Variablen sind ins. Berufsbezeichnung. Ort ebenfalls aber man sollte es am einfachsten auf „Deutschland“ beschränken.

* Für jede Berufsbezeichung eine eigene Excel Tabelle?
* Eine Spalte mit dem AKTUELLEN Datum hinzufügen: dadurch unterscheidet man Daten wenn man zu versch. Zeitpunkten scrapen will (über mehrere Monate z.B)
* Bedenken: Wie lange dauert eine Jobbezeichnung? -> insgesamt Zeit? Parallelisieren nötig?
* Fenster mit genauerer Jobbeschreibung: aktuell wird nicht die komplette jobbbeschreibung genommen
  + *~~Fehler könnte sein: man muss auf „mehr anzeigen“ KLICKEN und dann bekommen man den vollen Text? Fehler behoben~~*
  + *Gleiche Descriptions kommen mehrfach bei unterschiedlichen Firmen vor. Sleep Function hinzufügen, evtl wird zu schnell durch die Anzeigen gescrollt?*
  + *Df2 München Version: LinkedIn hat mich gesperrt, daher wahrscheinlich kein output im 2.Part*
  + *Also nicht überprüfbar ob sleep funktioniert hat… vlt im WIK testen.. oder mit anderen Job/ sleep kürzer setzen*
* Liste an Jobbezeichnungen zum Thema Digitalisierung finden
* ADD-ON: Auf die Unternnehmenslogo klicken (link aka href ist dort hinterlegt) und die Unternehmensgröße heraussuchen. Html/body/main/section/div/section/div/dl/div/dd 🡪 davon der inner Text z.B “ 10.001+ Beschäftigte „
  + Siemens Seite OHNE ( 😊) Login: [Siemens | LinkedIn](https://de.linkedin.com/company/siemens?trk=public_jobs_topcard_logo)

Stand 24.01.23

* Beispiel Solution Director in Deutschland: insgesamt 294 Jobs, Code lief 572 Sekunden
  + 🡪 572/294 =1,945 sec pro job; Wobei eine Wartezeit von 1,5 sec manuell eingefügt wurde 🡪 1,945-1,5 =0,445 sec pro job
  + 🡪 Bei 500.000 Stellenanzeigen bedeutet das eine Laufzeit von: 1,945\*500000=972.500 sec, das sind mehr als 11 Tage…. Nicht machbar
  + 181000 IT Jobs gibt es bereits 🡪 4 Tage
  + Komisch: die links stimmen jedoch?!!

Stand 14.02.2023

* Sehr hilfreicher Blog [(16) Scraping Job Requirements from LinkedIn | LinkedIn](https://www.linkedin.com/pulse/scraping-job-requirements-from-linkedin-arjun-singh/)
* LinkedIn always shows a job post's details when you call a job query. You would think that the HTML for what you are looking at is immediately available. How could you possibly see the content if it hasn't been pragmatically assigned to the screen? Normally, you would think that if you can see content then the HTML must have been passed through to the WebDriver. This is partially True. Some web pages load in batches because there is a lot of content that needs to be collected from the web host. This means the target element you want may take some time to pass through. Basically, we need to add WebDriverWait that listens for when you're target content is available.

So the target GET request has been called and the target content has been loaded... now what? We grab the markup by calling driver.page\_source and pass this through BeautifulSoup. The job requirements and company name are easily identified using a combination of soup.find and soup.find\_all. We now have what we're looking for from a single job. Half the battle is over. We just need to grab data from multiple jobs.

* [Waits | Selenium](https://www.selenium.dev/documentation/webdriver/waits/)
* [LinkedIn Scraping with Python. How to set up a Python script to scrape… | by Federico Haag | Medium](https://medium.com/federicohaag/linkedin-scraping-with-python-d8d14519602d)
* [LinkedInScraping/Scraper.py at master · federicohaag/LinkedInScraping · GitHub](https://github.com/federicohaag/LinkedInScraping/blob/master/Scraper.py)
* [Web Scraping Reference: A Simple Cheat Sheet for Web Scraping with Python (hartleybrody.com)](https://blog.hartleybrody.com/web-scraping-cheat-sheet/#useful-libraries)
* [Wait until page is loaded with Selenium WebDriver for Python - Stack Overflow](https://stackoverflow.com/questions/26566799/wait-until-page-is-loaded-with-selenium-webdriver-for-python)

from selenium import webdriver

from selenium.webdriver.support.ui import WebDriverWait

from selenium.webdriver.support import expected\_conditions as EC

from selenium.webdriver.common.by import By

from selenium.common.exceptions import TimeoutException

browser = webdriver.Firefox()

browser.get("url")

delay = 3 # seconds

try:

myElem = WebDriverWait(browser, delay).until(EC.presence\_of\_element\_located((By.ID, 'IdOfMyElement')))

print "Page is ready!"

except TimeoutException:

print "Loading took too much time!"

