PII Detection – by Nirav Painter and Harsh Shah

To run this code, you'll need to install libraries listed below, and set up an input folder/output folder. Change the path to input folder and output folder wherever mentioned in the code.

```
presidio_analyzer
presidio anonymizer
from presidio anonymizer.entities import OperatorConfig
glob
docx2txt
textract
from pathlib import Path
import stat
import time
import datetime
easyocr
from PIL import Image, ImageOps, ImageEnhance
from csv import reader
from abc import abstractmethod
from typing import List, Dict
from presidio_analyzer import RecognizerResult
from presidio analyzer.nlp engine import NlpArtifacts
```

## Input path:

## Path to fetch files

```
#Change this path to your folder containing files to detect PII from.
path = r"/Users/nirav/Documents/SPU/DS-630 Machine Learning/ProjectInputFiles/"
```

## Output Path:

## **Output Path**

```
final_op = (csv_info) + (numbers_results_csv + ["",""]) + (txt_info) + (numbers_results_txt + ["",""]) + (primal_op)

df = pd.DataFrame(final_op)

#Change the path and locate your ouput folder, this is where the CSV containing detected PII will export.

df.to_csv('/Users/nirav/Documents/SPU/DS-630 Machine Learning/ProjectOutputFiles/pii_output.csv')
```