336-7336.tif
11000/39000
/content/images/imagesp/p/u/c/puc00d00/50313290-3290.tif
/content/DocImages/validation/file folder/50313290-3290.tif
12000/39000
/content/images/imagesv/v/g/b/vgb79e00/2029148786.tif
/content/DocImages/validation/presentation/2029148786.tif
13000/39000
/content/images/imagesl/l/q/c/lqc15c00/2025623388.tif
```

```
/content/DocImages/validation/file folder/2025623388.tif
14000/39000
/content/images/imagesf/f/j/o/fjo93f00/0000349265.tif
/content/DocImages/validation/advertisement/0000349265.tif
15000/39000
/content/images/imagese/e/c/t/ect91a00/2063070687.tif
/content/DocImages/validation/file folder/2063070687.tif
16000/39000
/content/images/imagesx/x/x/x/xxx97c00/527795844+-5850.tif
/content/DocImages/validation/email/527795844+-5850.tif
17000/39000
/content/images/imagesn/n/h/l/nhl64e00/1003033270.tif
/content/DocImages/validation/memo/1003033270.tif
18000/39000
/content/images/imagest/t/i/i/tii02c00/2085029935_9958.tif
/content/DocImages/validation/presentation/2085029935_9958.tif
19000/39000
/content/images/imagesi/i/b/r/ibr62f00/tob11709.41.tif
/content/DocImages/validation/news article/tob11709.41.tif
20000/39000
/content/images/imagesd/d/w/j/dwj24f00/0000160498.tif
/content/DocImages/validation/advertisement/0000160498.tif
21000/39000
/content/images/imagest/t/b/x/tbx24e00/2023683453_2023683470.tif
/content/DocImages/validation/questionnaire/2023683453_2023683470.tif
22000/39000
/content/images/imagesy/y/s/q/ysq51a00/0011810523.tif
/content/DocImages/validation/resume/0011810523.tif
23000/39000
/content/images/imagesm/m/z/i/mzi36d00/50542317-2318.tif
/content/DocImages/validation/resume/50542317-2318.tif
24000/39000
/content/images/imagesg/g/w/u/gwu60f00/0011972925.tif
/content/DocImages/validation/presentation/0011972925.tif
25000/39000
/content/images/imagesw/w/y/b/wyb61c00/516785188+-5192.tif
/content/DocImages/validation/form/516785188+-5192.tif
26000/39000
/content/images/imagesc/c/e/j/cej23e00/1005103529-c.tif
/content/DocImages/validation/news article/1005103529-c.tif
27000/39000
/content/images/imagesw/w/w/i/wwi21c00/71953697.tif
/content/DocImages/validation/email/71953697.tif
28000/39000
/content/images/imagesr/r/o/m/rom03d00/514832233_514832233a.tif
/content/DocImages/validation/advertisement/514832233_514832233a.tif
29000/39000
/content/images/imagesh/h/e/x/hex61a00/2057402604_2057402605.tif
/content/DocImages/validation/specification/2057402604_2057402605.tif
30000/39000
/content/images/imagesc/c/s/d/csd12c00/2085113042.tif
/content/DocImages/validation/email/2085113042.tif
31000/39000
/content/images/imagesg/g/m/g/gmg80e00/89679233.tif
/content/DocImages/validation/memo/89679233.tif
32000/39000
/content/images/imagesv/v/d/s/vds35f00/0060061083.tif
/content/DocImages/validation/questionnaire/0060061083.tif
33000/39000
/content/images/imagesv/v/i/n/vin25c00/2505619008.tif
/content/DocImages/validation/scientific report/2505619008.tif
```

```
34000/39000
/content/images/imagesp/p/g/k/pgk88c00/00952491_2510.tif
/content/DocImages/validation/scientific report/00952491_2510.tif
35000/39000
/content/images/imagesn/n/x/p/nxp54f00/0060281221.tif
/content/DocImages/validation/file folder/0060281221.tif
36000/39000
/content/images/imagesm/m/z/s/mzs68e00/2001206054.tif
/content/DocImages/validation/questionnaire/2001206054.tif
37000/39000
/content/images/imagesm/m/q/c/mqc39e00/2501561459_2501561460.tif
/content/DocImages/validation/scientific report/2501561459_2501561460.tif
38000/39000
/content/images/imagese/e/f/a/efa67c00/2073440776.tif
/content/DocImages/validation/invoice/2073440776.tif
```

In [0]:

```
print(paths[1])
```

```
['/content/images/imagesi/i/y/k/iyk38c00/512015827+-5827.tif', '0']
```

In [0]:

```
make_doc_dirs('/content/DocumentImages/train/')
make_doc_dirs('/content/DocumentImages/test/')
make_doc_dirs('/content/DocumentImages/valid/')
```

In [0]:

```
%%time
path = '/content/DocImages/training/'
roots = []
for root, dirs, files in os.walk(path):
    roots.append(root)
print(roots[1:])
```

```
['/content/DocImages/training/email', '/content/DocImages/training/letter',
'/content/DocImages/training/resume', '/content/DocImages/training/file fold
er', '/content/DocImages/training/handwritten', '/content/DocImages/trainin
g/scientific publication', '/content/DocImages/training/invoice', '/content/
DocImages/training/specification', '/content/DocImages/training/budget', '/c
ontent/DocImages/training/scientific report', '/content/DocImages/training/p
resentation', '/content/DocImages/training/news article', '/content/DocImage
s/training/questionnaire', '/content/DocImages/training/form', '/content/Doc
Images/training/memo', '/content/DocImages/training/advertisement']
CPU times: user 244 ms, sys: 115 ms, total: 359 ms
Wall time: 381 ms
```

In [0]:

```
%%time
import random
path = '/content/DocImages/training/'
roots = []
for root, dirs, files in os.walk(path):
    roots.append(root)

random.seed(29)
i = 1
for root in roots[1:]:
    print("Subsetting images from class {}/16".format(i))
    print(root)
    images = []
    for root, dirs, files in os.walk(root):
        for file in files:
            images.append(str(root)+os.sep+str(file))
    random.shuffle(images)
    images_subset = images[:19500]
    print(len(images_subset))
    for image in images_subset:
        split_im = image.split("DocImages/training")
        os.rename(image, split_im[0]+'DocumentImages/train'+split_im[1])
    i += 1
```

```
Subsetting images from class 1/16
/content/DocImages/training/email
19500
Subsetting images from class 2/16
/content/DocImages/training/letter
19500
Subsetting images from class 3/16
/content/DocImages/training/resume
19500
Subsetting images from class 4/16
/content/DocImages/training/file folder
19500
Subsetting images from class 5/16
/content/DocImages/training/handwritten
19500
Subsetting images from class 6/16
/content/DocImages/training/scientific publication
19500
Subsetting images from class 7/16
/content/DocImages/training/invoice
19500
Subsetting images from class 8/16
/content/DocImages/training/specification
19500
Subsetting images from class 9/16
/content/DocImages/training/budget
19500
Subsetting images from class 10/16
/content/DocImages/training/scientific report
19500
Subsetting images from class 11/16
/content/DocImages/training/presentation
19500
Subsetting images from class 12/16
```

```
/content/DocImages/training/news article
19500
Subsetting images from class 13/16
/content/DocImages/training/questionnaire
19500
Subsetting images from class 14/16
/content/DocImages/training/form
19500
Subsetting images from class 15/16
/content/DocImages/training/memo
19500
Subsetting images from class 16/16
/content/DocImages/training/advertisement
19500
CPU times: user 1.27 s, sys: 6.8 s, total: 8.07 s
Wall time: 9.05 s
```

In [0]:

```
%%time
path = '/content/DocImages/testing/'
roots = []
for root, dirs, files in os.walk(path):
    roots.append(root)

random.seed(29)
i = 1
for root in roots[1:]:
    print("Subsetting images from class {}/16".format(i))
    print(root)
    images = []
    for root, dirs, files in os.walk(root):
        for file in files:
            images.append(str(root)+os.sep+str(file))
    random.shuffle(images)
    images_subset = images[:2300]
    print(len(images_subset))
    for image in images_subset:
        split_im = image.split("DocImages/testing")
        os.rename(image, split_im[0]+'DocumentImages/test'+split_im[1])
    i += 1
```

```
Subsetting images from class 1/16
/content/DocImages/testing/email
2300
Subsetting images from class 2/16
/content/DocImages/testing/letter
2300
Subsetting images from class 3/16
/content/DocImages/testing/resume
2300
Subsetting images from class 4/16
/content/DocImages/testing/file folder
2300
Subsetting images from class 5/16
/content/DocImages/testing/handwritten
2300
Subsetting images from class 6/16
/content/DocImages/testing/scientific publication
2300
Subsetting images from class 7/16
/content/DocImages/testing/invoice
2300
Subsetting images from class 8/16
/content/DocImages/testing/specification
2300
Subsetting images from class 9/16
/content/DocImages/testing/budget
2300
Subsetting images from class 10/16
/content/DocImages/testing/scientific report
2300
Subsetting images from class 11/16
/content/DocImages/testing/presentation
2300
Subsetting images from class 12/16
/content/DocImages/testing/news article
```

```
2300
Subsetting images from class 13/16
/content/DocImages/testing/questionnaire
2300
Subsetting images from class 14/16
/content/DocImages/testing/form
2300
Subsetting images from class 15/16
/content/DocImages/testing/memo
2300
Subsetting images from class 16/16
/content/DocImages/testing/advertisement
2300
CPU times: user 162 ms, sys: 844 ms, total: 1.01 s
Wall time: 1.02 s
```

In [0]:

```
%%time
path = '/content/DocImages/validation/'
roots = []
for root, dirs, files in os.walk(path):
    roots.append(root)

random.seed(29)
i = 1
for root in roots[1:]:
    print("Subsetting images from class {}/16".format(i))
    print(root)
    images = []
    for root, dirs, files in os.walk(root):
        for file in files:
            images.append(str(root)+os.sep+str(file))
    random.shuffle(images)
    images_subset = images[:2300]
    print(len(images_subset))
    for image in images_subset:
        split_im = image.split("DocImages/validation")
        os.rename(image, split_im[0]+'DocumentImages/valid'+split_im[1])
    i += 1
```

```
Subsetting images from class 1/16
/content/DocImages/validation/email
2300
Subsetting images from class 2/16
/content/DocImages/validation/letter
2300
Subsetting images from class 3/16
/content/DocImages/validation/resume
2300
Subsetting images from class 4/16
/content/DocImages/validation/file folder
2300
Subsetting images from class 5/16
/content/DocImages/validation/handwritten
2300
Subsetting images from class 6/16
/content/DocImages/validation/scientific publication
2300
Subsetting images from class 7/16
/content/DocImages/validation/invoice
2300
Subsetting images from class 8/16
/content/DocImages/validation/specification
2300
Subsetting images from class 9/16
/content/DocImages/validation/budget
2300
Subsetting images from class 10/16
/content/DocImages/validation/scientific report
2300
Subsetting images from class 11/16
/content/DocImages/validation/presentation
2300
Subsetting images from class 12/16
/content/DocImages/validation/news article
```

```
2300
Subsetting images from class 13/16
/content/DocImages/validation/questionnaire
2300
Subsetting images from class 14/16
/content/DocImages/validation/form
2300
Subsetting images from class 15/16
/content/DocImages/validation/memo
2300
Subsetting images from class 16/16
/content/DocImages/validation/advertisement
2300
CPU times: user 158 ms, sys: 768 ms, total: 925 ms
Wall time: 928 ms
```

In [0]:

```
1
```

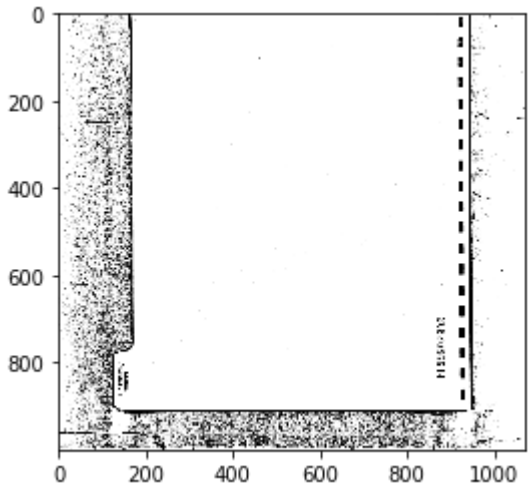
In [0]:

```
# Extract all image paths
def extract_file_paths(path):
    image_filenames = []
    for root, dirs, files in os.walk(path):
        if len(files) > 0:
            for file in files:
                if(file[-3:] == "tif" or file[-3:] == "Tif"):
                    image_filenames.append(str(root)+os.sep+str(file))
    print(len(image_filenames))
    return image_filenames
```

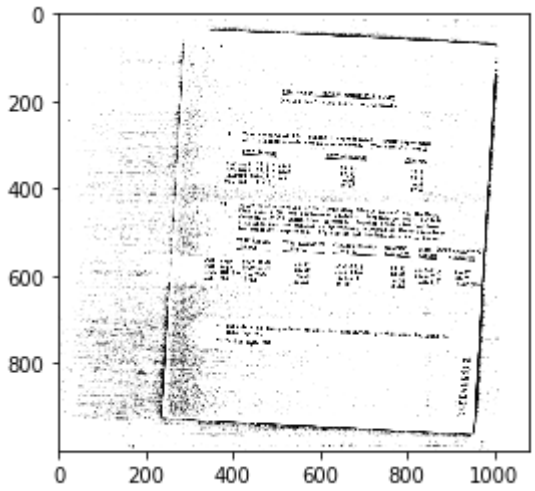
In [0]:

```
image_files = extract_file_paths("/content/DocumentImages/train")
#Get image dimensions
height = []
width = []
docs = []

for image in image_files:
    path = os.path.dirname(image)
    base = os.path.basename(path)
    docs.append(base)
    img = cv2.imread(image)
    height.append(img.shape[0])
    width.append(img.shape[1])
#Print images wider than taller
    if img.shape[1] > 1000:
        print(base)
        print(image)
        plt.imshow(img)
        plt.show()
```

```
312000
scientific report
/content/DocumentImages/train/scientific report/2058086644_2058086726.tif
```

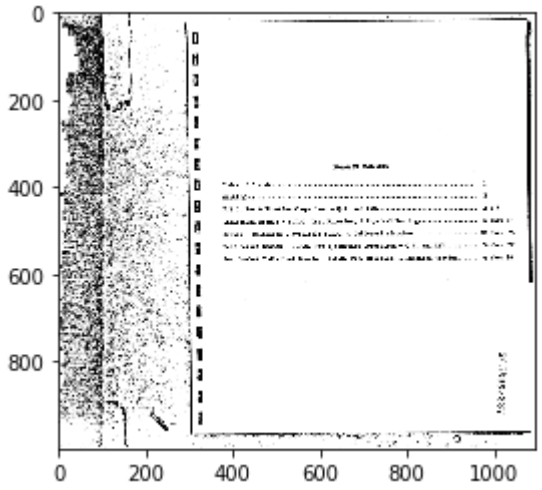



```
scientific report
/content/DocumentImages/train/scientific report/2058083372.tif
```

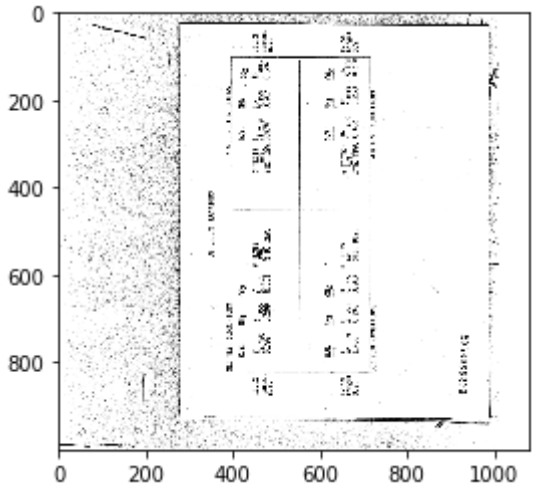



```
scientific report
```

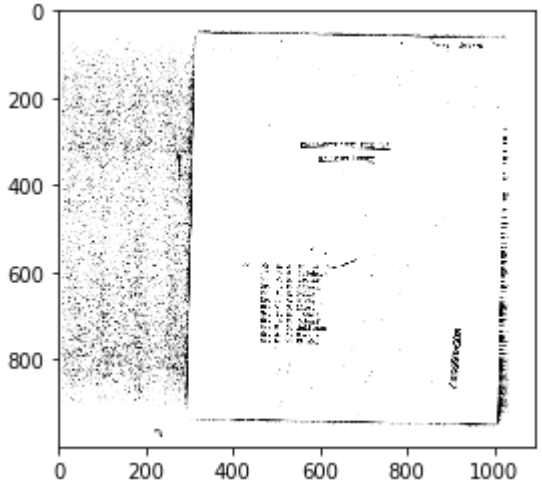
```
/content/DocumentImages/train/scientific report/2058087237_2058087320.tif
```



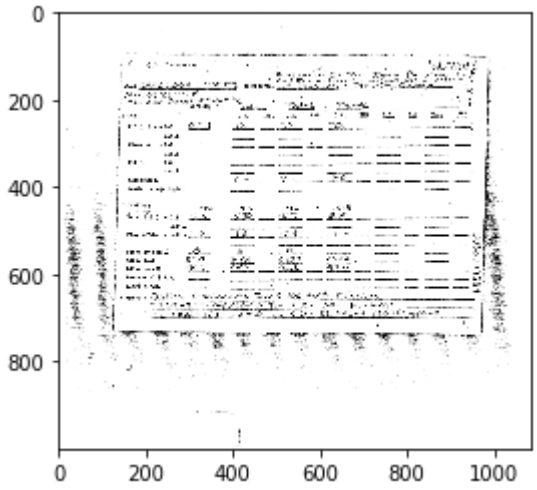

```
scientific report
/content/DocumentImages/train/scientific report/2058085273_2058085278.tif
```



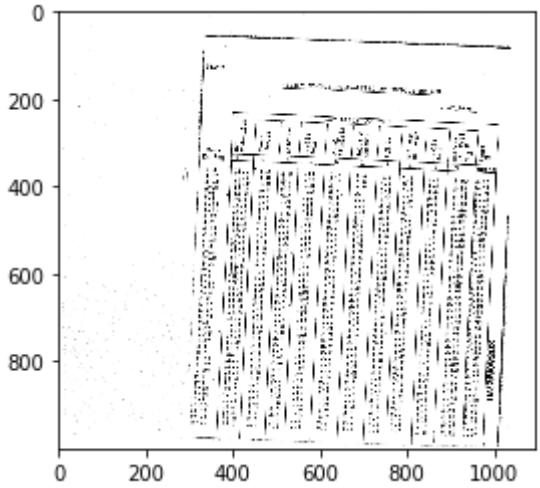

```
scientific report
/content/DocumentImages/train/scientific report/2058088527_2058088530.tif
```



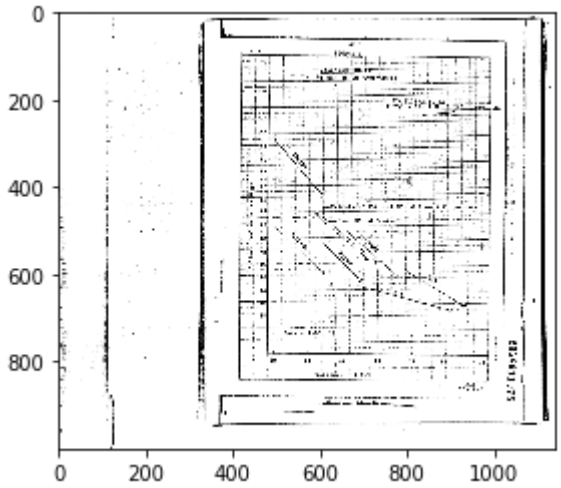

```
scientific report
/content/DocumentImages/train/scientific report/2058009564.tif
```



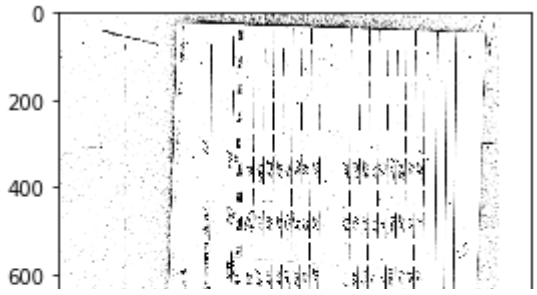
```
scientific report
/content/DocumentImages/train/scientific report/2058088570.tif
```



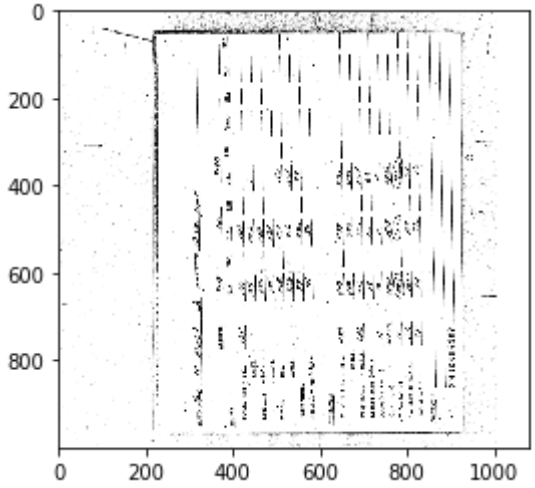
```
scientific report
/content/DocumentImages/train/scientific report/2058011975_2058011981.tif
```



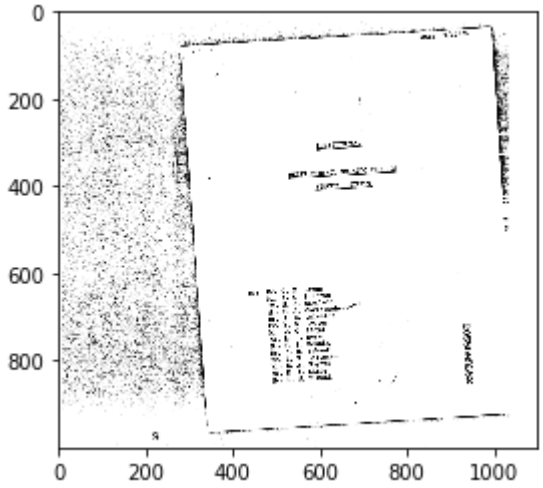
```
scientific report
/content/DocumentImages/train/scientific report/2058083882_2058083883.tif
```



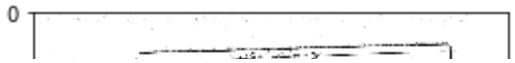

```
scientific report
/content/DocumentImages/train/scientific report/2058083963.tif
```



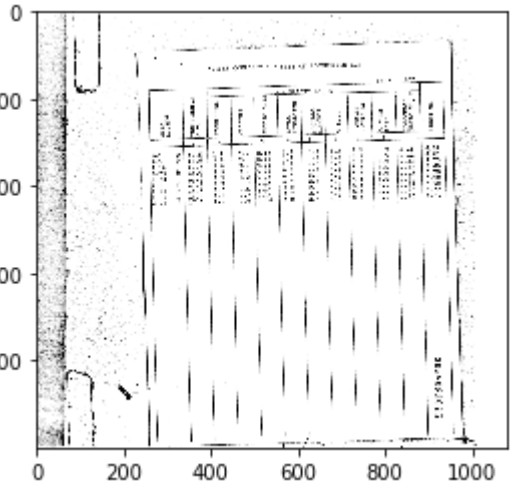

```
scientific report
/content/DocumentImages/train/scientific report/2058088382_2058088386.tif
```



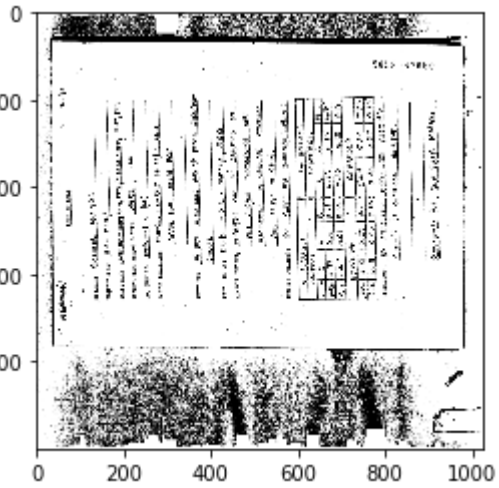

```
scientific report
/content/DocumentImages/train/scientific report/2058088712.tif
```



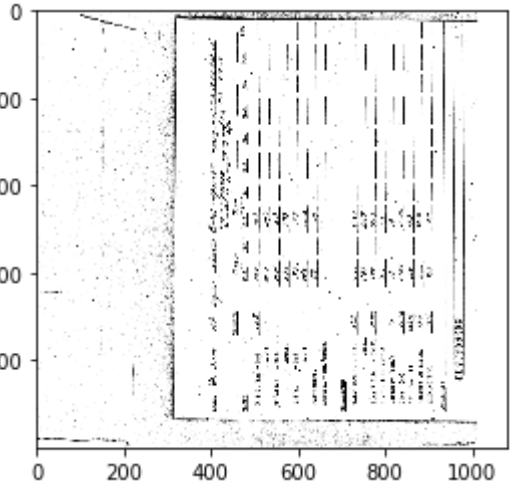
scientific report
/content/DocumentImages/train/scientific report/2058083626.tif



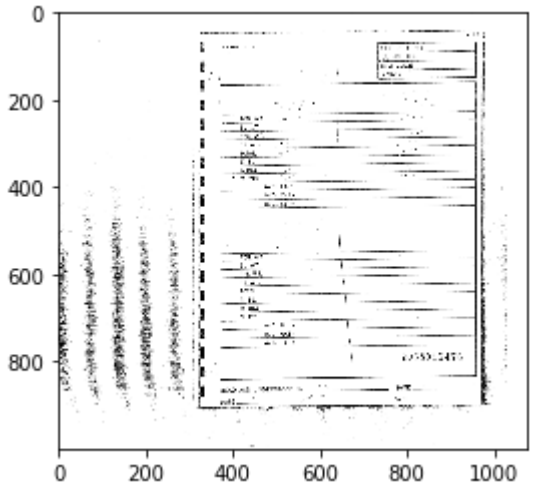
scientific report
/content/DocumentImages/train/scientific report/2058087650.tif



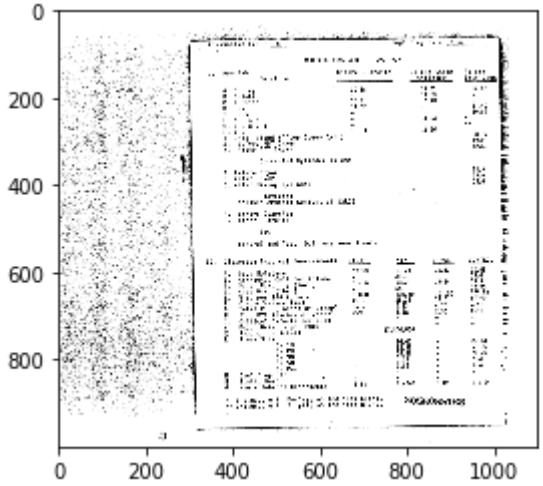
scientific report
/content/DocumentImages/train/scientific report/2058083833.tif



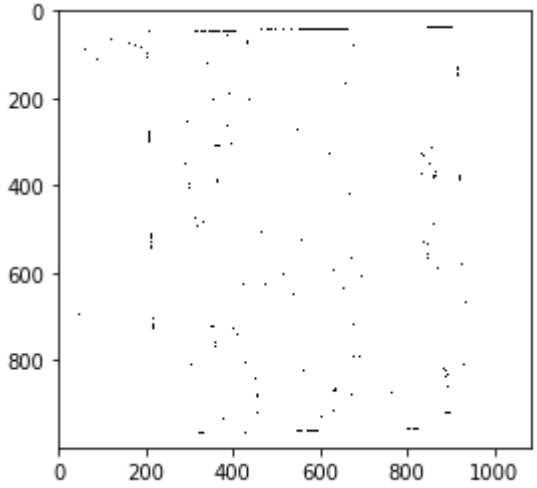
scientific report
/content/DocumentImages/train/scientific report/2058015476.tif



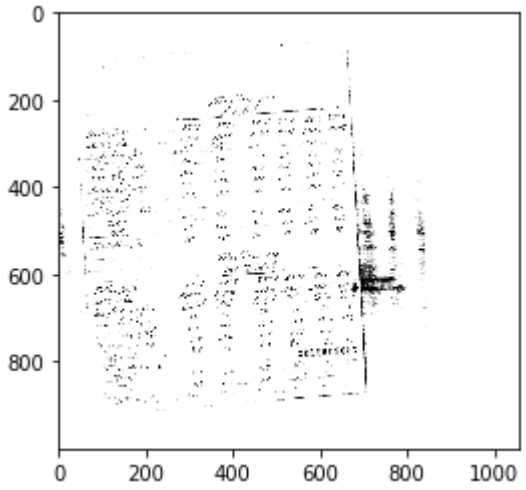

```
scientific report
/content/DocumentImages/train/scientific report/2058088165_2058088166.tif
```



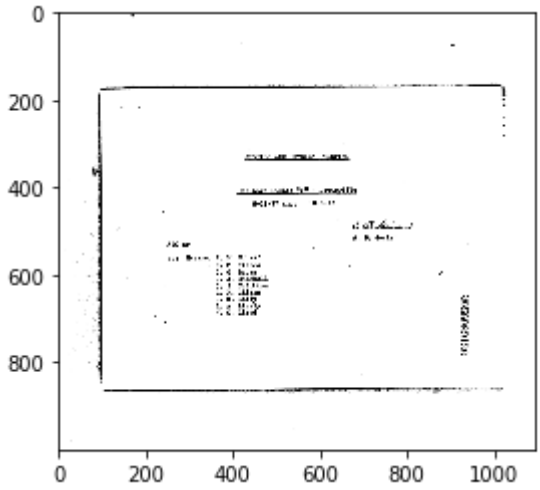

```
scientific report
/content/DocumentImages/train/scientific report/2058083841_2058083842.tif
```



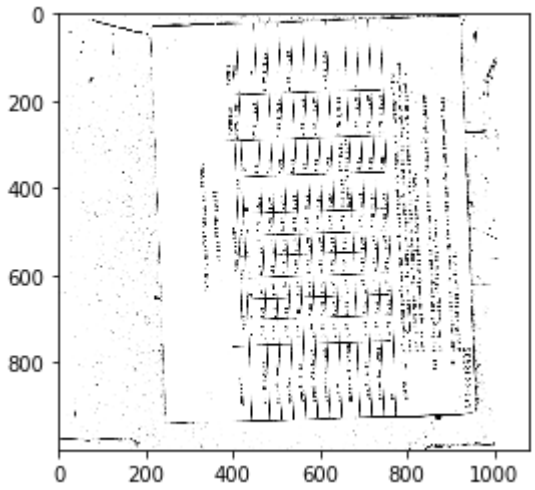

```
scientific report
/content/DocumentImages/train/scientific report/2058008087_2058008088.tif
```



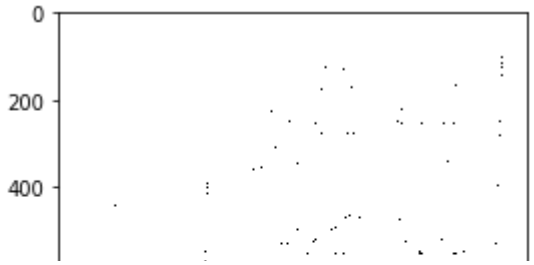

```
scientific report
/content/DocumentImages/train/scientific report/2058089436_2058089437.tif
```



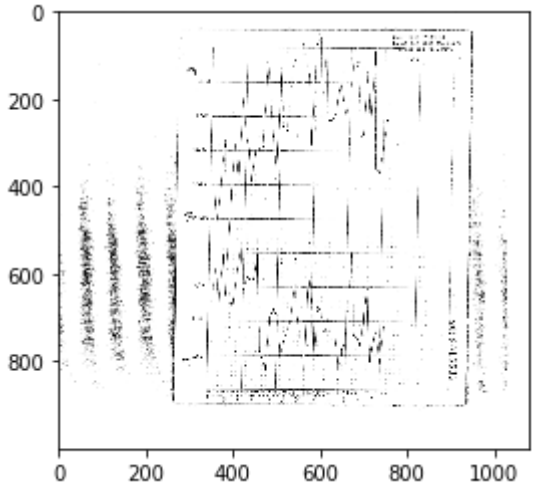

```
scientific report
/content/DocumentImages/train/scientific report/2058083654.tif
```



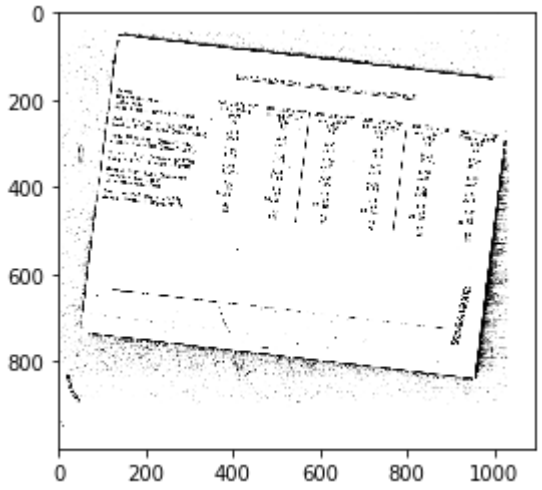

```
scientific report
/content/DocumentImages/train/scientific report/2058016380_2058016384.tif
```




```
scientific report
/content/DocumentImages/train/scientific report/2058015535.tif
```

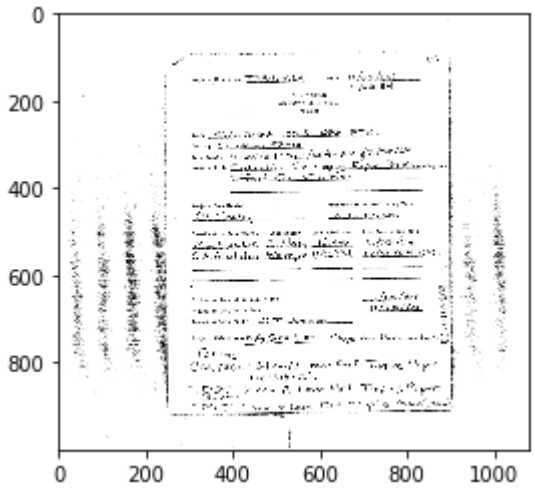



```
scientific report
/content/DocumentImages/train/scientific report/2058088026.tif
```

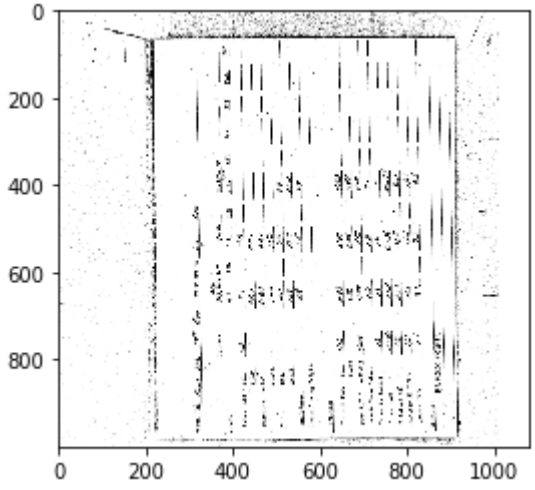