**Error Meldung 15.02 bei Versuch webdriverwait manually**

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webelement.py", line 93, in click

self.\_execute(Command.CLICK\_ELEMENT)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webelement.py", line 410, in \_execute

return self.\_parent.execute(command, params)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 444, in execute

self.error\_handler.check\_response(response)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\errorhandler.py", line 249, in check\_response

raise exception\_class(message, screen, stacktrace)

selenium.common.exceptions.ElementClickInterceptedException: Message: element click intercepted: Element <li>...</li> is not clickable at point (319, 66). Other element would receive the click: <p>...</p>

(Session info: chrome=110.0.5481.96)

Stacktrace:

Backtrace:

(No symbol) [0x011F37D3]

(No symbol) [0x01188B81]

(No symbol) [0x0108B36D]

(No symbol) [0x010C4E3B]

(No symbol) [0x010C26DB]

(No symbol) [0x010BFD0B]

(No symbol) [0x010BE4D8]

(No symbol) [0x010B3253]

(No symbol) [0x010DB41C]

(No symbol) [0x010B2B96]

(No symbol) [0x010DB774]

(No symbol) [0x010F1215]

(No symbol) [0x010DB216]

(No symbol) [0x010B0D97]

(No symbol) [0x010B253D]

GetHandleVerifier [0x0146ABF2+2510930]

GetHandleVerifier [0x01498EC1+2700065]

GetHandleVerifier [0x0149C86C+2714828]

GetHandleVerifier [0x012A3480+645344]

(No symbol) [0x01190FD2]

(No symbol) [0x01196C68]

(No symbol) [0x01196D4B]

(No symbol) [0x011A0D6B]

BaseThreadInitThunk [0x776E00F9+25]

RtlGetAppContainerNamedObjectPath [0x77A07BBE+286]

RtlGetAppContainerNamedObjectPath [0x77A07B8E+238]

Unternehmensprofile webscrapen

* Die Stellenanzeige muss angeklickt werden (mehr Details auf der rechten Seite)
* Im neuen Fenster ist oben immer das Logo, in welchem ein href verbunden ist: Man muss auf das Logo klicken, dann öffnet sich die Unternehmensbeschreibung
* Den Path Finden: rechtklick auf das gewünscht Objekt >> untersuchen, dann rechtsklick auf die markierte Stelle im HTML Code >> kopieren >> gesamten XPATH kopieren

1. Option
   1. Bei jedem Klick auf die Stellenausschreibung (SA) wird nach der Textextraktion auf den Link zu dem Profil des Unternehmens geklickt.
   2. Dann werden die Daten gesammelt
   3. Zurück auf die Hauptseite mit nächster Stellenanzeige (loop geht weiter); vorher URL zurücksetzen nötig?
2. Option
   1. Bei jedem Klick auf die SA wird der Link zum Profil gesammelt in einer List.
   2. Dann wird jeder Link der Liste aufgerufen und die Daten extrahiert
   3. Gefahr des gebannt warden größer da viele unterschiedliche URL

10.03.23

Man kann zwischen den geöffneten Tabs switchen:

parent= driver.current\_window\_handle

# parent # 1. Geöffneter Tab

listAllTabs= driver.window\_handles # list of all tabs, sorted

driver.switch\_to.window(listAllTabs[0])

**Wichtig**:

„Weitere Jobs anzeigen“ lädt bei sehr vielen Jobs nicht mehr weitere Stellenanzeigen. Der Button ist anklickbar, aber lädt die Jobs nicht.

Damit das Problem behoben wird, muss man einen Job weiter unten anklicken

24.03.2023

* Ab >1000 Stellenanzeigen tritt das Problem auf. Lösungsansatz: alle 100 Anzeigen klickt man darauf
* Wie umsetzen?

If (printcounter==100)

Do click on the current job description

Printcounter=0

Printcounter +=1

12.04.2023: check für die ersten 2 Jobs

[java - Selenium WebDriver: clicking on elements within an SVG using XPath - Stack Overflow](https://stackoverflow.com/questions/14592213/selenium-webdriver-clicking-on-elements-within-an-svg-using-xpath)

Das Pop-up Fenster kann bei einem Job geschlossen werden, und die Infos zu dem Unternehmen können extrahiert werden. Zudem wird die Anzahl der Mitarbeiter erfasst, um die Unternehmen in der späteren Datenanalyse in Größen zu clustern (Kleinst, Klein, mittleres und Großunternehmen)

Nächster Schritt: Die Hauptseite (get\_url) bloß nicht neu laden, da neue Stellen hinzukommen können! ☹ Das verschiebt die Liste und somit alle gesammelten Inhalte.

Idee: Die Haupt-url immer offen lassen 🡪 Unternehmen in neuen Tab öffnen?

* Danach zurück zur Haupt-URL wechseln und Unternehmenstab direkt schließen (um nicht am ende 1000 Tabs offen zu haben, welches die Laufzeit enorm verlangsamen würde)
* [How to open a new tab using Selenium WebDriver in Java? - Stack Overflow](https://stackoverflow.com/questions/17547473/how-to-open-a-new-tab-using-selenium-webdriver-in-java)

24.04.2023

- IDEE ML umsetzen

26.04.2023

Das Skript ist nun auf Github hochgeladen und wird regelmäßig über GIT in VSCode aktualisiert

* Der profile\_link\_path ändert sich in regelmäßigen Abständen!!

