ANVÄNDARGUIDE – Prisskrapa

Innehåll

[Inloggningsuppgifter 2](#_Toc137550558)

[Användarguide 2](#_Toc137550559)

[1. Tunnelbear 2](#_Toc137550560)

[2. Installera Python och Pycharm 2](#_Toc137550561)

[3. Github 3](#_Toc137550562)

[Vad är det och varför är det smidigt att använda? 3](#_Toc137550563)

[Guide för nedladdning av kod via git clone 3](#_Toc137550564)

[Github 3](#_Toc137550565)

[Git clone 5](#_Toc137550566)

[Tips och Trix för användning av terminalen 5](#_Toc137550567)

[4. Installera packages 5](#_Toc137550568)

[5. Köra koden – för respektive land 7](#_Toc137550569)

[a. Sverige. 8](#_Toc137550570)

[b. Finland 9](#_Toc137550571)

[c. Norge 10](#_Toc137550572)

[d. Danmark 11](#_Toc137550573)

[6. Efter körning, uppdatera excel-filer på github 12](#_Toc137550574)

[Gudie för användning git för update respektive push & commit 12](#_Toc137550575)

[Update project 13](#_Toc137550576)

[Push and commit 14](#_Toc137550577)

[7. Uppdatera excelfilerna på servern 16](#_Toc137550578)

[Felhantering med de vanligaste felen 16](#_Toc137550579)

[Vanligaste felen: Uppdatera HTML-tag/class eller xpath 17](#_Toc137550580)

[Vanligaste felen: Uppdatera Selenium och chromdriver.exe 18](#_Toc137550581)

[Större fel: Ta bort filer från main 19](#_Toc137550582)

# Inloggningsuppgifter

* Inloggningsuppgifter till Tunnelbear och Github finns i pathen nedan
* **G:\Mediavision projekt\+ MEDIAVISION INSTANT INSIGHTS (MII)**
* Lokalisera sedan word-dokumentet **lösenord** för att hitta uppgifterna.

# Användarguide

## 1. Tunnelbear

1. Ladda ner **Google Chrome** om det inte finns nedladdat på dator, se till att det är uppdaterat till senaste versionen.

2. Ladda ner Tunnelbear-appen och logga in med inloggningsuppgifterna

a. [Download a VPN (tunnelbear.com)](https://www.tunnelbear.com/download)

b. Logga in på Tunnelbear och prova att byta så att du surfar från ett annat land. Bilden nedan visar hur det ska se ut när man t ex. surfar med VPN från Finland

Map

Description automatically generated

## 2. Installera Python och Pycharm

3. Installera Python på din dator:[Download Python | Python.org](https://www.python.org/downloads/)

4. Installera Pycharm på din dator: <https://www.jetbrains.com/pycharm/download/#section=windows>

Det går bra med Pycharm community.

5. Öppna Pycharm appen och öppna ett nytt projekt. Det ska se ut som på bilden nedan när du öppnat ett nytt projekt. OBS vikitgt att venv-mappen är inkluderarad, som syns på bilden.

Text

Description automatically generated

## 3. Github

## Vad är det och varför är det smidigt att använda?

* Git och git clone och bidrar till att genom att skriva en rad i terminalen kan få åtkomst till din kod. Detta kan underlätta för uppladdning av t ex. specifika koder eller om flera arbetar med koden samtidigt.

## Guide för nedladdning av kod via git clone

1. Ladda ner ”git” till din dator <https://git-scm.com/download/win>. Se till att git läggs in i din mapp i pycharm, för att kopiera hela din path till pycharm, vänsterklicka och tryck sedan ”Copy Path/Reference…”> ”Absolute Path”. Här finns en guide för att ladda ner git till windows: <https://phoenixnap.com/kb/how-to-install-git-windows>

2. Öppna ett terminalfönster i Pycharm genom att trycka View>Tool Windows”>”Terminal”. Ett terminalfönster öppnas då och din nuvarande path kan ses. Alternativ högerklicka och tryck ”open in”>”Terminal”

3. Stäng ner och starta om pycharm för att se till att alla programmen laddas in och git finns!

## Github

3. Gå till <https://github.com/> för att hämta koden och logga in med inloggningsuppgifterna. OBS! Det är tvåstegsautentisiering/verifiering. En kod skickas till mediavisions mobiltelefon som behövs för att logga in.

