Optical Character Recognition from Image using PyTesseract in Python on Colab

In [ ]:

```
#Libraries Installation
!sudo apt install tesseract-ocr
!pip install pytesseract
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
tesseract-ocr is already the newest version (4.00~git2288-10f4998a-2).
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
Requirement already satisfied: pytesseract in /usr/local/lib/python3.7/dist-packages (0.3.8)
Requirement already satisfied: Pillow in /usr/local/lib/python3.7/dist-packages (from pytesseract) (7.1.2)
```

In [ ]:

```
#Import Libraries

import pytesseract
import shutil
import os
import random
try:
 from PIL import Image
except ImportError:
 import Image

import matplotlib.image as mpimg
import matplotlib.pyplot as plt
```

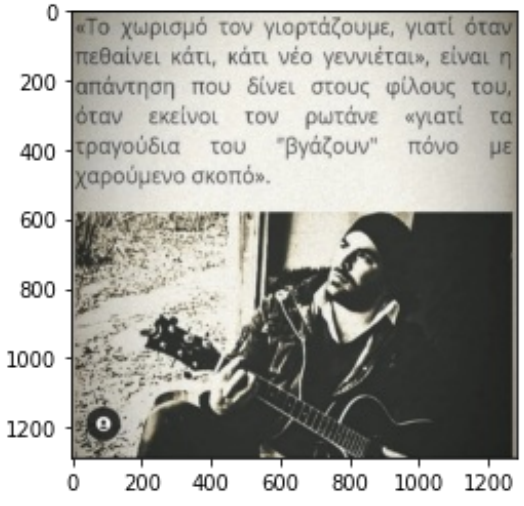
In [ ]:

```
# Read Images
img = mpimg.imread('/content/drive/MyDrive/Pantelis.jpg')

# Output Images
plt.imshow(img)
```

Out[ ]:

```
<matplotlib.image.AxesImage at 0x7f557498ae90>
```



Download source language from "https://github.com/tesseract-ocr/tessdata"

Since the image displays greek text, we'll download the grc.traineddata from https://github.com/tesseract-ocr/tessdata/blob/master/grc.traineddata

In [ ]:

```
# filenames = os.listdir('/usr/share/tesseract-ocr/4.00/tessdata/')
# print(filenames)


# import shutil
# src = '/content/drive/MyDrive/grc.traineddata'
# dest = '/usr/share/tesseract-ocr/4.00/tessdata/'
# shutil.copy(src, dest)

filenames = os.listdir('/usr/share/tesseract-ocr/4.00/tessdata/')
print(filenames)
```

```
['tessconfigs', 'pdf.ttf', 'configs', 'grc.traineddata', 'eng.traineddata', 'osd.traineddata']
['tessconfigs', 'pdf.ttf', 'configs', 'grc.traineddata', 'eng.traineddata', 'osd.traineddata']
```

In [ ]:

```
image_path_in_colab='/content/drive/MyDrive/Pantelis.jpg'
```

```
extractedInformation = pytesseract.image_to_string(Image.open(image_path_in_colab), lang='grc')
print(extractedInformation)
```

```
«Τὸ χωρισμό τον γιορτάζουμε, γιατί ὁταν
πεθαίνει κάτι, κάτι νέο γεννιέται», είναι ἡ
απάντηση που δίνει στοὺς φίλους τοῦ,
οἶαν εκείνοι τοὸν ρωτάνε «γιατί τᾶ
Πραγούδια τοῦυ "βγάζουν' πόνο Π
χαρούμενο σκοπό».
```

In [ ]:

```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('GreekTextExtractionOCR.ipynb')
```

```
File ‘colab_pdf.py’ already there; not retrieving.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

[NbConvertApp] Converting notebook /content/drive/MyDrive/Colab Notebooks/GreekTextExtractionOCR.ipynb to pdf
[NbConvertApp] Support files will be in GreekTextExtractionOCR_files/
[NbConvertApp] Making directory ./GreekTextExtractionOCR_files
[NbConvertApp] Writing 27259 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: [u'xelatex', u'./notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: [u'bibtex', u'./notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 107830 bytes to /content/drive/My Drive/GreekTextExtractionOCR.pdf
```

Out[ ]:

```
'File ready to be Downloaded and Saved to Drive'
```