Den Code für Robotics laufen lassen:

Probleme:

1. Bei dem Runter scrollen hört es bei 1000 Einträgen auf und der button ist nicht mehr anklickbar
2. Bei den Detailinformationen + Website: kommt nach ein Duzend Abläufen immer die Meldung „Loading took too much time“ 🡪 Problem beim Anklicken der Jobs??? Kam vorher noch nie

**09.05.2023**

selenium.common.exceptions.ElementClickInterceptedException: Message: element click intercepted: Element <li>...</li> is not clickable at point (319, 92). Other element would receive the click: <p>...</p>

neue Fehler im profile\_link\_path

beim letzten Durchgang werden alle allgemeine Informationen gesammelt(<1k) aber detaillierte Infos nur ein Duzend, und keine Infos aus der extra seite

warum wird prof0 nicht an prof angehangen, obwohl er den link richtig gefunden und sogar geprintet hat? 🡪 Antwort: es wurde schon vorher ein fehler ausgelöst und der code springt direkt in den except modus

**19.05.2023**

- Anstatt direkt auf die Seite zu springen und zurück auf die Hauptseite:

1. Erst alle Links + weiterführende Informationen sammeln

2. die Links (Datentyp List) als loop durchgehen und anklicken

Ziel : Die Seite soll nicht plötzlich weiß werden und nichts mehr laden

**Probleme aktuell:**

1. Ab 1000 Jobs kann man auf der Seite weder automatisch noch manuell runter scrollen um weitere Jobs zu laden

LinkedIn search displays a **max** of 100 results pages. For searches resulting in thousands of results, you **cannot** scrape more than the first 1000 results with a regular LinkedIn account. Sales Navigator displays 25 results per page, so with these accounts you can scrape a **max** of 2500 results. To get every result, make more specific searches and scrape each batch of results separately. 🡪 [LinkedIn automation rate limits and best practices [2023 edition] | PhantomBuster](https://phantombuster.com/blog/guides/linkedin-automation-rate-limits-2021-edition-5pFlkXZFjtku79DltwBF0M)

**Search result extraction** – LinkedIn search displays thousands of results. You can scrape up to 1000 results using a regular LinkedIn account in batches. A scraper will manage to scrape up to 2500 results by making more specific searches and scraping each batch separately. 🡪 [What is LinkedIn Scraping? Challenges and 4 Use Cases (proxyscrape.com)](https://proxyscrape.com/blog/linkedin-scraping)

 As mentioned above, you can only squeeze out a maximum of 999 jobs from a single query. 🡪 [Scaling and Scraping (linkedin.com)](https://www.linkedin.com/pulse/scaling-scraping-arjun-singh)

The current last page is start=975. It goes to a 404 page when hitting 1000.

[How to Scrape LinkedIn Legally: 15k Job Listing in Seconds (scraperapi.com)](https://www.scraperapi.com/blog/how-to-build-a-linkedin-scraper/)

[LinkedIn Search Limits (dux-soup.com)](https://www.dux-soup.com/blog/linkedin-search-limits)

1. Weiterführende Informationen werden nach ein duzend nicht geladen „took too much time“
   1. Vorschlag auf Stackoverflow [python - Web scraping LinkedIn job posts using Selenium gives repeated or empty results - Stack Overflow](https://stackoverflow.com/questions/75132137/web-scraping-linkedin-job-posts-using-selenium-gives-repeated-or-empty-results)

job\_time\_experience .append( WebDriverWait(driver, 10).until(EC.visibility\_of\_element\_located((By.CSS\_SELECTOR, 'li.jobs-unified-top-card\_\_job-insight:nth-child(1)'))).text )

job\_employees\_sector.append( WebDriverWait(driver, 10).until(EC.visibility\_of\_element\_located((By.CSS\_SELECTOR, 'li.jobs-unified-top-card\_\_job-insight:nth-child(2)'))).text )

job\_description .append( WebDriverWait(driver, 10).until(EC.visibility\_of\_element\_located((By.CSS\_SELECTOR, 'div.jobs-description'))).text )

1. Fehler im Aufbau der Loop bei den Seitenswitch: Meist (NICHT IMMER) muss man einen Kasten wegklicken, um an die Infos zu komen.

ZU 1. Das Problem kann nicht gelöst oder umgangen werden. Es tritt immer auf.