```
scientific report
/content/DocumentImages/train/scientific report/2058083929.tif
```

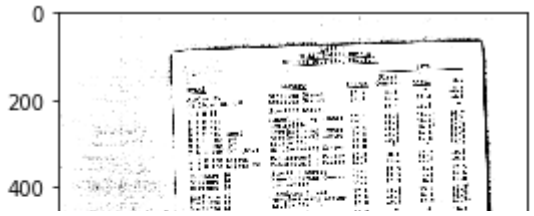
scientific report
/content/DocumentImages/train/scientific report/2058010470_2058010471.tif



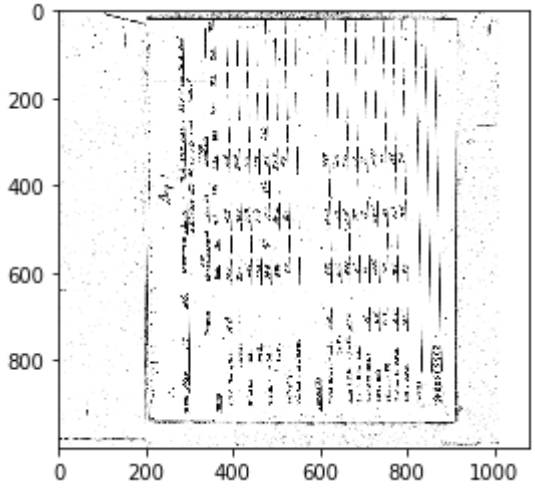
scientific report
/content/DocumentImages/train/scientific report/2058083949_2058083950.tif



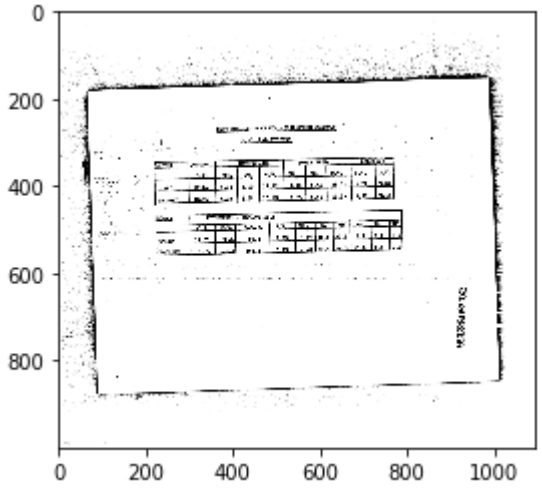
scientific report
/content/DocumentImages/train/scientific report/2058083409_2058083412.tif



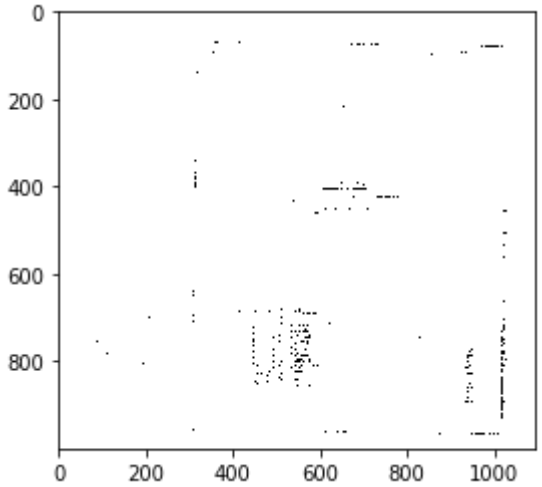
```
scientific report
/content/DocumentImages/train/scientific report/2058083943.tif
```



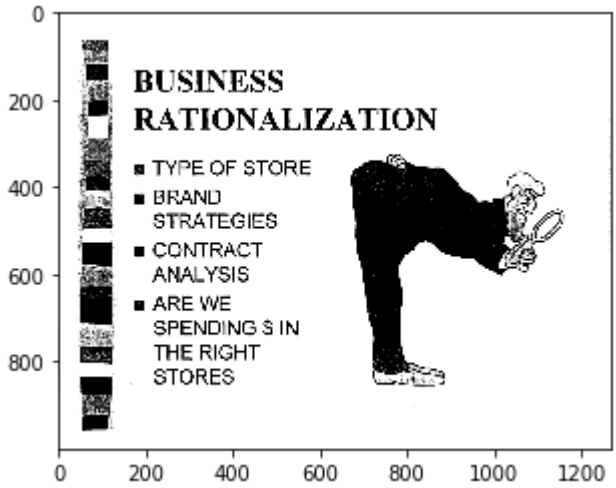
```
scientific report
/content/DocumentImages/train/scientific report/2058088958.tif
```



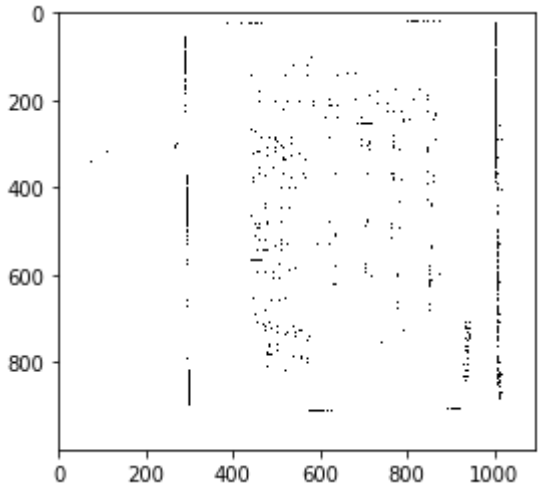
```
scientific report
/content/DocumentImages/train/scientific report/2058088387_2058088391.tif
```



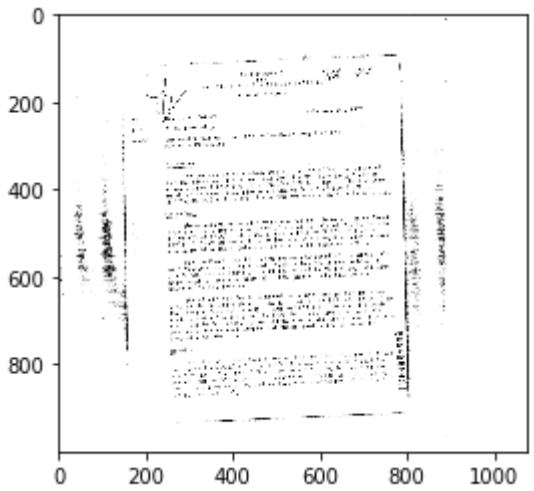
```
presentation
/content/DocumentImages/train/presentation/0013404138.tif
```




```
memo
/content/DocumentImages/train/memo/2058088060.tif
```

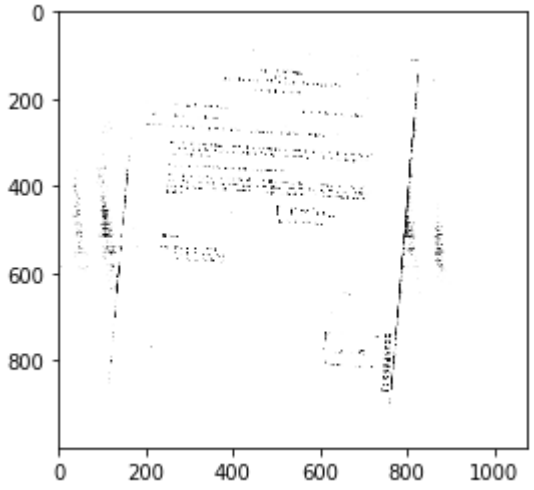


```
memo
/content/DocumentImages/train/memo/2058008434_2058008447.tif
```

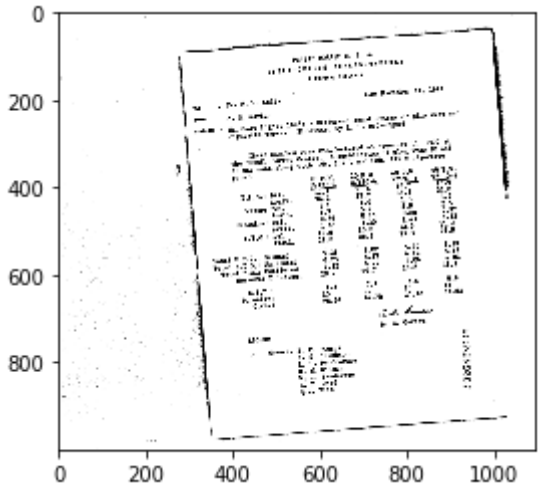


```
memo
```

```
/content/DocumentImages/train/memo/2058008311.tif
```

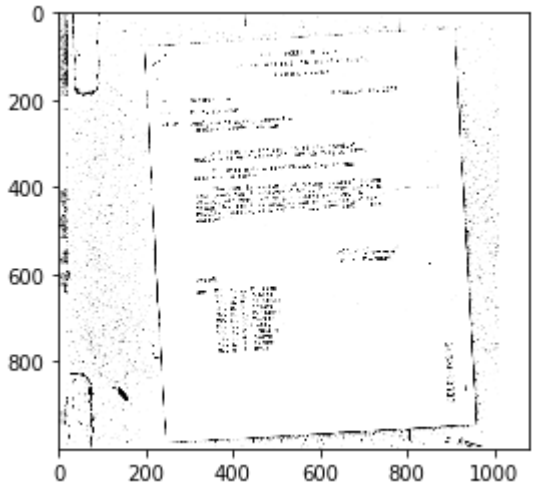



```
memo
/content/DocumentImages/train/memo/2058089267.tif
```

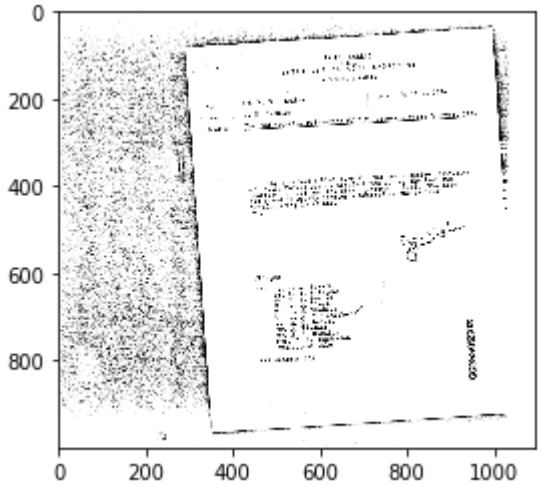



```
memo
/content/DocumentImages/train/memo/2058015468.tif
```

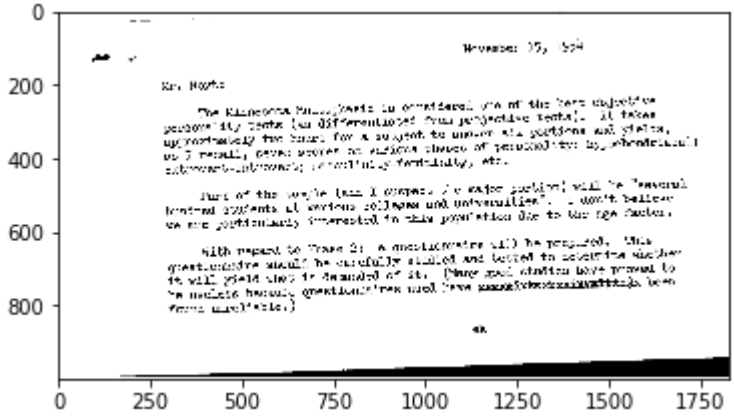
memo
/content/DocumentImages/train/memo/2058083637.tif



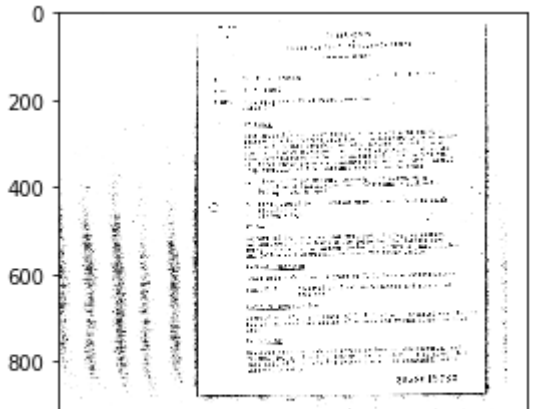
memo
/content/DocumentImages/train/memo/2058088299.tif



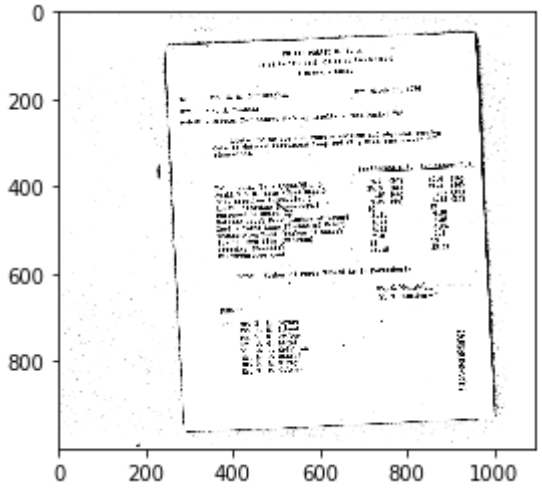
memo
/content/DocumentImages/train/memo/50072044.tif



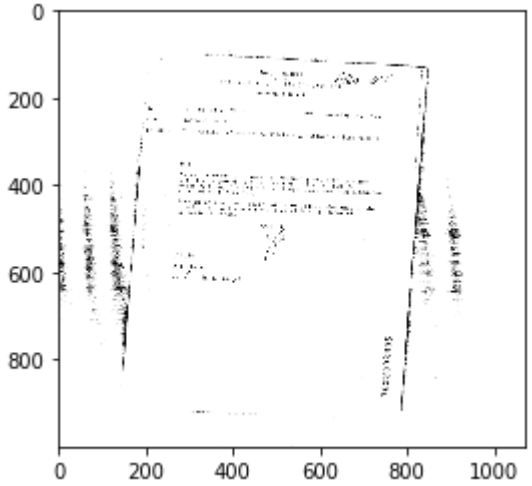
memo
/content/DocumentImages/train/memo/2058015783.tif



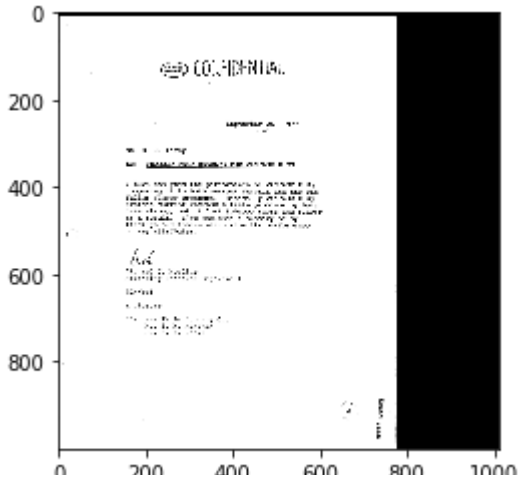
memo
/content/DocumentImages/train/memo/2058088795.tif



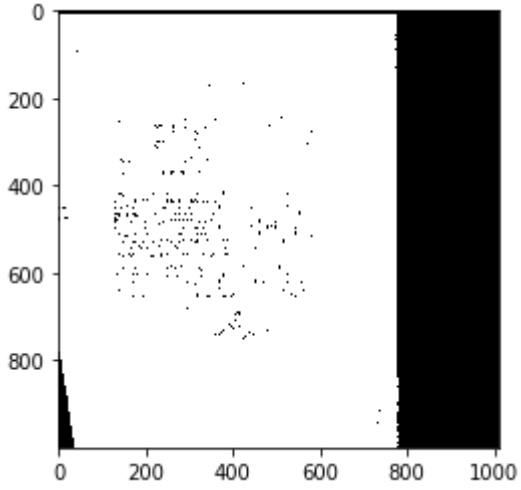
memo
/content/DocumentImages/train/memo/2058008391.tif



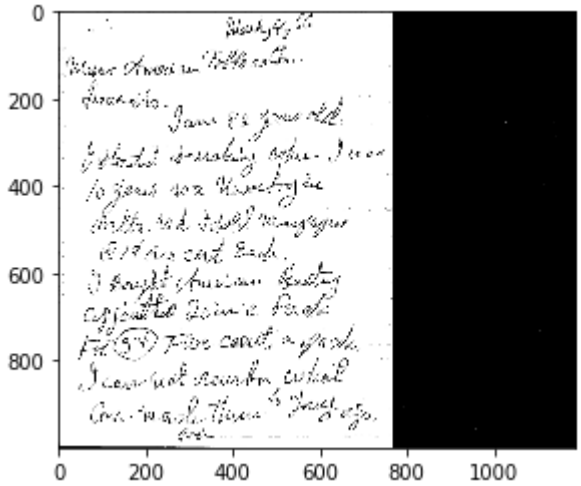
letter
/content/DocumentImages/train/letter/500853668.tif



letter
/content/DocumentImages/train/letter/500858976.tif

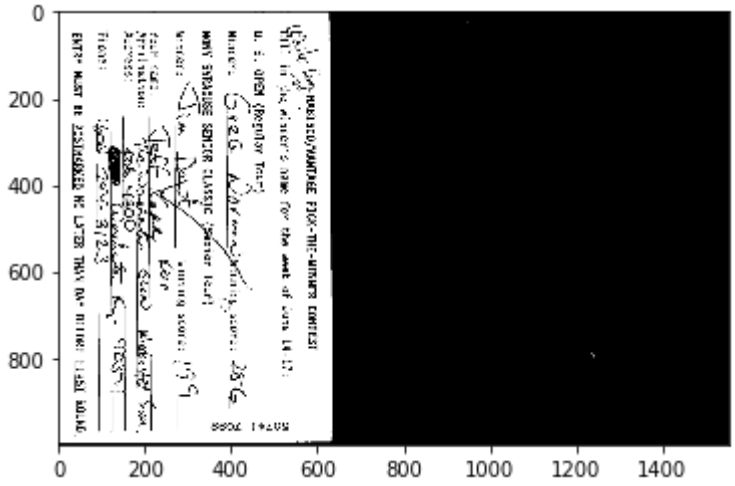


letter
/content/DocumentImages/train/letter/0013248032.tif

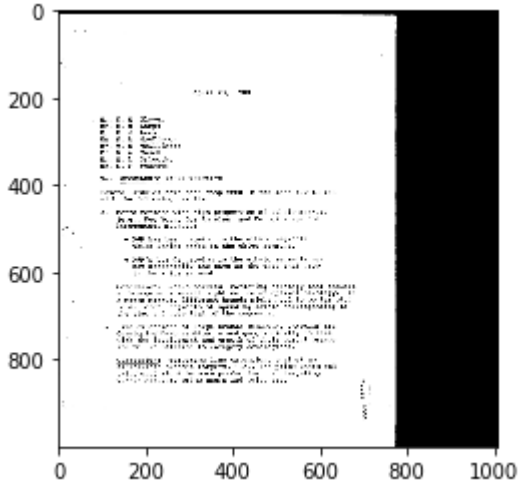


letter
/content/DocumentImages/train/letter/500858906.tif

letter
/content/DocumentImages/train/letter/507417099_507417100.tif

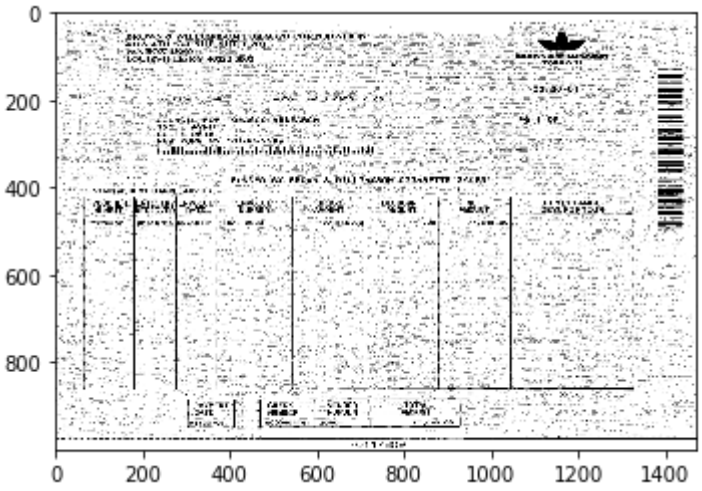


letter
/content/DocumentImages/train/letter/500768535_500768536.tif

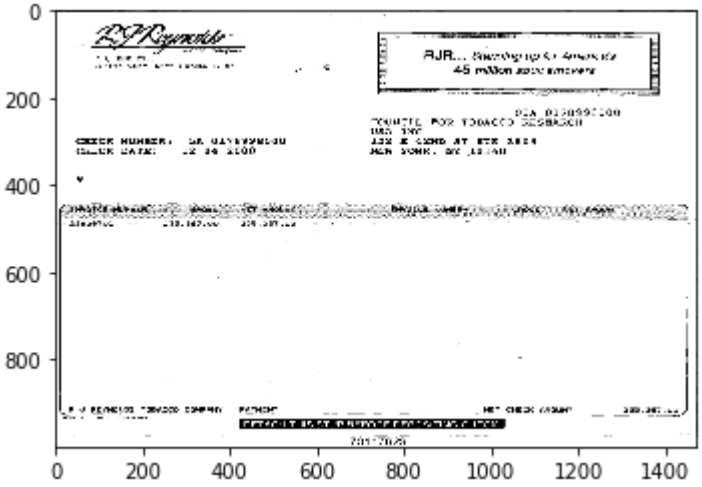


letter
/content/DocumentImages/train/letter/50125639-5639.tif

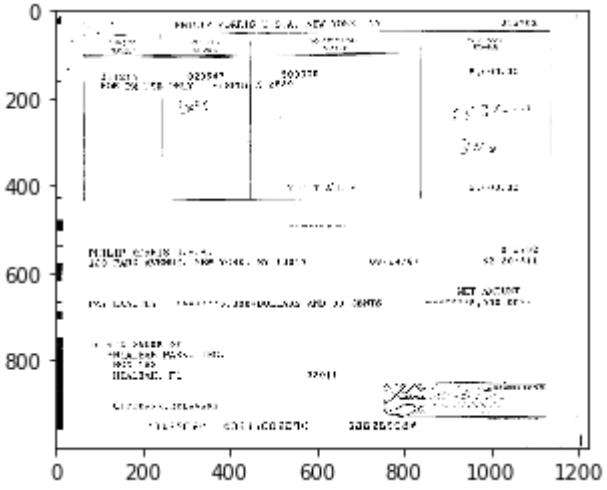
invoice
/content/DocumentImages/train/invoice/70117809-7809.tif



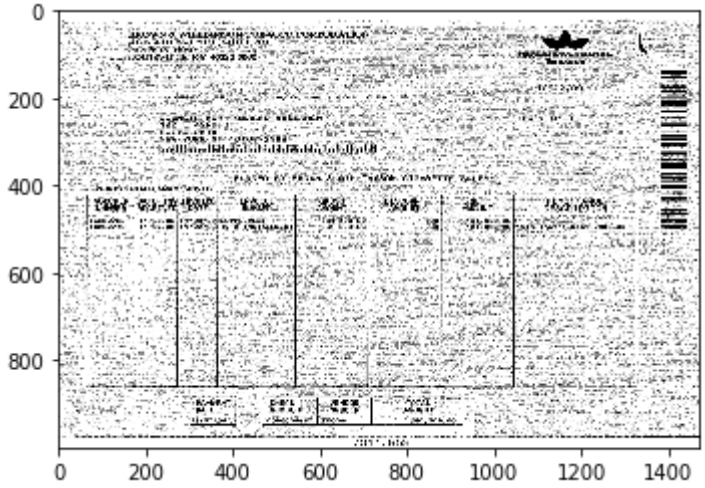
invoice
/content/DocumentImages/train/invoice/70117823-7823.tif



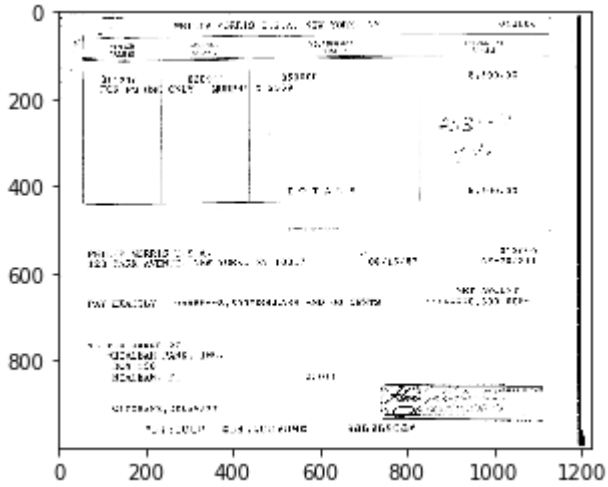
invoice
/content/DocumentImages/train/invoice/2040879785.tif



```
invoice
/content/DocumentImages/train/invoice/70117853-7853.tif
```

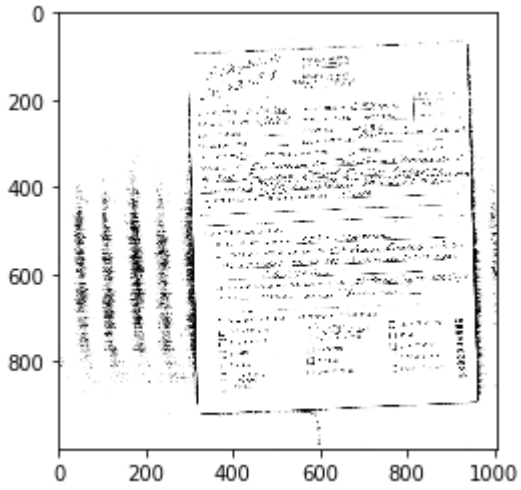



```
invoice
/content/DocumentImages/train/invoice/2040879786.tif
```

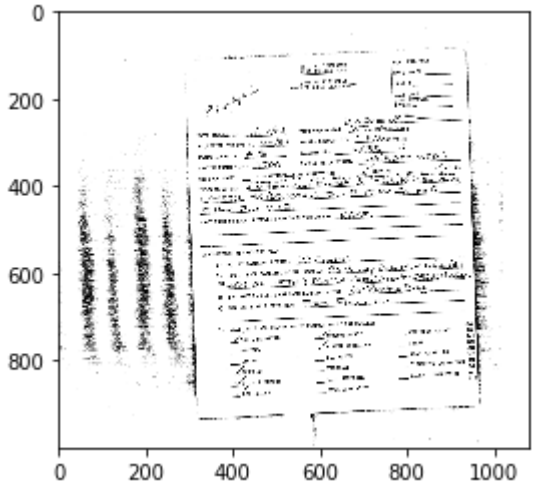



```
form
/content/DocumentImages/train/form/0030041076.tif
```

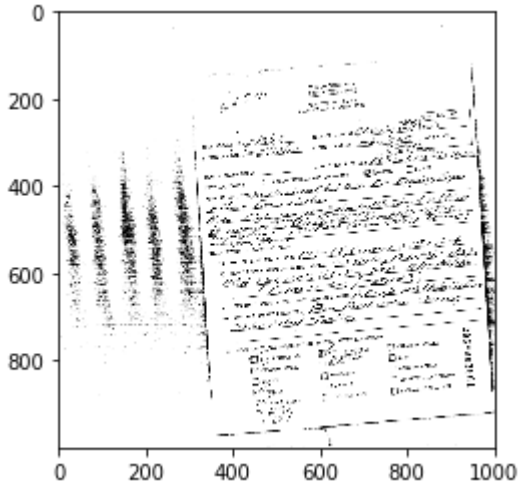
form
/content/DocumentImages/train/form/2058003688.tif



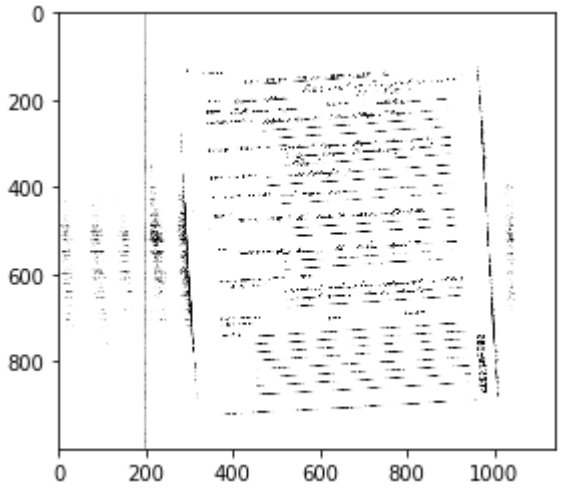
form
/content/DocumentImages/train/form/2058010373_2058010381.tif



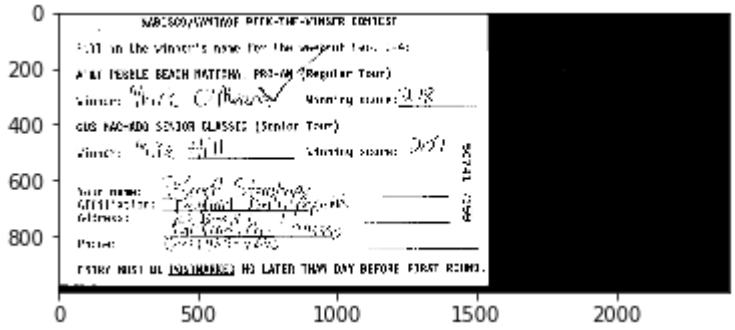
form
/content/DocumentImages/train/form/2058003961.tif



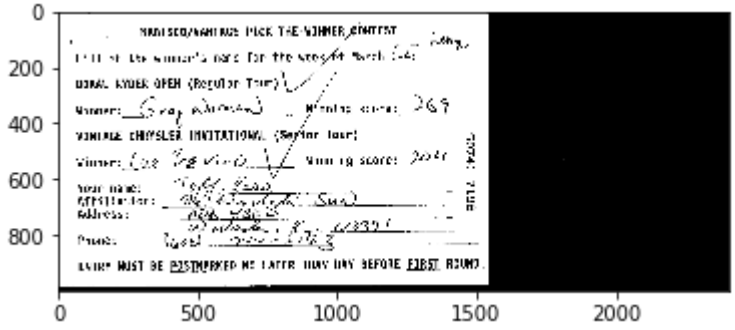
form
/content/DocumentImages/train/form/2058012537_2058012539.tif



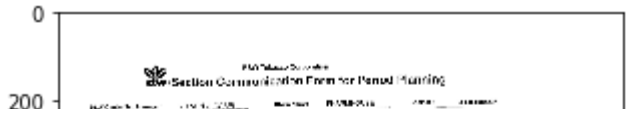
form
/content/DocumentImages/train/form/507417159_507417160.tif



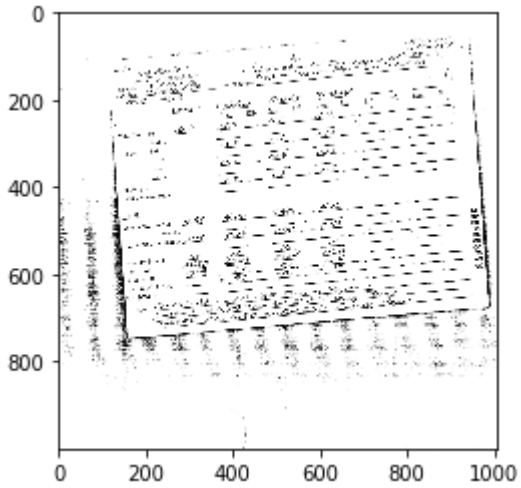
form
/content/DocumentImages/train/form/507417198_507417199.tif



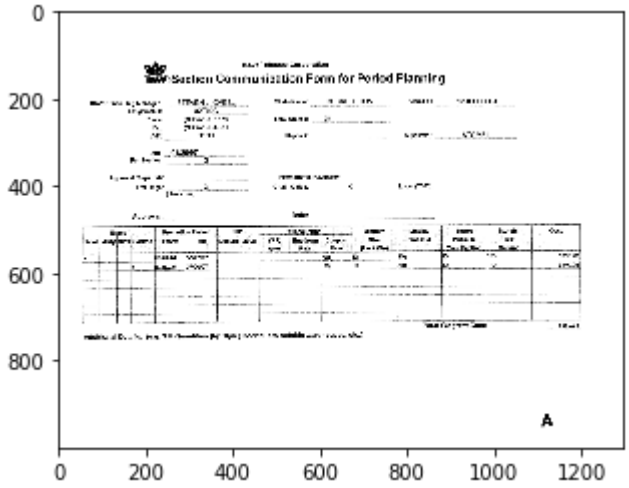
form
/content/DocumentImages/train/form/0030041180.tif



form
/content/DocumentImages/train/form/2058003673.tif

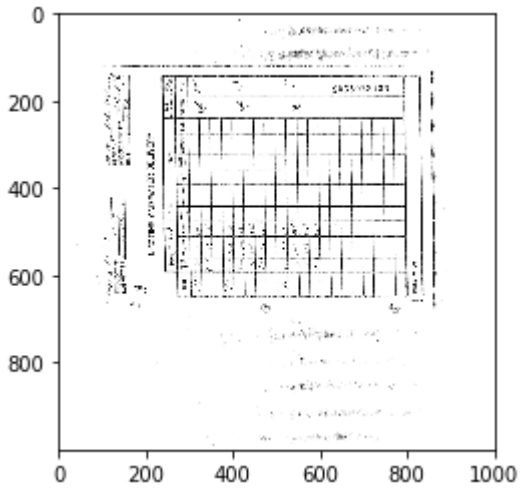


form
/content/DocumentImages/train/form/0030041222.tif

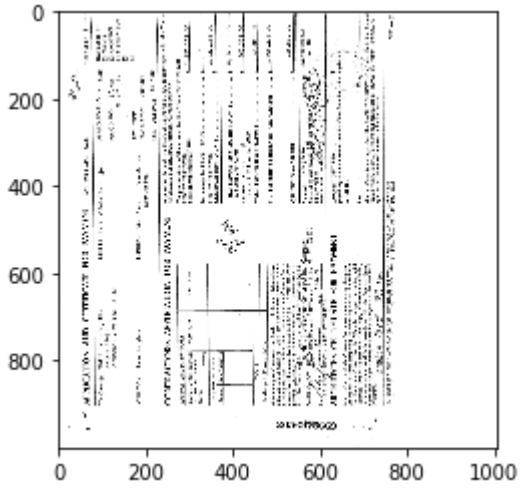


form
/content/DocumentImages/train/form/0030030842.tif

```
form
/content/DocumentImages/train/form/2058033198.tif
```

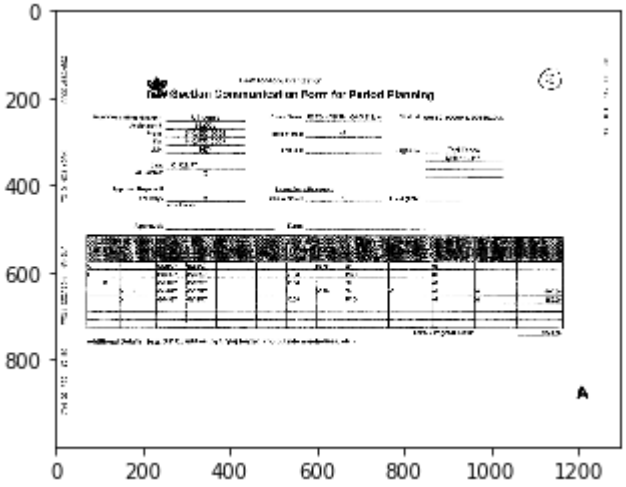



```
form
/content/DocumentImages/train/form/2056975669.tif
```

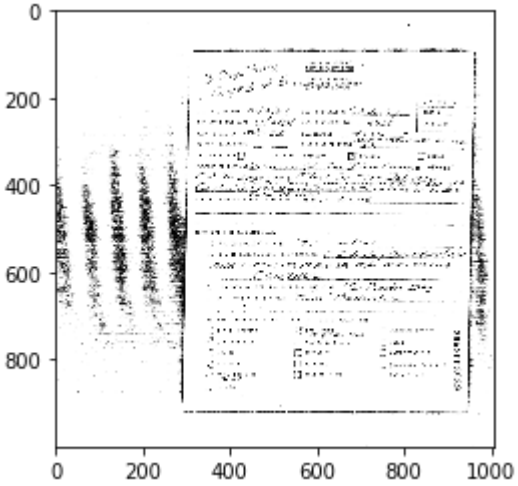



