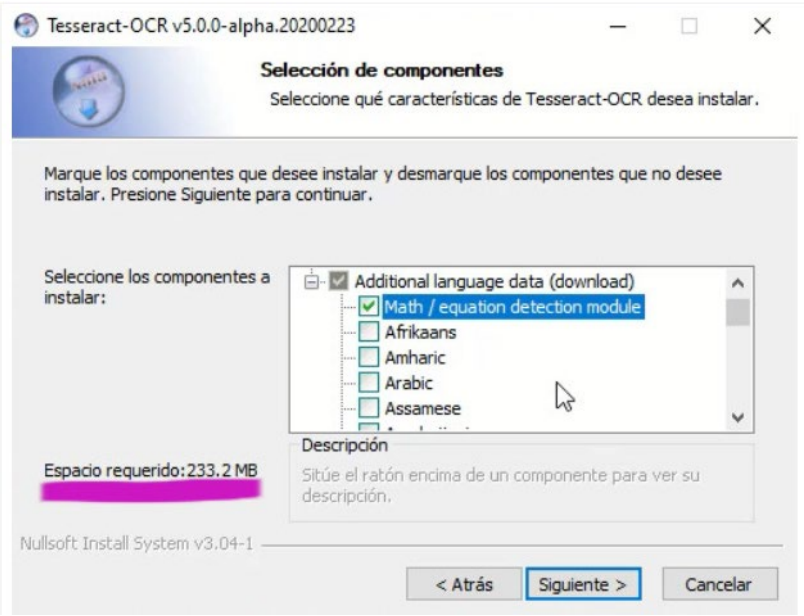
TRABAJO 1: Búsqueda de una determinada palabra

El objetivo es contar el número de veces que aparece una determinada palabra en la imagen “texto_delfines.png” con texto. Para lograrlo vamos a usar Tesseract-OCR, un motor de reconocimiento óptico de caracteres de código abierto financiado por Google que podría ser útil en un futuro.

Instalación de las librerías

Para iniciar con la instalación de Tesseract nos dirigimos a (https://github.com/UB-Mannheim/tesseract/wiki) y seleccionamos la versión que corresponda con vuestro PC (64 o 32 bits). Una vez descargado, ejecutamos el instalador. Escogemos un idioma para realizar el proceso, aquí por ejemplo me aparece el español. Damos clic en OK y continuamos con la instalación.

En la sección de *Selección de componentes* vamos a ver que por defecto se instalan los datos del idioma inglés. Si deseamos otros idiomas vamos a expandir *Aditional language data*, en donde podremos elegir aquellos que deseemos instalar.



A la izquierda vas a poder ver la cantidad de espacio requerido según el número de idiomas o componentes que elijas. Finalizamos la instalación dándole a “siguiente” varias veces y después terminar.

Instalar Pytesseract

Para que podamos emplear Reconocimiento Óptico de Caracteres a través de Tesseract desde Python necesitamos instalar pytesseract. En esta página tenemos la información sobre esta herramienta.

https://pypi.org/project/pytesseract/

Tecleamos en cmd: pip install pytesseract

Después en tus programas python deberás incluir la siguiente línea:

pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract'

Ya tienes las instalaciones hechas.

Búsqueda de una palabra en una imagen de un texto

Los pasos que podrías seguir son:

1) Leer la imagen a color.
2) Convertir a nivel de gris.
3) Mejorar la imagen en caso de ser necesario.
4) Umbralizar, resulta una imagen con fondo negro y con las letras en blanco.
5) Invertir para tener fondo blanco y letras en negro.
6) Buscar la palabra que desees usando OCR y opencv. En el trozo de código para uso de OCR se ha usado la palabra “delfines”. Puedes probar con otras palabras.
7) Probar con la imagen castellano_antiguo. Buscar alguna de sus palabras. ¿Qué tal funciona?
1. Sacar conclusiones.
2. Escribir memoria.

CÓDIGO DE USO DE OCR

```
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract'

from pytesseract import Output
data_image= pytesseract.image_to_data(invert_image, output_type=Output.DICT)

# Encontrar determinado texto
color = (0, 255, 0)
n_boxes = len(data_image['text'])
num_palabras=0
for i in range(n_boxes):
    if int(data_image['conf'][i]) > 40:
       match= re.match('delfines',data_image['text'][i])
       if match:
           num_palabras=num_palabras+1
           (x, y, w, h) = (data_image['left'][i], data_image['top'][i],data_image['width'][i], data_image['height'][i])
           img = cv2.rectangle(img, (x, y), (x + w, y + h), (255,0,0),2)
plt.imshow(img)
plt.show()

print('El número de palabras encontrado es: ', num_palabras)
```