Filter einsetzen:

1. Berufserfahrung
   1. Berufseinstieg

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_E=2&position=1&pageNum=0>

+ehrenamt

https://de.linkedin.com/jobs/search?keywords=Analyst&location=Deutschland&locationId=&geoId=101282230&f\_TPR=&f\_JT=V&f\_E=2&position=1&pageNum=0

* 1. Berufserfahren

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_E=3&position=1&pageNum=0>

* 1. Management

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_E=4&position=1&pageNum=0>

* 1. Direktor:in VP

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_E=5&position=1&pageNum=0>

Vor Ort

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_WT=1&position=1&pageNum=0>

Hybrid

<https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_WT=3&position=1&pageNum=0>

Remote

https://de.linkedin.com/jobs/search?keywords=Data%20Science&location=Deutschland&locationId=&geoId=101282230&f\_TPR=&f\_WT=2&position=1&pageNum=0

Die Meisten Jobs sind:

* Vor Ort
* Management
* Vollzeit

Sehr wenig bei:

* Remote

Ort wählen – Filter Schwierig da bei jeder suche anders

<https://de.linkedin.com/jobs/search?keywords=Analyst&location=Deutschland&locationId=&geoId=101282230&f_TPR=&f_PP=100477049&position=1&pageNum=0>

<https://de.linkedin.com/jobs/search?keywords=Analyst&location=Deutschland&geoId=101282230&f_PP=106967730&currentJobId=3592873412&position=9&pageNum=0>

Filtern mit den 3 Möglichkeiten ergeben 3x5x5 =75 Kombinationen

Diese kann man reduzieren indem man ungewöhnliche Berufskombinationen im Vorfeld auslässt: z.B Direktor:in als Ehrenamt oder Ausbildung in Zeitarbeit.

Im -Bereich Ort wird nichts rausgenommen aber bei Art der Beschäftigung und Berufserfahrung:

* Es folgt aus der Tabelle, dass von den 5x5=25 Kombinationen
* 11 können rausgenommen werden, also nur **14**
* Oder Ehrenamt rausnehmen 🡪 aus 20 kombis nimmt man 7 raus, also **13** kombis

Letzteres ergibt also 13\*3=39 Kombinationen

ODER: Nur Vollzeit… Dann 3\*5=15 Kombinationen

Techliste=[cloud , it, robotics, blabla]

Ortliste= [1,2,3]

experienceListe=[1,2,3,4,5]

For i in techListe:

For j in ortListe:

For k in experienceListe:

**Alle Code-Parts in Funktionen schreiben 🡪 meist schnellere Laufzeit, wenn man so for-loops verhindern kann**

**25.05.2023**

* **Basic informationen als Code – Done**
* **Detailed information 🡪 write as function, then excecute fun(0:10) and fun(10:)**
* **Profil description 🡪 Check for duplicates before looping, because one firm might have multiple job offers**

Authentification wall wenn ich auf die Profile will…. Driver.back und wieder profil.click() hilft nicht wirklich… was tun? Neues Tab öffnen?

**30.05.2023**

Basic information +detailed information läuft mit gut… ggf. den Gesamtcode umstellen sodass das Skript in einem durchläuft.

Profilbeschreibung: Neues Tab öffnen löst das Problem nicht 🡪 Authentification WALL

Notiz: Händisch lassen sich die Links in neuen Tabs öffnen (mit Rechsklick in neuem Tab öffnen) ohne dass man geblockt wird.

01.06.2023

🡪 neuen Browser benutzen?

* Scrollen + 10 neue Tabs über rrechtsklick öffnen…

**29.06.2023**

* Strg + click auf den Link, öffnet einen Link im neuen Tab -> Testen

from selenium.webdriver.common.keys import Keys

from selenium.webdriver.common.action\_chains import ActionChains

profile\_click = WebDriverWait(driver= driver, timeout=10).until(EC.presence\_of\_element\_located((By.XPATH, '/html/body/div[1]/div/section/div[2]/section/div/a')))

    actions = ActionChains(driver)

    actions.key\_down(Keys.CONTROL)

    actions.click(on\_element=profile\_click)

    actions.perform()

    tabs = driver.window\_handles

    driver.switch\_to.window(tabs[1])

* Umsetzung klappt; Nun schauen ob die Authentification Wall auch hier auftritt

**03.07.2023**

**🡪** Wenn detailed Information ausgeführt werden; kommt die Meldung „Loading took too much time“ ; Meldung verschwindet, wenn man auf der Seite einige Stellenanzeigen „händisch“ antippt, sodass LinkedIn erkennt, dass es sich nicht um einen Roboter/Code handelt..

Wie kann man das Phänomen als Code lösen, sodass man nicht dauerhaft überwachen muss?

* Job Profilseiten:

1. Testen des Codeblocks auf Funktionalität
   1. Done: Funktioniert für 10 Unternehmen, keine Dopplungen oder leere Einträge
2. Testen auf Skalierung: Kommen Fehler auf, wenn viele Tabs geöffnet werden? Authentification Wall?
   1. 🡪 erster Test sah gut aus!
3. Ab ca. 10 Tabs werden alle außer das erste geschlossen. Dann wieder 10 neue nacheinander geöffnet etc.
4. Wie erreiche ich jedes Unternehmen nur einmal?! Link unique passt, aber selenium webelement auch gleich??