```
form
/content/DocumentImages/train/form/0030041221.tif
```

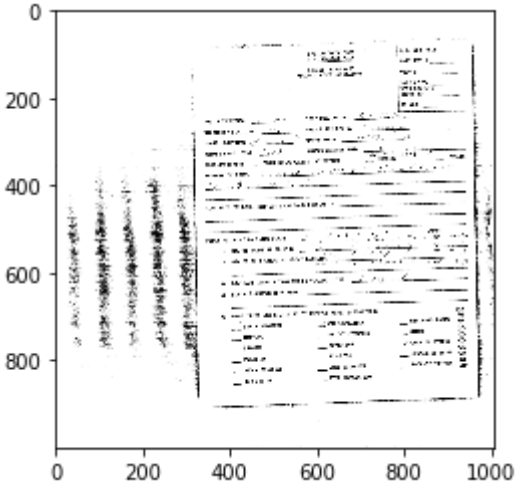
form
/content/DocumentImages/train/form/0030041244.tif



form
/content/DocumentImages/train/form/2058003689_2058003690.tif

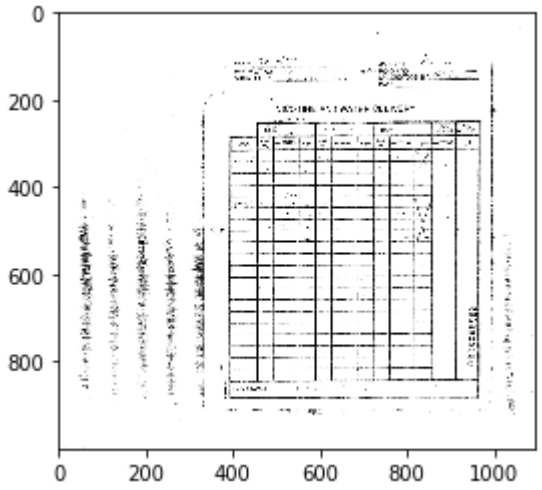


form
/content/DocumentImages/train/form/2058003598.tif

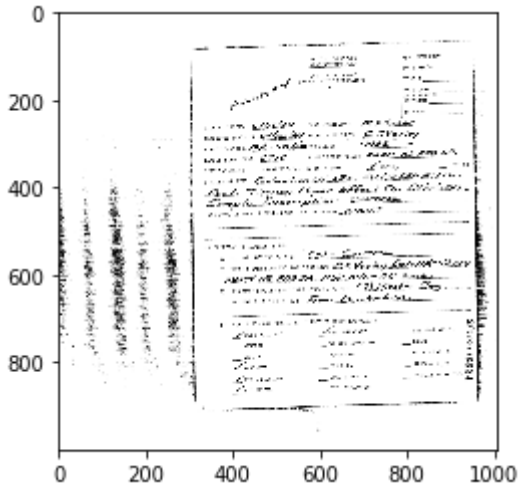


form

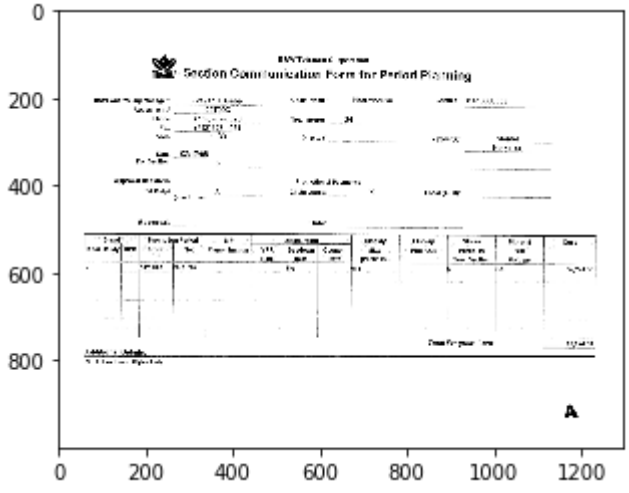
/content/DocumentImages/train/form/2058032800.tif



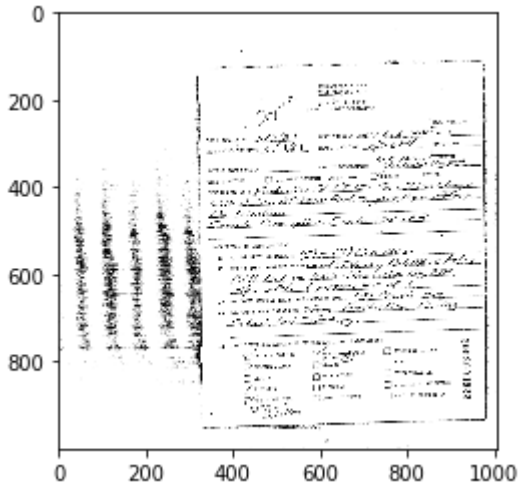
form
/content/DocumentImages/train/form/2058003684_2058003685.tif



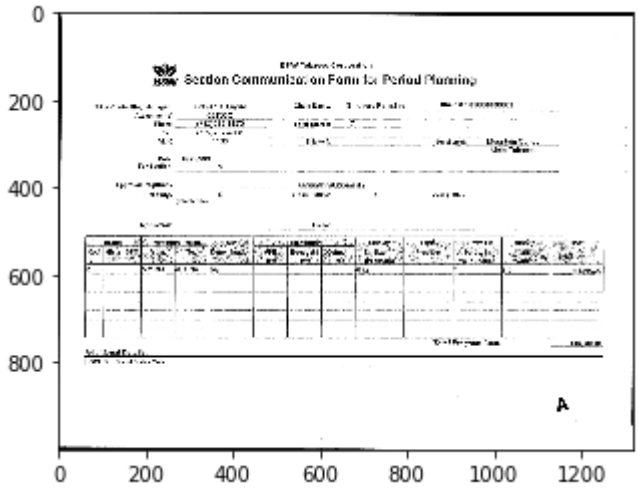
form
/content/DocumentImages/train/form/0030041089.tif



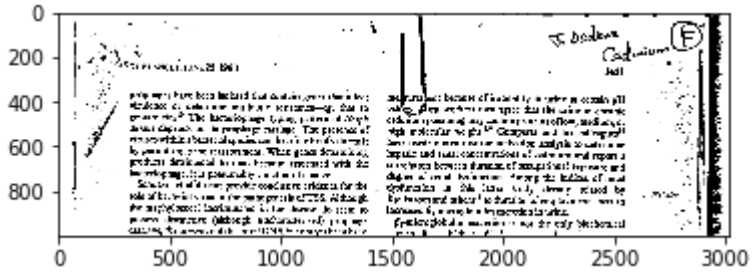
form
/content/DocumentImages/train/form/2058003922.tif



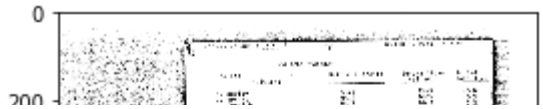
form
/content/DocumentImages/train/form/0030030844.tif



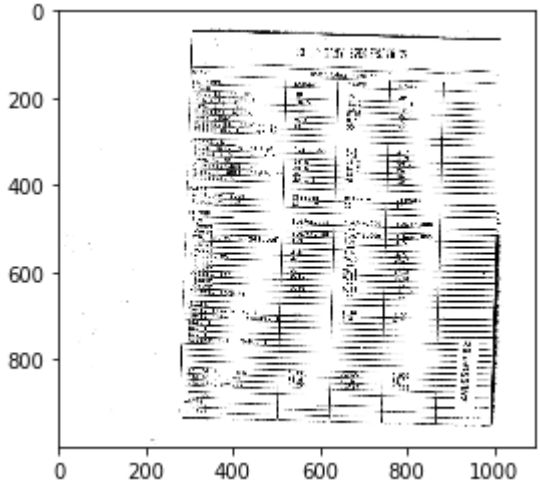
scientific publication
/content/DocumentImages/train/scientific publication/1000080993_1000080996.tif



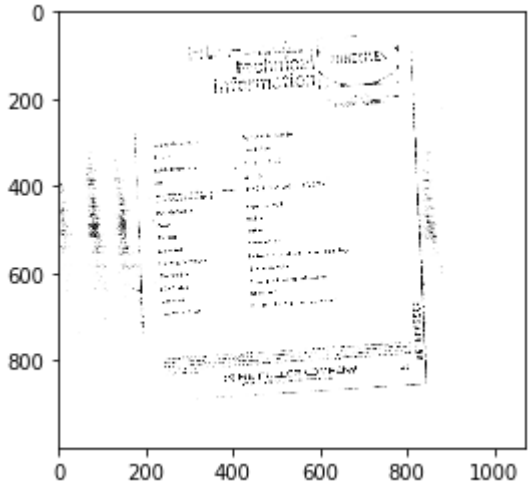
specification
/content/DocumentImages/train/specification/2058088177_2058088178.tif



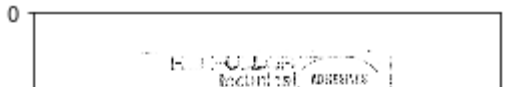
specification
/content/DocumentImages/train/specification/2058089487.tif



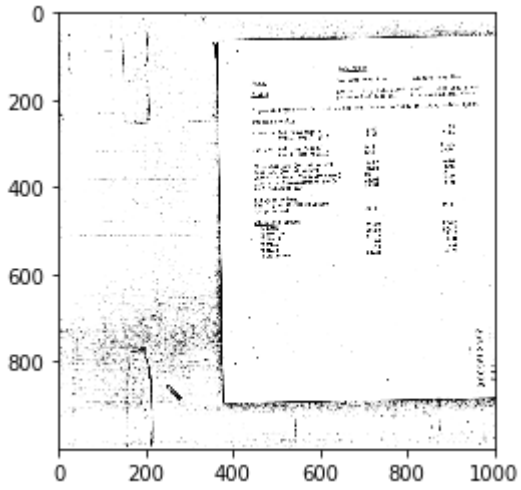
specification
/content/DocumentImages/train/specification/2058008190.tif



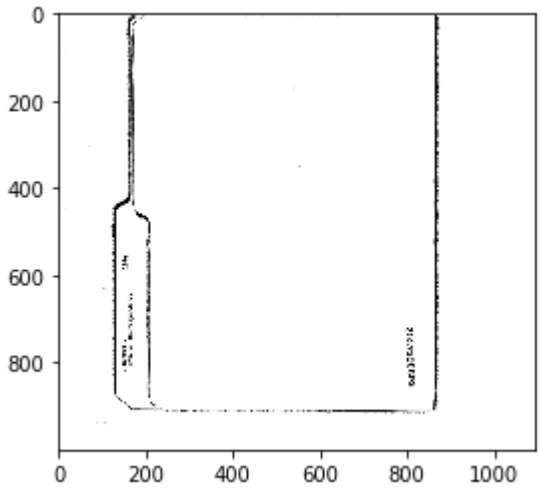
specification
/content/DocumentImages/train/specification/2058008194.tif



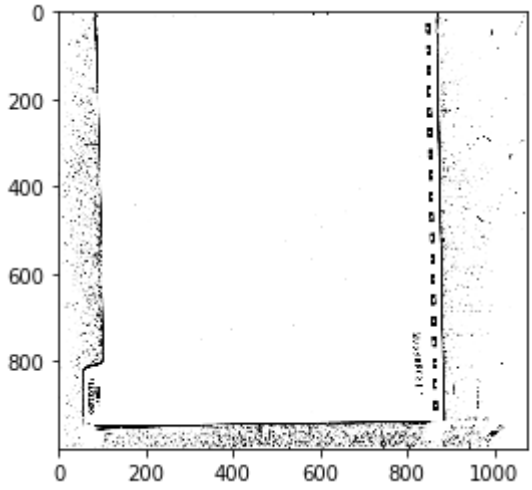
```
specification
/content/DocumentImages/train/specification/2058096006.tif
```



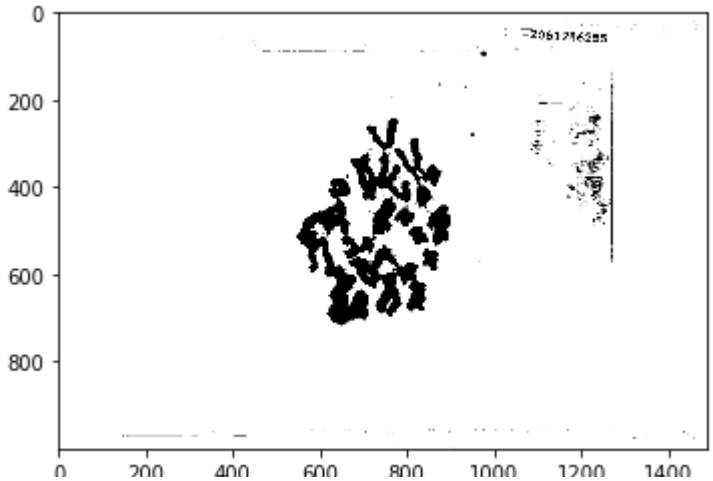

```
file folder
/content/DocumentImages/train/file folder/2057996603.tif
```



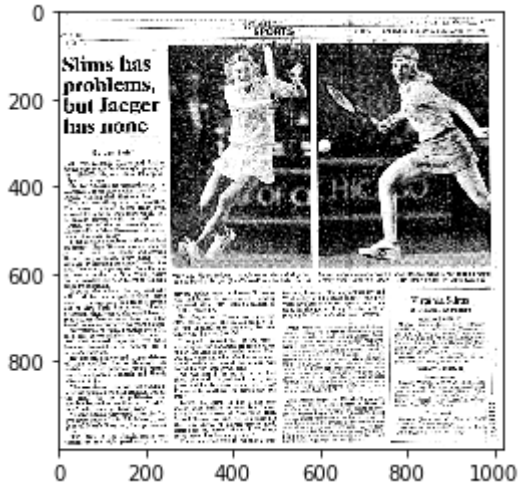

```
file folder
/content/DocumentImages/train/file folder/2058086812.tif
```



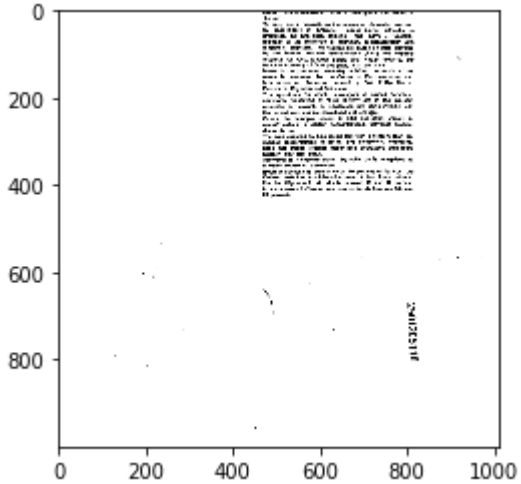

```
file folder
/content/DocumentImages/train/file folder/2061986255.tif
```



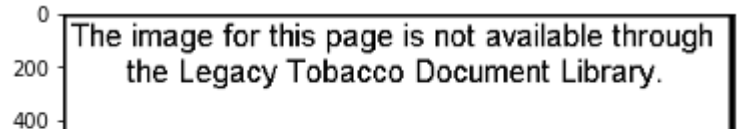
news article
/content/DocumentImages/train/news article/2042824722.tif



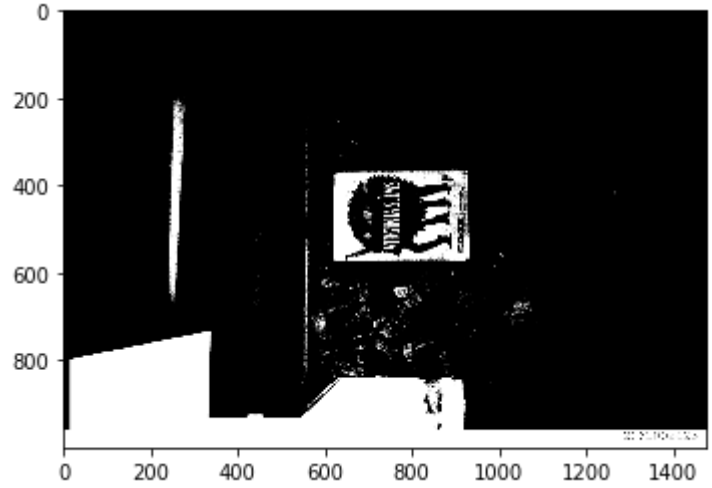
news article
/content/DocumentImages/train/news article/2501205110.tif



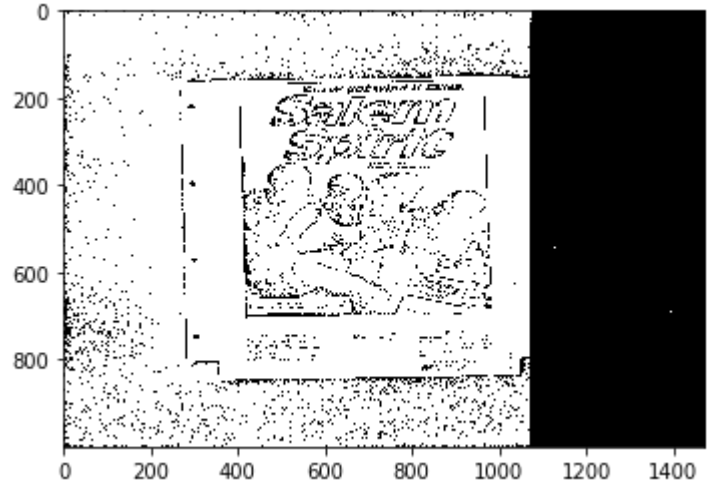
news article
/content/DocumentImages/train/news article/1003042957.tif



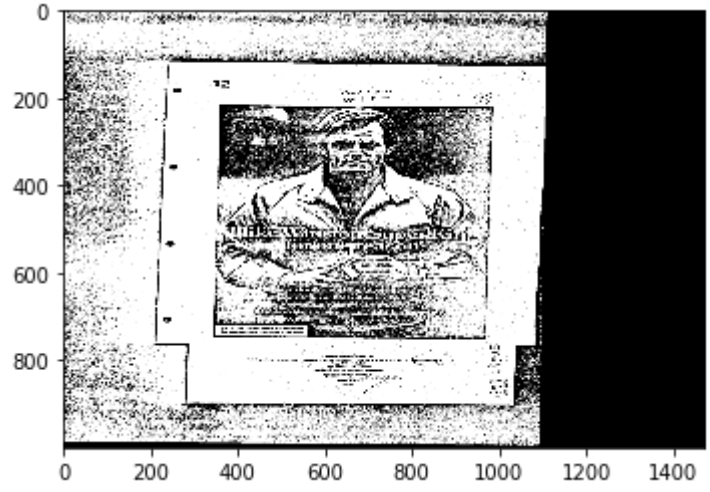
advertisement
/content/DocumentImages/train/advertisement/2061002423.tif



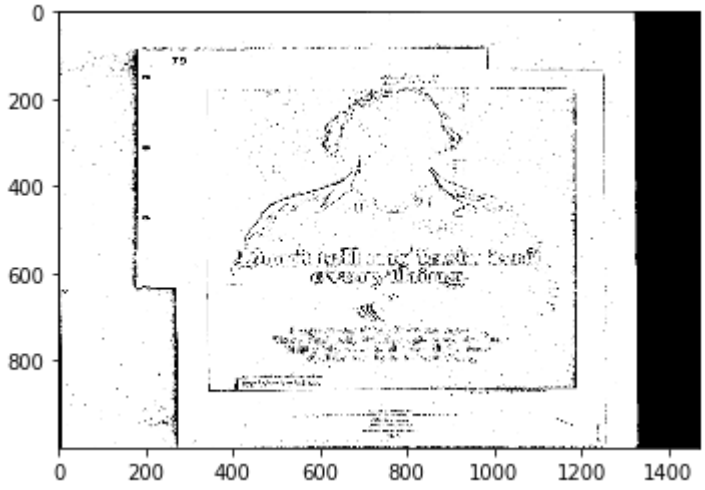
advertisement
/content/DocumentImages/train/advertisement/502610793+-0793.tif



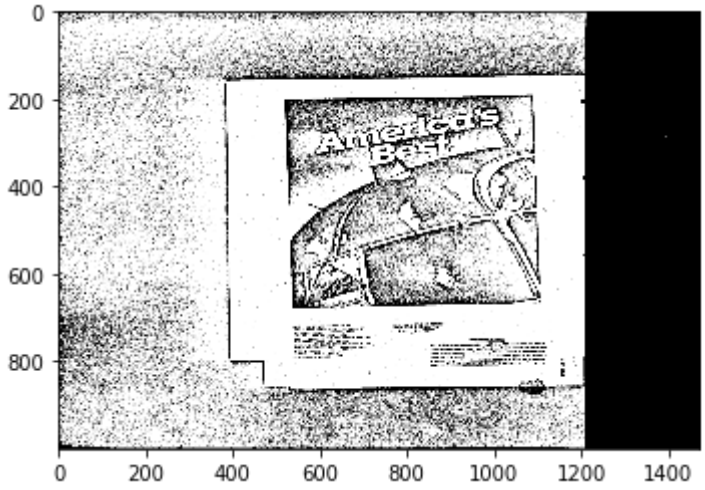
advertisement
/content/DocumentImages/train/advertisement/502606631+-6631.tif



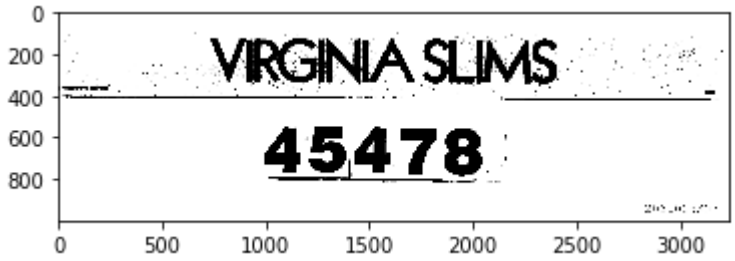
advertisement
/content/DocumentImages/train/advertisement/502605775+-5775.tif



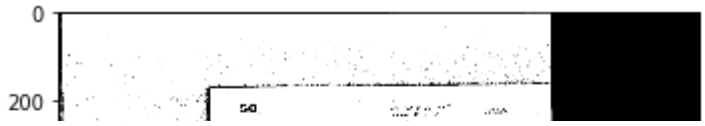
advertisement
/content/DocumentImages/train/advertisement/502610823+-0823.tif



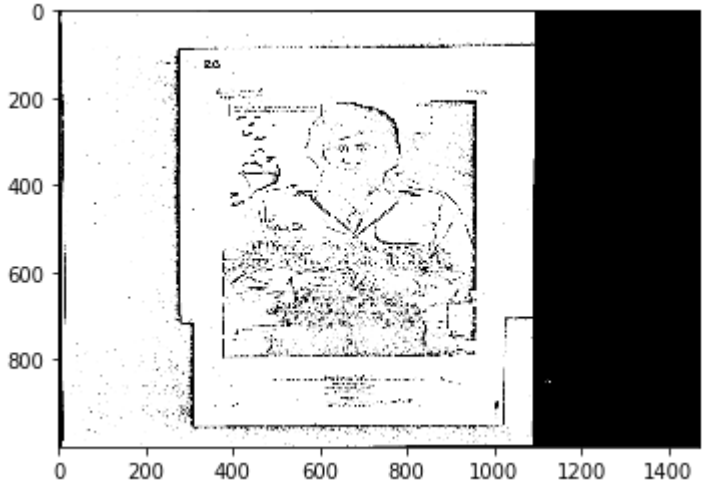
advertisement
/content/DocumentImages/train/advertisement/2061003307.tif



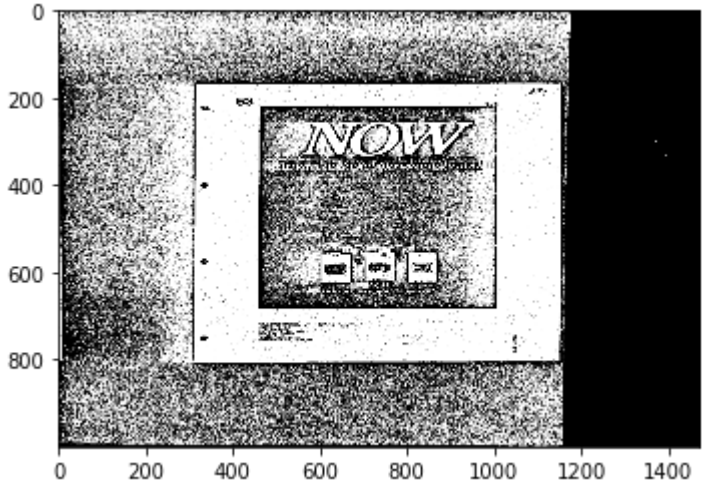
advertisement
/content/DocumentImages/train/advertisement/502605734+-5734.tif



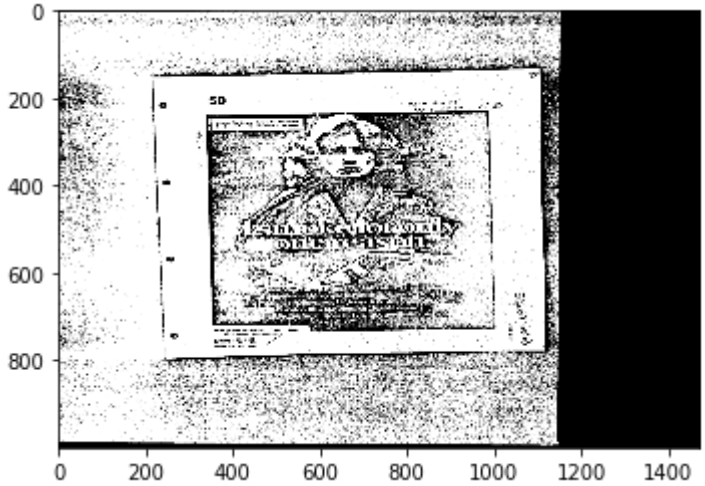
advertisement
/content/DocumentImages/train/advertisement/502605692+-5692.tif



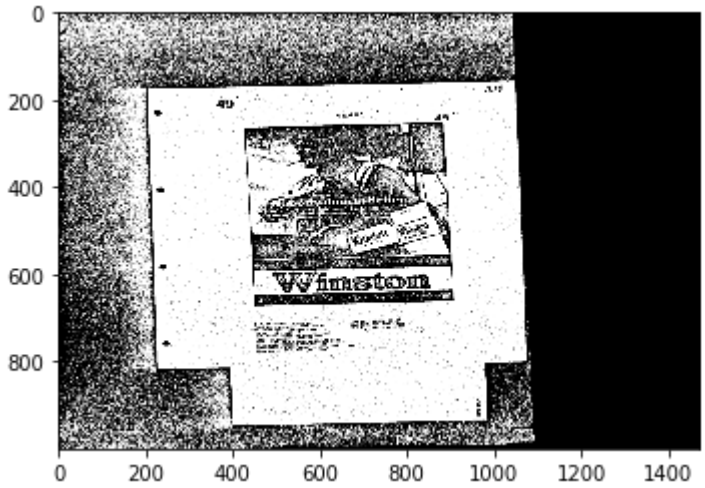
advertisement
/content/DocumentImages/train/advertisement/502610513+-0516.tif



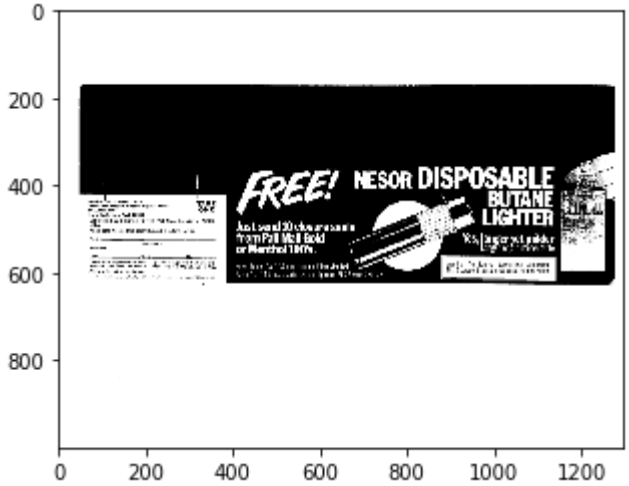
advertisement
/content/DocumentImages/train/advertisement/502606688+-6688.tif



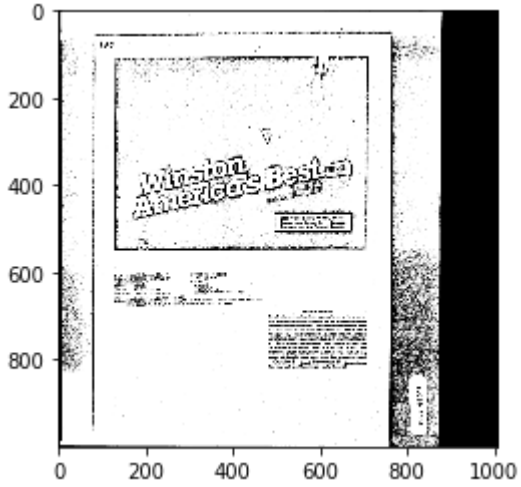
advertisement
/content/DocumentImages/train/advertisement/502610509+-0510.tif



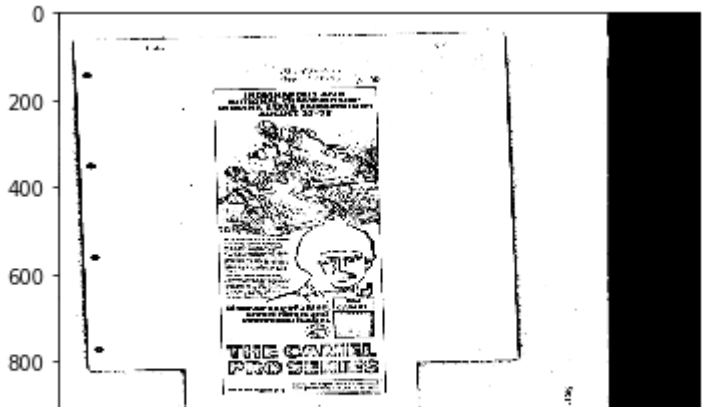
advertisement
/content/DocumentImages/train/advertisement/0071087562.tif



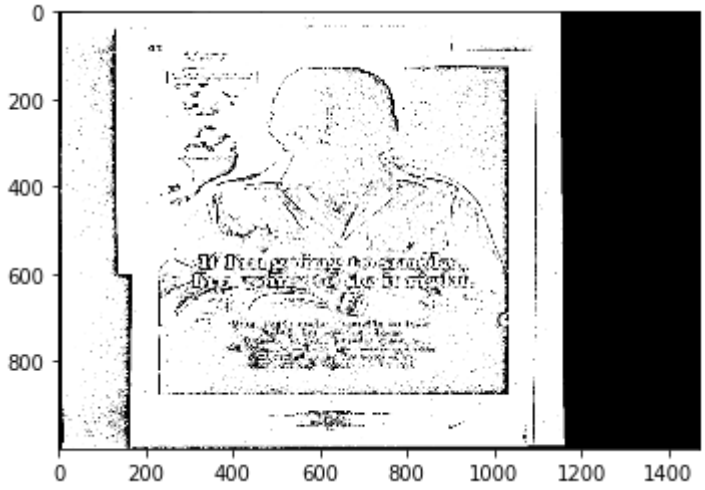
advertisement
/content/DocumentImages/train/advertisement/507806534.tif



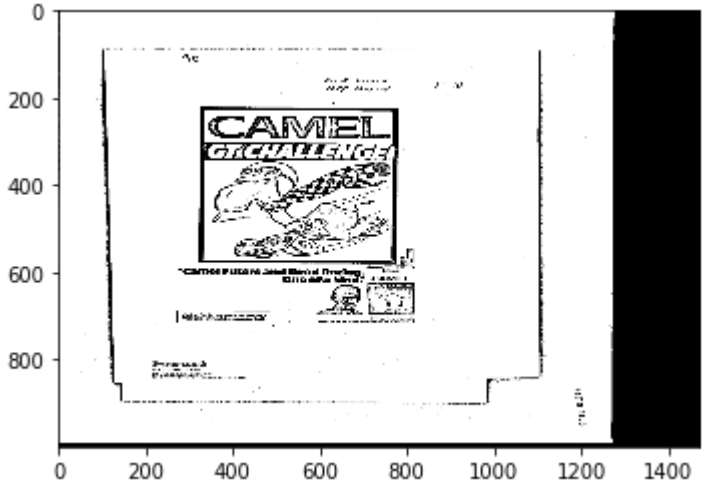
advertisement
/content/DocumentImages/train/advertisement/502100393+-0393.tif




```
advertisement
/content/DocumentImages/train/advertisement/502605717+-5717.tif
```

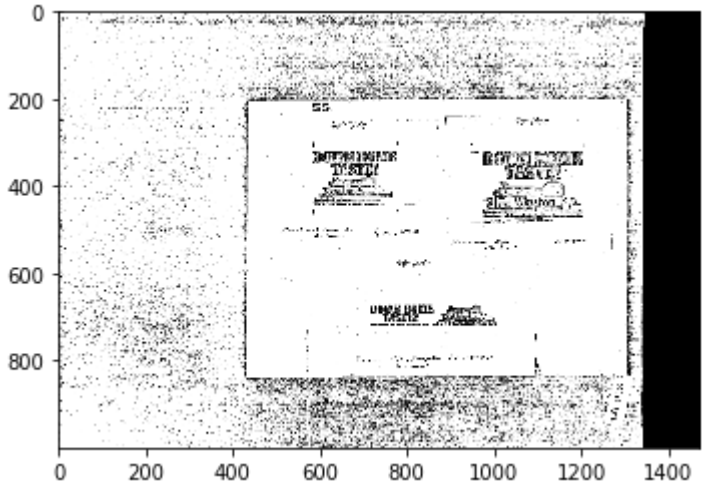


```
advertisement
/content/DocumentImages/train/advertisement/502100635+-0635.tif
```

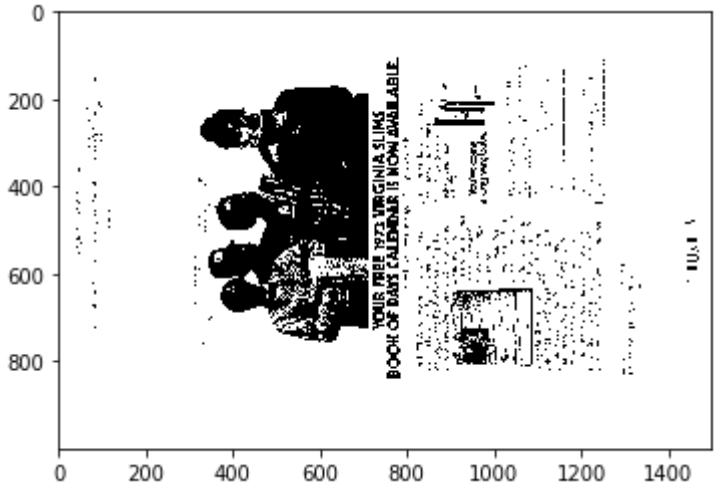



