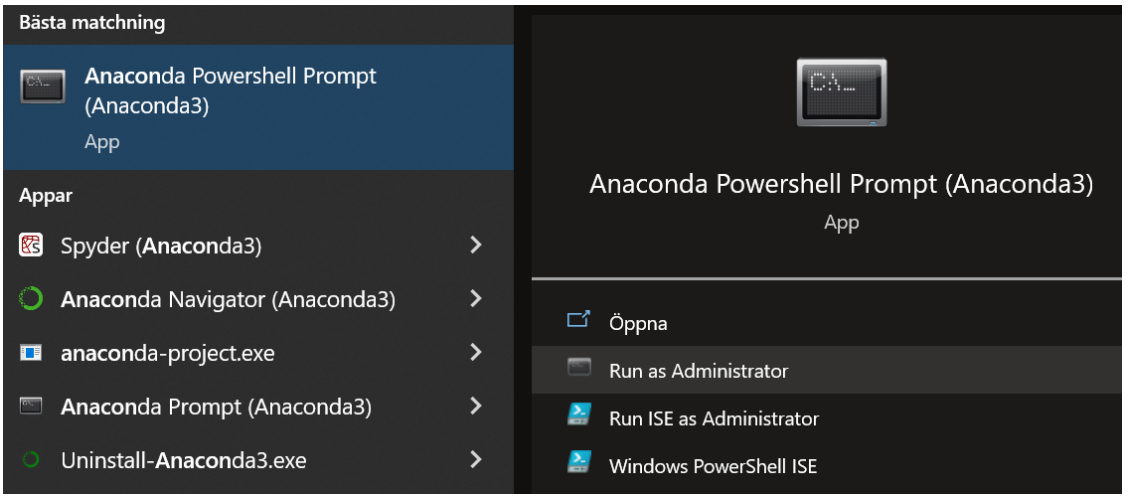
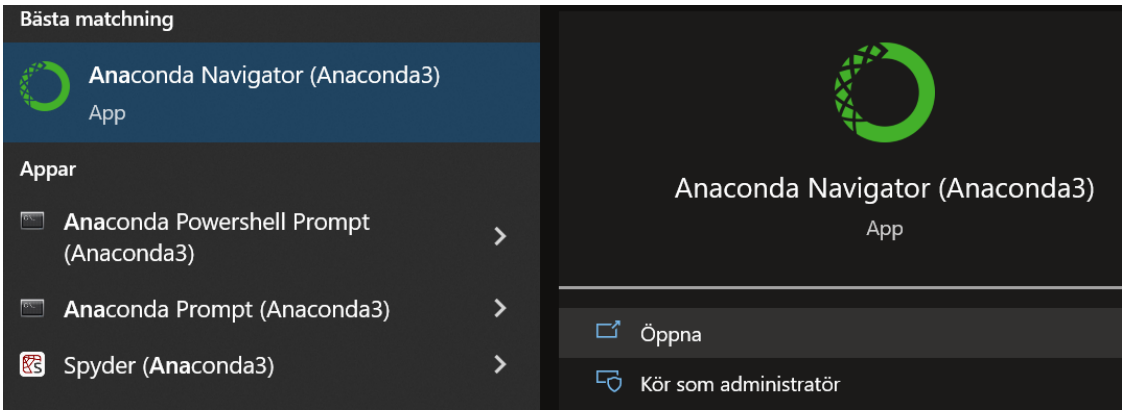
OCR-tesseract-SE

Installation

1. Ladda ned Anaconda från: https://www.anaconda.com/products/individual
2. Installera Anaconda genom att dubbelklicka på den nedladdade filen med alla standardinställningar.
3. Sök efter *Anaconda Powershell Prompt* (även kallad Anaconda terminalen i denna manual) i Windows sökfönster, och välj starta i admin-läge (*Run as Administrator)*