**04.07.2023**

* Tabs lassen sich korrekt öffnen, jedoch wurden Tabs nicht begrenzt (Tabs nach max 10 Stück werden geschlossen), wodurch eine Fehlermeldung kam
* Vermutung: fehlender Verweis zurück zum Jobposting Tab
* Test Codeblock für das schließen der Tabs in dem Loop
* Matrix erstellt. Webelemente sind alle unique, Links haben beim gleichen Unternehmen Dopplungen. Daher Duplikate (alle Zeilen außer das erste Vorkommen des Duplikats) entfernt

pop-up

Traceback (most recent call last):

File "<stdin>", line 28, in <module>

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 861, in find\_element

return self.execute(Command.FIND\_ELEMENT, {"using": by, "value": value})["value"]

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 444, in execute

self.error\_handler.check\_response(response)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\errorhandler.py", line 249, in check\_response

raise exception\_class(message, screen, stacktrace)

selenium.common.exceptions.NoSuchElementException: Message: no such element: Unable to locate element: {"method":"xpath","selector":"//\*[@id="main-content"]/section[1]/div/section[1]/div/p"}

(Session info: chrome=114.0.5735.199)

During handling of the above exception, another exception occurred:

Traceback (most recent call last):

File "<stdin>", line 39, in <module>

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 861, in find\_element

return self.execute(Command.FIND\_ELEMENT, {"using": by, "value": value})["value"]

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 444, in execute

self.error\_handler.check\_response(response)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\errorhandler.py", line 249, in check\_response

raise exception\_class(message, screen, stacktrace)

selenium.common.exceptions.NoSuchElementException: Message: no such element: Unable to locate element: {"method":"xpath","selector":"//\*[@id="main-content"]/section[1]/div/section[1]/div/p"}

(Session info: chrome=114.0.5735.199)

* Fehlermeldung trat auf, da Unternehmen keine Beschreibung auf seiner Seite hat 🡪 HTML Snippet existiert nicht und es wird ein Fehler ausgegeben
  + **Daraus folgt:** Exception einbauen, wenn keine Beschreibung vorhanden – Done
  + //\*[@id="main-content"]/section[1]/div/section[1]
* Das Webscrapen der Unternehmensbeschreibungen funktioniert-🡪 aber beim Durchlaufen passiert das:
* Ein Bild, das Text, Screenshot, Software, Computersymbol enthält.

  Automatisch generierte Beschreibung
  + Seite Neu Laden hilft nicht
* Nicht alle Unternehmen haben eine eigene Profilseite auf LinkedIn: Klick auf das Logo führt zur Fehlermeldung
  + Exception für diesen Fall einführen
  + Prüfen wieviele es im Vergleich zur Gesamtheit sind
* **Wichtig: Webelement und die Results( Beschreibung & Company Size) müssen zusammen gespeichert werden, um die Daten später mit dem anderen Dataframe zu joinen**
* **Neuer Fehler**

File "<stdin>", line 79, in <module>

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\common\action\_chains.py", line 78, in perform

self.w3c\_actions.perform()

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\common\actions\action\_builder.py", line 91, in perform self.driver.execute(Command.W3C\_ACTIONS, enc)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webdriver.py", line 444, in execute

self.error\_handler.check\_response(response)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\errorhandler.py", line 249, in check\_response

raise exception\_class(message, screen, stacktrace)

**selenium.common.exceptions.ElementNotInteractableException: Message: element not interactable: https://de.linkedin.com/company/topjobs-aktuell?trk=public\_jobs\_topcard\_logo has no size and location**

(Session info: chrome=114.0.5735.199)

* Error is at action.perform(), because the right side of the website with the detailed information is not loaded. Only appear with manually movements around the page, refresh does not work
* Immernoch viele Dopplungen bei der detailed description und dadurch auch bei den Firmen, weil die URL aus der detailed description extrahiert werden

**12.07.2023**

**🡪 Tracking:** die temporären Variablen wie jd0, emp\_type0 etc werden nach jedem job posting gelöscht. Dadurch werden Dopplungen im Datensatz vermieden und diese Fehlerhaften Stellen werden Schneller entdeckt.

* Nun kommen bei nicht geladenen Detailed Description ein “None” 🡪 bessere Sichtbarkeit zum Cleanen der Daten
* Check wieviel der Daten das ausmacht

**🡪 Performance:** Detailed Description braucht 1h! IDEEN

* Mehrere Fenster gleichzeitig ausführen
* Multithreading

*# PIPELINE*

*# 1. DEFINE YOUR FULL LIST OF DATA*

list\_of\_data = []

*# 2. DEFINE FUNCTION*

def **your\_function**(data):

*# Write your function code here*

*# 3. EXECUTE THE FUNCTION ON THREADS*

with concurrent.futures.ThreadExecutor(max\_worker=NUM\_THREADS) as executor:

*# This function with map each chunk on available threads and apply the function on each chunk*

executor.map(your\_function, list\_of\_data)

**13.07.2023**

ToDo:

1. Test Multithreading 🡪 Maybe implement for all 3 parts (Basic\_info;detailed\_indo and company profiles

2. Split the file into multiple files by functions + one main file where the functions are loaded and the code is executed

>>> import threading

>>> import time

>>> def mythread():

... time.sleep(1000)

...

>>> def main():

... threads = 0 #thread counter

... y = 1000000 #a MILLION of 'em!

... for i in range(y):

... try:

... x = threading.Thread(target=mythread, daemon=True)

... threads += 1 #thread counter

... x.start() #start each thread

... except RuntimeError: #too many throws a RuntimeError

... break

... print("{} threads created.\n".format(threads))

...