```
advertisement
```

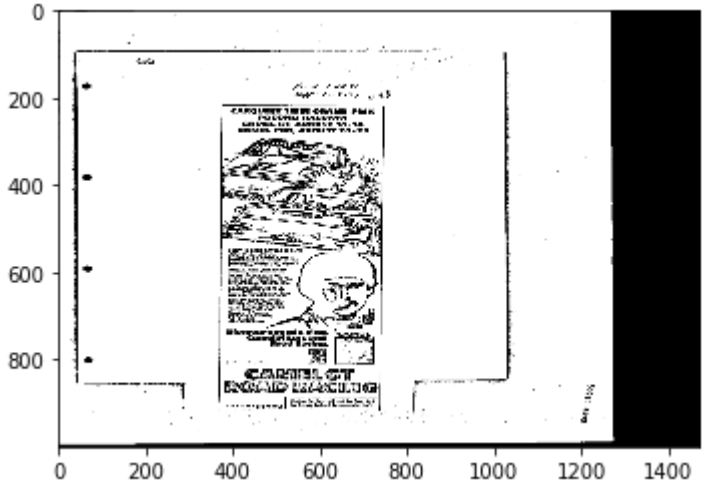
/content/DocumentImages/train/advertisement/502061214+-1214.tif



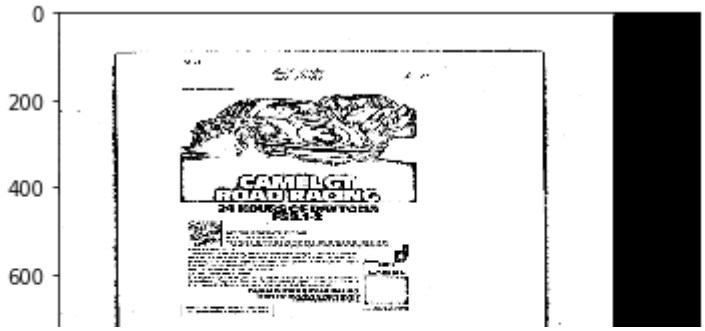
advertisement
/content/DocumentImages/train/advertisement/2058504140.tif



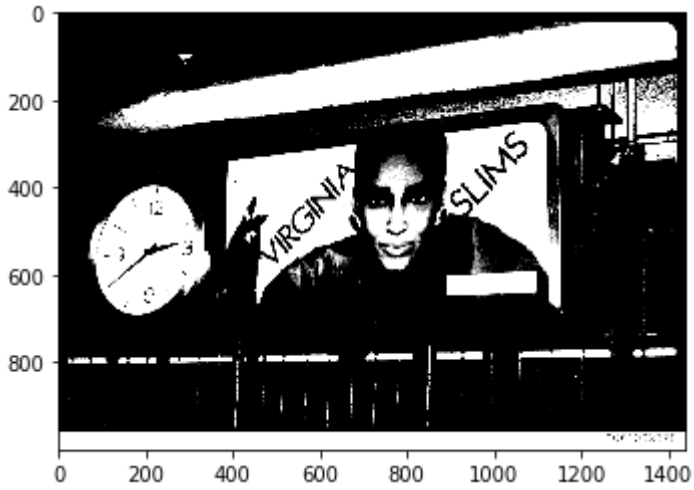
advertisement
/content/DocumentImages/train/advertisement/502100590+-0590.tif



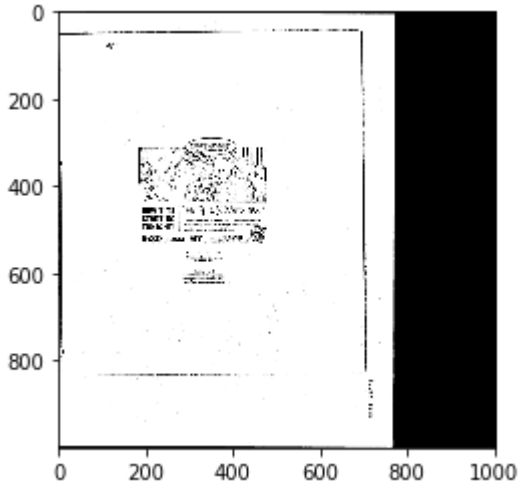
advertisement
/content/DocumentImages/train/advertisement/502100510+-0510.tif



advertisement
/content/DocumentImages/train/advertisement/2061002598.tif

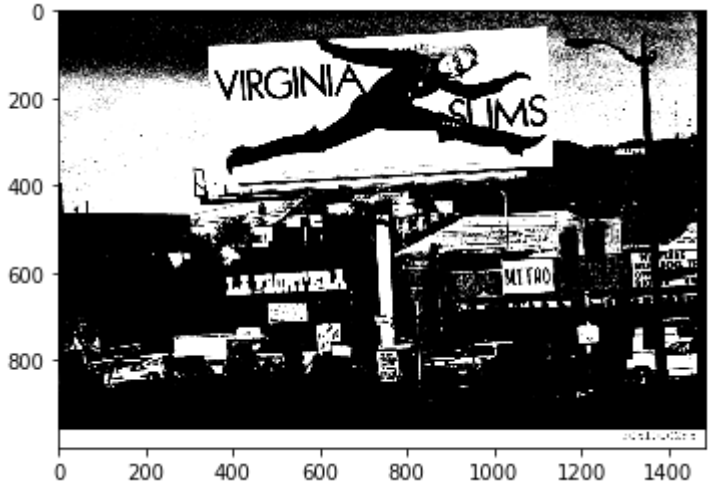


advertisement
/content/DocumentImages/train/advertisement/502474085.tif

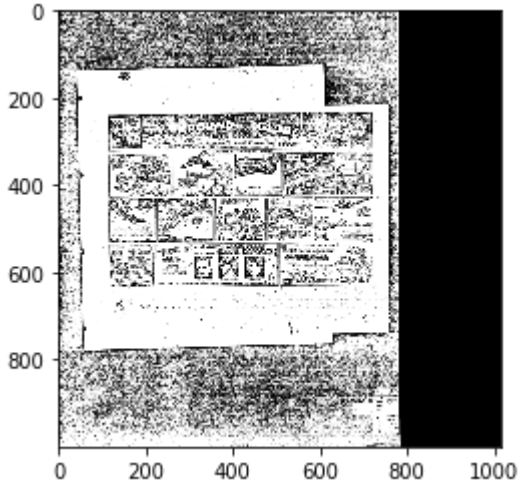


advertisement
/content/DocumentImages/train/advertisement/502606719+-6719.tif

advertisement
/content/DocumentImages/train/advertisement/2061000285.tif

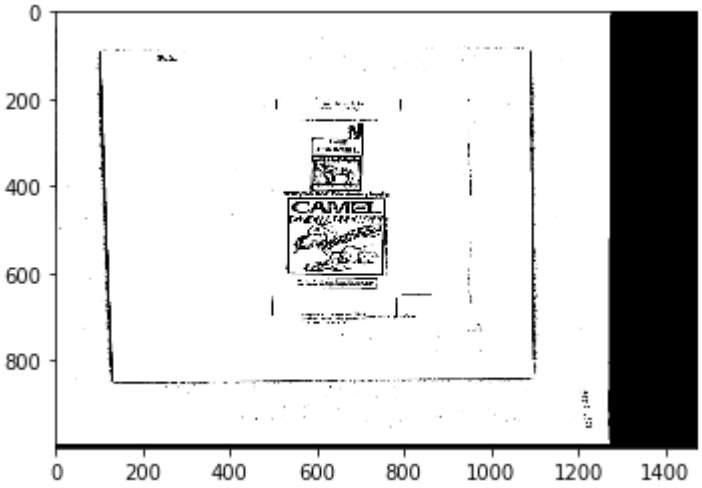


advertisement
/content/DocumentImages/train/advertisement/501947887.tif

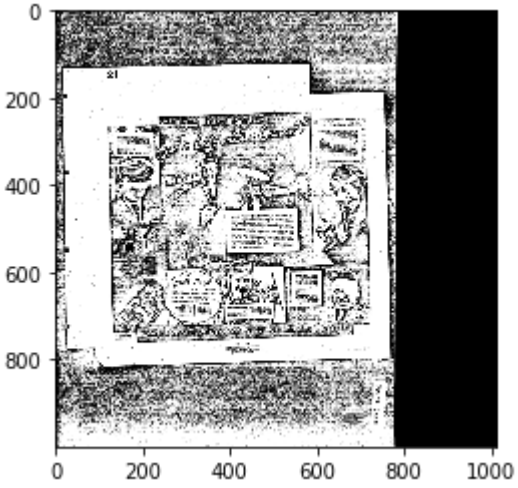


advertisement
/content/DocumentImages/train/advertisement/503961297+-1297.tif

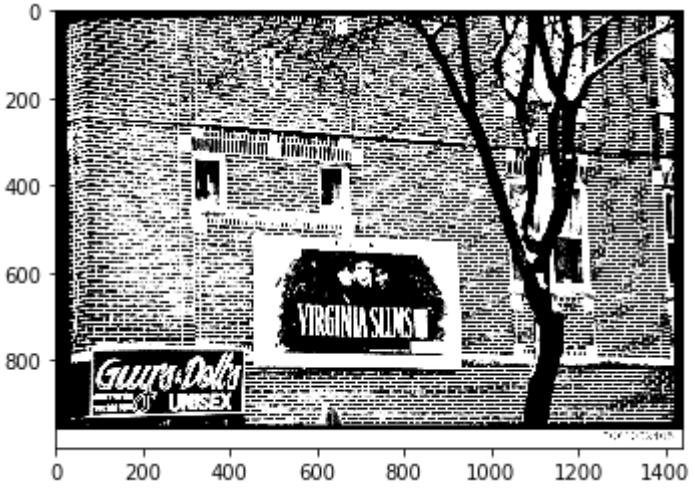
advertisement
/content/DocumentImages/train/advertisement/502100627+-0627.tif



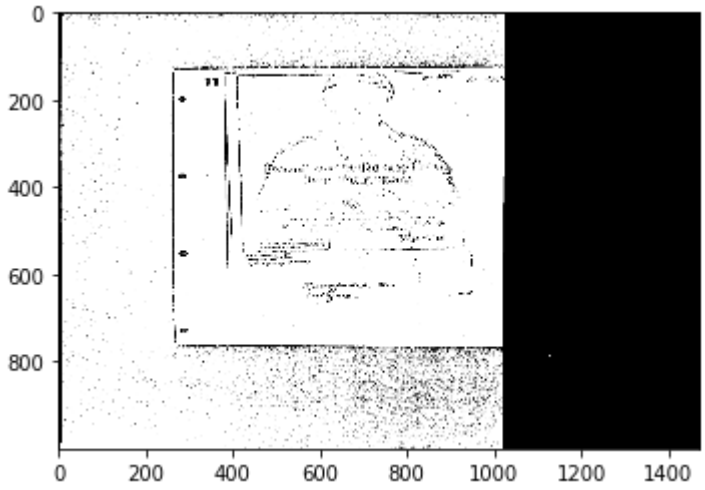
advertisement
/content/DocumentImages/train/advertisement/501947860.tif



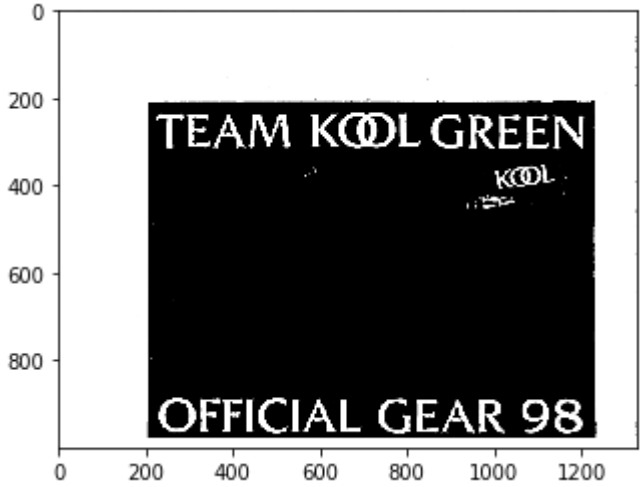
advertisement
/content/DocumentImages/train/advertisement/2061002405.tif



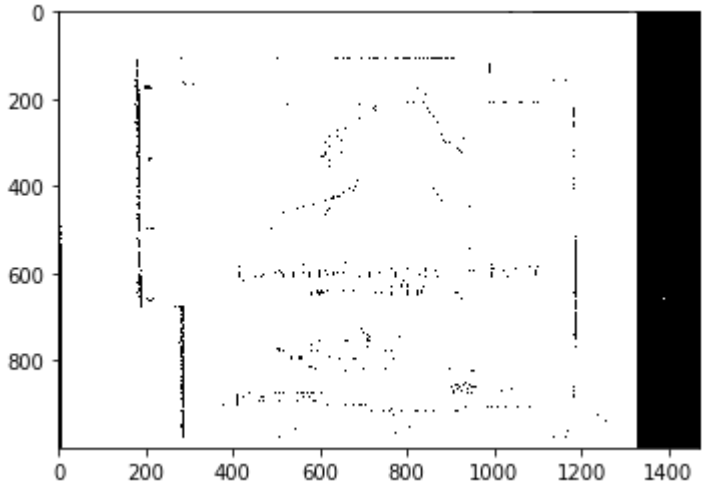
advertisement
/content/DocumentImages/train/advertisement/502605833+-5833.tif



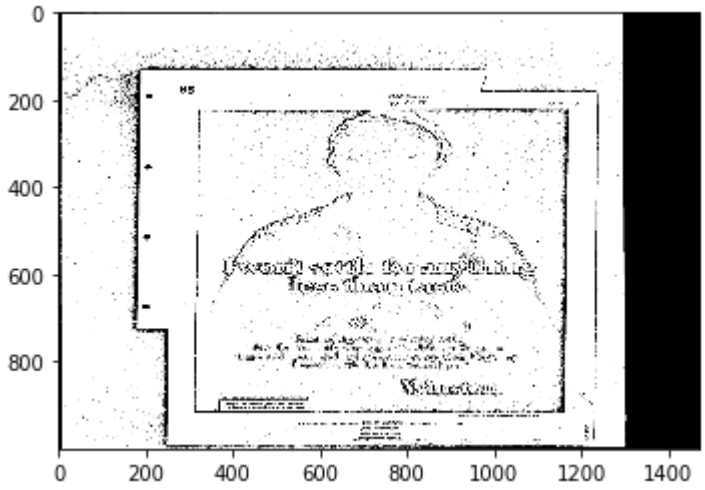
advertisement
/content/DocumentImages/train/advertisement/0030001056.tif



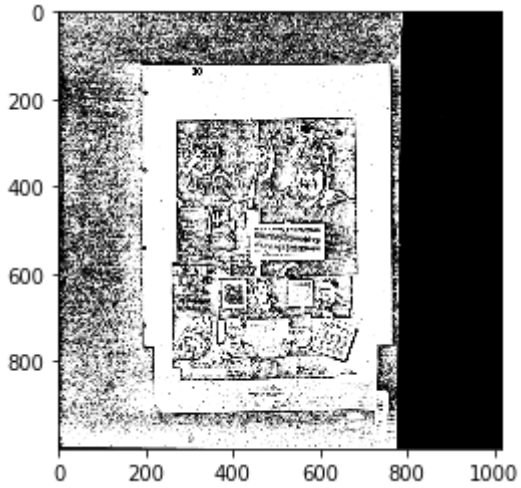
advertisement
/content/DocumentImages/train/advertisement/502605779+-5779.tif



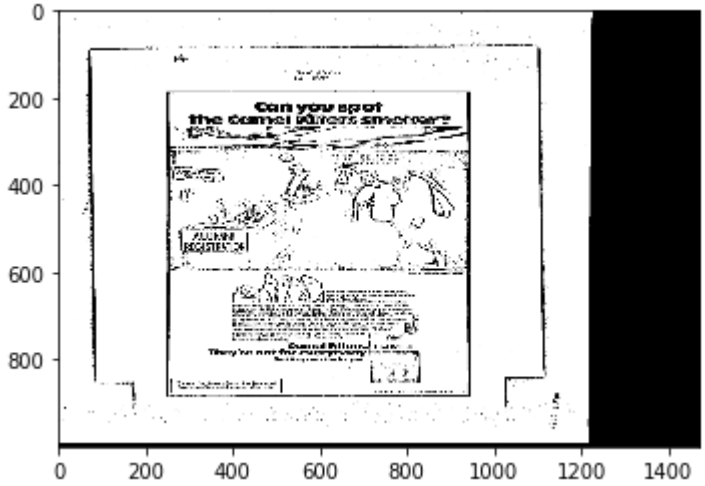
advertisement
/content/DocumentImages/train/advertisement/502605781+-5781.tif



advertisement
/content/DocumentImages/train/advertisement/501947869.tif

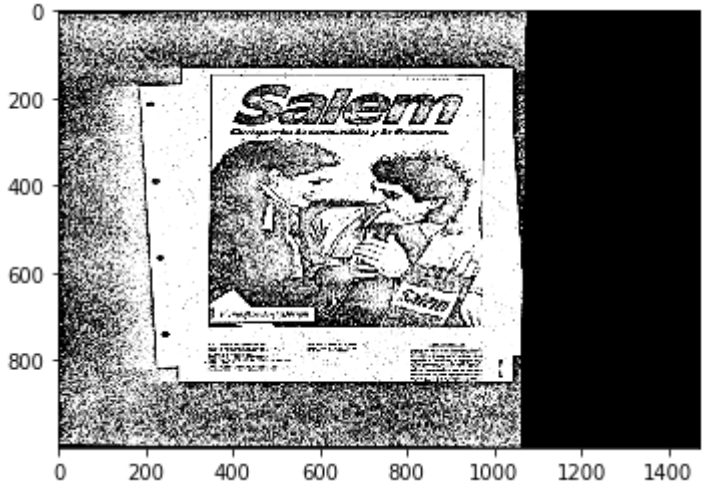


advertisement
/content/DocumentImages/train/advertisement/502100231+-0231.tif

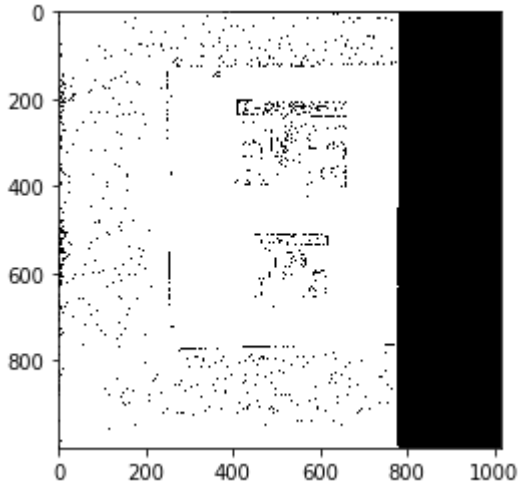


advertisement
/content/DocumentImages/train/advertisement/503944263+-4263.tif

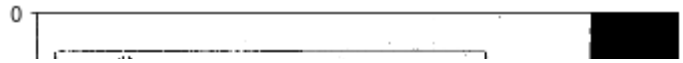
advertisement
/content/DocumentImages/train/advertisement/502610801+-0801.tif



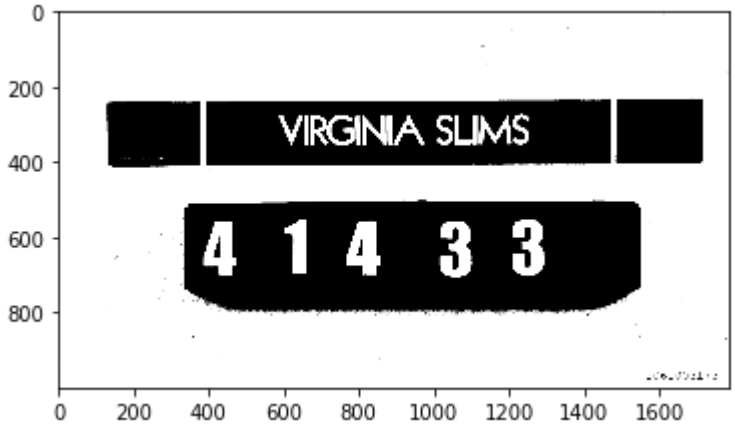
advertisement
/content/DocumentImages/train/advertisement/501947877.tif



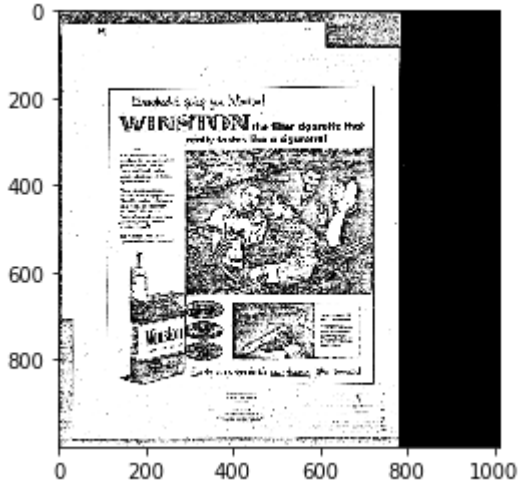
advertisement
/content/DocumentImages/train/advertisement/502100555+-0555.tif



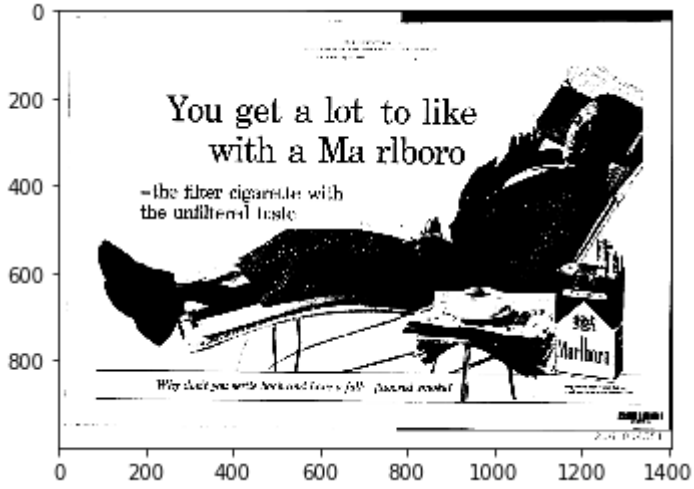
advertisement
/content/DocumentImages/train/advertisement/2061003175.tif



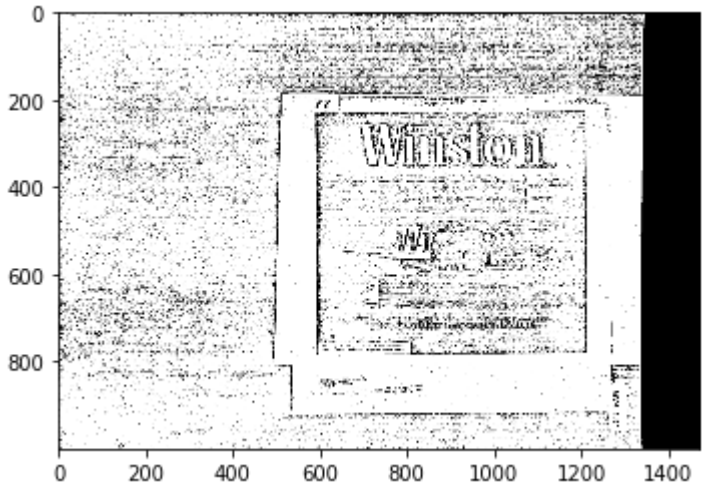
advertisement
/content/DocumentImages/train/advertisement/502218066.tif



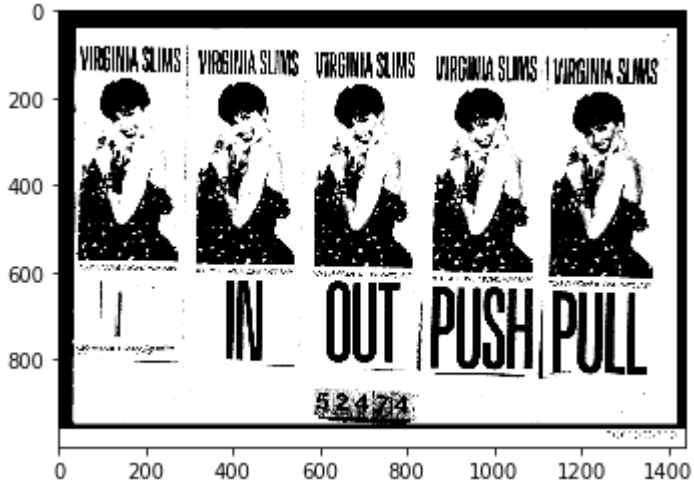
advertisement
/content/DocumentImages/train/advertisement/2061004551.tif



advertisement
/content/DocumentImages/train/advertisement/503961607+-1607.tif

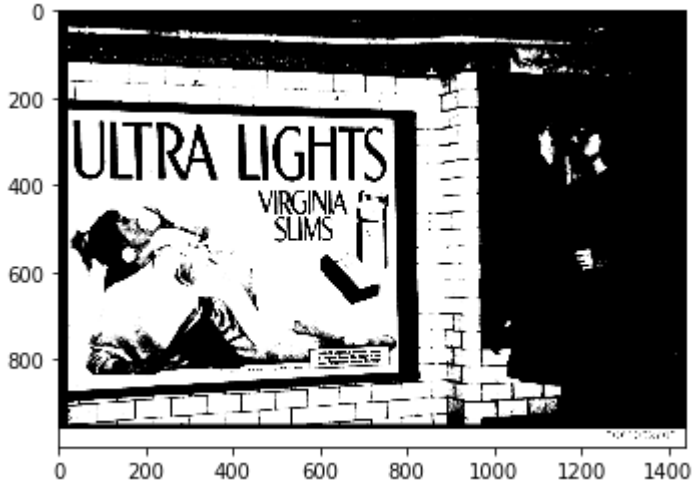


advertisement
/content/DocumentImages/train/advertisement/2061003310.tif

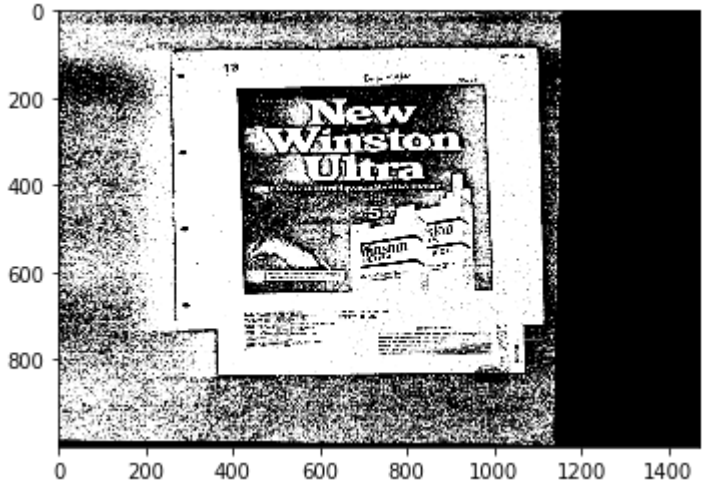


advertisement
/content/DocumentImages/train/advertisement/502610858+-0858.tif

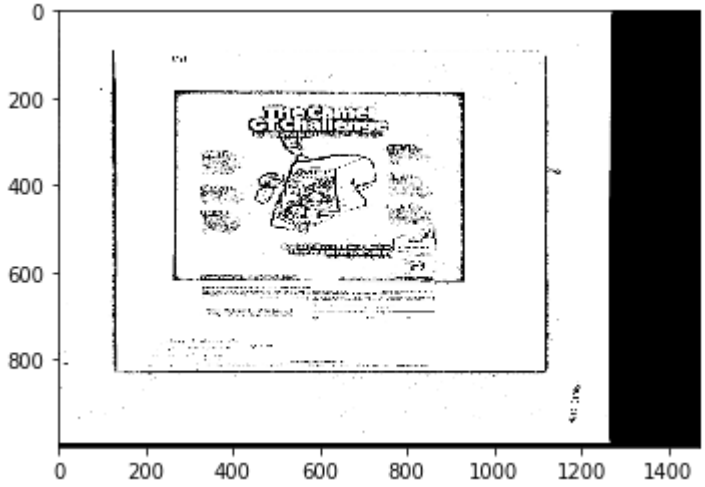
advertisement
/content/DocumentImages/train/advertisement/2061002602.tif



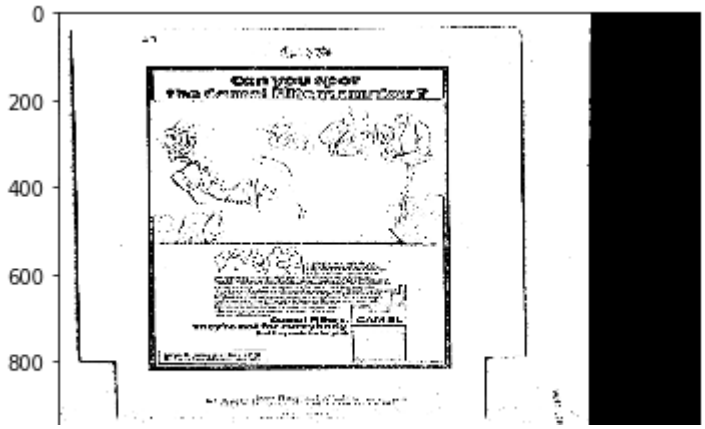
advertisement
/content/DocumentImages/train/advertisement/502607273+-7273.tif



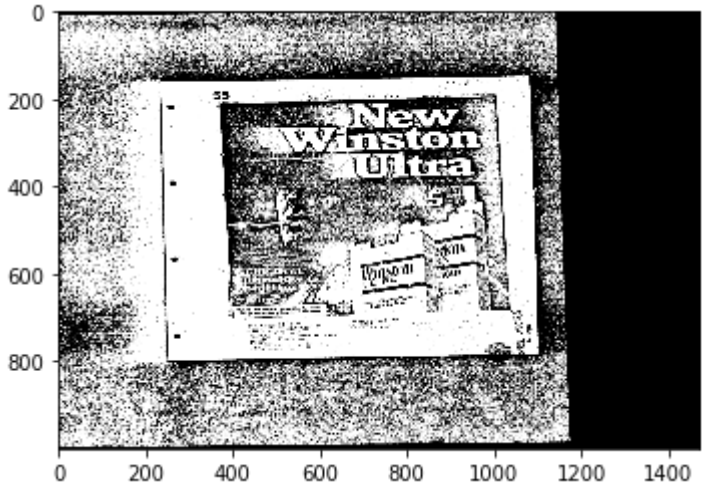
advertisement
/content/DocumentImages/train/advertisement/502100626+-0626.tif



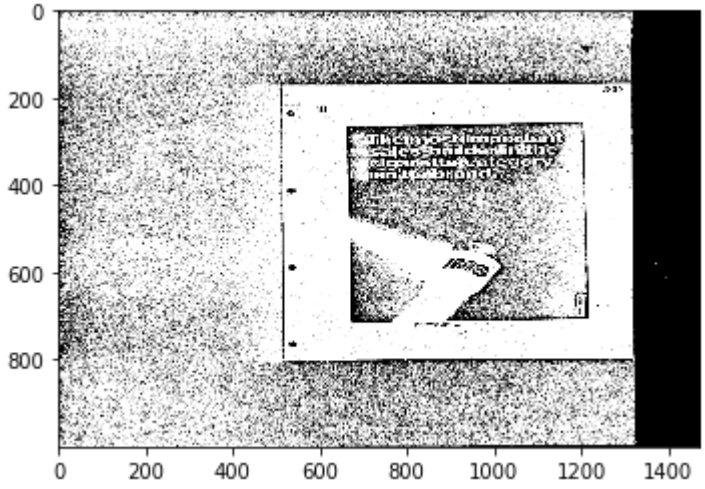
advertisement
/content/DocumentImages/train/advertisement/502100244+-0244.tif



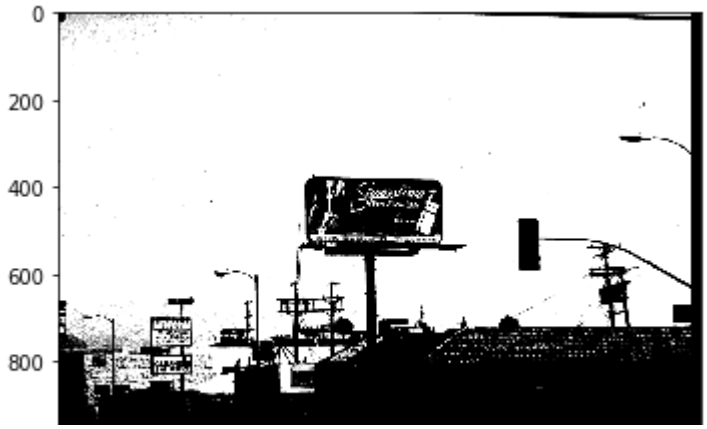
advertisement
/content/DocumentImages/train/advertisement/502607260+-7260.tif



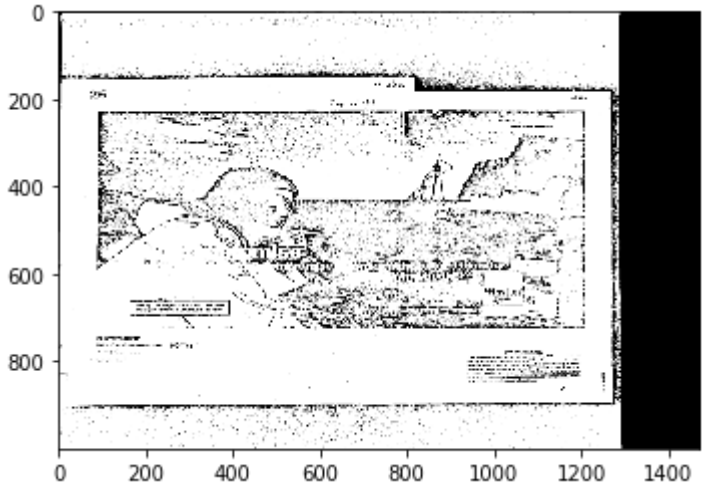
advertisement
/content/DocumentImages/train/advertisement/502610456+-0460.tif



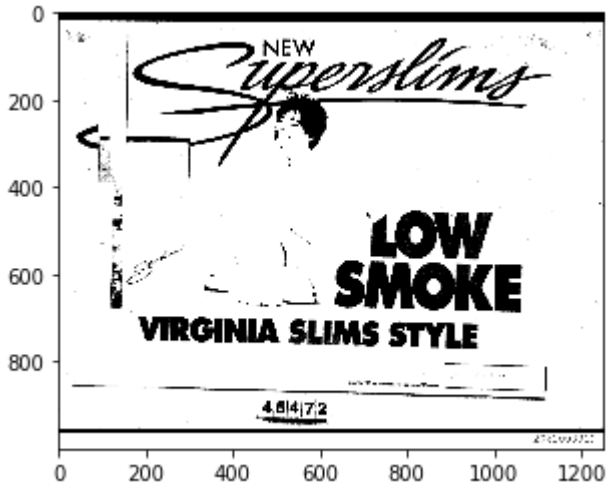
advertisement
/content/DocumentImages/train/advertisement/2061000249.tif



advertisement
/content/DocumentImages/train/advertisement/502607082+-7082.tif

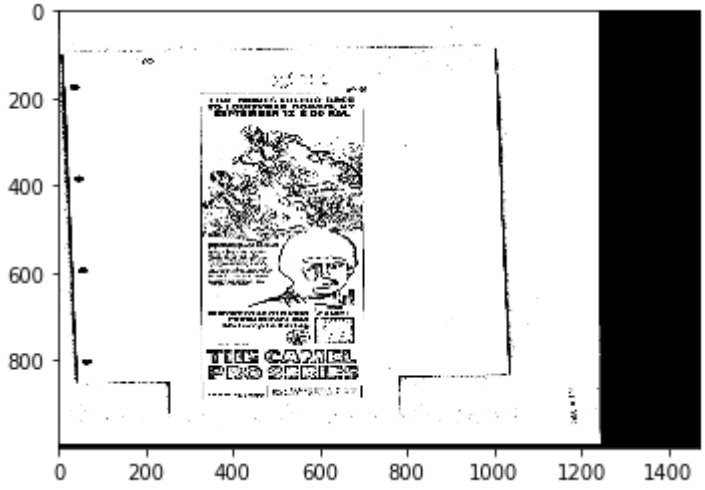


advertisement
/content/DocumentImages/train/advertisement/2061003319.tif

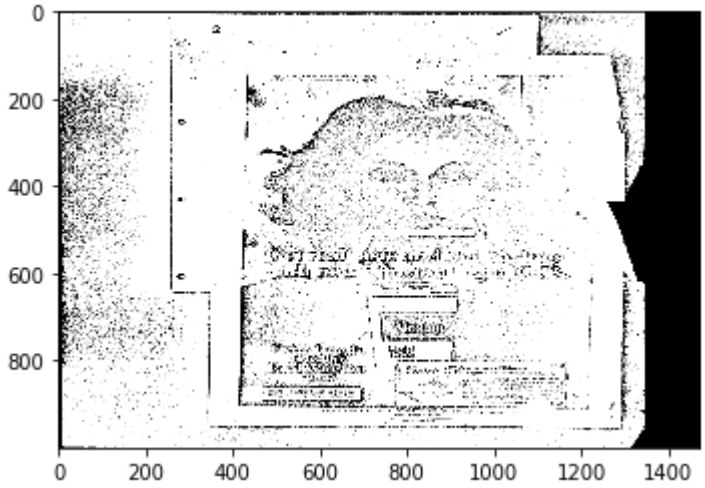


advertisement
/content/DocumentImages/train/advertisement/2061003179.tif

advertisement
/content/DocumentImages/train/advertisement/502100397+-0397.tif

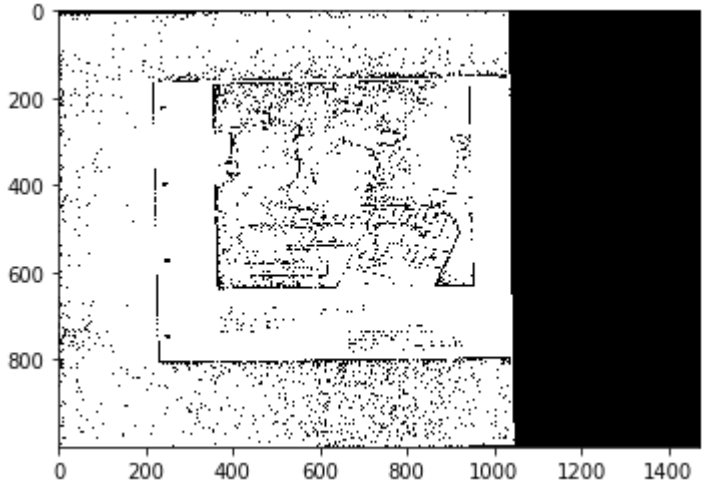


advertisement
/content/DocumentImages/train/advertisement/502607187+-7187.tif

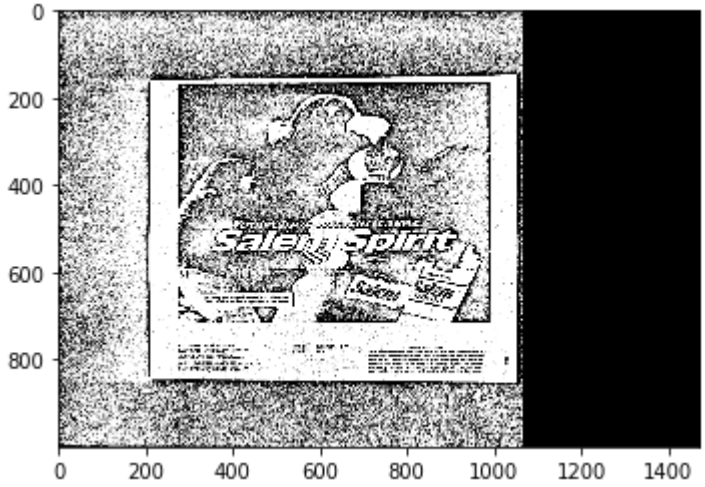


advertisement
/content/DocumentImages/train/advertisement/502605768+-5768.tif

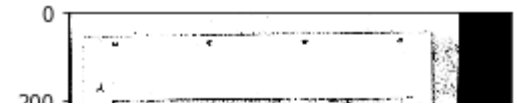
advertisement
/content/DocumentImages/train/advertisement/502610767+-0767.tif



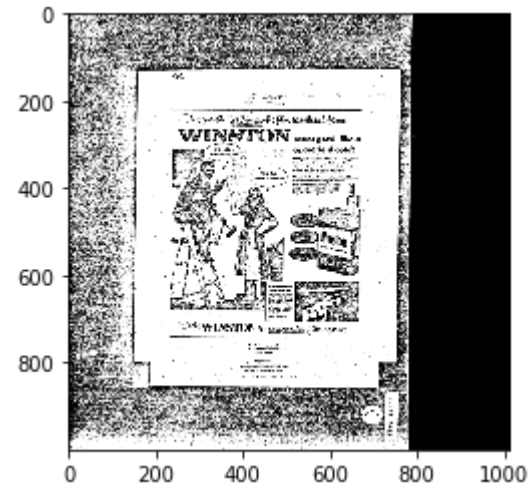
advertisement
/content/DocumentImages/train/advertisement/502610712+-0712.tif



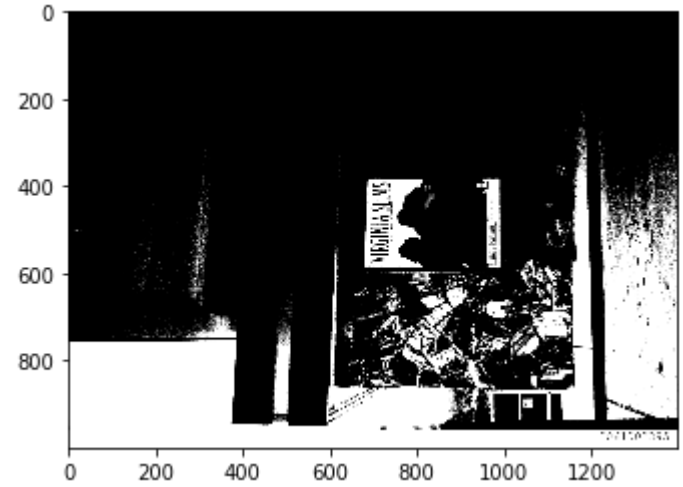
advertisement
/content/DocumentImages/train/advertisement/507806591.tif



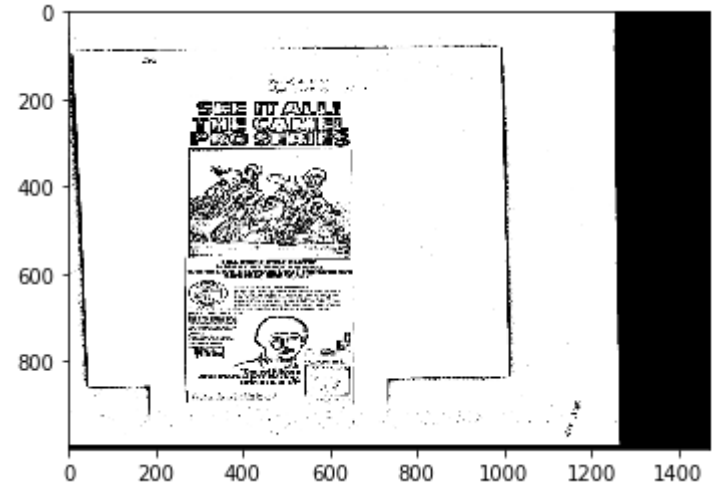
advertisement
/content/DocumentImages/train/advertisement/502218069.tif



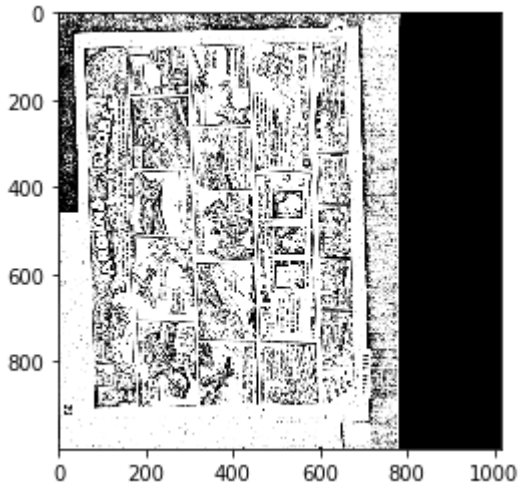
advertisement
/content/DocumentImages/train/advertisement/2061002395.tif



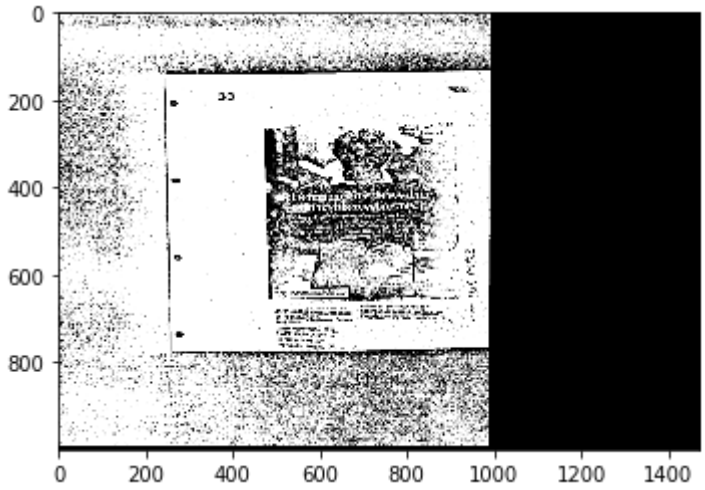
advertisement
/content/DocumentImages/train/advertisement/502100336+-0336.tif



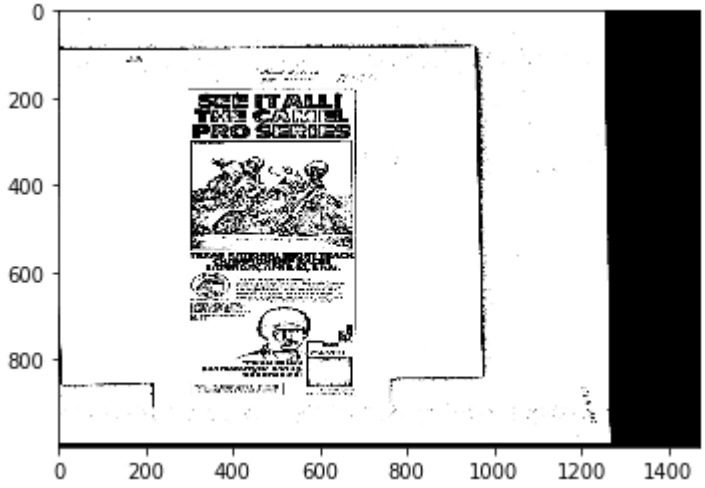
advertisement
/content/DocumentImages/train/advertisement/501947895.tif



advertisement
/content/DocumentImages/train/advertisement/502606747+-6747.tif

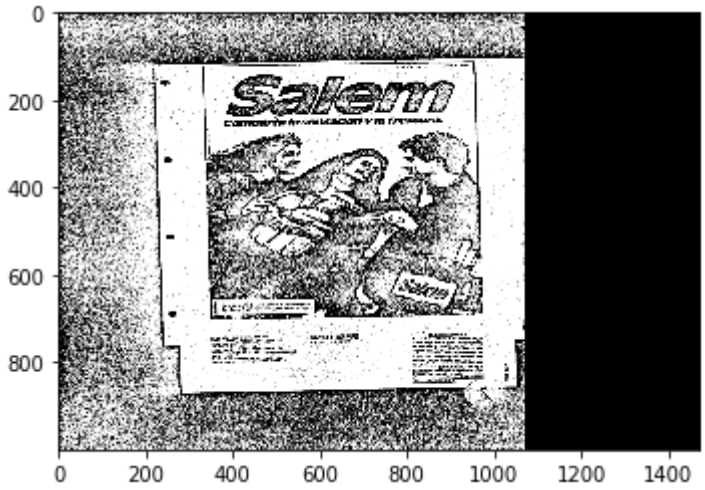


advertisement
/content/DocumentImages/train/advertisement/502100310+-0310.tif

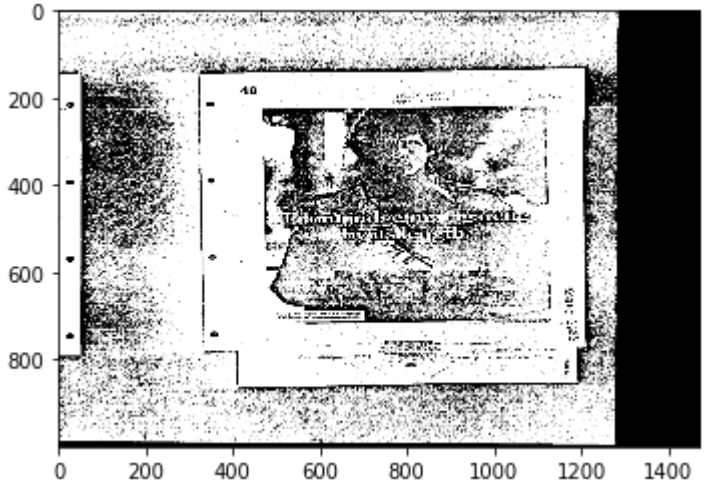


advertisement
/content/DocumentImages/train/advertisement/502100580+-0580.tif

