# The Project

1. This is a project with minimal scaffolding. Expect to use the the discussion forums to gain insights! It's not cheating to ask others for opinions or perspectives!
2. Be inquisitive, try out new things.
3. Use the previous modules for insights into how to complete the functions! You'll have to combine Pillow, OpenCV, and Pytesseract
4. There are hints provided in Coursera, feel free to explore the hints if needed. Each hint provide progressively more details on how to solve the issue. This project is intended to be comprehensive and difficult if you do it without the hints.

## The Assignment

Take a ZIP file (https://en.wikipedia.org/wiki/Zip_(file_format)) of images and process them, using a library built into python (https://docs.python.org/3/library/zipfile.html) that you need to learn how to use. A ZIP file takes several different files and compresses them, thus saving space, into one single file. The files in the ZIP file we provide are newspaper images (like you saw in week 3). Your task is to write python code which allows one to search through the images looking for the occurrences of keywords and faces. E.g. if you search for "pizza" it will return a contact sheet of all of the faces which were located on the newspaper page which mentions "pizza". This will test your ability to learn a new (library (https://docs.python.org/3/library/zipfile.html)), your ability to use OpenCV to detect faces, your ability to use tesseract to do optical character recognition, and your ability to use PIL to composite images together into contact sheets.

Each page of the newspapers is saved as a single PNG image in a file called images.zip (./readonly/images.zip). These newspapers are in english, and contain a variety of stories, advertisements and images. Note: This file is fairly large (~200 MB) and may take some time to work with, I would encourage you to use small_img.zip (./readonly/small_img.zip) for testing.

Here's an example of the output expected. Using the small_img.zip (./readonly/small_img.zip) file, if I search for the string "Christopher" I should see the following image:

```
Results found in file a-0.png
```

```
Results found in file a-3.png
```

If I were to use the images.zip (./readonly/images.zip) file and search for "Mark" I should see the following image (note that there are times when there are no faces on a page, but a word is found!):

```
Results found in file a-0.png
```

```
Results found in file a-1.png
```

```
Results found in file a-10.png
```

```
But there were no faces in that file!
Results found in file a-13.png
```

```
Results found in file a-2.png
```

```
Results found in file a-3.png
```

```
Results found in file a-8.png
But there were no faces in that file!
```

Note: That big file can take some time to process - for me it took nearly ten minutes! Use the small one for testing.

In [1]:
```
import zipfile

from PIL import Image
import pytesseract
import cv2 as cv
import numpy as np


face_cascade = cv.CascadeClassifier('readonly/haarcascade_frontalface_
```

In [2]:
```
def extract_news_info(zip_path):
    news_info = {}
    with zipfile.ZipFile(zip_path) as zip_file:
        fnames = zip_file.namelist()
        for name in fnames:
            with zip_file.open(name) as file:

                cv_image = cv.imdecode(np.frombuffer(file.read(), np.u

                gray = cv.cvtColor(cv_image, cv.COLOR_BGR2GRAY)
                faces = face_cascade.detectMultiScale(gray, 1.3, 5)


                pil_image = Image.open(file)

                text = pytesseract.image_to_string(pil_image)


                news_info[name] = [pil_image, text, faces]

    return news_info
```

In [3]:
```
def search_images(name, info):
    keys = list(info.keys())
    for key in keys:
        if name in info[key][1]:
            image = info[key][0]
            faces = info[key][2]


            if len(faces) > 0:

                contact_sheet_w = 100
                contact_sheet_h = 100
                rows = int(np.ceil(len(faces) / 5))
                contact_sheet = PIL.Image.new(image.mode, (contact_she



                r = 0
                c = 0
                for x, y, w, h in faces:
                    crop = image.crop((x, y, x+w, y+h))
                    if crop.width > contact_sheet_w:
                        crop = crop.resize((100, 100))
                        contact_sheet.paste(crop, (r, c))

                    else:
                        contact_sheet.paste(crop, (r, c))

                    if crop.width + r == contact_sheet.width:
                        r = 0
                        c += crop.height

                    else:
                        r += contact_sheet_w


                contact_sheet = contact_sheet.resize((int(contact_shee

                print("Results found in file {}".format(key))
                display(contact_sheet)


            else:
                print("Results found in file {}".format(key))
                print("But there were no faces in that file!")
```

In [*]:
```
small_test = extract_news_info("readonly/small_img.zip")
```

In [ ]:
```
big_test = extract_news_info("readonly/images.zip")
```

In [ ]:
```
search_images("Christopher", small_test)
```

In [ ]:
```
search_images("Mark", big_test)
```

In [ ]:
```

```

In [ ]:
```

```