>>> if \_\_name\_\_ == "\_\_main\_\_":

... main()

...

**161264 threads created. 🡪 Laptop daraufhin abgestürzt**

**Zwischenfazit:**

-Daten werden fehlerhaft gesammelt. Grund: Die Seite springt durch das Parallelisieren zu schnell.

- Threading sogar für eine deutlich kürzere Laufzeit (5 Minuten anstatt 60 bei 30 Threads)

**18.07.2023**

**-** Multithreading macht auf bei einem Chrome Tab keinen Sinn, da auf einer Seite nicht parallel die Aufgaben abgearbeitet werden können, sondern nur sequentiell (da man immer auf die jobs klicken muss)

- Neuer Ansatz: Mehrere Fenster öffnen (2-3) über mehrere Treiber und dort die Anzahl an Jobs aufteilen;

Arbeitspakete im code

1. Scroll through jobs
2. Basic Information from left side panel
3. Detailed Information
4. Profile information

Vorgehen mit mehreren Fenstern:

1. Die unveränderbaren Variablen werden festgelegt (location etc.) CHECK
2. Beide Treiber werden initialisiert CHECK- beliebig viele, MAX 8 (da 8 Kerne)
3. Beide Treiber finden die jobs\_num und scrollen durch ALLE Jobs bis zum Ende
4. Jeder Treiber findet die Var „jobs“ in seinem Tab bzw Fenster
5. Ich habe 2 verschiedene Listen an „jobs“ 🡪 Check ob beide identical==True 🡪 JA CHECK ERFOLGREICH, Sollte so sein, sodass Reihenfolge bei gleicher URL identisch istt.
6. ~~Eine Var Jobs löschen~~,da beide gleich; Scrambled function auf Var Jobs anwenden
7. Basic Info für einen Driver ausführen
8. Detailed\_function: die alte Function nehmen (detail\_info); start und ende so definieren
9. Rand\_jobs aufteilen: chunks = [data[x:x+100] for x in range(0, len(data), 100)]

19.07.2023

Um zu Parallelisieren: Input : die verschiedenen Driver

Output: Die Datasets die anschließend zusammengeführt werden

* Öffnen der Browser und Scrolle bis zum Ende SEQUENTIELL
* Basic Info (linke Spalte) wird nur für ein Fenster komplett ausgeführt. (dauert 1min)
* Detailed Info (rechte Spalte) wird auch die Anzahl an Fenstern/Driver aufgeteilt

Zeile 186

        num= jobs\_ordered[0].index(item) # not rand\_jobs, because the order changed there!

Wie ruft die Parallelize Function die Tuples richtig auf?

Webelements als input in das Parallisieren einsetzten?? Webelement können nicht „unpickled“ werden, was voraussetzung für das Modul ist!

21.07.2023

Große Änderung – Neue Herangehensweise:

* Kein Parallelisieren der Detailed Job Function
* Multithreading der Suchanfragen +Berufserfahrung filter

d.h.

tuples =[(Cloud, Vollzeit), (Cloud, Teilzeit),…,(AI, Vollzeit),…]

orte=[remote, hybrid, office]

For ( tuple in tuples):

For ort in orte:

Scrape like before

21.07.2023

Excecution time for profiles: 0:13:39.701188

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "C:\Users\km\Desktop\repo\functions\company\_profiles.py", line 167, in scrape\_profiles

df\_profiles= pd.DataFrame({

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\frame.py", line 636, in \_\_init\_\_

mgr = dict\_to\_mgr(data, index, columns, dtype=dtype, copy=copy, typ=manager)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 502, in dict\_to\_mgr

return arrays\_to\_mgr(arrays, columns, index, dtype=dtype, typ=typ, consolidate=copy)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 120, in arrays\_to\_mgr

index = \_extract\_index(arrays)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 688, in \_extract\_index

raise ValueError(msg)

ValueError: array length 155 does not match index length 158

Bootstrapping

1. Deutschlandkarte digitaljobs

* Für KMU
* Für Vollzeit

1. Topic modeling /NLP , übersetzen , wenn englisch

Next Steps:

1. Marie schaut nach passenden Wörter für Mitarbeiterfähigkeitn

31.07.2023

Basic und detail\_info laufen nun, die Zwischenergebnisse werden gespeichert.