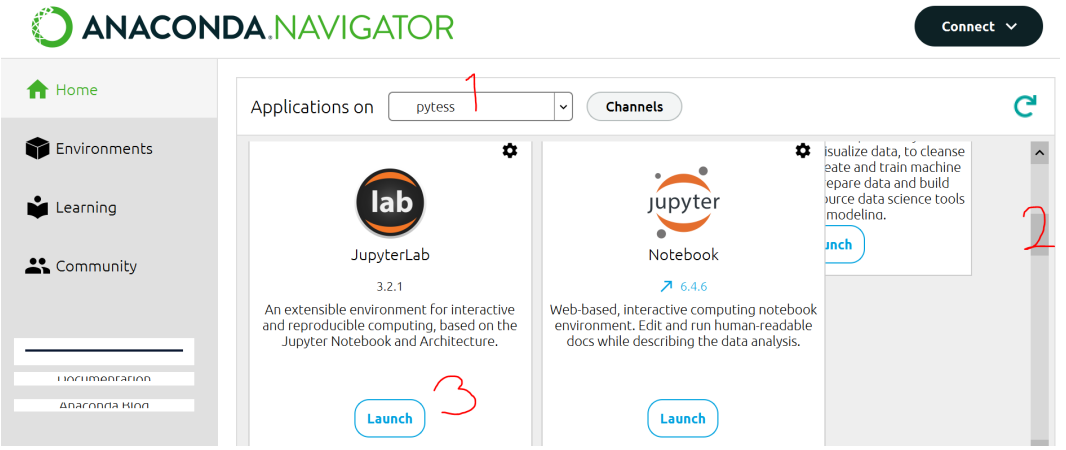


4. I Anaconda-terminalen som nu öppnas så anger du följande kommandon:
    1. conda create --name pytess
    2. conda activate pytess
5. Sök efter Anaconda Navigator i Windows sökfönster, och öppna programmet via öppna (admin-läge inte nödvändigt)

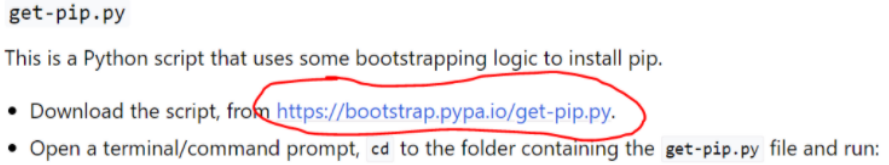


6. När Anaconda Navigator öppnats i din webbläsare så gör du följande tre saker:
    1. Under *Applications on* väljer du *pytess* (din virtuella värld)
    2. Bläddra sedan ned via rullisten till höger tills du hittar JupterLab

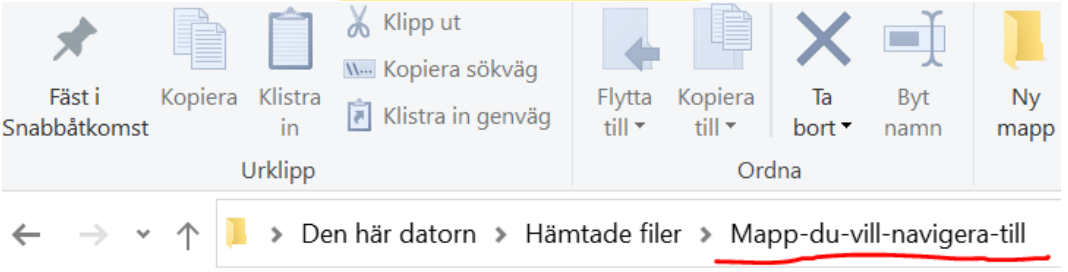
3. Klicka på installations ikonen (Install) för JupterLab (det står *Launch* i bilden nedan för att programmet installerades innan denna manual gjordes...)



7. Gå till baka till Anaconda-terminalen och säkerställ att du Python är installerat på din dator. Detta görs helt enkelt genom att skriva python i terminalen. Om Python är installerat visas nu info om vilken version m.m. Ange sedan exit() för att komma tillbaka till Anacondas "vanliga" terminalläge.
    - Om det skulle visa sig att Python inte är installerat är det nu läge att installera programmet. Hur detta går tillväga har vi dock inte fått någon information om.
8. Det är nu dags att installera pip (ett pakethanteringsprogam för Python).
    1. Surfa till https://pip.pypa.io/en/stable/installation/ och bläddra ned till "rubriken" *get-pip.py* och högerklicka på länken som går till https://bootstrap.pypa.io/get-pip.py och välj *Spara länk som* och spara ned till valfri mapp (t.ex. vanliga nedladdningsmappen).