```
advertisement
/content/DocumentImages/train/advertisement/502610784+-0784.tif
```

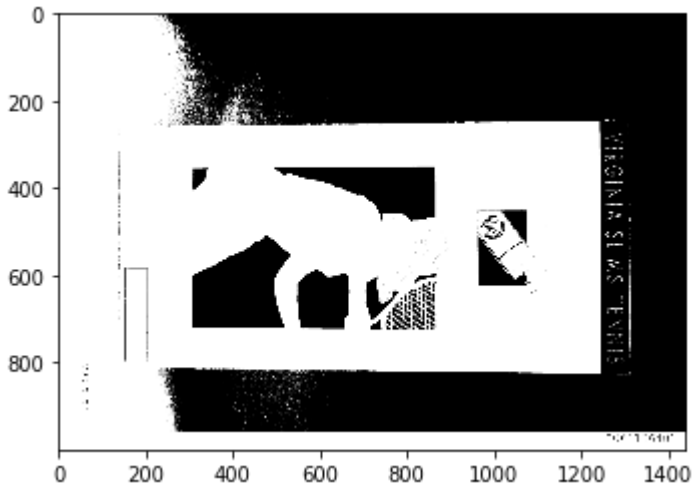



```
advertisement
/content/DocumentImages/train/advertisement/502606686+-6686.tif
```

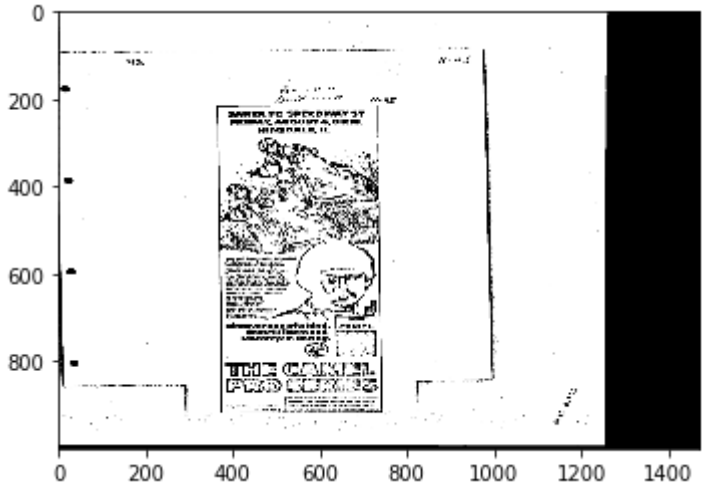


```
advertisement
```

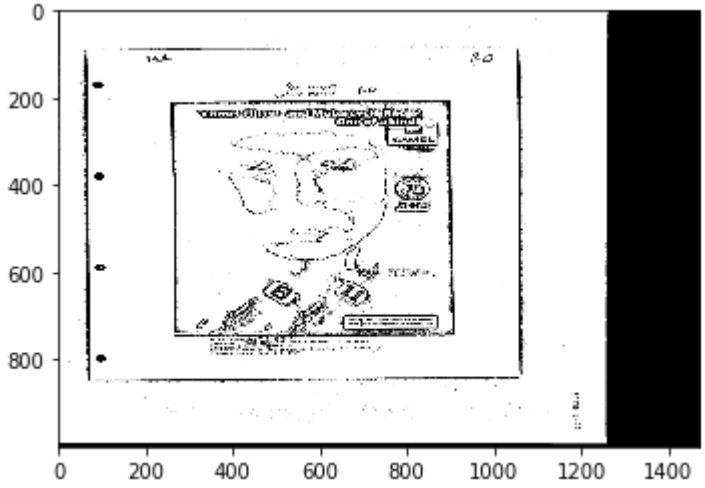
/content/DocumentImages/train/advertisement/2061186481.tif



advertisement
/content/DocumentImages/train/advertisement/502100438+-0438.tif

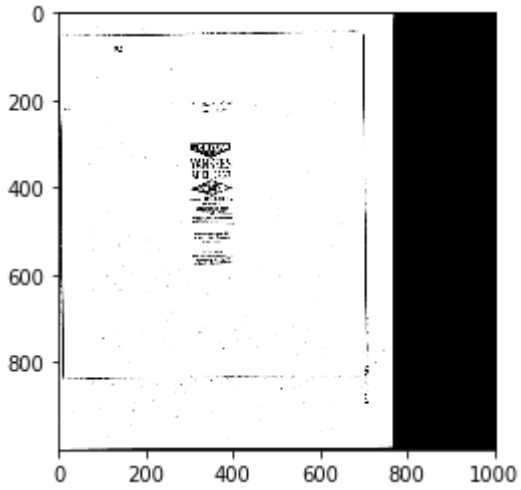


advertisement
/content/DocumentImages/train/advertisement/502100472+-0472.tif

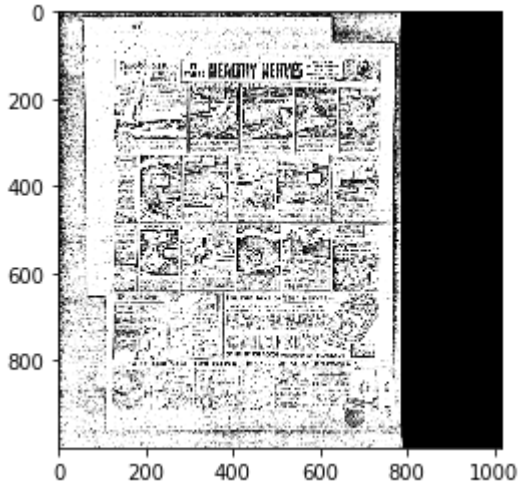


advertisement
/content/DocumentImages/train/advertisement/503944440+-4440.tif

advertisement
/content/DocumentImages/train/advertisement/502474201.tif

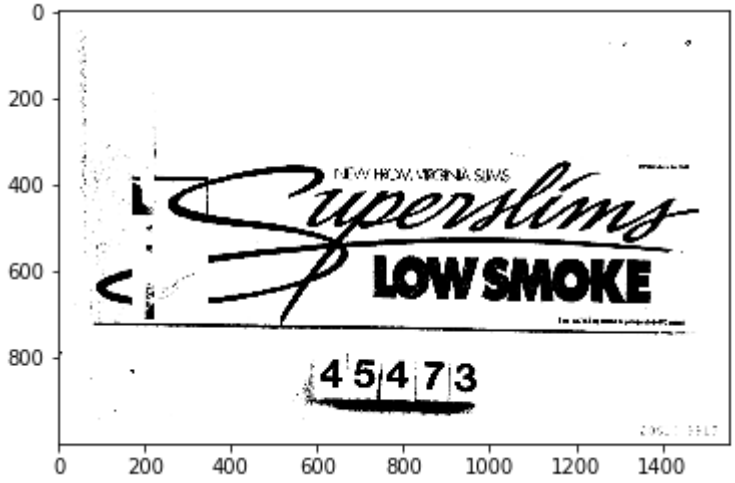


advertisement
/content/DocumentImages/train/advertisement/501949667.tif



advertisement
/content/DocumentImages/train/advertisement/501947846.tif

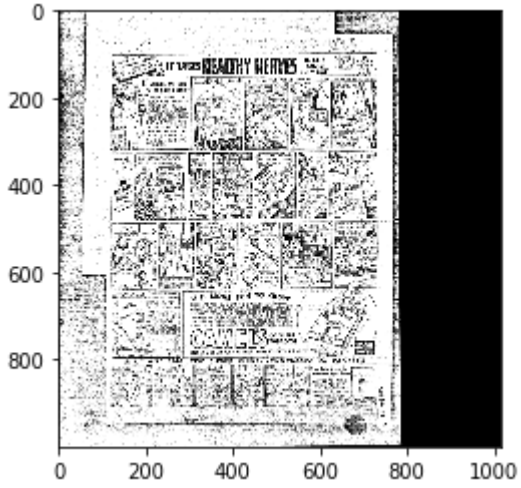
advertisement
/content/DocumentImages/train/advertisement/2061003315.tif



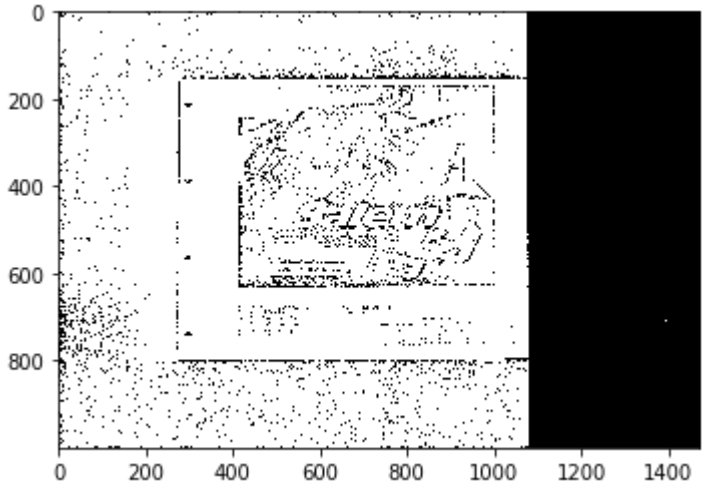
advertisement
/content/DocumentImages/train/advertisement/2061002461.tif

advertisement
/content/DocumentImages/train/advertisement/501947872.tif

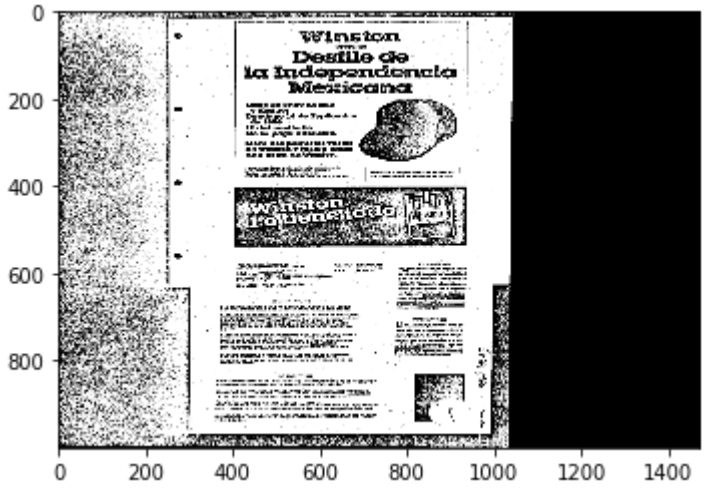
advertisement
/content/DocumentImages/train/advertisement/501949640.tif



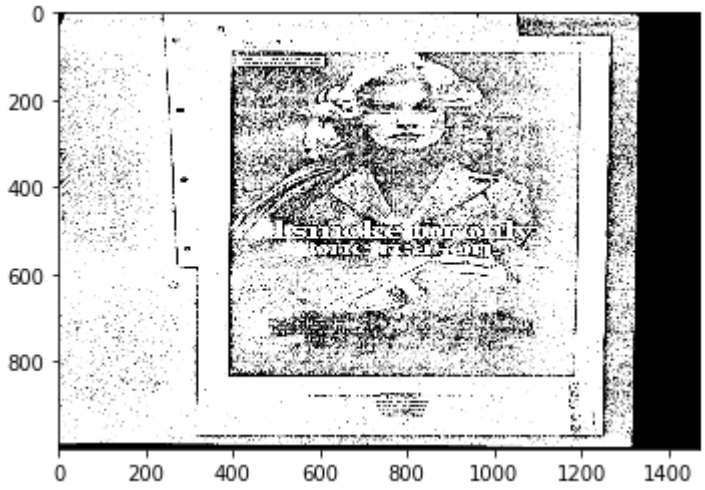
advertisement
/content/DocumentImages/train/advertisement/502610714+-0714.tif



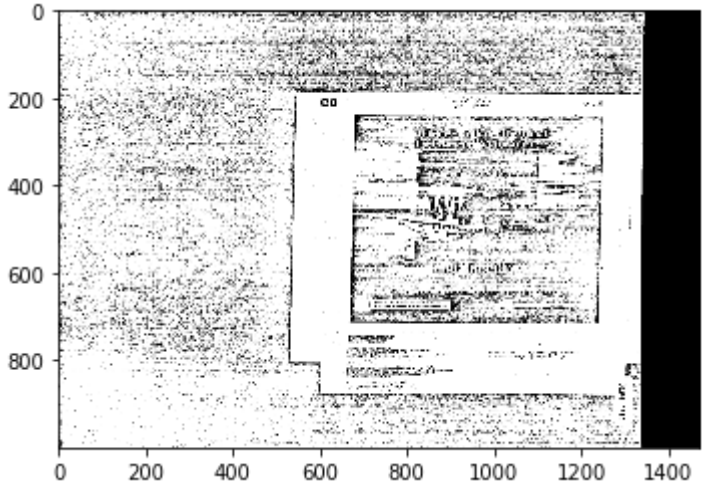
advertisement
/content/DocumentImages/train/advertisement/502610866+-0866.tif



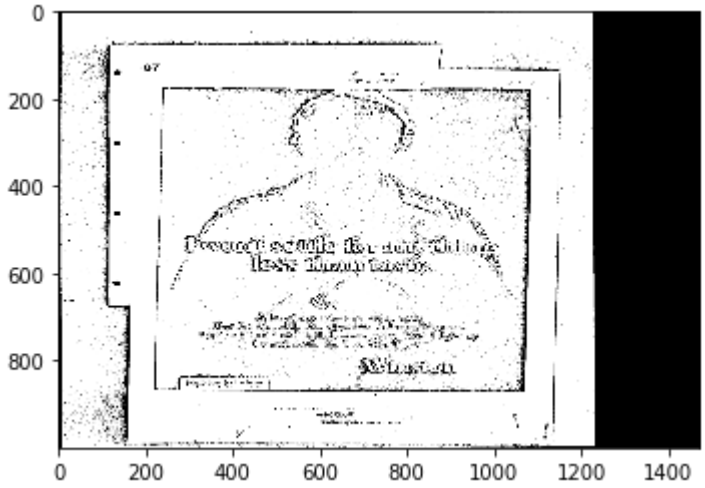
advertisement
/content/DocumentImages/train/advertisement/502606628+-6628.tif



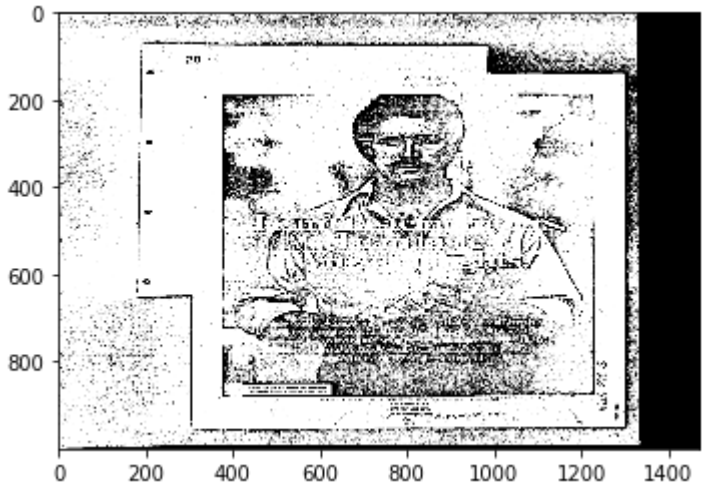
advertisement
/content/DocumentImages/train/advertisement/503961610+-1610.tif



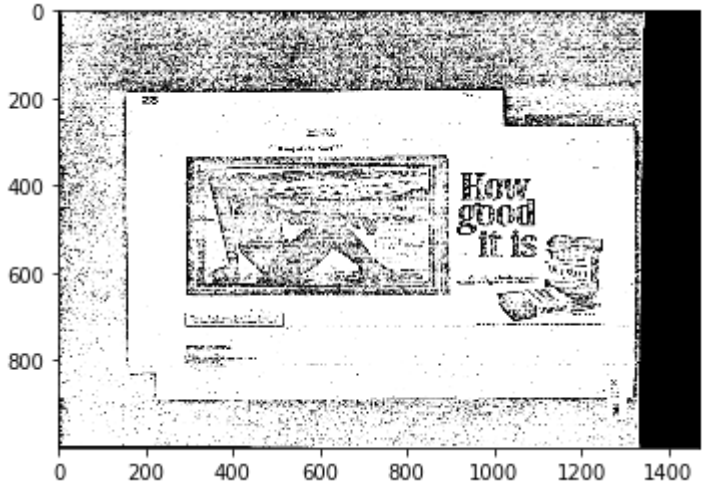
advertisement
/content/DocumentImages/train/advertisement/502605783+-5783.tif



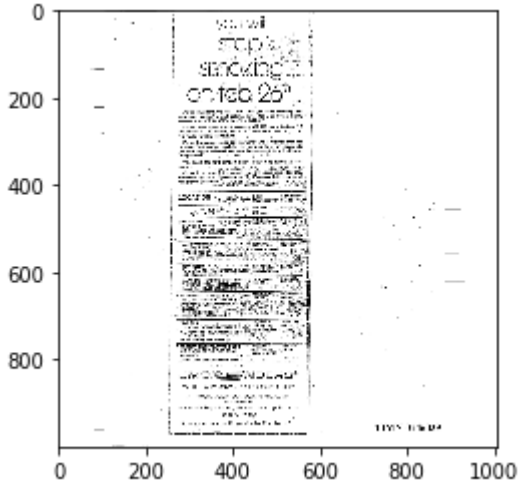
advertisement
/content/DocumentImages/train/advertisement/502606734+-6734.tif



advertisement
/content/DocumentImages/train/advertisement/503961552+-1552.tif

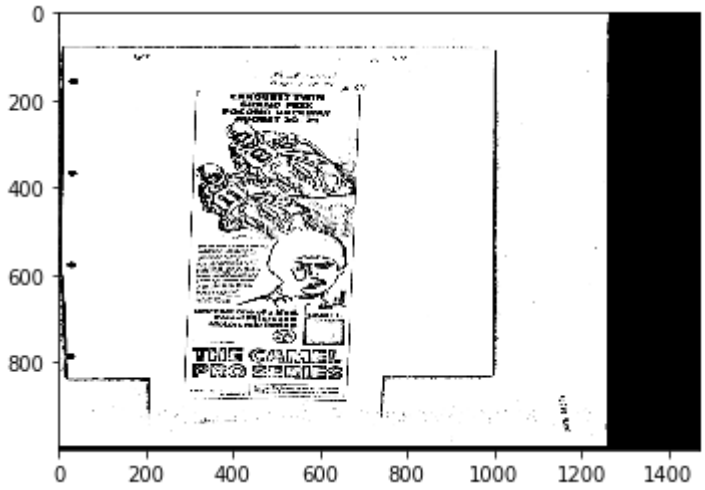


advertisement
/content/DocumentImages/train/advertisement/tob06520.84.tif

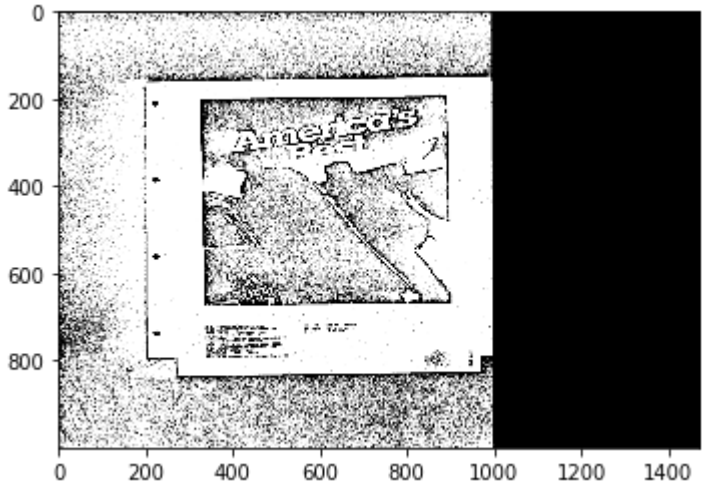


advertisement
/content/DocumentImages/train/advertisement/2040993101.tif

advertisement
/content/DocumentImages/train/advertisement/502100394+-0394.tif

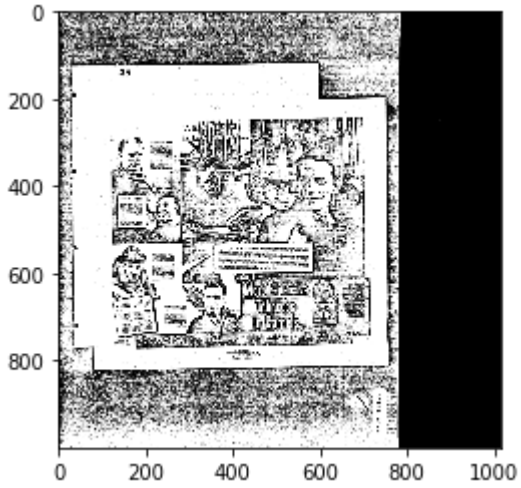


advertisement
/content/DocumentImages/train/advertisement/502610825+-0825.tif

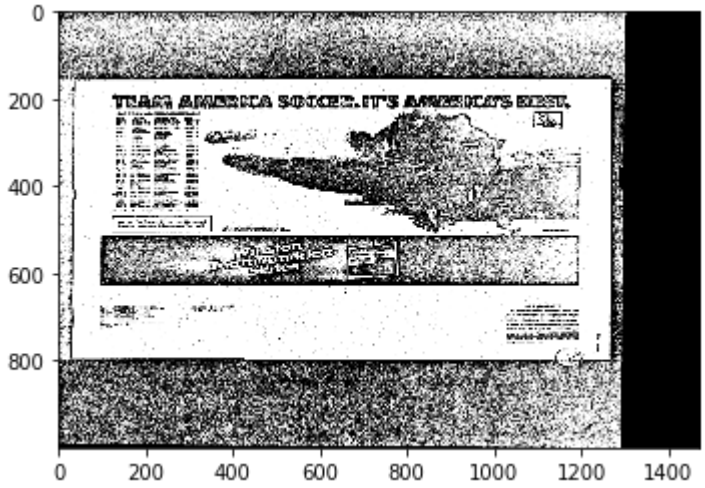


advertisement
/content/DocumentImages/train/advertisement/502100232+-0232.tif

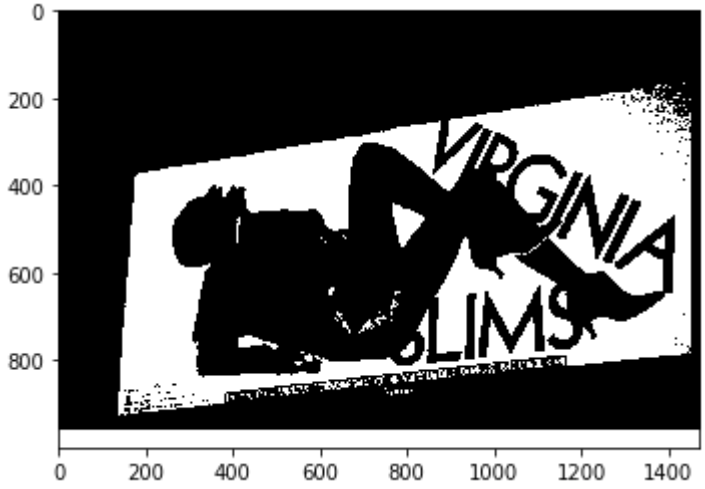
advertisement
/content/DocumentImages/train/advertisement/501947874.tif



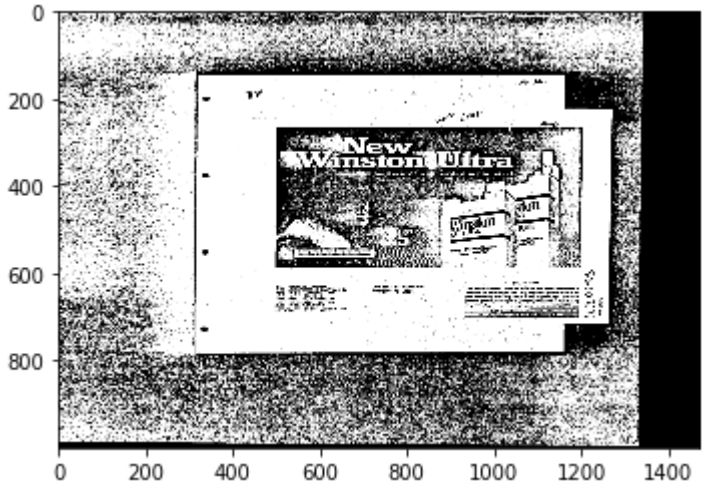
advertisement
/content/DocumentImages/train/advertisement/502610884+-0884.tif



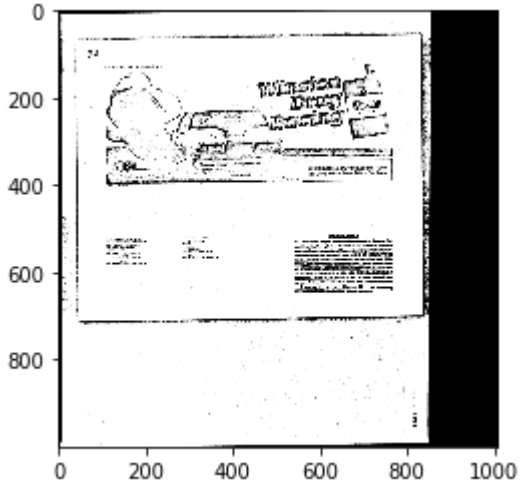
advertisement
/content/DocumentImages/train/advertisement/2061000284.tif



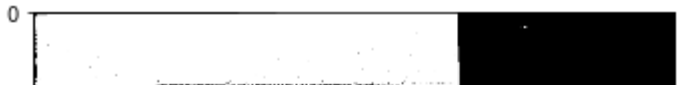
```
advertisement
/content/DocumentImages/train/advertisement/502607277+-7277.tif
```



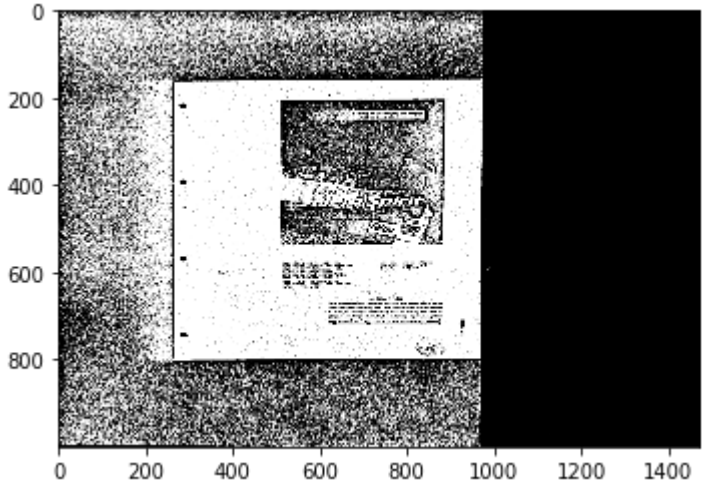