Ein Fehler kam 7-8 Mal zwischendrin auf, als die Unternehmensprofile gewebscrapt wurden: Die Seite war blockiert und es kam folgende Meldung am Terminal: (der Code wurde dadurch nicht unterbrochen)

[12292:4576:0731/111651.667:ERROR:interface\_endpoint\_client.cc(702)] Message 0 rejected by interface blink.mojom.WidgetHost

Excecution time for profiles: 0:34:58.504872

Traceback (most recent call last):

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\indexes\base.py", line 3621, in get\_loc

return self.\_engine.get\_loc(casted\_key)

File "pandas\\_libs\index.pyx", line 136, in pandas.\_libs.index.IndexEngine.get\_loc

File "pandas\\_libs\index.pyx", line 163, in pandas.\_libs.index.IndexEngine.get\_loc

File "pandas\\_libs\hashtable\_class\_helper.pxi", line 5198, in pandas.\_libs.hashtable.PyObjectHashTable.get\_item

File "pandas\\_libs\hashtable\_class\_helper.pxi", line 5206, in pandas.\_libs.hashtable.PyObjectHashTable.get\_item

KeyError: 0

The above exception was the direct cause of the following exception:

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "C:\Users\km\Desktop\repo\functions\company\_profiles.py", line 170, in scrape\_profiles

links= unique[0]

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\frame.py", line 3505, in \_\_getitem\_\_

indexer = self.columns.get\_loc(key)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\indexes\base.py", line 3623, in get\_loc

raise KeyError(key) from err

KeyError: 0

Während des scrollen:

[19880:19884:0731/151148.509:ERROR:device\_event\_log\_impl.cc(222)] [15:11:48.509] USB: usb\_service\_win.cc:307 SetupDiEnumDeviceInterfaces: Es sind keine Daten mehr verf³gbar. (0x103)

[19880:19884:0731/151149.896:ERROR:device\_event\_log\_impl.cc(222)] [15:11:49.896] USB: usb\_service\_win.cc:104 SetupDiGetDeviceProperty({{A45C254E-DF1C-4EFD-8020-67D146A850E0}, 6}) failed: Element nicht gefunden. (0x490)

01.08.2023

swesen\n ', '\n Geschäftsentwicklung und Vertrieb\n ', '\n

Marketing und Vertrieb\n ', '\n Marketing und Vertrieb\n ', '\n

Marketing und Vertrieb\n ', '\n Produktmanagement und Marketing\n

', '\n Vertrieb und Geschäftsentwicklung\n ', '\n Design, Art/reative und IT\n ', '\n Sonstiges\n ', '\n Projektmanagemnt, IT und Produktmanagement\n ', '\n Sonstiges\n ', '\n arketing und Vertrieb\n ', '\n Strategie & Planung\n ', '\n

Projektmanagement\n ', '\n Geschäftsentwicklung und Unternehmensberatung\n

', '\n Produktmanagement und Marketing\n ', '\n Vertrieb, erbung und Geschäftsentwicklung\n ', '\n Marketing und Vertrieb\n ',

'\n Öffentlichkeitsarbeit/PR, Schreiben/Redaktion und Marketing\n ', '\n

Ingenieurwesen und IT\n ', '\n Marketing und Vertrieb\n ', '\n

Geschäftsentwicklung\n ', '\n Sonstiges\n ', '\n

Marketing und Vertrieb\n ', '\n Personalwesen\n ', '\n eschäftsentwicklung und Unternehmensberatung\n ', '\n Vertrieb, Werbung und Gschäftsentwicklung\n ', '\n Vertrieb\n ', '\n Finanzwesen

und Wirtschaftsprüfung/Buchhaltung\n ', '\n Marketing und Vertrieb\n

', '\n Logistik\n ', '\n IT\n ', '\n Produktmaagement und Marketing\n ', '\n Marketing\n ', '\n Marketig, Management und Sonstiges\n ', '\n Sonstiges\n ', '\n Mrketing und Vertrieb\n ', '\n Art/Creative und Strategie & Planung\n

', '\n Personalwesen\n ', '\n Marketing und Wissenschaft und Bildug\n ', '\n Marketing\n ', '\n Projektmanagement\n

', '\n IT\n ', '\n Vertrieb, Werbung und Geschäftsentwicklung\n

', '\n Marketing und Vertrieb\n ', '\n IT\n ', '\n

Geschäftsentwicklung und Vertrieb\n ', '\n Marketing\n ',

'\n Werbung und Marketing\n ', '\n Vertrieb und Kundenservice\n

', '\n Produktmanagement und Marketing\n ', '\n Design\n

', '\n Sonstiges\n ', '\n Marketing und Vertrieb\n ' '\n Ingenieurwesen\n ', '\n Sonstiges\n ', '\n

Geschäftsentwicklung und Vertrieb\n ', '\n Vertrieb, Werbung und Geschäftsenwicklung\n ', '\n Sonstiges, IT und Management\n ', '\n Lgistik\n ', '\n Vertrieb und Geschäftsentwicklung\n ', '\n

Finanzwesen\n ', '\n Management und Fertigung\n ', '\n Pojektmanagement und IT\n ', '\n Sonstiges\n ', '\n Qualittssicherung, Strategie & Planung und Produktmanagement\n ', '\n Sonstiges\n

', '\n Design, Art/Creative und IT\n ', '\n Geschäftsentwiclung und Produktmanagement\n ', '\n Qualitätssicherung, Strategie & Planung ud Produktmanagement\n ', '\n Sonstiges\n ', '\n Sonstigesn ', '\n Finanzwesen und Wirtschaftsprüfung/Buchhaltung\n ', '\n