4. För videoskrapan klicka på ”repositoryn”: /mediavision21/prisskrapa. För ljudskrapan är det ”repositoryn”: /mediavision21/prisskrapa\_ljud

En bild som visar text, Teckensnitt, programvara, Webbsida

Automatiskt genererad beskrivning

5. När du klickat på ”repositoryn” kommer du till sidan där all kod ligger. Klicka på den gröna knappen ”code”. Se till att det är ett orange sträck under HTTPS och klicka på de två fyrkanterna bredvid URLen för att kopiera den.

En bild som visar text

Automatiskt genererad beskrivning

## Git clone

6. Se till att du är inloggad på github-kontot via Pycharm. För att göra detta gå till Settings>Version Control>Github och se till att mediavision ligger inlagt som ett konto. Det ska se ut som bilden.En bild som visar text, skärmbild, programvara, Multimedieprogram

Automatiskt genererad beskrivning

7. Gå tillbaka till pycharm, öppna ett terminalfönster och skriv ”git clone XXX”, där XXX står för urlen som kopierades från github och som skall klistras in. Autentisera dig via github och vänta en kort stund medan koden från github laddas in i pycharm. Ytterligare guidning för Git clone: <https://git-scm.com/book/en/v2/Git-Basics-Getting-a-Git-Repository>

8. När alla filer laddats ner. Stäng ner pycharm en gång och starta om det, vilket kan behövas för att allt ska fungera sen.

# Tips och Trix för användning av terminalen

Några användbara kommandon i terminalen:

* ls: för att se vad som finns i din nuvarande path t ex. mappar eller filer
* cd XX: för att gå in i t ex en mapp och då skrivs cd prisskrapa om prisskrapa är namnet på mappen
* cd .. för att gå tillbaka ett steg så att du kommer en mapp längre bak

En bild som visar text

Automatiskt genererad beskrivning



En bild som visar text

Automatiskt genererad beskrivning

## 4. Installera packages

Dessa packages krävs för att köra programmet

* *requests*
* *bs4*
* *selenium*
* *pandas*
* *XlsxWriter*
* *openpyxl*

a. Om detta är första gången du kör koden så måste en del packages installeras.

b. Vissa paket syns inte direkt i koden t ex. openpyxl. Alla paket kan installeras på detta sätt dock

- I Pycharm: File>Settings> Project:pythonproject>Python Interpreter> Klicka på “+” som den röda pilen pekar på i bild 2>skriv in namnet på paketet t ex. ”openpyxl” som i bild 3 och tryck på ”Install package”

A screenshot of a computer

Description automatically generated with medium confidence

*Visar hur ”Settings” ser ut inuti PyCharm*

Graphical user interface, text

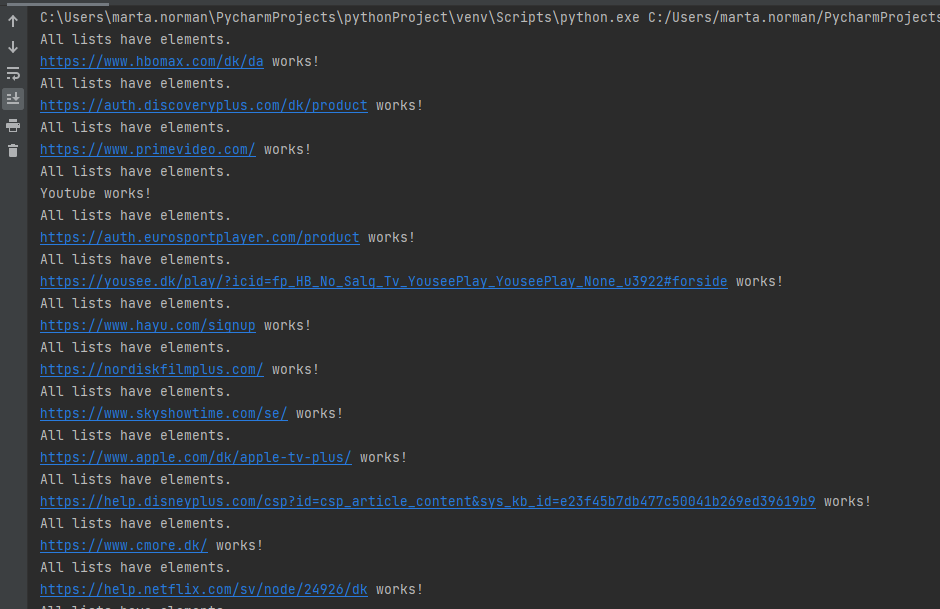
Description automatically generated

*Visar hur det ser ut när man installerar ett nytt paket inuti PyCharm.*

## 5. Köra koden – för respektive land

**Generellt:** Oavsett land så körs den samlade scrapingen via *main.py-filen.* Men då Finland, Danmark och Norge kräver speciella VPN så kan man bara köra ett land i taget.

