**LAPORAN**

**RENCANA TUGAS MANDIRI (RTM) Ke-2**

**MATA KULIAH BIG DATA**

**“Web Scraping/Web Crawling dari Suatu Situs”**



**DISUSUN OLEH:**

Chelsea Ayu Adhigiadany ( 21083010028 )

**DOSEN PENGAMPU:**

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

**PROGRAM STUDI SAINS DATA**

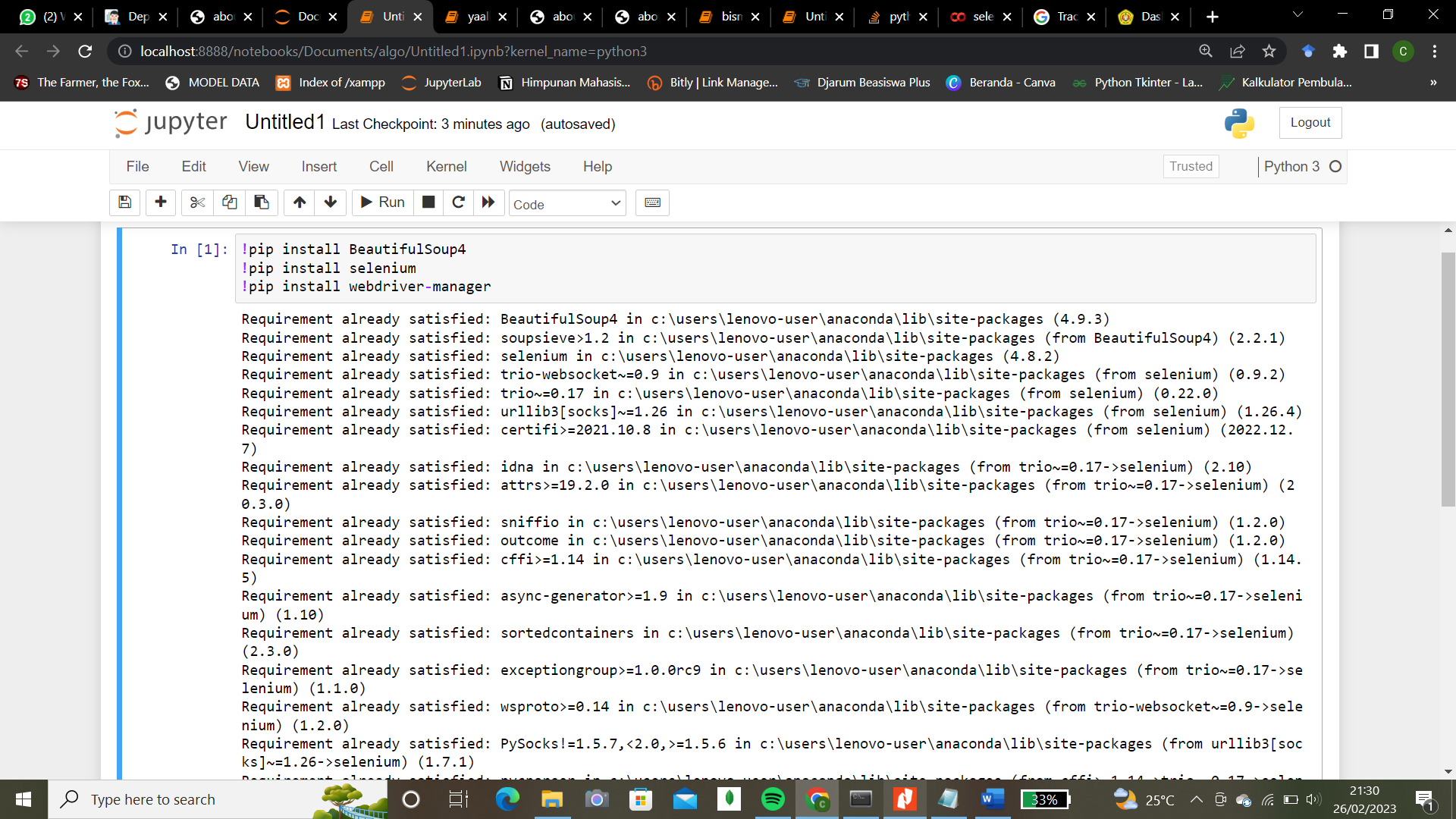
**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR**

**2022**

**Dokumentasi Web Scraping**

1. Install package



1. Tulis script dibawah untuk membuat modul, kemudian run

Script :

from bs4 import BeautifulSoup

from selenium import webdriver

from webdriver\_manager.chrome import ChromeDriverManager

chrome\_options = webdriver.ChromeOptions()

chrome\_options.add\_argument("--headless")

driver = webdriver.Chrome(ChromeDriverManager().install(), chrome\_options=chrome\_options)

#query to obtain links

query = 'Joko Widodo'

#links = ..