Produktmanagement und Marketing\n ', '\n Marketing und Vertrieb\n

', '\n Marketing\n ', 'Marketing, Öffentlichkeitsarbeit/PR und Schreiben/edaktion', '\n Marketing, Öffentlichkeitsarbeit/PR und Schreiben/Redaktion\n , '\n Marketing und Vertrieb\n ', '\n Marketing und Vertrieb\n

', '\n Vertrieb\n ', '\n Produktmanagement und Marketing\n

', '\n Marketing\n ', '\n Personalwesen\n ', '\n

Geschäftsentwicklung und Vertrieb\n ', '\n Marketing und Vertrieb\n

', '\n Marketing und Vertrieb\n ', '\n Personalwesen\n

', '\n Projektmanagement\n ', '\n Vertrieb, Werbung und Geschätsentwicklung\n ', '\n Verwaltung\n ', '\n Schreiben/Redation\n ', '\n Vertrieb\n ', '\n Produktmanagement und Mareting\n ', '\n Produktmanagement und Marketing\n ', '\n Poduktmanagement und Marketing\n ', '\n Ingenieurwesen und IT\n ', '\ Marketing, Öffentlichkeitsarbeit/PR und Schreiben/Redaktion\n ', '\n

Sonstiges\n ', '\n Sonstiges\n ', '\n Vertrieb und Geshäftsentwicklung\n ', '\n Werbung und Marketing\n ', '\n irtschaftsprüfung/Buchhaltung\n ', '\n Projektmanagement, IT und Produktmanagment\n ', '\n Finanzwesen und Wirtschaftsprüfung/Buchhaltung\n ', '\ Marketing und Vertrieb\n ', '\n Marketing und Vertrieb\n

', '\n IT\n ', '\n Qualitätssicherung, Strategie & Planung und Prouktmanagement\n ', '\n Marketing und Vertrieb\n ', '\n Fianzwesen und Wirtschaftsprüfung/Buchhaltung\n ', '\n Distribution\n , '\n Sonstiges\n ', '\n Marketing\n ', '\n Vewaltung und Sonstiges\n ', '\n Einkauf und Logistik\n ', '\n

Produktmanagement und Marketing\n ', '\n IT\n ', '\n Srategie & Planung, Produktmanagement und Projektmanagement\n ', '\n Geschäftsntwicklung und Vertrieb\n ', '\n Sonstiges\n ', '\n Vertreb, Werbung und Geschäftsentwicklung\n ', '\n Management und Fertigung\n

', '\n Management und Fertigung\n ', '\n IT\n ', '\n

Personalwesen und Verwaltung\n ', '\n Produktmanagement und Marketig\n ', '\n Sonstiges\n ', '\n Finanzwesen und Wirtschaftsrüfung/Buchhaltung\n ', '\n Unternehmensberatung, IT und Personalwesen\n

', '\n Distribution und Logistik\n ', '\n IT\n ', '\n

Marketing und Vertrieb\n ', '\n Geschäftsentwicklung und Vertrieb\ ', '\n Schreiben/Redaktion, Rechtswesen und Projektmanagement\n ']

Traceback (most recent call last):

File "<stdin>", line 3, in <module>

File "<stdin>", line 60, in page\_webscraping

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\frame.p", line 636, in \_\_init\_\_

mgr = dict\_to\_mgr(data, index, columns, dtype=dtype, copy=copy, typ=manager)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internas\construction.py", line 502, in dict\_to\_mgr

return arrays\_to\_mgr(arrays, columns, index, dtype=dtype, typ=typ, consolidate=copy)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 120, in arrays\_to\_mgr

index = \_extract\_index(arrays)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 688, in \_extract\_index

raise ValueError(msg)

ValueError: array length 655 does not match index length 656

--------------------------------------------------------------------\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Guest Laptop :

* Pop up case ist aufgetreten 🡪 ist auf dem jeweiligen Tab geblieben und nicht zurück zm ersten Tab –> Loading too much errors im Anschluss

02.08.2023

8:48 GuestKM

1h 45 min für eine Kombination

1,75\*240= 420h entspricht 52,8 Tage

gement und Fertigung\n ', '\n Verwaltung\n ', '\n Verwaltung\n ', '\n Vertrieb und Geschäftsentwicklung\n ']

Traceback (most recent call last):

File "<stdin>", line 2, in <module>

File "<stdin>", line 60, in page\_webscraping

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\frame.py", line 636, in \_\_init\_\_

mgr = dict\_to\_mgr(data, index, columns, dtype=dtype, copy=copy, typ=manager)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 502, in dict\_to\_mgr

return arrays\_to\_mgr(arrays, columns, index, dtype=dtype, typ=typ, consolidate=copy)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 120, in arrays\_to\_mgr

index = \_extract\_index(arrays)

File "C:\Users\km\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\core\internals\construction.py", line 688, in \_extract\_index

raise ValueError(msg)

ValueError: array length 500 does not match index length 656

>>> [4672:1636:0802/092326.961:ERROR:ssl\_client\_socket\_impl.cc(980)] handshake failed; returned -1, SSL error code 1, net\_error -101

[4672:1636:0802/092326.972:ERROR:ssl\_client\_socket\_impl.cc(980)] handshake failed; returned -1, SSL error code 1, net\_error -101