```
advertisement
/content/DocumentImages/train/advertisement/507806506.tif
```



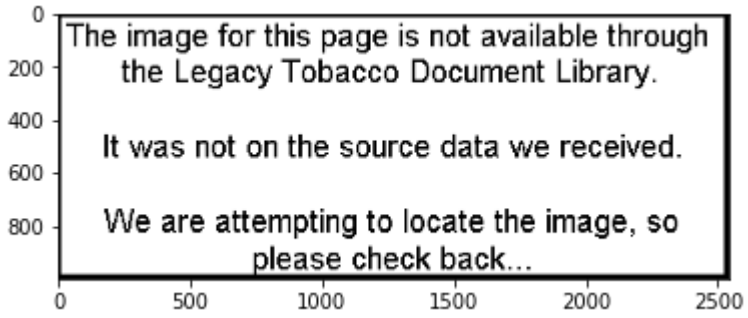
```
advertisement
/content/DocumentImages/train/advertisement/502605400+-5400.tif
```



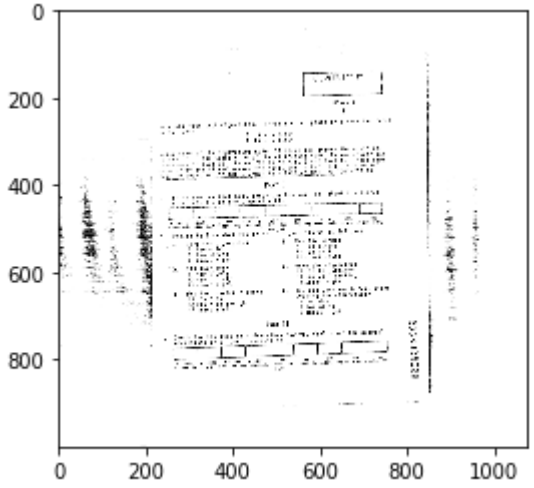
advertisement
/content/DocumentImages/train/advertisement/502610752+-0752.tif



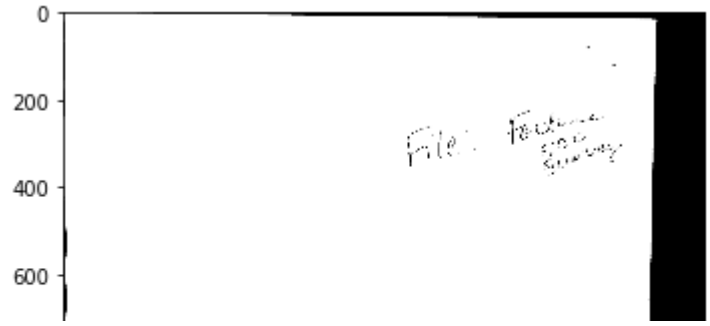
questionnaire
/content/DocumentImages/train/questionnaire/1002481794_1799.tif



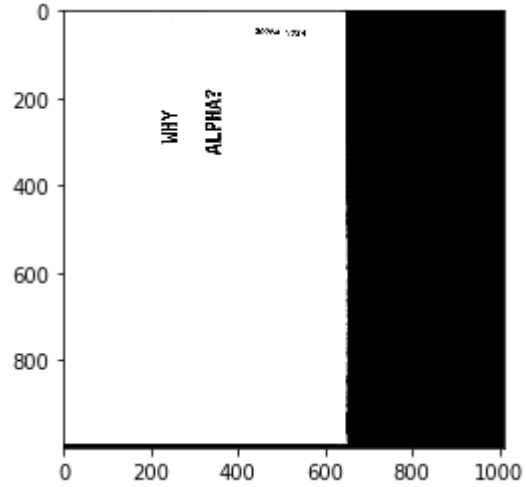
questionnaire
/content/DocumentImages/train/questionnaire/2058008535_2058008536.tif



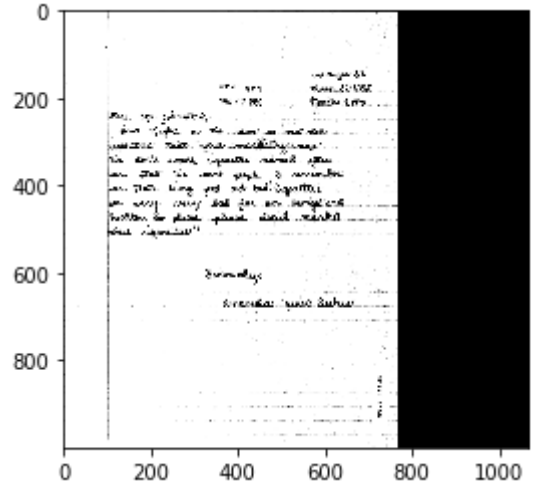
questionnaire
/content/DocumentImages/train/questionnaire/507781198+-1204.tif



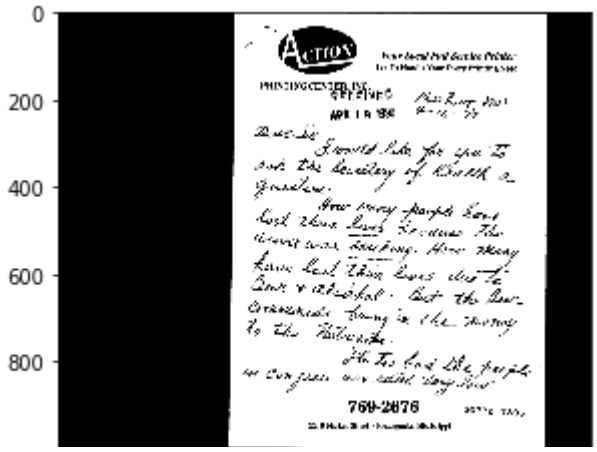
questionnaire
/content/DocumentImages/train/questionnaire/507543024_507543025.tif



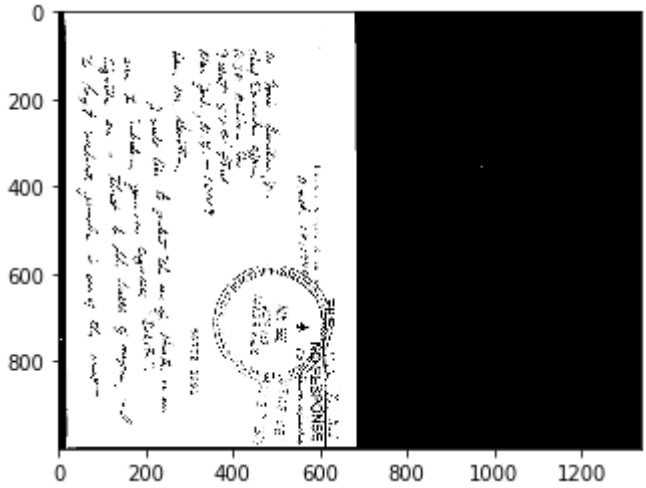
handwritten
/content/DocumentImages/train/handwritten/507706166.tif



handwritten
/content/DocumentImages/train/handwritten/507707896_507707898.tif



handwritten
/content/DocumentImages/train/handwritten/507703993_507703994.tif



In [0]:

```
1 image_dim = {"Document": docs, "Height": height, "Width": width}
2 image_dim_df = pd.DataFrame(image_dim, columns = ["Document", "Height", "Width"])
```

In [0]:

```
#Quantiles by class
image_dim_df.groupby('Document').quantile([0, 0.25, 0.5, 0.75, 1])
```

Out[35]:

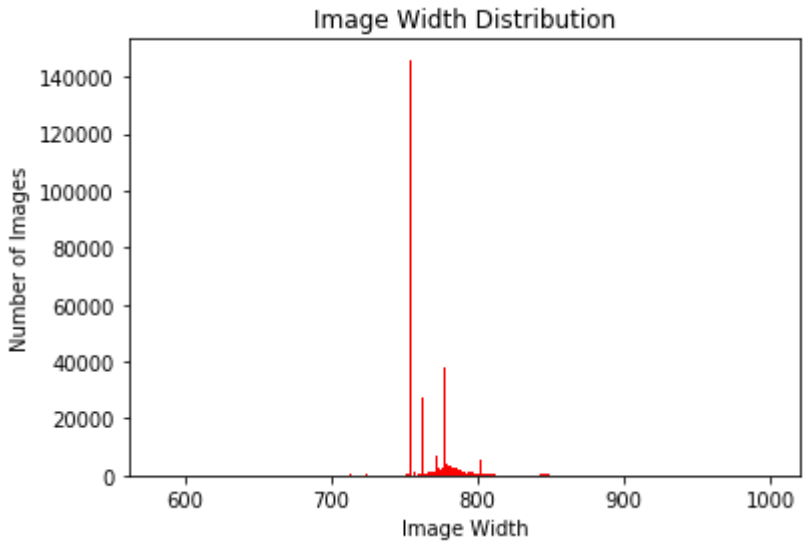
		Height	Width
Document			
	0.00	1000.0	596.0
	0.25	1000.0	754.0
advertisement	**0.50**	1000.0	757.0
	0.75	1000.0	777.0
	1.00	1000.0	3235.0
...	**...**	...	...
	0.00	1000.0	611.0
	0.25	1000.0	754.0
specification	**0.50**	1000.0	775.0
	0.75	1000.0	784.0
	1.00	1000.0	1097.0

80 rows × 2 columns

In [0]:

```
#Image Width Distribution
print(len(width))
width1 = [w for w in width if w <= 1000]
plt.hist(width, bins = range(min(width1), max(width1) + 1, 1), color = 'red')
plt.title('Image Width Distribution')
plt.xlabel('Image Width')
plt.ylabel('Number of Images')
plt.show()
```

```
312000
```

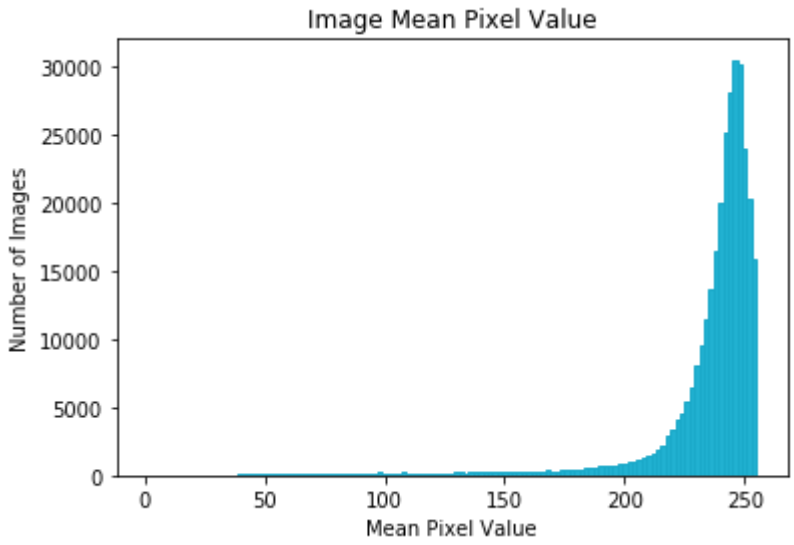


In [0]:

```
#Getting mean pixel value of each image
pixel_means = []
for image in image_files:
    img = cv2.imread(image)
    pixel_mean = np.mean(np.array(img))
    pixel_means.append(pixel_mean)
```

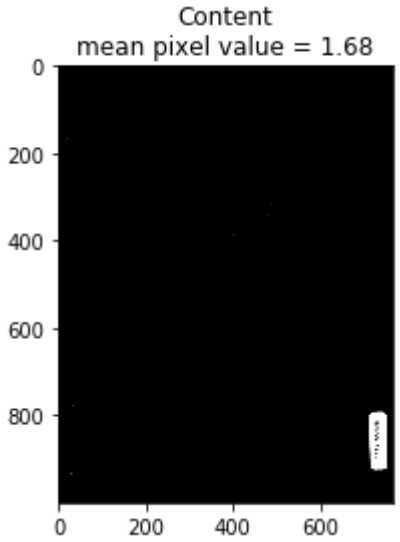
In [0]:

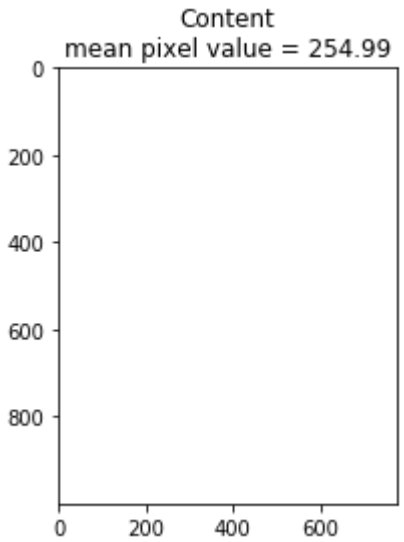
```
pixel_means_int = [int(x) for x in pixel_means]
plt.hist(pixel_means,bins = range(min(pixel_means_int),max(pixel_means_int) + 2, 2),co
plt.title('Image Mean Pixel Value')
plt.xlabel('Mean Pixel Value')
plt.ylabel('Number of Images')
plt.show()
```



In [0]:

```
#Identifying the images with Highest (Whitest) mean pixel value and Lowest (Blackest) m
max_index = pixel_means.index(max(pixel_means))
min_index = pixel_means.index(min(pixel_means))
max_image = image_files[max_index]
min_image = image_files[min_index]
max_img = cv2.imread(max_image, 0)
min_img = cv2.imread(min_image, 0)
max_path = os.path.dirname(max_image)
min_path = os.path.dirname(min_image)
max_doctype = max_path.split('/')[1]
min_doctype = min_path.split('/')[1]
plt.imshow(min_img, cmap = 'gray')
plt.title(min_doctype.title() + "\nmean pixel value = {}".format(min(pixel_means).roun
plt.show()
plt.imshow(max_img, cmap = 'gray')
plt.title(max_doctype.title() + "\nmean pixel value = {}".format(max(pixel_means).roun
plt.show()
```





In [0]:

```
#Creating images at half size of the previous for visual examination
full_image = cv2.imread('/content/DocumentImages/train/letter/00001807_00001809.tif', 
height, width = full_image.shape
print(full_image.shape)
cv2.imwrite('/content/Report/Images/fullimage.tif', full_image)
half_image = cv2.resize(full_image, (int(width/2), int(height/2)))
print(half_image.shape)
cv2.imwrite('/content/Report/Images/halfimage.tif', half_image)
quarter_image = cv2.resize(full_image, (int(width/4), int(height/4)))
print(quarter_image.shape)
cv2.imwrite('/content/Report/Images/quarterimage.tif', quarter_image)
eighth_image = cv2.resize(full_image, (int(width/8), int(height/8)))
print(eighth_image.shape)
cv2.imwrite('/content/Report/Images/eighthimage.tif', eighth_image)
sixteenth_image = cv2.resize(full_image, (int(width/16), int(height/16)))
print(sixteenth_image.shape)
cv2.imwrite('/content/Report/Images/sixteenthimage.tif', sixteenth_image)
```

In [0]:

```
random.seed(29)
def random_photo_per_class(path):
    random_images = []
    for root, dirs, files in os.walk(path):
        if root[-5:] != "train":
            image = random.choice(os.listdir("{}".format(root)))
            random_images.append(str(root)+os.sep+str(image))
    return random_images

random_photo_per_class('/content/DocumentImages/train//')
```

Out[29]:

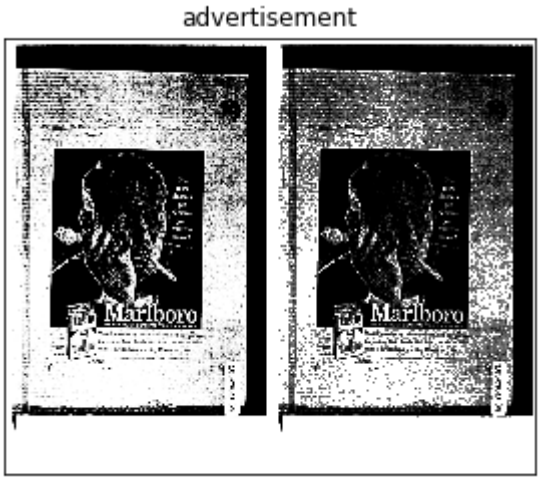
```
['/content/DocumentImages/train///form',
 '/content/DocumentImages/train//memo/2041794703.tif',
 '/content/DocumentImages/train//specification/2053453701.tif',
 '/content/DocumentImages/train//form/2026528544.tif',
 '/content/DocumentImages/train//file folder/2063071581.tif',
 '/content/DocumentImages/train//letter/0071035877.tif',
 '/content/DocumentImages/train//news article/2048785571_5575.tif',
 '/content/DocumentImages/train//budget/2040786789_2040786796.tif',
 '/content/DocumentImages/train//advertisement/502613194a-3195.tif',
 '/content/DocumentImages/train//presentation/0060240890.tif',
 '/content/DocumentImages/train//invoice/87149101_87149104.tif',
 '/content/DocumentImages/train//scientific publication/PUBLICATIONS030212-
0.tif',
 '/content/DocumentImages/train//scientific report/87733690.tif',
 '/content/DocumentImages/train//handwritten/518220449+-0450.tif',
 '/content/DocumentImages/train//resume/50640305-0306.tif',
 '/content/DocumentImages/train//email/528806787+-6789.tif',
 '/content/DocumentImages/train//questionnaire/2024748477_2024748479.tif']
```

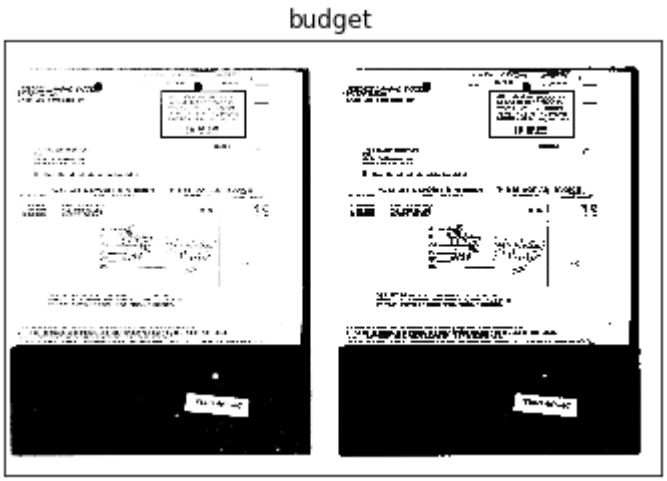
In [0]:

```
examples = random_photo_per_class('/content/DocumentImages/train/')
```

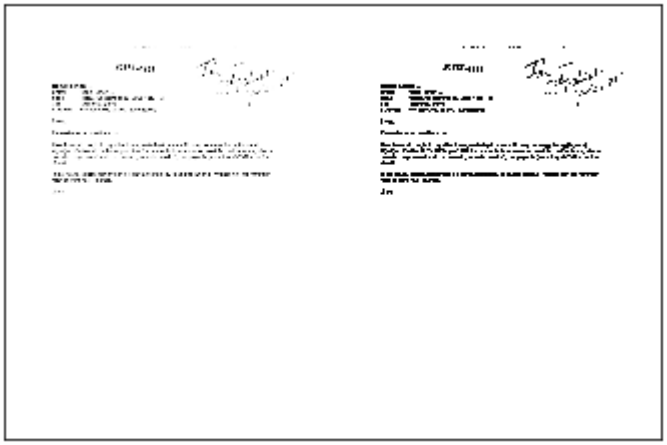
In [0]:

```
for image in examples:
    directory = os.path.dirname(image)
    doctype = directory.rsplit('\\', 1)[-1]
    img = cv2.imread(image, 0)

    equ = cv2.equalizeHist(img)
    res = np.hstack((img,equu))

    plt.subplot(111)
    plt.imshow(res, cmap='Greys_r')
    plt.title('{}'.format(doctype))
    plt.xticks([])
    plt.yticks([])
    plt.show()
```

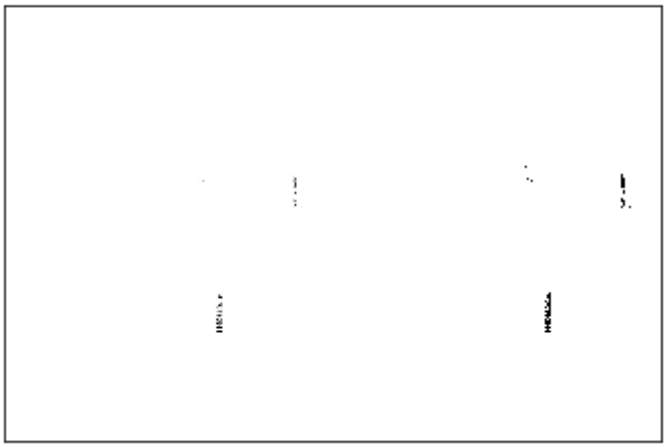




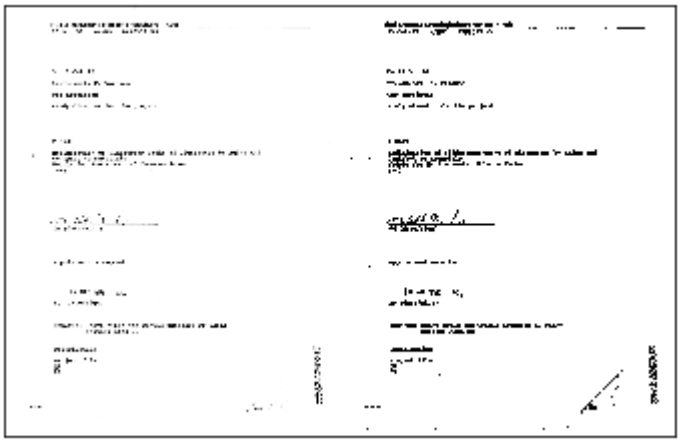
email



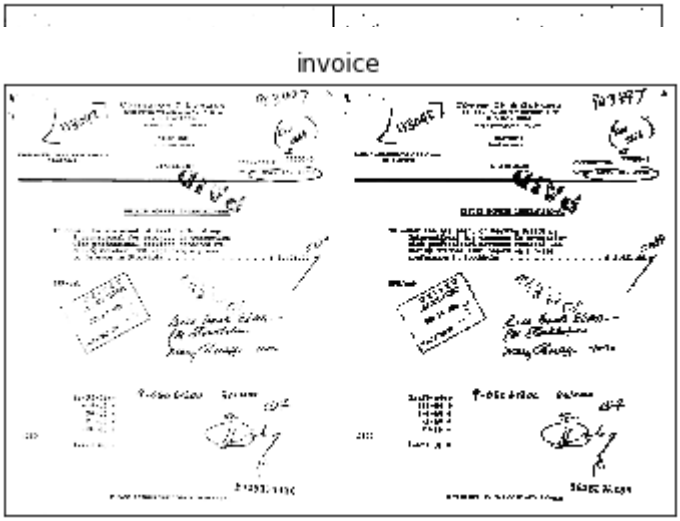
file folder



form

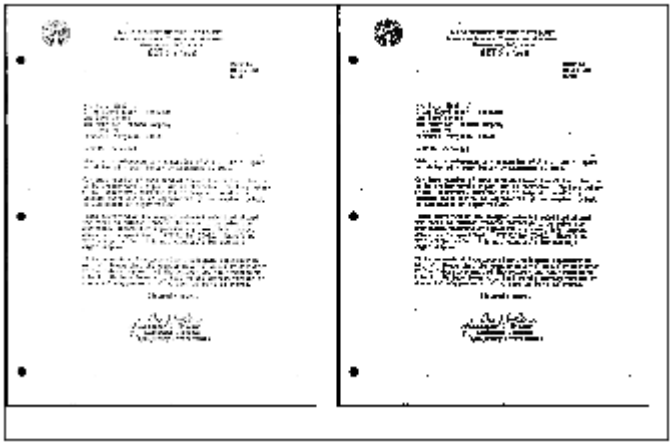


handwritten

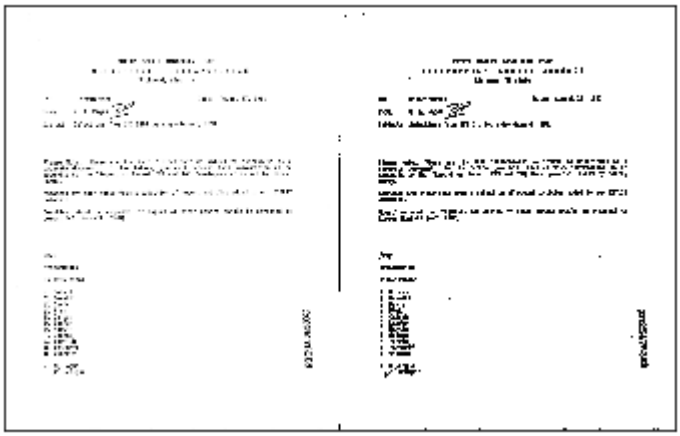


invoice

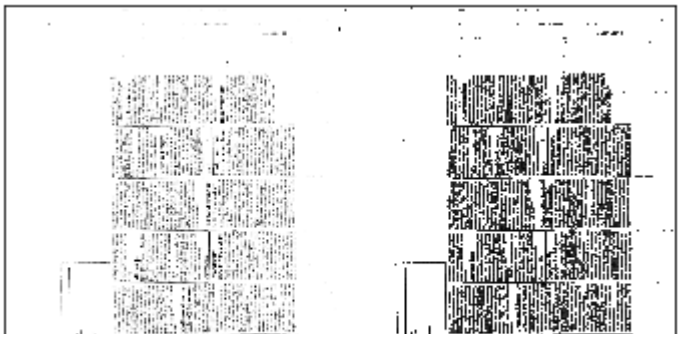
letter



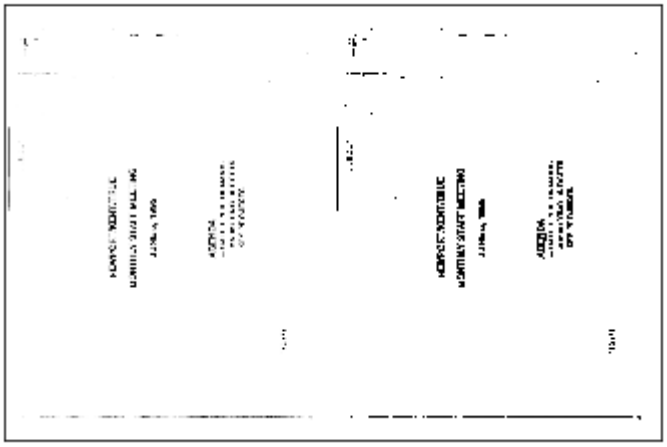
memo



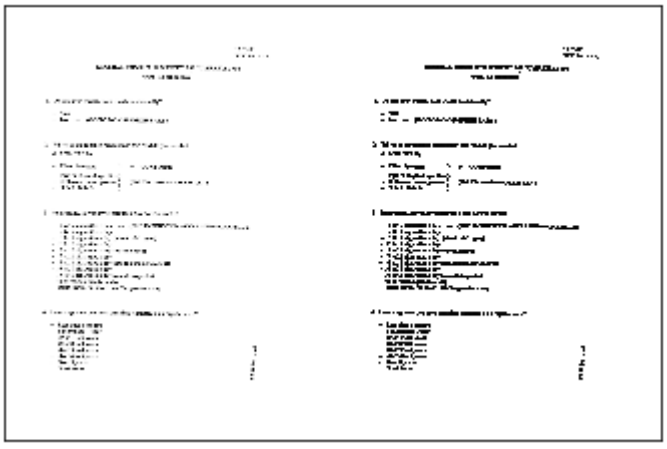
news article



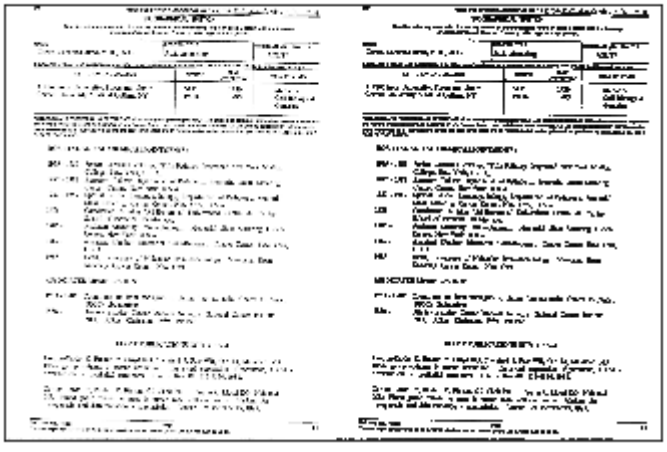
presentation



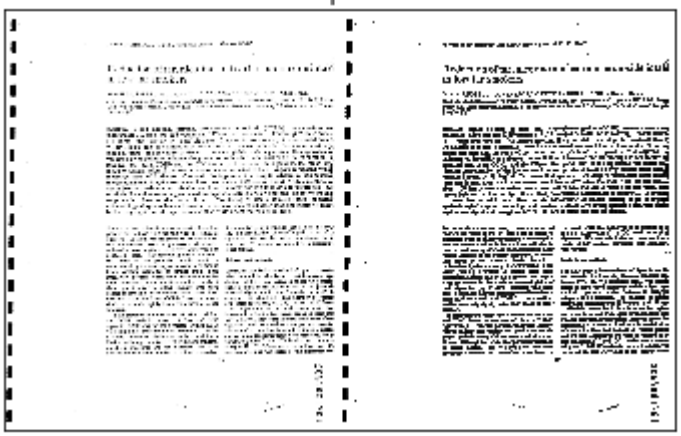
questionnaire



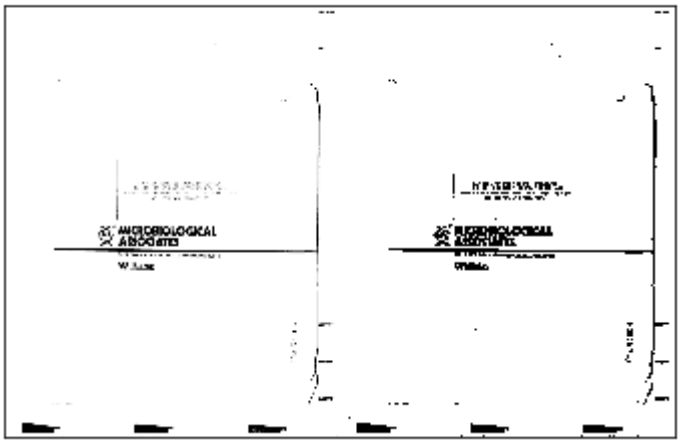
resume



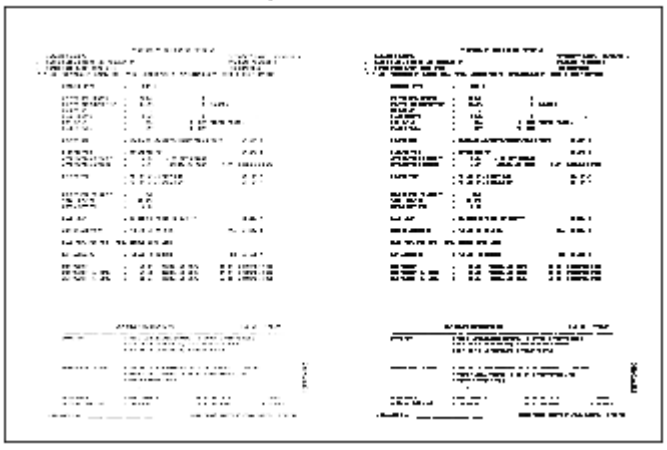
scientific publication



scientific report

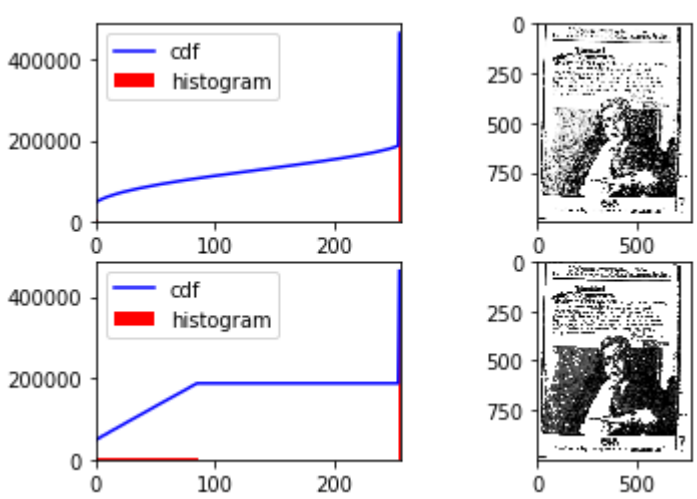


specification



In [0]:

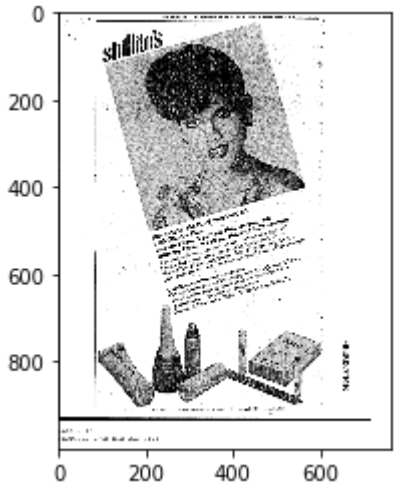
```
#https://github.com/jpcrum/Final_Project_Group5
#Histogram Equalization

img = cv2.imread('/content/DocumentImages/train/advertisement/00002210.tif',0)