2. Gå till baka till Anaconda terminalen och navigera till den mapp vari du sparade *bootstrap*-skriptet i föregående steg. För att byta mapp anger man följande kommando: cd /väg/till/mapp Information om vägen till mappen får du enklast fram genom att i navigera till mappen i fråga via Windows vanliga utforskare, för att sedan högerklicka på mappens namn i navigatorfönstret (se bild nedan) och välj *Kopiera adress*. Återgå sedan till Anaconda terminalen och klistra in din kopierade adress såsom i cd klistra-in-din-adress-här och klicka på retur.



   - Av okänd anledning funkar denna teknik för alla mappar förutom datorns vanliga nedladdningsmapp på min dator. Om ni har samma problematik löses detta enkelt genom att öppna en mapp (skapa en om ingen finns redan) i nedladdningsmappen, och kopiera adressen till den. Kopiera in den adressen i Anacondas terminal för att sedan radera den del av "adressen" som är efter nedladdningsmappen. Ett praktiskt exempel på detta är att en adress såsom C:\Users\Theo\Downloads\~~mapp-vars-adress-du-kopierar~~ förkortas till C:\Users\Theo\Downloads På så vis får du relativt smidigt korrekt adress till datorns nedladdningsmapp.
3. När du väl står i mappen du laddat ned bootstrap-skriptet till installerar du *pip* via följande kommando: python get-pip.py (Intressant nog överensstämmer detta inte med kommandot på installationssidan för pip, https://pip.pypa.io/en/stable/installation/, där *pyton* är förkortat till *py*, vilket dock inte funkar i praktiken...)

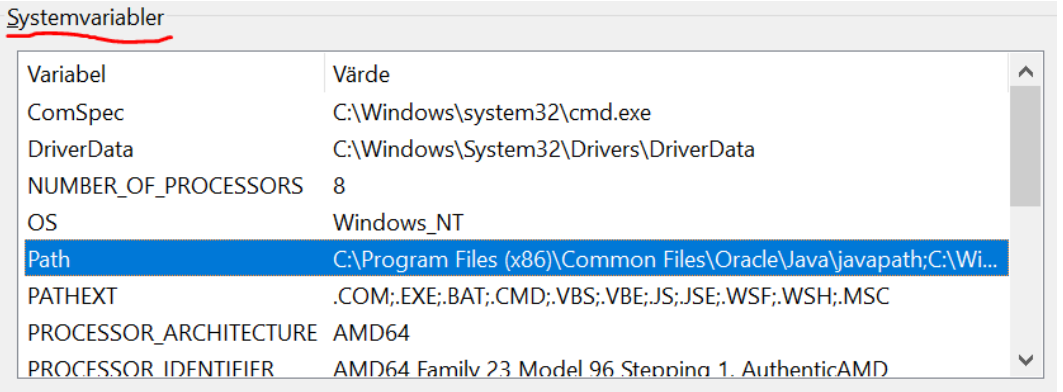
9. pip-installera sedan pytesseract genom att ange: pip install pytesseract i Anaconda terminalen
10. Det är nu dags att installera Tesseract, själva OCR programmet. Detta görs genom att surfa till https://github.com/UB-Mannheim/tesseract/wiki och ladda ned 64-bitars versionen (om nu din dator körs i 64-bit läge) av Tesserect

    The latest installers can be downloaded here:

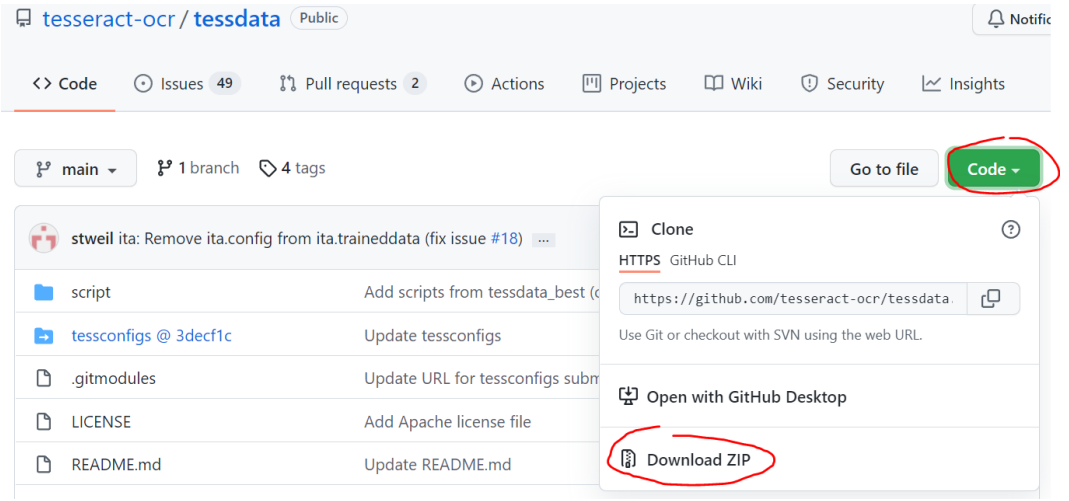
    - tesseract-ocr-w32-setup-v5.0.0.20211201.exe (32 bit) and
    - tesseract-ocr-w64-setup-v5.0.0.20211201.exe (64 bit) resp.