* Viss felhantering är inkluderad som kollar att det finns data i alla listor. Om det finns data för respektive tjänst ser det ut som på bilden nedan med ’URLn’/namnet på tjänsten + works och ’All lists have elements’!
* Om någon tjänst inte har data t ex. printas outputen ’URLn/namnet på tjänsten + has no data och ett meddelande ”Some lists are empty”/”All lists are empty”/”Not all lists have the same length”



### a. Sverige.

För att köra Sverige se till att endast *Sweden\_VPN(merge\_output)* inte är utkommenterad. Det ska se ut som på bilden nedan. För att ”köra koden” för Sverige tryck på den gröna play-knappen bredvid *if \_\_name\_\_ == ”\_\_main\_\_”* eller vänsterklicka och tryck ”Run main(1)”*.* Koden körs då för alla tjänster i Sverige och om inte den redan körts för dagen läggs dagens skrapning in i EXCEL-filen **video\_SE\_2023.xlsx.**

**Text

Description automatically generated**

Efter en lyckad körning, om dagens datum redan finns med ska det så ut som följande i terminalen efter körningen.



Kolla excel-filen så att dagens datum finns med och att datan ser bra ut.

A picture containing background pattern

Description automatically generated

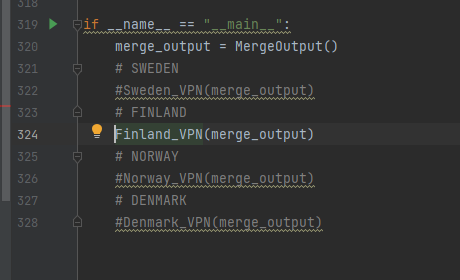
### b. Finland

1. Först så måste man ändra VPN till Finland. Detta görs genom att trycka på Finland i Tunnelbear-appen. Med finsk VPN ska det se ut som nedan i Tunnelbear appen.

Map

Description automatically generated

2. För att köra Finland se till att endast *Finland\_VPN(merge\_output)* inte är utkommenterad. Det ska se ut som på bilden nedan. För att ”köra koden” för Finland tryck på den gröna play-knappen bredvid *if \_\_name\_\_ == ”\_\_main\_\_”* eller vänsterklicka och tryck ”Run main(1)”*.* Koden körs då för alla tjänster i Finland och om inte den redan körts för dagen läggs dagens skrapning in i EXCEL-filen **video\_FI\_2023.xlsx.**

****

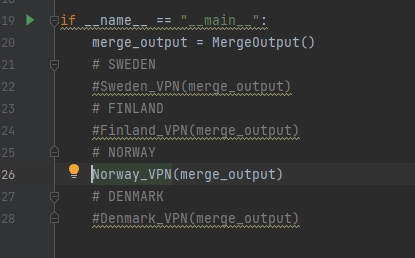
### c. Norge

1. Först så måste man ändra VPN till Norge. Detta görs genom att trycka på Norge i Tunnelbear-appen. Med norsk VPN ska det se ut som nedan i Tunnelbear appen.

Map

Description automatically generated

2. För att köra Norge se till att endast *Norway\_VPN(merge\_output)* inte är utkommenterad. Det ska se ut som på bilden nedan. För att ”köra koden” för Norge tryck på den gröna play-knappen bredvid *if \_\_name\_\_ == ”\_\_main\_\_”* eller vänsterklicka och tryck ”Run main(1)”*.* Koden körs då för alla tjänster i Norge och om inte den redan körts för dagen läggs dagens skrapning in i EXCEL-filen **video\_NO\_2023.xlsx.**

**2**

### d. Danmark

1. Först så måste man ändra VPN till Danmark. Detta görs genom att trycka på Danmark i Tunnelbear-appen. Med dansk VPN ska det se ut som nedan i Tunnelbear appen.