#get histogram (count of each value) values and bin edges for all pixel values
hist,bins = np.histogram(img.flatten(),256,[0,256])

#cumulative sum to get cumulative distribution function
cdf = hist.cumsum()
#Normalize cdf by multiplying by most common pixel value divided by sum of pixel values
cdf_normalized = cdf * hist.max()/ cdf.max()


cdf_m = np.ma.masked_equal(cdf,0)
#min-max normalizing to [0,255] range
cdf_m = (cdf_m - cdf_m.min())*255/(cdf_m.max()-cdf_m.min())
cdf = np.ma.filled(cdf_m,0).astype('uint8')

img2 = cdf[img]

hist2,bins2 = np.histogram(img2.flatten(),256,[0,256])

cdf2 = hist2.cumsum()
cdf_normalized2 = cdf2 * hist2.max()/ cdf2.max()

plt.subplot(221)
plt.plot(cdf_normalized, color = 'b')
plt.hist(img.flatten(),256,[0,256], color = 'r')
plt.xlim([0,256])
plt.legend(('cdf','histogram'), loc = 'upper left')

plt.subplot(222)
plt.imshow(img, cmap='Greys_r')

plt.subplot(223)
plt.plot(cdf_normalized2, color = 'b')
plt.hist(img2.flatten(),256,[0,256], color = 'r')
plt.xlim([0,256])
plt.legend(('cdf','histogram'), loc = 'upper left')

plt.subplot(224)
plt.imshow(img2, cmap='Greys_r')

plt.show()
print(len(cdf))
print(len(cdf_m))
```



256
256

In [0]:

```
image = cv2.imread('/content/DocumentImages/test/advertisement/604659.tif', 0)
plt.imshow(image, cmap='gray')
plt.show()
```



Here we are Breaking the Images into 5 different Sub-categories like

1.Header Images

2.Footer Images

3.Left Body Images

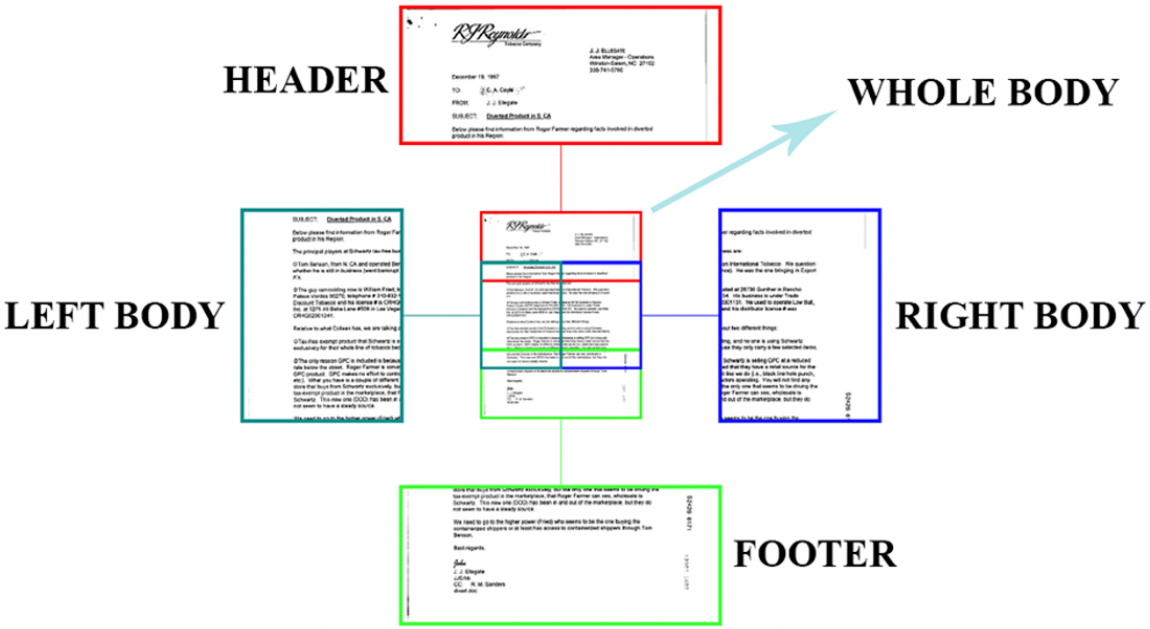
4.Right Body images

5.Whole Images (Header Images + Footer Images + Left Body Images + Right Body images)

In [0]:

```
from IPython.display import Image
Image("REGION_BASED_SPLITTING.png", width=800, height=600)
```

Out[13]:



Whole Model

In [0]:

```
#https://keras.io/preprocessing/image/
train_datagen = ImageDataGenerator(rescale = 1./255)
test_datagen = ImageDataGenerator(rescale = 1./255)
valid_datagen = ImageDataGenerator(rescale = 1./255)
```

In [0]:

```
train_whole = train_datagen.flow_from_directory('/content/DocumentImages/train', batch
test_whole = test_datagen.flow_from_directory('/content/DocumentImages/test', batch_si
valid_whole = valid_datagen.flow_from_directory('/content/DocumentImages/valid', batch
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
#https://keras.io/callbacks/
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=2, min_lr=0.000
mcp_save = ModelCheckpoint('model_whole.hdf5', save_best_only=True, monitor='acc', mod
```

In [0]:

```
1 model_whole = VGG16(weights = "imagenet", include_top = False, input_shape = (256, 256
```

In [0]:

```
for layer in model_whole.layers:
    layer.trainable=True

x = model_whole.output
x = Flatten()(x)
x = Dense(256, activation="relu")(x)
x = Dropout(0.5)(x)
x = Dense(128, activation="relu")(x)
predictions = Dense(16, activation="softmax")(x)

model_whole_final = Model(input = model_whole.input, output = predictions)

model_whole_final.compile(loss = "categorical_crossentropy", optimizer=Adam(lr=0.0001)

model_whole_final.summary()
```

```
Model: "model_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         (None, 256, 256, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
_________________________________________________________________
```

```
flatten_2 (Flatten)          (None, 32768)             0
_________________________________________________________________
dense_4 (Dense)              (None, 256)               8388864
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_5 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_6 (Dense)              (None, 16)                2064
=================================================================
Total params: 23,138,512
Trainable params: 23,138,512
Non-trainable params: 0
_________________________________________________________________
```

In [0]:

```
1 tbCallBack = callbacks.TensorBoard(log_dir='./Graph', histogram_freq=0, write_graph=Tru
```

In [0]:

```
1 history = model_whole_final.fit_generator(train_whole, steps_per_epoch = 312000/128, e
```

```
Epoch 1/10
2438/2437 [==============================] - 767s 315ms/step - loss: 1.1941
- acc: 0.6427 - val_loss: 0.7395 - val_acc: 0.7772
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/callbac
ks.py:1265: The name tf.Summary is deprecated. Please use tf.compat.v1.Summa
ry instead.

Epoch 2/10
2438/2437 [==============================] - 766s 314ms/step - loss: 0.7662
- acc: 0.7774 - val_loss: 0.6474 - val_acc: 0.8082
Epoch 3/10
2438/2437 [==============================] - 765s 314ms/step - loss: 0.6616
- acc: 0.8078 - val_loss: 0.5324 - val_acc: 0.8421
Epoch 4/10
2438/2437 [==============================] - 768s 315ms/step - loss: 0.6039
- acc: 0.8248 - val_loss: 0.5005 - val_acc: 0.8538
Epoch 5/10
2438/2437 [==============================] - 757s 311ms/step - loss: 0.5170
- acc: 0.8486 - val_loss: 0.4802 - val_acc: 0.8582
Epoch 6/10
2438/2437 [==============================] - 765s 314ms/step - loss: 0.5088
- acc: 0.8516 - val_loss: 0.4507 - val_acc: 0.8657
Epoch 7/10
2438/2437 [==============================] - 765s 314ms/step - loss: 0.4833
- acc: 0.8592 - val_loss: 0.4295 - val_acc: 0.8737
Epoch 8/10
2438/2437 [==============================] - 766s 314ms/step - loss: 0.4693
- acc: 0.8634 - val_loss: 0.4330 - val_acc: 0.8758
Epoch 9/10
2438/2437 [==============================] - 757s 311ms/step - loss: 0.3977
- acc: 0.8818 - val_loss: 0.4111 - val_acc: 0.8798
Epoch 10/10
2438/2437 [==============================] - 765s 314ms/step - loss: 0.4012
- acc: 0.8825 - val_loss: 0.3991 - val_acc: 0.8797
```

In [0]:

```
score = model_whole_final.evaluate_generator(test_whole,steps=36800/128,workers=2, use
```

In [0]:

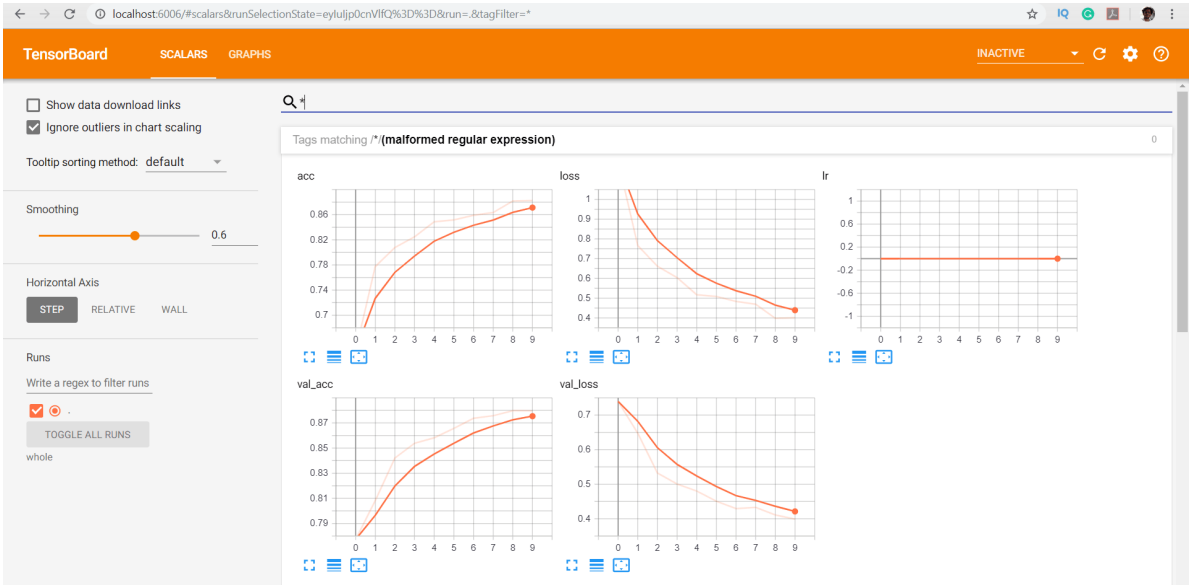
```
print("The Accuracy for the whole image model:",score[1]*100,"%")
```

```
The Accuracy for the whole image model: 87.61935763888889 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/whole_scalar graph.png")
```

Out[1]:



Header Image

In [0]:

```
model_header = VGG16(weights = "imagenet", include_top=False, input_shape = (256, 256,
```

In [0]:

```
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=2, min_lr=0.000
mcp_save = ModelCheckpoint('model_header.hdf5', save_best_only=True, monitor='acc', mo
```

In [0]:

```
#https://github.com/samblshikhar/Document-Image-Classification-with-Intra-Domain-Transf
for layer in model_header.layers:
    layer.trainable=True

x = model_header.output
x = Flatten()(x)
x = Dense(256, activation="relu")(x)
x = Dropout(0.5)(x)
x = Dense(128, activation="relu")(x)
predictions = Dense(16, activation="softmax")(x)

model_header_final = Model(input = model_header.input, output = predictions)

model_header_final.compile(loss = "categorical_crossentropy", optimizer =Adam(lr=0.000
```

In [0]:

```
1 model_header_final.summary()
```

```
Model: "model_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_4 (InputLayer)         (None, 256, 256, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
_________________________________________________________________
flatten_3 (Flatten)          (None, 32768)             0
_________________________________________________________________
dense_7 (Dense)              (None, 256)               8388864
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_8 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_9 (Dense)              (None, 16)                2064
=================================================================
Total params: 23,138,512
Trainable params: 23,138,512
Non-trainable params: 0
_________________________________________________________________
```

In [0]:

```
model_header_final.load_weights('/content/drive/My Drive/model_whole.hdf5')
```

In [0]:

```
def height_crop_generator(batches):
    while True:
        batch_X, label = next(batches)
        batch_Xcrops = np.zeros((batch_X.shape[0],256,256,3))
        for i in range(batch_X.shape[0]):
            batch_Xcrops[i] = batch_X[i][:256,:,:]
        yield (batch_Xcrops,label)
```

In [0]:

```
#https://keras.io/preprocessing/image/
train_datagen = ImageDataGenerator(rescale = 1./255)
test_datagen = ImageDataGenerator(rescale = 1./255)
valid_datagen = ImageDataGenerator(rescale = 1./255)
```

In [0]:

```
train_header = train_datagen.flow_from_directory('/content/DocumentImages/train',batch
test_header = test_datagen.flow_from_directory('/content/DocumentImages/test',batch_si
valid_header = valid_datagen.flow_from_directory('/content/DocumentImages/valid', batc
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
tbCallBack = callbacks.TensorBoard(log_dir='./Graph', histogram_freq=0, write_graph=Tr
```

In [0]:

```
history = model_header_final.fit_generator(height_crop_generator(train_header), steps_
```

```
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tensorflow_co
re/python/ops/math_grad.py:1424: where (from tensorflow.python.ops.array_op
s) is deprecated and will be removed in a future version.
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backen
d/tensorflow_backend.py:1033: The name tf.assign_add is deprecated. Please u
se tf.compat.v1.assign_add instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backen
d/tensorflow_backend.py:1020: The name tf.assign is deprecated. Please use t
f.compat.v1.assign instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/callbac
ks.py:1122: The name tf.summary.merge_all is deprecated. Please use tf.compa
t.v1.summary.merge_all instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/callbac
ks.py:1125: The name tf.summary.FileWriter is deprecated. Please use tf.comp
at.v1.summary.FileWriter instead.

Epoch 1/10
1219/1218 [==============================] - 429s 352ms/step - loss: 0.8161
- acc: 0.7608 - val_loss: 0.6550 - val_acc: 0.8090
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/callbac
ks.py:1265: The name tf.Summary is deprecated. Please use tf.compat.v1.Summa
ry instead.

Epoch 2/10
1219/1218 [==============================] - 429s 352ms/step - loss: 0.6835
- acc: 0.8003 - val_loss: 0.5752 - val_acc: 0.8340
Epoch 3/10
1219/1218 [==============================] - 425s 349ms/step - loss: 0.6371
- acc: 0.8145 - val_loss: 0.5774 - val_acc: 0.8320
Epoch 4/10
1219/1218 [==============================] - 426s 349ms/step - loss: 0.6058
- acc: 0.8241 - val_loss: 0.5686 - val_acc: 0.8394
Epoch 5/10
1219/1218 [==============================] - 426s 349ms/step - loss: 0.5900
- acc: 0.8285 - val_loss: 0.4997 - val_acc: 0.8516
Epoch 6/10
1219/1218 [==============================] - 429s 352ms/step - loss: 0.5678
- acc: 0.8367 - val_loss: 0.5144 - val_acc: 0.8509
Epoch 7/10
1219/1218 [==============================] - 429s 352ms/step - loss: 0.5742
- acc: 0.8351 - val_loss: 0.5144 - val_acc: 0.8509
Epoch 8/10
1219/1218 [==============================] - 432s 354ms/step - loss: 0.5109
- acc: 0.8538 - val_loss: 0.4722 - val_acc: 0.8585
Epoch 9/10
1219/1218 [==============================] - 401s 329ms/step - loss: 0.4094
- acc: 0.8789 - val_loss: 0.4655 - val_acc: 0.8681
Epoch 10/10
1219/1218 [==============================] - 416s 341ms/step - loss: 0.4073
- acc: 0.8828 - val_loss: 0.4479 - val_acc: 0.8696
```

In [0]:

```
score = model_header_final.evaluate_generator(height_crop_generator(test_header),steps
```

In [0]:

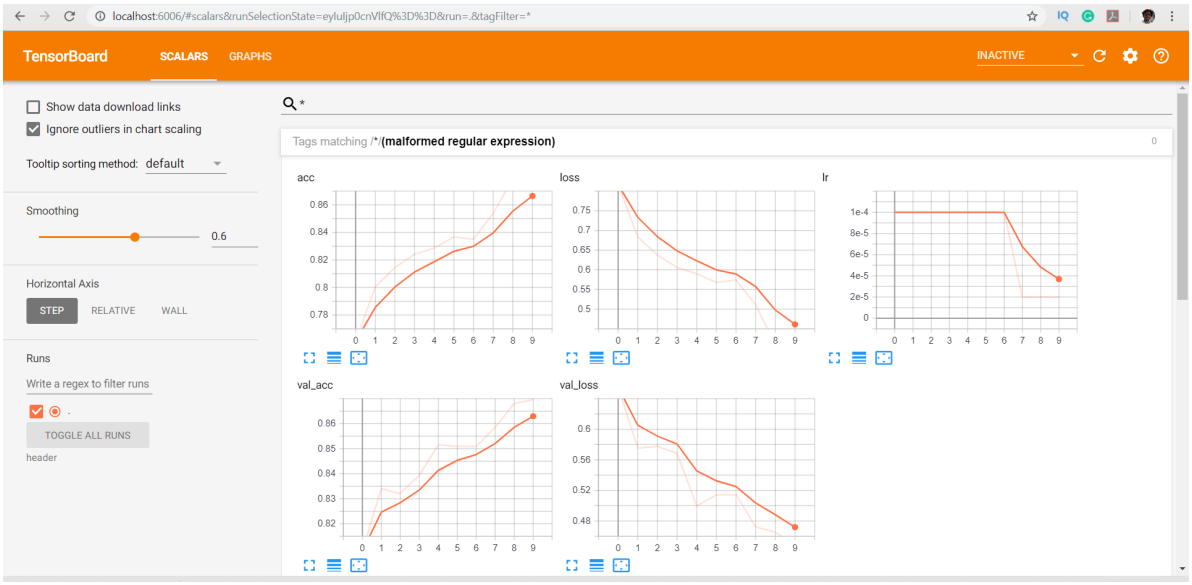
```
print("The Accuracy for the header image model:",score[1]*100,"%")
```

```
The Accuracy for the header image model: 86.84895833333334 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/header_scalar graph.png")
```

Out[2]:



Footer Image

In [0]:

```
model_footer = VGG16(weights = "imagenet", include_top=False, input_shape = (256, 256,
```

In [0]:

```
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=2, min_lr=0.000
mcp_save = ModelCheckpoint('model_footer.hdf5', save_best_only=True, monitor='acc', mo
```

In [0]:

```
for layer in model_footer.layers:
    layer.trainable=True

x = model_footer.output
x = Flatten()(x)
x = Dense(256, activation="relu")(x)
x = Dropout(0.5)(x)
x = Dense(128, activation="relu")(x)
predictions = Dense(16, activation="softmax")(x)

model_footer_final = Model(input = model_footer.input, output = predictions)

model_footer_final.compile(loss = "categorical_crossentropy", optimizer =Adam(lr=0.000
```

In [0]:

```
1 model_footer_final.summary()
```

```
Model: "model_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_5 (InputLayer)         (None, 256, 256, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
_________________________________________________________________
flatten_4 (Flatten)          (None, 32768)             0
_________________________________________________________________
dense_10 (Dense)             (None, 256)               8388864
_________________________________________________________________
dropout_4 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_11 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_12 (Dense)             (None, 16)                2064
=================================================================
Total params: 23,138,512
Trainable params: 23,138,512
Non-trainable params: 0
_________________________________________________________________
```

In [0]:

```
model_footer_final.load_weights('/content/drive/My Drive/model_whole.hdf5')
```

In [0]:

```
#https://github.com/sambalshikhar/Document-Image-Classification
def bottom_crop_generator(batches):
    while True:
        batch_X, label = next(batches)
        batch_Xcrops = np.zeros((batch_X.shape[0],256,256,3))
        for i in range(batch_X.shape[0]):
            batch_Xcrops[i] = batch_X[i][-256:,:,:]
        yield (batch_Xcrops,label)
```

In [0]:

```
#https://keras.io/preprocessing/image/
train_datagen = ImageDataGenerator(rescale = 1./255)
test_datagen = ImageDataGenerator(rescale = 1./255)
valid_datagen = ImageDataGenerator(rescale = 1./255)
```

In [0]:

```
train_footer = train_datagen.flow_from_directory('/content/DocumentImages/train',batch
test_footer = test_datagen.flow_from_directory('/content/DocumentImages/test',batch_si
valid_footer = valid_datagen.flow_from_directory('/content/DocumentImages/valid', batc
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
tbCallBack = callbacks.TensorBoard(log_dir='./Graph', histogram_freq=0, write_graph=Tr
```

In [0]:

```
history = model_footer_final.fit_generator(bottom_crop_generator(train_footer), steps_
```

```
Epoch 1/10
1219/1218 [==============================] - 411s 337ms/step - loss: 0.9550
- acc: 0.7121 - val_loss: 0.7584 - val_acc: 0.7700
Epoch 2/10
1219/1218 [==============================] - 415s 341ms/step - loss: 0.7914
- acc: 0.7604 - val_loss: 0.7036 - val_acc: 0.7815
Epoch 3/10
1219/1218 [==============================] - 426s 349ms/step - loss: 0.7458
- acc: 0.7786 - val_loss: 0.6748 - val_acc: 0.7945
Epoch 4/10
1219/1218 [==============================] - 431s 353ms/step - loss: 0.7155
- acc: 0.7866 - val_loss: 0.6537 - val_acc: 0.7967
Epoch 5/10
1219/1218 [==============================] - 433s 355ms/step - loss: 0.7005
- acc: 0.7923 - val_loss: 0.6340 - val_acc: 0.8049
Epoch 6/10
1219/1218 [==============================] - 430s 353ms/step - loss: 0.6775
- acc: 0.7991 - val_loss: 0.6398 - val_acc: 0.8064
Epoch 7/10
1219/1218 [==============================] - 456s 374ms/step - loss: 0.6669
- acc: 0.7985 - val_loss: 0.6176 - val_acc: 0.8149
Epoch 8/10
1219/1218 [==============================] - 441s 361ms/step - loss: 0.6663
- acc: 0.8004 - val_loss: 0.6018 - val_acc: 0.8147
Epoch 9/10
1219/1218 [==============================] - 410s 336ms/step - loss: 0.5710
- acc: 0.8270 - val_loss: 0.6028 - val_acc: 0.8179
Epoch 10/10
1219/1218 [==============================] - 421s 345ms/step - loss: 0.5884
- acc: 0.8248 - val_loss: 0.5837 - val_acc: 0.8257
```

In [0]:

```
score = model_footer_final.evaluate_generator(bottom_crop_generator(test_footer),steps
```

In [0]:

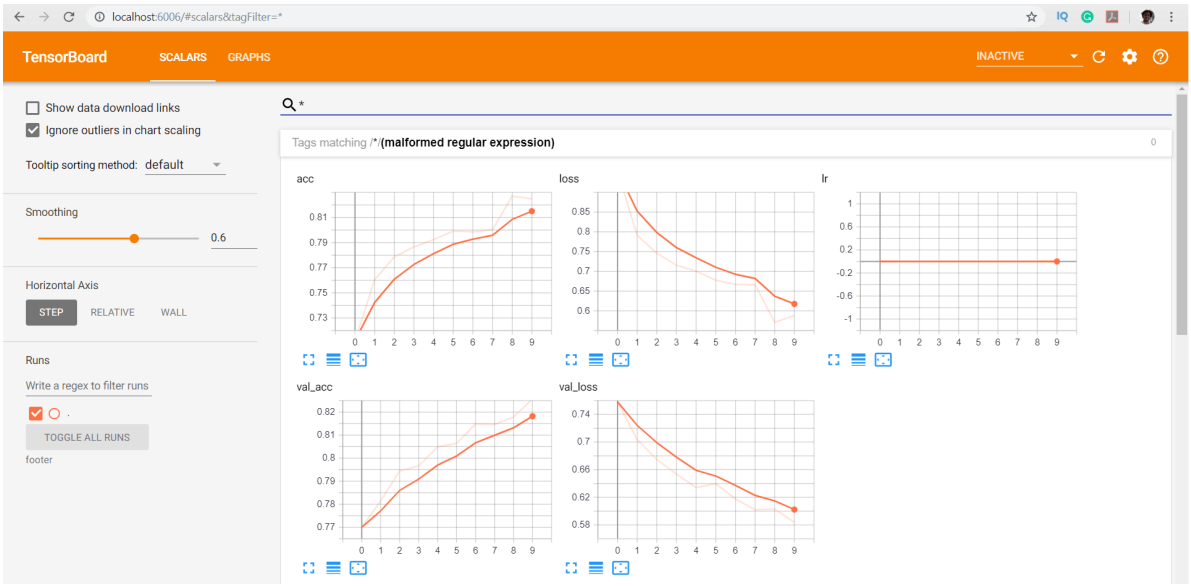
```
print("The Accuracy for the footer image model:",score[1]*100,"%")
```

```
The Accuracy for the footer image model: 82.13975694444444 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/footer_scalar graph.png")
```

Out[3]:



Left Body Image

In [0]:

```
model_left_body = VGG16(weights = "imagenet", include_top=False, input_shape = (256,25
```

In [0]:

```
for layer in model_left_body.layers:
    layer.trainable=True

x = model_left_body.output
x = Flatten()(x)
x = Dense(256, activation="relu")(x)
x = Dropout(0.5)(x)
x = Dense(128, activation="relu")(x)
predictions = Dense(16, activation="softmax")(x)

model_left_body_final = Model(input = model_left_body.input, output = predictions)

model_left_body_final.compile(loss = "categorical_crossentropy", optimizer =Adam(lr=0.(
```

In [0]:

```
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=2, min_lr=0.000(
mcp_save = ModelCheckpoint('model_left_body.hdf5', save_best_only=True, monitor='acc',
```

In [0]:

```
model_left_body_final.load_weights('/content/drive/My Drive/model_whole.hdf5')
```

In [0]:

```
def left_crop_generator(batches):
    while True:
        batch_X, label = next(batches)
        batch_Xcrops = np.zeros((batch_X.shape[0],256,256,3))
        for i in range(batch_X.shape[0]):
            batch_Xcrops[i] = batch_X[i][:,:256,:]
        yield (batch_Xcrops,label)
```

In [0]:

```
#https://keras.io/preprocessing/image/
train_datagen = ImageDataGenerator(rescale = 1./255)
test_datagen = ImageDataGenerator(rescale = 1./255)
valid_datagen = ImageDataGenerator(rescale = 1./255)
```

In [0]:

```
train_left_body = train_datagen.flow_from_directory('/content/DocumentImages/train',ba
test_left_body = test_datagen.flow_from_directory('/content/DocumentImages/test',batch_
valid_left_body = valid_datagen.flow_from_directory('/content/DocumentImages/valid', b
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
history = model_left_body_final.fit_generator(left_crop_generator(train_left_body), st
```

```
Epoch 1/10
1219/1218 [==============================] - 415s 340ms/step - loss: 0.8189
- acc: 0.7586 - val_loss: 0.6619 - val_acc: 0.8079
Epoch 2/10
1219/1218 [==============================] - 426s 349ms/step - loss: 0.6667
- acc: 0.8056 - val_loss: 0.6199 - val_acc: 0.8199
Epoch 3/10
1219/1218 [==============================] - 433s 355ms/step - loss: 0.6299
- acc: 0.8176 - val_loss: 0.5821 - val_acc: 0.8325
Epoch 4/10
1219/1218 [==============================] - 442s 363ms/step - loss: 0.6319
- acc: 0.8170 - val_loss: 0.5620 - val_acc: 0.8364
Epoch 5/10
1219/1218 [==============================] - 437s 358ms/step - loss: 0.5877
- acc: 0.8298 - val_loss: 0.5035 - val_acc: 0.8533
Epoch 6/10
1219/1218 [==============================] - 436s 358ms/step - loss: 0.5749
- acc: 0.8352 - val_loss: 0.5447 - val_acc: 0.8392
Epoch 7/10
1219/1218 [==============================] - 435s 357ms/step - loss: 0.5597
- acc: 0.8381 - val_loss: 0.5221 - val_acc: 0.8424
Epoch 8/10
1219/1218 [==============================] - 433s 356ms/step - loss: 0.5194
- acc: 0.8508 - val_loss: 0.4671 - val_acc: 0.8637
Epoch 9/10
1219/1218 [==============================] - 405s 332ms/step - loss: 0.4219
- acc: 0.8777 - val_loss: 0.4624 - val_acc: 0.8622
Epoch 10/10
1219/1218 [==============================] - 418s 343ms/step - loss: 0.4082
- acc: 0.8801 - val_loss: 0.4508 - val_acc: 0.8715
```

In [0]:

```
score = model_left_body_final.evaluate_generator(left_crop_generator(test_left_body),s
```

In [0]:

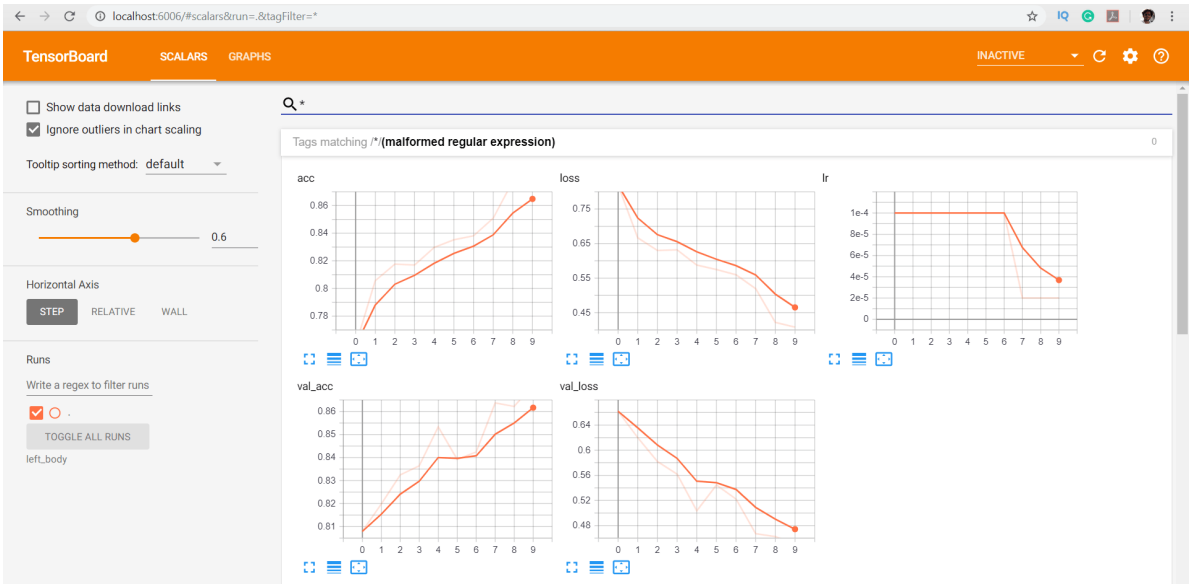
```
print("The Accuracy for the left body image model:",score[1]*100,"%")
```

```
The Accuracy for the left body image model: 87.00086805555556 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/left_body_scalar graph.png")
```

Out[4]:



Right Body Image

In [0]:

```
model_right_body = VGG16(weights = "imagenet", include_top=False, input_shape = (256,2
```

In [0]:

```
for layer in model_right_body.layers:
    layer.trainable=True

x = model_right_body.output
x = Flatten()(x)
x = Dense(256, activation="relu")(x)
x = Dropout(0.5)(x)
x = Dense(128, activation="relu")(x)
predictions = Dense(16, activation="softmax")(x)

model_right_body_final = Model(input = model_right_body.input, output = predictions)

model_right_body_final.compile(loss = "categorical_crossentropy", optimizer =Adam(lr=0
```

In [0]:

```
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=2, min_lr=0.000
mcp_save = ModelCheckpoint('model_right_body.hdf5', save_best_only=True, monitor='acc'
```

In [0]:

```
model_right_body_final.load_weights('/content/drive/My Drive/model_whole.hdf5')
```

In [0]:

```
1 model_right_body_final.summary()
```

```
Model: "model_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_7 (InputLayer)         (None, 256, 256, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
_________________________________________________________________
flatten_6 (Flatten)          (None, 32768)             0
_________________________________________________________________
dense_16 (Dense)             (None, 256)               8388864
_________________________________________________________________
dropout_6 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_17 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_18 (Dense)             (None, 16)                2064
=================================================================
Total params: 23,138,512
Trainable params: 23,138,512
Non-trainable params: 0
_________________________________________________________________
```

In [0]:

```
def right_crop_generator(batches):
    while True:
        batch_X, label = next(batches)
        batch_Xcrops = np.zeros((batch_X.shape[0],256,256,3))
        for i in range(batch_X.shape[0]):
            batch_Xcrops[i] = batch_X[i][:,-256:,:]
        yield (batch_Xcrops,label)
```

In [0]:

```
#https://keras.io/preprocessing/image/
train_datagen = ImageDataGenerator(rescale = 1./255)
test_datagen = ImageDataGenerator(rescale = 1./255)
valid_datagen = ImageDataGenerator(rescale = 1./255)
```

In [0]:

```
train_right_body = train_datagen.flow_from_directory('/content/DocumentImages/train',b
test_right_body = test_datagen.flow_from_directory('/content/DocumentImages/test',batc
valid_right_body = valid_datagen.flow_from_directory('/content/DocumentImages/valid', 
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
history = model_right_body_final.fit_generator(right_crop_generator(train_right_body),
```

```
Epoch 1/10
1219/1218 [==============================] - 414s 340ms/step - loss: 0.8891
- acc: 0.7305 - val_loss: 0.6971 - val_acc: 0.7906
Epoch 2/10
1219/1218 [==============================] - 424s 348ms/step - loss: 0.7244
- acc: 0.7807 - val_loss: 0.6168 - val_acc: 0.8155
Epoch 3/10
1219/1218 [==============================] - 429s 352ms/step - loss: 0.6769
- acc: 0.7985 - val_loss: 0.5938 - val_acc: 0.8194
Epoch 4/10
1219/1218 [==============================] - 433s 355ms/step - loss: 0.6480
- acc: 0.8053 - val_loss: 0.5995 - val_acc: 0.8160
Epoch 5/10
1219/1218 [==============================] - 431s 353ms/step - loss: 0.6317
- acc: 0.8132 - val_loss: 0.5513 - val_acc: 0.8331
Epoch 6/10
1219/1218 [==============================] - 436s 358ms/step - loss: 0.6165
- acc: 0.8156 - val_loss: 0.5531 - val_acc: 0.8314
Epoch 7/10
1219/1218 [==============================] - 451s 370ms/step - loss: 0.5851
- acc: 0.8232 - val_loss: 0.5451 - val_acc: 0.8381
Epoch 8/10
1219/1218 [==============================] - 446s 366ms/step - loss: 0.5767
- acc: 0.8293 - val_loss: 0.5051 - val_acc: 0.8442
Epoch 9/10
1219/1218 [==============================] - 410s 336ms/step - loss: 0.5038
- acc: 0.8495 - val_loss: 0.4892 - val_acc: 0.8507
Epoch 10/10
1219/1218 [==============================] - 428s 351ms/step - loss: 0.5105
- acc: 0.8467 - val_loss: 0.5012 - val_acc: 0.8468
```

In [0]:

```
score = model_right_body_final.evaluate_generator(right_crop_generator(test_right_body
```

In [0]:

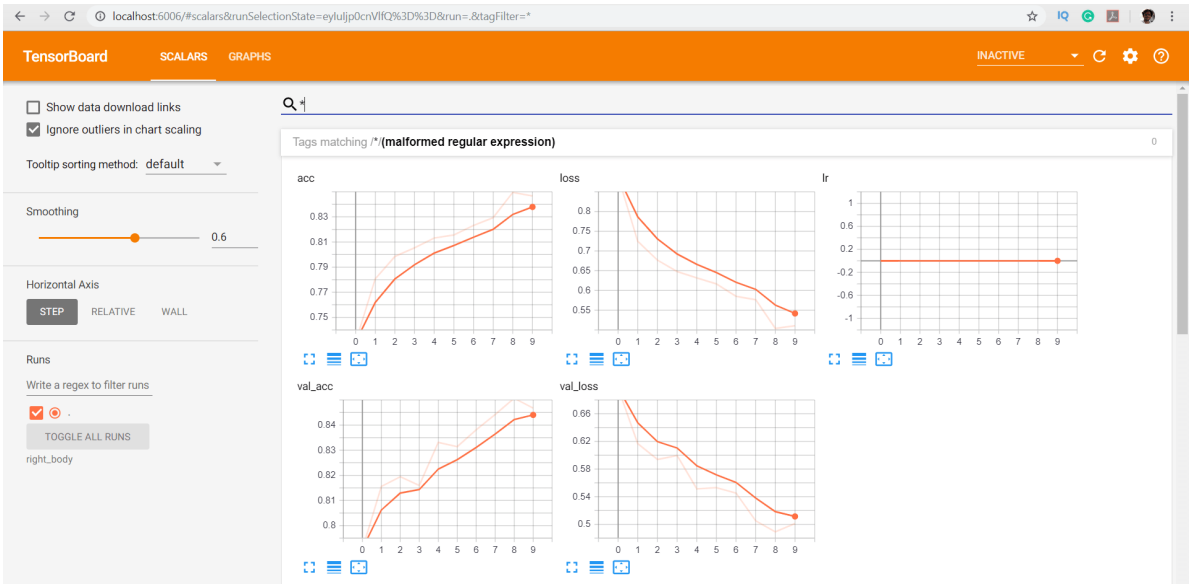
```
print("The Accuracy for the right body image model:",score[1]*100,"%")
```

```
The Accuracy for the right body image model: 85.17795138888889 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/right_body_scalar graph.png")
```

Out[5]:



Loading, Dumping and Stacking the Data

In [0]:

```
model_header_final.load_weights('/content/drive/My Drive/model_header.hdf5')
model_footer_final.load_weights('/content/drive/My Drive/model_footer.hdf5')
model_left_body_final.load_weights('/content/drive/My Drive/model_left_body.hdf5')
model_right_body_final.load_weights('/content/drive/My Drive/model_right_body.hdf5')
```

In [0]:

```
train_header_footer = train_datagen.flow_from_directory('/content/DocumentImages/train
test_header_footer = test_datagen.flow_from_directory('/content/DocumentImages/test',ba
valid_header_footer = valid_datagen.flow_from_directory('/content/DocumentImages/valid
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
train_header_footer.reset()
test_header_footer.reset()
valid_header_footer.reset()
```

In [0]:

```
feature_top = model_header_final.predict_generator(height_crop_generator(valid_header_
```

In [0]:

```
train_header_footer.reset()
test_header_footer.reset()
valid_header_footer.reset()
```

In [0]:

```
feature_bottom = model_footer_final.predict_generator(bottom_crop_generator(valid_head
```

In [0]:

```
train_header_footer.reset()
test_header_footer.reset()
valid_header_footer.reset()
```

In [0]:

```
train_left_right.reset()
test_left_right.reset()
valid_left_right.reset()
```

In [0]:

```
train_left_right = train_datagen.flow_from_directory('/content/DocumentImages/train',b
test_left_right = test_datagen.flow_from_directory('/content/DocumentImages/test',batc
valid_left_right = valid_datagen.flow_from_directory('/content/DocumentImages/valid',
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
train_left_right.reset()
test_left_right.reset()
valid_left_right.reset()
```

In [0]:

```
feature_left = model_left_body_final.predict_generator(left_crop_generator(valid_left_
```

In [0]:

```
train_left_right.reset()
test_left_right.reset()
valid_left_right.reset()
```

In [0]:

```
feature_right = model_right_body_final.predict_generator(right_crop_generator(valid_le
```

In [0]:

```
train_left_right.reset()
test_left_right.reset()
valid_left_right.reset()
```

In [0]:

```
joblib.dump(feature_top,'top')
joblib.dump(feature_bottom,'bottom')
joblib.dump(feature_left,'left_body')
joblib.dump(feature_right,'right_body')
```

Out[89]:

```
['right_body']
```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
train_total = train_datagen.flow_from_directory('/content/DocumentImages/train',batch_
test_total = test_datagen.flow_from_directory('/content/DocumentImages/test',batch_siz
valid_total = valid_datagen.flow_from_directory('/content/DocumentImages/valid', batch_
```

```
Found 312000 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
Found 36800 images belonging to 16 classes.
```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
model_whole_final.load_weights('/content/drive/My Drive/model_whole.hdf5')
```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
holistic = model_whole_final.predict_generator(valid_total,steps = 36800)
```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
joblib.dump(holistic,'holistic')
```

Out[98]:

```
['holistic']
```

In [0]:

```
top = joblib.load('/content/top')
bottom = joblib.load('/content/bottom')
left_body = joblib.load('/content/left_body')
right_body = joblib.load('/content/right_body')
holistic = joblib.load('/content/holistic')
```

In [0]:

```
total_features_valid = np.hstack((top, bottom, left_body, right_body, holistic))
```

In [0]:

```
total_features_valid.shape
```

Out[101]:

```
(36800, 80)
```

In [0]:

```

```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
holistic_test = model_whole_final.predict_generator(test_total,steps = 36800)
```

In [0]:

```
train_total.reset()
test_total.reset()
valid_total.reset()
```

In [0]:

```
test_header_footer.reset()
```

In [0]:

```
top_test = model_header_final.predict_generator(height_crop_generator(test_header_foot
```

In [0]:

```
test_header_footer.reset()
```

In [0]:

```
bottom_test = model_footer_final.predict_generator(bottom_crop_generator(test_header_f
```

In [0]:

```
test_header_footer.reset()
```

In [0]:

```
test_left_right.reset()
```

In [0]:

```
left_test = model_left_body_final.predict_generator(left_crop_generator(test_left_righ
```

In [0]:

```
test_left_right.reset()
```

In [0]:

```
right_test = model_right_body_final.predict_generator(right_crop_generator(test_left_r
```

In [0]:

```
test_left_right.reset()
```

In [0]:

```
joblib.dump(holistic_test,'holistic_test')
joblib.dump(top_test,'top_test')
joblib.dump(bottom_test,'bottom_test')
joblib.dump(left_test,'left_body_test')
joblib.dump(right_test,'right_body_test')
```

Out[115]:

```
['right_body_test']
```

In [0]:

```
header_test=joblib.load('/content/top_test')
footer_test=joblib.load('/content/bottom_test')
left_body_test=joblib.load('/content/left_body_test')
right_body_test=joblib.load('/content/right_body_test')
holistic_test=joblib.load('/content/holistic_test')
```

In [0]:

```
total_features_test=np.hstack((header_test, footer_test, left_body_test, right_body_te
```

In [0]:

```
total_features_test.shape
```

Out[118]:

```
(36800, 80)
```

In [0]:

```

```

In [0]:

```
train_total.reset()
```

In [0]:

```
holistic_train = model_whole_final.predict_generator(train_total,steps = 312000)
joblib.dump(holistic_train,'holistic_train')
```

Out[120]:

```
['holistic_train']
```

In [0]:

```
train_total.reset()
train_header_footer.reset()
```

In [0]:

```
top_train = model_header_final.predict_generator(height_crop_generator(train_header_fo
joblib.dump(top_train,'top_train')
```

Out[122]:

```
['top_train']
```

In [0]:

```
train_header_footer.reset()
```

In [0]:

```
bottom_train = model_footer_final.predict_generator(bottom_crop_generator(train_header_
joblib.dump(bottom_train,'bottom_train')
```

Out[124]:

```
['bottom_train']
```

In [0]:

```
train_header_footer.reset()
train_left_right.reset()
```

In [0]:

```
left_body_train = model_left_body_final.predict_generator(left_crop_generator(train_le
joblib.dump(left_body_train,'left_train')
```

Out[126]:

```
['left_train']
```

In [0]:

```
train_left_right.reset()
```

In [0]:

```
right_body_train = model_right_body_final.predict_generator(right_crop_generator(train_
joblib.dump(right_body_train,'right_train')
```

Out[130]:

```
['right_train']
```

In [0]:

```
train_left_right.reset()
```

In [0]:

```
holistic_train = model_whole_final.predict_generator(train_total,steps = 312000)
top_train = model_header_final.predict_generator(height_crop_generator(train_header_fo
bottom_train = model_footer_final.predict_generator(bottom_crop_generator(train_header_
left_body_train = model_left_body_final.predict_generator(left_crop_generator(train_le
right_body_train = model_right_body_final.predict_generator(right_crop_generator(train_
```

In [0]:

```
joblib.dump(holistic_train,'holistic_train')
joblib.dump(top_train,'top_train')
joblib.dump(bottom_train,'bottom_train')
joblib.dump(left_body_train,'left_train')
joblib.dump(right_body_train,'right_train')
```

In [0]:

```
header_train = joblib.load('/content/top_train')
footer_train = joblib.load('/content/bottom_train')
left_body_train = joblib.load('/content/left_train')
right_body_train = joblib.load('/content/right_train')
holistic_train = joblib.load('/content/holistic_train')
```

In [0]:

```
total_features_train = np.hstack((header_train, footer_train, left_body_train, right_b
```

In [0]:

```
total_features_train.shape
```

Out[134]:

```
(312000, 80)
```

In [0]:

```
train_labels=[]
for i in range(312000):
   batch=next(train_total)
   train_labels.append(batch[1])
```

In [0]:

```
valid_labels=[]
for i in range(36800):
  batch=next(valid_total)
  valid_labels.append(batch[1])
```

In [0]:

```
test_labels=[]
for i in range(36800):
   batch=next(test_total)
   test_labels.append(batch[1])
```

In [0]:

```
joblib.dump(train_labels,'train_labels')
joblib.dump(valid_labels,'valid_labels')
joblib.dump(test_labels,'test_labels')
```

Out[138]:

```
['test_labels']
```

In [0]:

```
train_labels = joblib.load('/content/train_labels')
valid_labels = joblib.load('/content/valid_labels')
test_labels = joblib.load('/content/test_labels')
```

Final Model

In [0]:

```
final_model = Sequential()
final_model.add(Dense(512, activation='relu', input_shape=(total_features_valid.shape[
final_model.add(Dropout(0.50))
final_model.add(BatchNormalization())
final_model.add(Dense(256, activation='relu'))
final_model.add(Dropout(0.50))
final_model.add(Dense(16, activation='softmax'))
```

In [0]:

```
tbCallBack = callbacks.TensorBoard(log_dir='./Graph', histogram_freq=0, write_graph=Tr
```

In [0]:

```
final_model.compile(optimizer=Adam(lr=0.0001), loss='categorical_crossentropy', metric
history = final_model.fit(total_features_train, train_labels , validation_data= (total
```

```
Train on 312000 samples, validate on 36800 samples
Epoch 1/10
312000/312000 [==============================] - 26s 84us/step - loss: 0.373
6 - acc: 0.9092 - val_loss: 0.3621 - val_acc: 0.9053
Epoch 2/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.301
5 - acc: 0.9228 - val_loss: 0.3471 - val_acc: 0.9056
Epoch 3/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.288
8 - acc: 0.9237 - val_loss: 0.3388 - val_acc: 0.9061
Epoch 4/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.281
2 - acc: 0.9244 - val_loss: 0.3355 - val_acc: 0.9060
Epoch 5/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.276
2 - acc: 0.9249 - val_loss: 0.3321 - val_acc: 0.9070
Epoch 6/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.272
2 - acc: 0.9251 - val_loss: 0.3323 - val_acc: 0.9066
Epoch 7/10
312000/312000 [==============================] - 26s 82us/step - loss: 0.269
7 - acc: 0.9256 - val_loss: 0.3310 - val_acc: 0.9064
Epoch 8/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.267
4 - acc: 0.9258 - val_loss: 0.3313 - val_acc: 0.9068
Epoch 9/10
312000/312000 [==============================] - 25s 80us/step - loss: 0.265
4 - acc: 0.9262 - val_loss: 0.3293 - val_acc: 0.9067
Epoch 10/10
312000/312000 [==============================] - 25s 81us/step - loss: 0.263
5 - acc: 0.9263 - val_loss: 0.3290 - val_acc: 0.9071
```

In [0]:

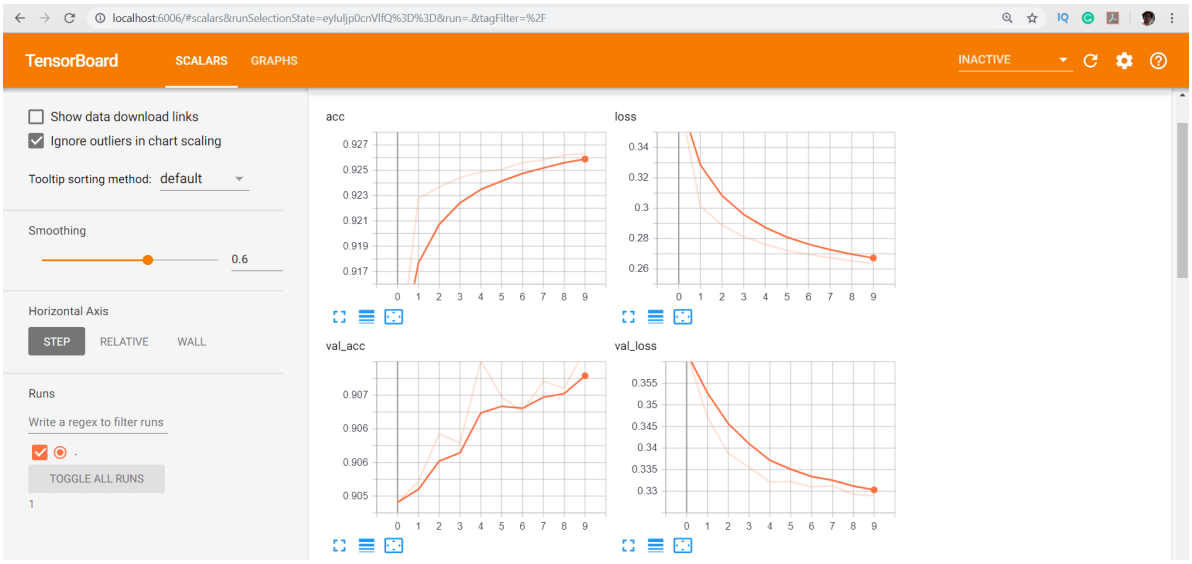
```
score_test = final_model.evaluate(total_features_test, test_labels, batch_size=64)
print("Accuracy on the test data:",score_test[1]*100,'%')
```

```
36800/36800 [==============================] - 1s 23us/step
Accuracy on the test data: 90.6288043478261 %
```

In [0]:

```
from IPython.display import Image
Image("tensorboard/final_scalar_graph.png")
```

Out[6]:



Testing the images

In [0]:

```
classes = ("advertisement", "budget", "email", "file folder", "form", "handwritten", "
```

In [0]:

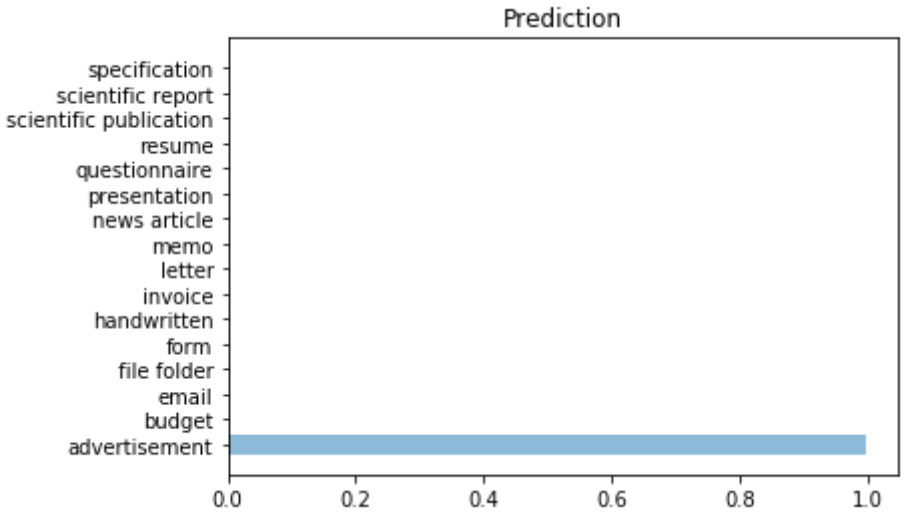
```
cmap = cm.get_cmap('gray')

def height_crop(path,type):
    image = cv2.imread(path)
    image = cv2.resize(image, (512, 256))
    if type =='bottom':
        image = image[-256:,:,:]
    else:
        image = image[:256,:,:]
    return image

def width_crop(path,type):
    image = cv2.imread(path)
    image = cv2.resize(image, (256, 512))
    if type == 'right':
        image = image[:,:256,:]
    else:
        image = image[:,-256:,:]
    return image

def full_image(path):
    image = cv2.imread(path)
    image = cv2.resize(image, (256, 256))
    return image

def preprocess(image):
    image = image/255
    image = np.expand_dims(image, axis=0)
    return image
```

In [0]:

```
img_path='0000176244.tif' #enter the image url/path you want to test
img = image.load_img(img_path, target_size = (256, 256))
x = image.img_to_array(img)
x = np.expand_dims(img,axis=0)
x = preprocess_input(x)

top_pred=model_header.predict(x)
bottom_pred=model_footer.predict(x)
left_pred=model_left_body.predict(x)
right_pred=model_right_body.predict(x)
holistic_pred=model_whole.predict(x)
total_features=np.hstack((top_pred,bottom_pred,left_pred,right_pred,holistic_pred))
prediction=final_model.predict(total_features)
```

In [0]:

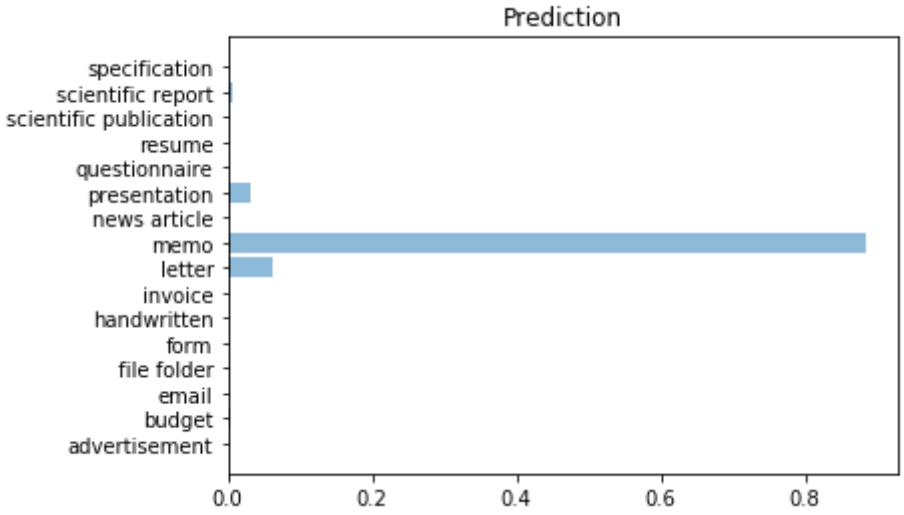
```
#For plotting the probabilities

y_pos = np.arange(len(classes))
score = prediction
score_up = np.ravel(score)
plt.barh(y_pos, score_up, align='center', alpha=0.5)
plt.yticks(y_pos, classes)
plt.title('Prediction')
plt.show()
plt.imshow(img)
plt.show()
print('Predicted:', score)
```

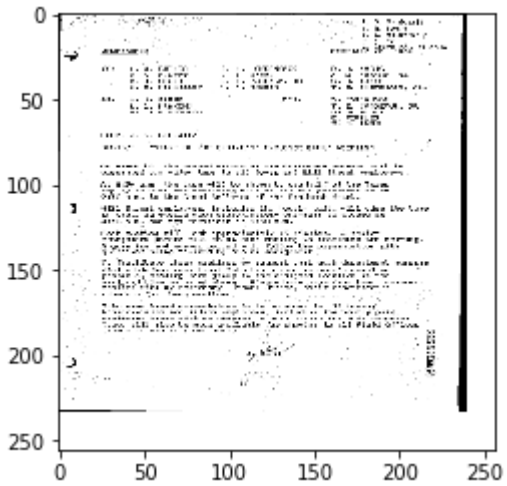


```
Predicted: [[9.9907327e-01 1.9776717e-05 2.0249520e-06 1.2867279e-04 6.79353
18e-05
  6.3766885e-05 6.3295051e-06 1.6552991e-05 1.9140631e-05 2.6307214e-04
  4.5207056e-05 1.6421723e-04 9.8421024e-06 5.9455353e-05 4.5557317e-05
  1.5104232e-05]]
```

In [0]:

```
#For plotting the probabilities

y_pos = np.arange(len(classes))
score = prediction
score_up = np.ravel(score)
plt.barh(y_pos, score_up, align='center', alpha=0.5)
plt.yticks(y_pos, classes)
plt.title('Prediction')
plt.show()
plt.imshow(img)
plt.show()
print('Predicted:', score)
```

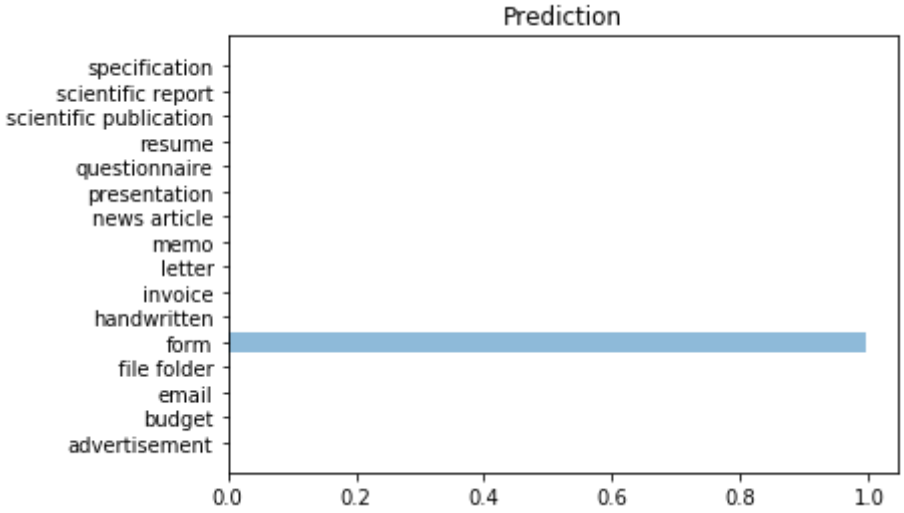


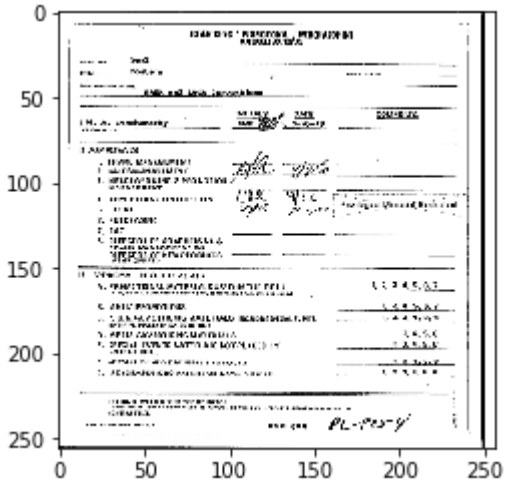



```
Predicted: [[8.1521028e-04 4.9587404e-03 1.5371690e-04 1.2334320e-03 1.70644
42e-03
  6.7390800e-05 5.4901419e-04 6.1257679e-02 8.8570374e-01 1.8910767e-03
  3.0695038e-02 2.3587374e-03 1.0129750e-03 7.9793608e-05 7.0529990e-03
  4.6394055e-04]]
```

In [0]:

```
#For plotting the probabilities

y_pos = np.arange(len(classes))
score = final_model.predict(total_features)
score_up = np.ravel(score)
plt.barh(y_pos, score_up, align='center', alpha=0.5)
plt.yticks(y_pos, classes)
plt.title('Prediction')
plt.show()
plt.imshow(img)
plt.show()
print('Predicted:', score)
```

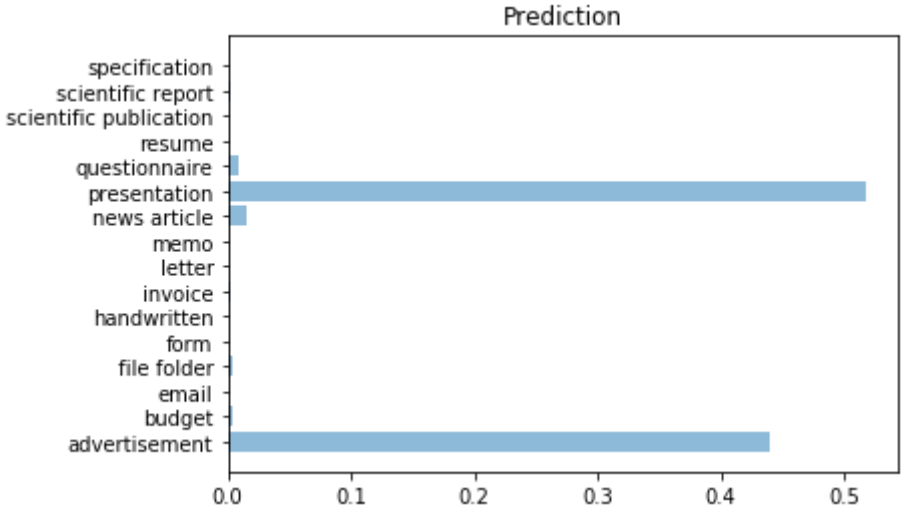


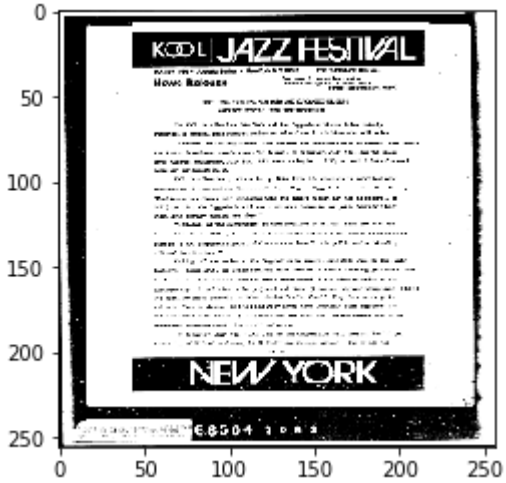


```
Predicted: [[5.2086009e-05 6.5495180e-05 3.0655590e-06 1.6898646e-05 9.97131
11e-01
  4.6097877e-05 6.6983677e-04 2.4601378e-04 3.1958561e-04 6.7149808e-06
  1.5328365e-05 2.6096855e-04 9.3445287e-06 2.4760773e-06 7.4978423e-04
  4.0505306e-04]]
```

In [0]:

```
#For plotting the probabilities

y_pos = np.arange(len(classes))
score = final_model.predict(total_features)
score_up = np.ravel(score)
plt.barh(y_pos, score_up, align='center', alpha=0.5)
plt.yticks(y_pos, classes)
plt.title('Prediction')
plt.show()
plt.imshow(img)
plt.show()
print('Predicted:', score)
```

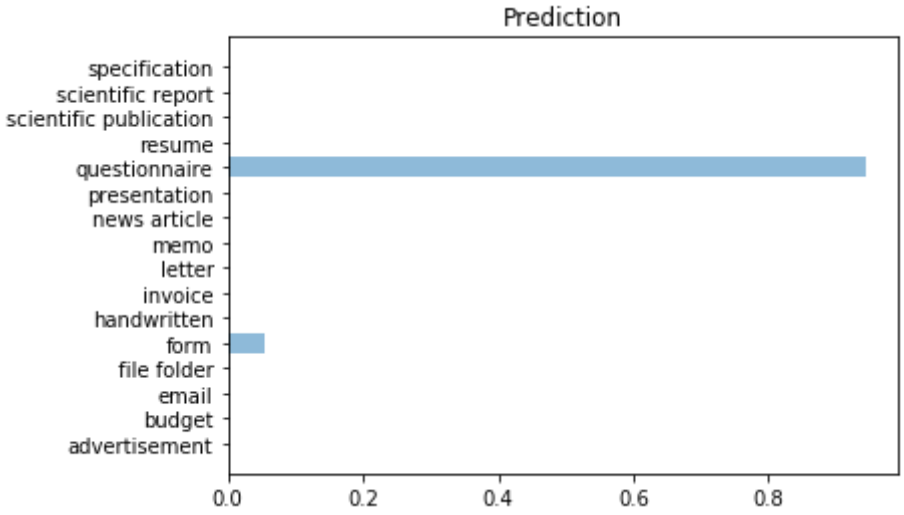


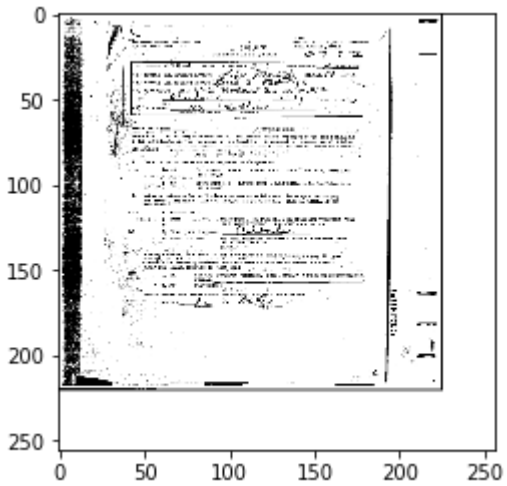



```
Predicted: [[4.3897784e-01 3.1364327e-03 9.6771670e-05 3.4926641e-03 1.35705
76e-03
  1.0697749e-03 3.0701491e-03 9.6566317e-04 4.5748119e-04 1.4858071e-02
  5.1819396e-01 9.0237353e-03 1.4644145e-03 1.0137364e-03 2.1998514e-03
  6.2231440e-04]]
```

In [0]:

```
#For plotting the probabilities

y_pos = np.arange(len(classes))
score = final_model.predict(total_features)
score_up = np.ravel(score)
plt.barh(y_pos, score_up, align='center', alpha=0.5)
plt.yticks(y_pos, classes)
plt.title('Prediction')
plt.show()
plt.imshow(img)
plt.show()
print('Predicted:', score)
```






```
Predicted: [[1.00699166e-04 3.87446817e-05 1.77937329e-06 4.70001942e-05
  5.39738387e-02 3.34770826e-04 1.73669850e-05 4.02496953e-05
  2.99407566e-06 1.95080884e-06 9.31568211e-05 9.45190370e-01
  1.58458431e-06 5.17170565e-06 6.31634612e-05 8.70753356e-05]]
```

Conclusion

1.In this Case Study we have to solve the Document Image Classification challenge. Here we took the RVL-CDIP Dataset.

2.Because the dataset is very big and needs GPU to computer so we use Google Colab for the free GPU. So, we download the data via curl and mounted the Colab Notebook with Google Drive.

We download the data via curl and mounted the Colab Notebook with Google Drive.

3.After collecting the data we need to sort it because unlinke the other datasets RVL-CDIP dose not have different data from different folders. All the classes and the dats are jumble up. So we have to correctly separate the 16 classes.

4.The the next step is to read the data from the folders with help of labels (.txt files)

5.Now we will take the data and sort then inro Train, Test and Validation. The distribution is 312000 training images, 36800 testing images and 38600 validation images.

6.Then we will perfrom the Exploratory Data Analysis like Each section of data print, Quantilies by class, Image width distribution, Image mean pixel value, Image contrast etc.

7.After that we will be Creating images at half size of the previous for visual examination like (1000, 762), (500, 381), (250, 190), (125, 95), (62, 47)

8.Then we will print the Histogram Equalization.

9.The next task is to break the Images into different regions like Whole, Footer, Header, Right Body, Left Body for the optimum accuracy.

10.For every region there will be train validation and test sets.

11.Then we will perform VGG16 Netork on every region separately. After doing this we will record the accuracies.

12.After that we will stack the features and save the models. One of the most important step is to reset the generators otherwise the result will be incorrect.

13.Then we will calculate train labels, test labels and validation labels respectively.

14.Finally we will calculate the final training, testing and validation accuracies for the final model.

15.Then we will check the model accuracy by taking a testing image and our model should be able to tell the class in which the image should belong. If we can classify the images correctly then we can say that our model is working properly.