    - Installera sedan programmet genom att dubbelklicka på den nedladdade filen och säg ja till alla standardinställningar.

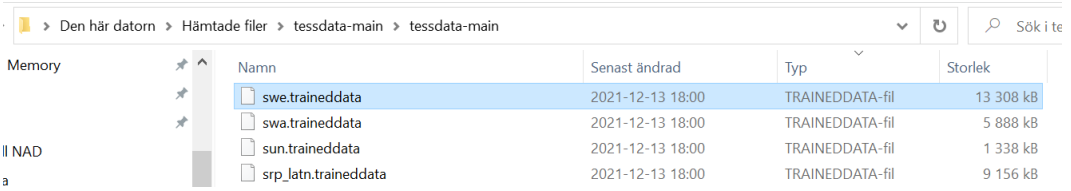
11. Det är nu dags att addera tesserect path till datorns systemvariabler. Detta görs genom att:
    1. Skriv in *miljö* i Windows sökfält, och klicka på *Redigera systemets miljövariabler*
    2. Klicka på *Miljövariabler* längst ned till höger i fönstret som öppnas
    3. Dubbelklicka på *Path* i fönstret för *Systemvariabler*



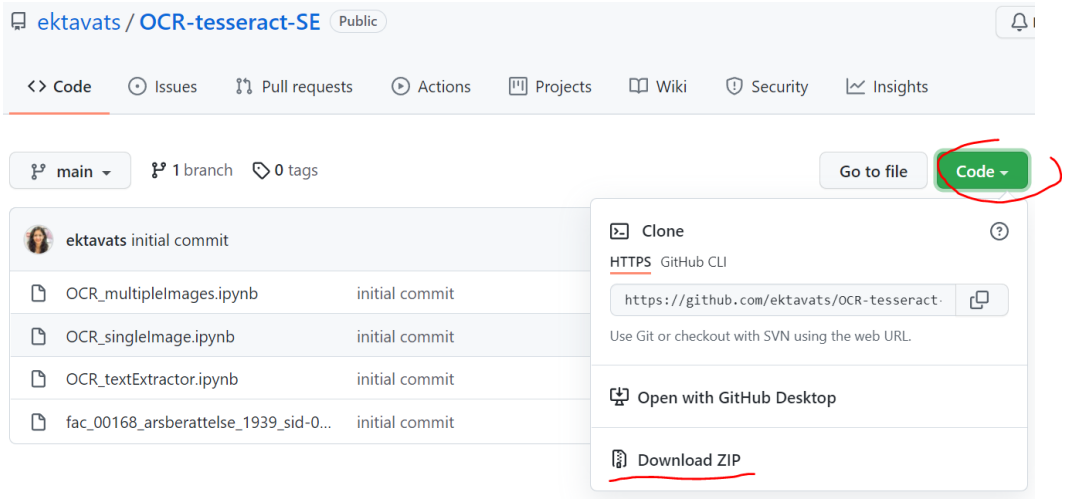
    4. Klicka på ny i fönstret som öppnas, och klistra in vägen (path) till mappen som Tesseract installerats till, varefter du klickar på OK. Om du inte gjort några ändringar under installationsproessen bör denna plats vara: **C:\Program Files\Tesseract-OCR** (men säkerhetskontrollera alltid att detta stämmer, och korrigera adressen vid behov)
12. [Valfritt steg] Kontrollera sedan vilka språkmodeller som finns med i din installation av Tesserect genom att ange kommandot: tesseract --list-langs i din Anaconda terminal
13. Om du vill komplettera din installation med fler språkmodeller så:
    1. Surfar du till https://github.com/tesseract-ocr/tessdata och klickar på *Code* följt av *Download ZIP* (se bild) för att ladda ned alla tillgängliga språkmodeller



2. Packa upp den just nedladdade mappen, och kopiera över ett urval eller alla språkmodeller (alla TRAINEDDATA-filer) till tessdata-mappen som ligger i mappen för din Tesseract installation (vanligt hittas denna mapp här: C:\Program Files\Tesseract-OCR\tessdata).
3. Kontrollera att Tesseract kommer åt språkmodellerna genom att i Ancondaterminalen återigen köra tesseract --list-langs kommandot, som nu skall lista de just tillagda språkmodellerna (plus de som var där från början).