Map

Description automatically generated

2. För att köra Danmark se till att endast *Denmark\_VPN(merge\_output)* inte är utkommenterad. Det ska se ut som på bilden nedan. För att ”köra koden” för Danmark tryck på den gröna play-knappen bredvid *if \_\_name\_\_ == ”\_\_main\_\_”* eller vänsterklicka och tryck ”Run main(1)”*.* Koden körs då för alla tjänster i Danmark och om inte den redan körts för dagen läggs dagens skrapning in i EXCEL-filen **video\_DK\_2023.xlsx.**

**Text

Description automatically generated**

## 6. Efter körning, uppdatera excel-filer på github

Efter att koden har körts så har excelfilerna uppdaterats med ny data. Om du har gjort ändringar i koden behöver även detta uppdateras på github. Detta görs genom push & commit. Nedan finns även en guide för hur du uppdaterar projektet om någon annan har gjort uppdateringar i tex. av kod eller filer sedan senaste gången då klona repot.

## Guide för användning git för update respektive push & commit

**En bild som visar text

Automatiskt genererad beskrivning**

**En bild som visar skärmbild, Teckensnitt

Automatiskt genererad beskrivning**

När repot har blivit klonat behövs genvägar till Git. Tryck VCS>Enable Version Control Integration välj Git som Control System och tryck OK.

Här finns ytterligare dokumentation för hur Git används till Pycharm: <https://www.jetbrains.com/help/pycharm/using-git-integration.html>

Det ska se ut som bilderna ovan med ”Git:” och sedan en blå pil nedåt, en grön checkbox och en blå pil uppåt(som ses på bilderna ovan). Dessa kan användas för att uppdatera koden på din dator med koden på git respektive att uppdatera koden på git om tex. Du gjort förändringar i koden eller gjort en daglig skrapning och därav uppdaterat excelfiler.

## Update project

1. För att uppdatera din kod så att den matchar den på git. Tryck på den blåa nedåt-pilen.

2. Om detta är första gången kommer en liten ruta upp, se till att ”Merge incoming changes into current branch” är ifyllt och gör även en bock på ”Don’t show again” innan du trycker ok. Som på bilden nedan. Projektet uppdateras sedan och ändrar/lägger till filer utifrån vad som finns på github.

En bild som visar text, skärmbild, Multimedieprogram, Teckensnitt

Automatiskt genererad beskrivning

## Push and commit

**En bild som visar text

Automatiskt genererad beskrivning**

En bild som visar text

Automatiskt genererad beskrivning1. För att uppdatera github med eventuella kodförädringar eller uppdatering av excelfiler som du har gjort kan man använda ”push & commit”. Tryck på den gröna checkboxen mellan de två pilarna som ses på bilden ovan.

2. Då kommer bilden ovan upp till vänster. Här ses vilka filer som har ändrats, en kommentar kan läggas till t ex. ” Dagens skrapning 2023-06-13” och tryck sedan på ”push and commit…” som den röda pilen pekar på. Om det är första gången kan man behöva definiera användaren enligt bilden nedan

En bild som visar text

Automatiskt genererad beskrivning

3. Tryck sedan på Push knappen som den röda pilen pekar på i bilden nedan.

En bild som visar text

Automatiskt genererad beskrivning

4. Dubbelkolla att förändringen har kommit upp på github genom att gå in på github och kolla att en ny commit kommit in inom den senaste tiden och t ex. med din kommentar tex.

En bild som visar text

Automatiskt genererad beskrivning

## 7. Uppdatera excelfilerna på servern

* Avslutningsvis, lägg in de uppdaterade filerna på den gemensamma servern. Kopiera Excel-filerna från python-projektet till servern.
* För 2023 samlas data för servern här:
  + G:\Mediavision projekt\Resursmapp\Internprojekt\Kartläggning av SVOD-tjänster över tid\Prisskrapa\2023
* Högerklicka på projektet i Pycharm och tryck: Open in>explorer
* Ersätt filerna i mappen Prisskrapa\2023 med de från Pycharm som innehåller ny data!

En bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning

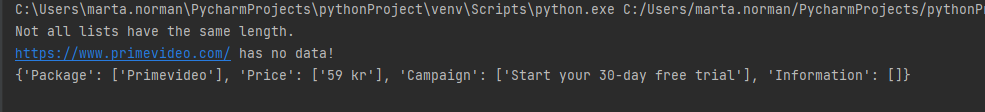
# Felhantering med de vanligaste felen

## Allmänt

* Börja med att prova att köra igen. Ibland kraschar programmet bara en gång utan att ett större problem uppstått och så fungerar det nästa gång man kör. Kan bara vara att internet är segt och att en hemsida inte hinner laddas in.