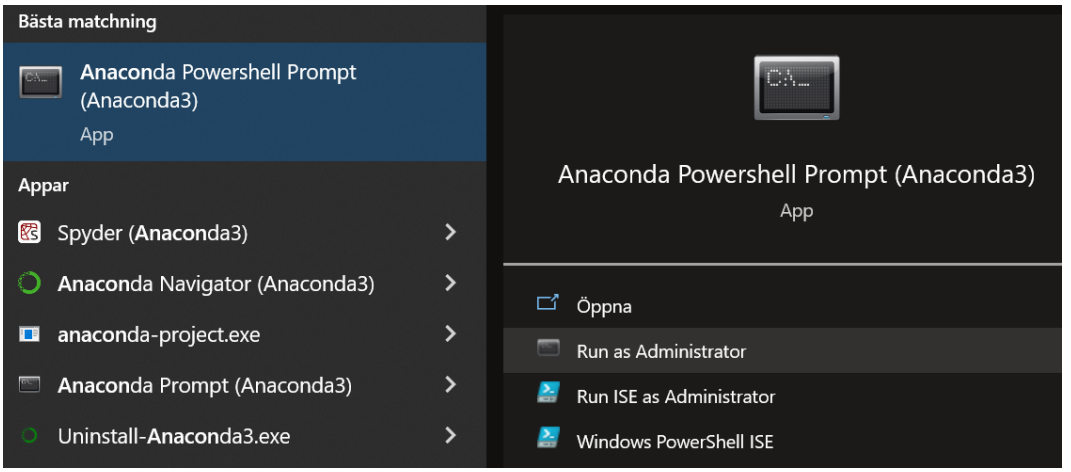
14. Återgå sedan till Anaconda-terminalen och installera sedan följande två paket via pip:
    1. pip install matplotlib
    2. pip install opencv-python
15. Du har nu installerat den tekniska infrastruktur som behöver vara på plats för använda programmet OCR-tesseract-SE-main. I dagsläget är det inte fullt ut bestämt vart koden kommer att förvaras långsiktigt, men i skrivande stund (2021-12-14) finns koden tillgänglig för nedladdning via Ekta Vats githubsida: https://github.com/ektavats/OCR-tesseract-SE
    1. Klicka på länken ovan för att komma till Ekta Vats githubsida, varefter du klickar på *Code* följt av *Download ZIP.*



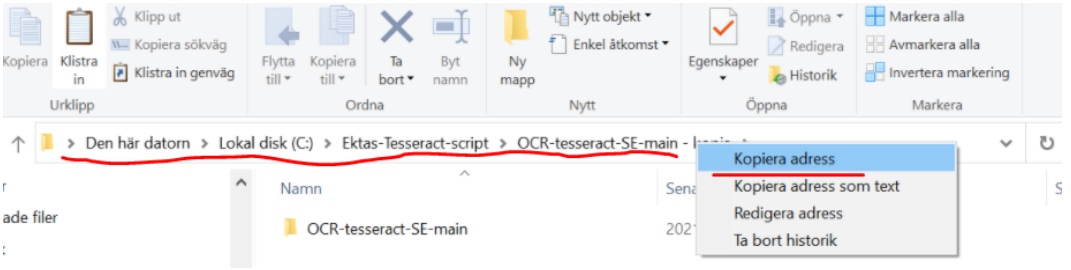
    2. Efter att du laddat ned koden extraherar du ZIP-filen och placerar du mappen innehållande programkoden (som vid skrivande stund heter *OCR-tesseract-SE-main*) i valfri lokal (ej server) mapp på din dator.
16. Grattis, installationen är nu slutförd, och det är nu dags att börja använda dig av programvaran.

Manual till OCR-tesseract-SE

1. Vid varje uppstart av OCR-tesseract-SE behöver du genomföra följande moment:
    1. Börja med att öppna *Anaconda Powershell Prompt* (även kallad Anaconda terminalen i denna manual) genom att söka på Anaconda i Windows sökfönster, och klicka på *Run as Administrator*.



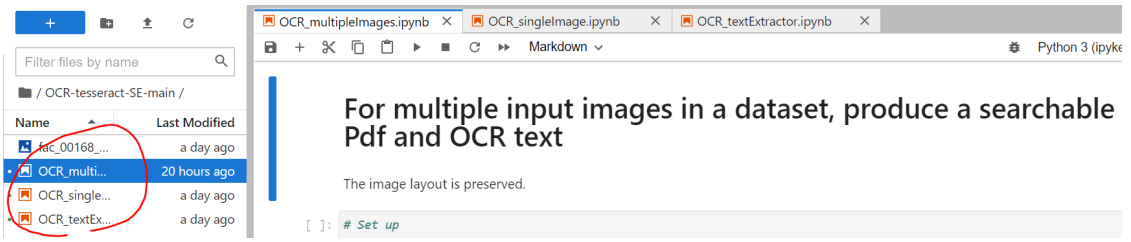
    2. Aktivera sedan den virtuella python miljö, pytess, som skapades i samband med installationen av mjukvaran. Detta göras genom att ange följande kommando i Anaconda terminalen: conda activate pytess
    3. För att kunna starta *OCR-tesseract-SE* behöver du först navigera via Anaconda terminalen till den mapp som innehåller programkoden. Mappens namn är *OCR-tesseract-SE-main*, och det är samma mapp som laddades ned under steg 15:2 i installationsprocessen (se ovan). I praktiken görs detta enklast genom att först navigera fram till mappen i fråga via Windows vanliga utforskare, varefter du kopierar adressen till mappen genom att högerklicka i utforskarens "adressfönster" (se bild nedan) och välj *Kopiera adress*. Återgå sedan till Anaconda terminalen och ange *cd* (cange directory) varefter du klistrar in (ctrl+v) den adress du just kopierade, varefter du bör få fram ett kommando som ser ut på liknande vis: cd C:\Ektas-Tesseract-script\OCR-tesseract-SE-main (OBS, din adress kommer vara unik för ordningen in din dator), varefter du klickar på retur.



    4. När du väl navigerat till *OCR-tesseract-SE-main* mappen i Anaconda terminalen anger du följande kommando för att starta programmet: jupyter lab

5. Programmet kommer nu att öppnas per automatik i din dators webbläsare under adressen http://localhost:8888/lab

2. I "menyn" längst till vänster i programmet går det att öppna tre olika "arbetsmallar", vid namn: *OCR_multipleimages.ipnb; OCR_singleimages.ipnb* och *OCR_textExtractor.ipnb.* Av dessa tre mallar är det främst *OCR_multipleimages.ipnb* som är av intresse då *OCR_singleimages.ipnb* främst är tänk att testa olika programinställningar på enskilda bilder, och *OCR_textExtractor.ipnb* mer är ett verktyg för att visa varifrån programmet hämta in text-information i dokumenten snarare än ett verktyg för att extrahera OCR:ad text. Med anledning av det fokuserar resten av denna "manual" på *OCR_textExtractor.ipnb* mallen.