## Vanligaste felen: Uppdatera HTML-tag/class eller xpath/CSS-selector

* Hemsidan har uppdaterats och class name/xpath har bytts ut, vilket gör att skrapan inte hittar elementen och inte skrapar någon data. Detta syns genom felhanteringen då ett meddelande, tex som nedan ”Not all lists have the same length”.



* **Lösning:**
* **CSS-selector:**
  + Gå till den specifika tjänstens fil, i detta fall Amazon. Prova först att bara köra den och printa ut content för hela den Dictionary som på bilden ovan. Då syns det att ”Information” saknar data.

En bild som visar text, skärmbild, Teckensnitt, programvara

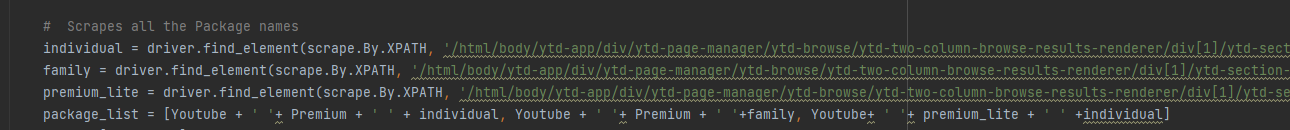
Automatiskt genererad beskrivning

* Metoden **information(driver)** använder selenium och en ”driver” som läser in hemsidan tas in som ett argument. Här skrapas information om hemsidan
* Information om paketen hittas via CSS-SELECTOR på HTML-koden på hemsidan. I detta fall ligger informationen i ’.dv-copy-body’, där ”.” framför namnet står för att det är en klass.
* Om listan är tom, kan klassen på hemsidan ändrats för att uppdatera taggen. Gå in på URL(använd den som ligger högst upp i filen skriven som ”self.URL = <https://www.primevideo.com/>”

En bild som visar text, skärmbild

Automatiskt genererad beskrivning

* Högerklicka>inspektera på hemisdan, lokalisera elementet du vill skrapa och kopiera klassen, taggen eller xpathen och klistra in i driver.find\_elements (scrape.By.CSS\_SELECTOR, ””) i detta fall, kom ihåg att ”div.dv-copy-body p” står för taggen ”<div> </div>” med klassen ”dv-copy-body” och sedan elementet som finns innanför taggen ”<p> </p>”.
* **Xpath:**
* Vissa hemsidor såsom Youtube används find\_element(scrape.By.XPATH, ””) istället för class name, css-selector eller tag name.
* Dessa kan också behövas uppdateras och hittas genom att kopiera ’full xpath’ i inspektormode på hemsidan. Se bilder nedan.

**

*Bilden ovan visar hur det ser ut när man kopierat ett xpath-uttryck till koden*

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning

*Bilden ovan visar hur det ser ut i inspektionsmode. Genom att klicka på det element som önskas kopieras och sedan ’kopiera hela xpath-uttrycket’. Kan xpath-uttryck uppdateras om hemsidan skulle förändras eller uppdateras.*

## Vanligaste felen: Uppdatera Selenium och chromdriver.exe

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning

* + Chromedriver är en fil som är definierad som ’chromedriver.exe’ och krävs för att köra selenium.
  + Denna filen kan behövas uppdateras i samband med uppdateringar av Google chrome, ladda ner en uppdaterad version av chromedriver då från följande länk: <https://chromedriver.chromium.org/>
  + Ersätt den nuvarande ’chromedriver.exe’ med den nya versionen
  + Detta märks genom att koden kraschar och ett **rött meddelande** kommer upp på skärmen där det står att chrome inte är kompatibelt med versionen.

## Större fel: Ta bort filer från main

* Fler problem kan uppstå om en hemsidan förändras i större utsträckning och byter design. Om t ex. Youtube inte skulle fungera men man vill köra det övriga tjänsterna kan denna kommenteras bort i Pythonfilen main() för att exkludera från körningen
* Kommentera ut först objektet som ses på bild 1 och ta sedan bort ”YoutubePremium\_SE\_obj.SE” från merge\_output.combine\_dicts([…]) som ses under bild 2.
* Då kan övriga tjänster köras och bara exkludera t ex. Youtube om den inte skulle fungera.

**En bild som visar text, skärmbild, Teckensnitt, programvara

Automatiskt genererad beskrivning**

1. **En bild som visar text, skärmbild, Teckensnitt

   Automatiskt genererad beskrivning**