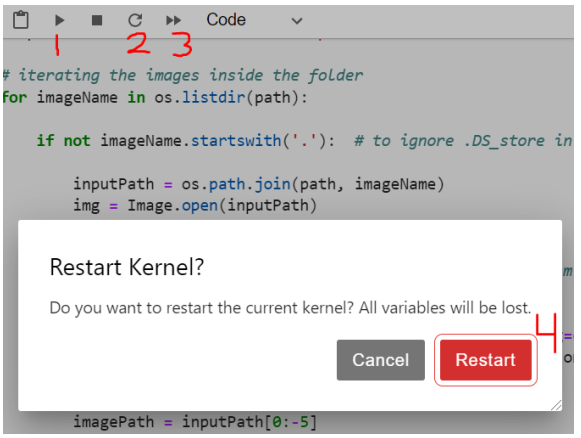


3. I bilden nedan visar jag på de fyra olika inställningar som är av intresse vid användning av OCR-programmet
    1. Ersätt den röda texten som står mellan citattecken (path ="**väg-till-mapp**") med adressen/vägen till den mapp som innehåller de bilder du vill OCR:a. Viktigt är att alla slashtecken " / " lutar framåt, annars hittar programmet inte till mappen i fråga. Om man kopierar en adress via Windows utforskare (som vi gjorde i steg 1:3, fast med den skillnaden att vi nu måste välja *Kopiera adress som text*) så är nämligen alla slashtecken bakåtlutande " \ ". I praktiken innebär detta att en adress som *C:\OCR-tesseract-SE-main\obehandladeFiler*måste ändras till *C:/OCR-tesseract-SE-main/obehandladeFiler* för att det hela skall fungera.
    2. Ersätt den rörda texten mellan citattecken med adressen till den mapp vartill du vill spara de OCR:ade filerna i PDF-format.
    3. Ersätt den rörda texten mellan citattecken med adressen till den mapp vartill du vill spara de OCR:ade filerna i TXT-format.

4. Om du OCR:ar dokument på annat språk än svenska behöver du byta vilken språkmodell programmet använder sig av. Dels behöver du se till att Tesseract har tillgång till den språkmodell du önskar använda (se steg 13:1-3 i "installationsmanualen" ovan), och dels behöver du ange i OCR-tesseract-SE vilken språkmodell du vill använda dig av (se nr 4 i bilden nedan). Man behöver ange språkmodell för båda TXT- och PDF-filer då skapandet av dessa olika utdataformat i praktiken görs via två olika arbetskedjor. Vilken eller vilka språkmodell som programmet skall använda sig av anges i OCR-tesseract-SE med s.k. kortkoder, där t.ex. **swe** står för *swedish*. En lista med alla tillgängliga kortkoder för olika språk går att finna här: https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html Om en dokument innehåller flera språk är det fullt möjligt att simultant använda sig av flera språkmodeller simultant genom att addera ett plustecken " + " mellan de olika språkkoderna, såsom i **lang ="swe+eng"**

```
def main():
    # path for the folder for getting the raw images
    path ="C:/Ektas-Tesseract-script/OCR-tesseract-SE-main/Test/obehandladeFiler"

    # path for the folder for getting the output
    tempPath ="C:/Ektas-Tesseract-script/OCR-tesseract-SE-main/Test/utPdf"
    tempPath2 ="C:/Ektas-Tesseract-script/OCR-tesseract-SE-main/Test/utTxt"

    # iterating the images inside the folder
    for imageName in os.listdir(path):

        if not imageName.startswith('.'):  # to ignore .DS_store in Mac

            inputPath = os.path.join(path, imageName)
            img = Image.open(inputPath)

            #configure parameters for pytesseract
            custom_config = r'--oem 3 --psm 6' #oem 3: Default; psm 6: Assume a single uniform block of text.

            # applying ocr using pytesseract
            text = pt.image_to_string(img, lang ="swe", config=custom_config)       # for text file output
            text2 = pt.image_to_pdf_or_hocr(img, lang='swe', config=custom_config)  # for searchable Pdf
```

4. När man är nöjd med hur man ställt in mappadresser för in- och utdata såväl som vilken språkmodell som skall användas är det bara att klicka på *play* (1) (se bild nedan) knappen högst upp i mitten för att sätta igång OCR processen. Om fått ett error t.ex. med anledning av att man glömt att andra " \ " till " / " i mappadresserna kan man behöva "ladda om" programmet genom att klicka på cirkel-pil (2) tecknet eller spolingsknappen (3) följt av *Restart* (4) när programmet frågar dig om du vill *Restart Kernel?*



5. När programmet OCR:at alla dokument i mappen du specificerat för indata (se steg 3:1) hittar du alla OCR:ade PDF och TXT filer i de mappar du angivit för utdata (steg 3:2-3).
6. Mallen *OCR_singleimage* följer i stort samma logik som den som angetts ovan, med den skillnad att vägen/adressen som man anger måste ledadirekt till den specifika bild man vill OCR:a (dvs. inte till mappen den ligger i).
7. När du vill avsluta programmet stänger du först ned webbläsaren vari programmet körs, varefter du växlar över till Anaconda terminalen och trycker på **ctrl+c** vilket avslutar programmet. Du kan sedan stänga Anaconda terminalen.
8. När du nästa gång vill använda dig av detta OCR-program startar du återigen upp det genom att börja från punkt 1 i denna användarmanual.