#specifi

n\_pages = 10

for page in range(1, n\_pages):

url = "https://www.google.com/search?q=" + query + "&start=" + str((page-1) \* 10)

driver.get(url)

soup = BeautifulSoup(driver.page\_source, 'html.parser')

#soup...

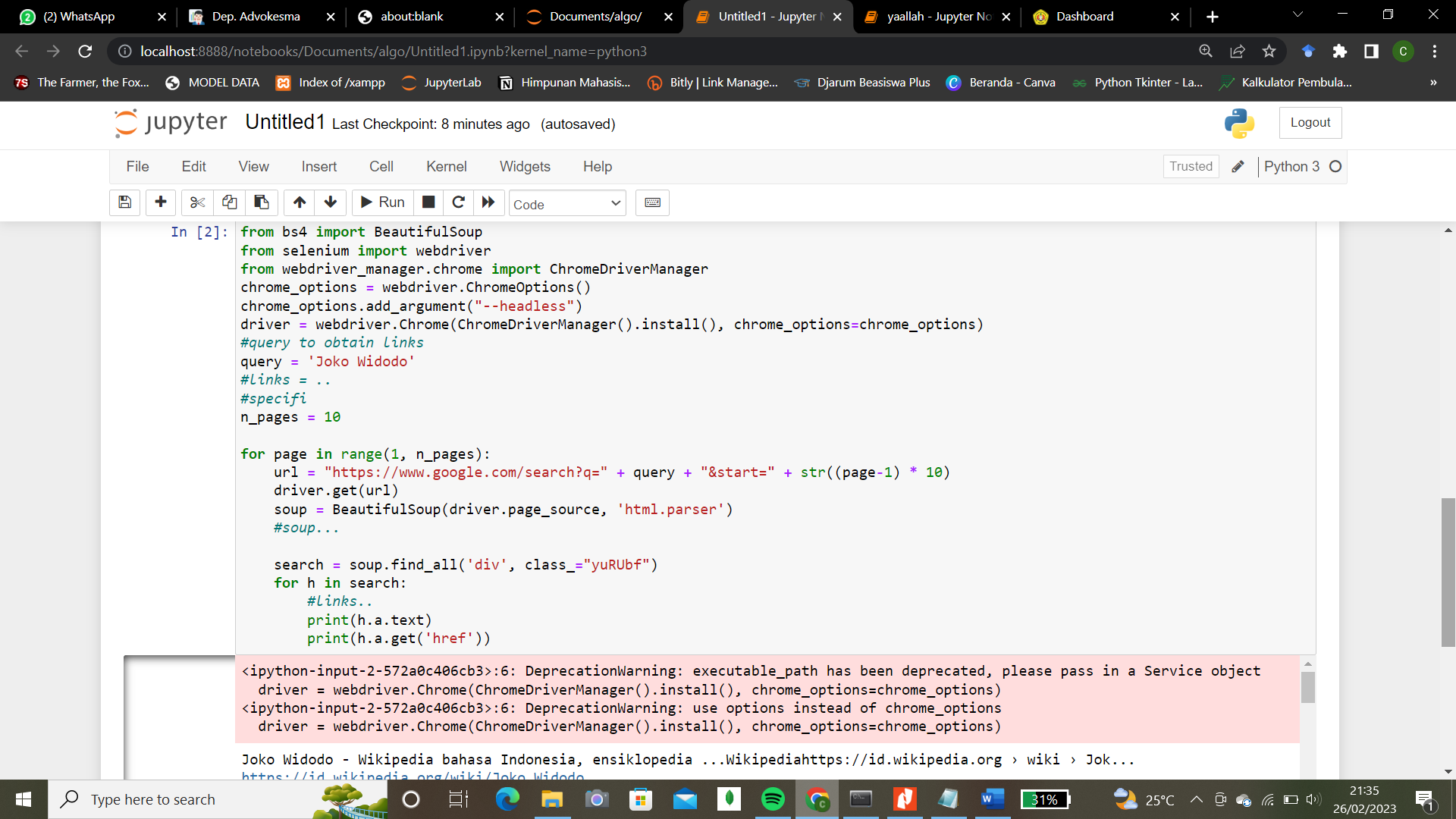
search = soup.find\_all('div', class\_="yuRUbf")

for h in search:

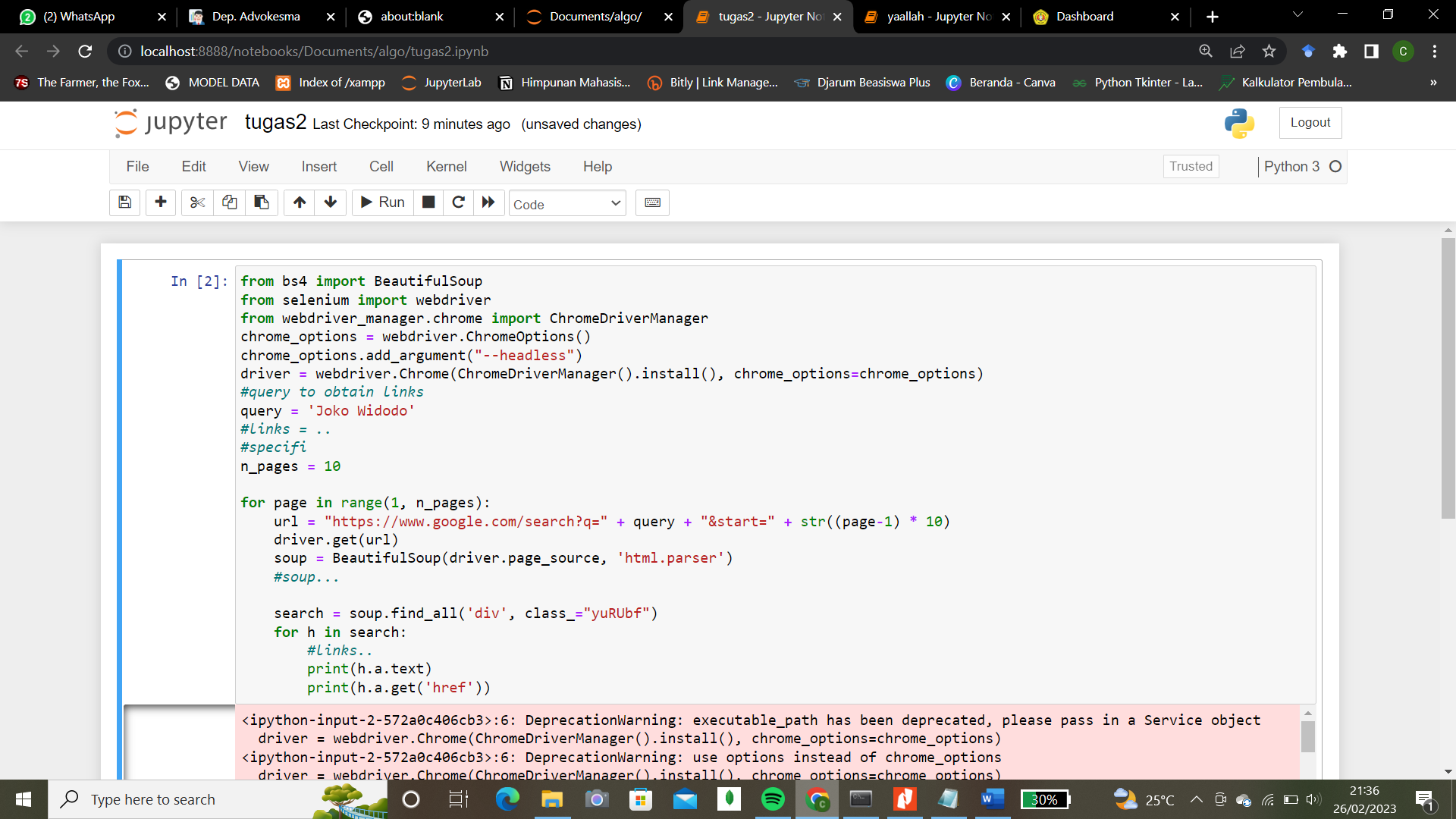
#links..

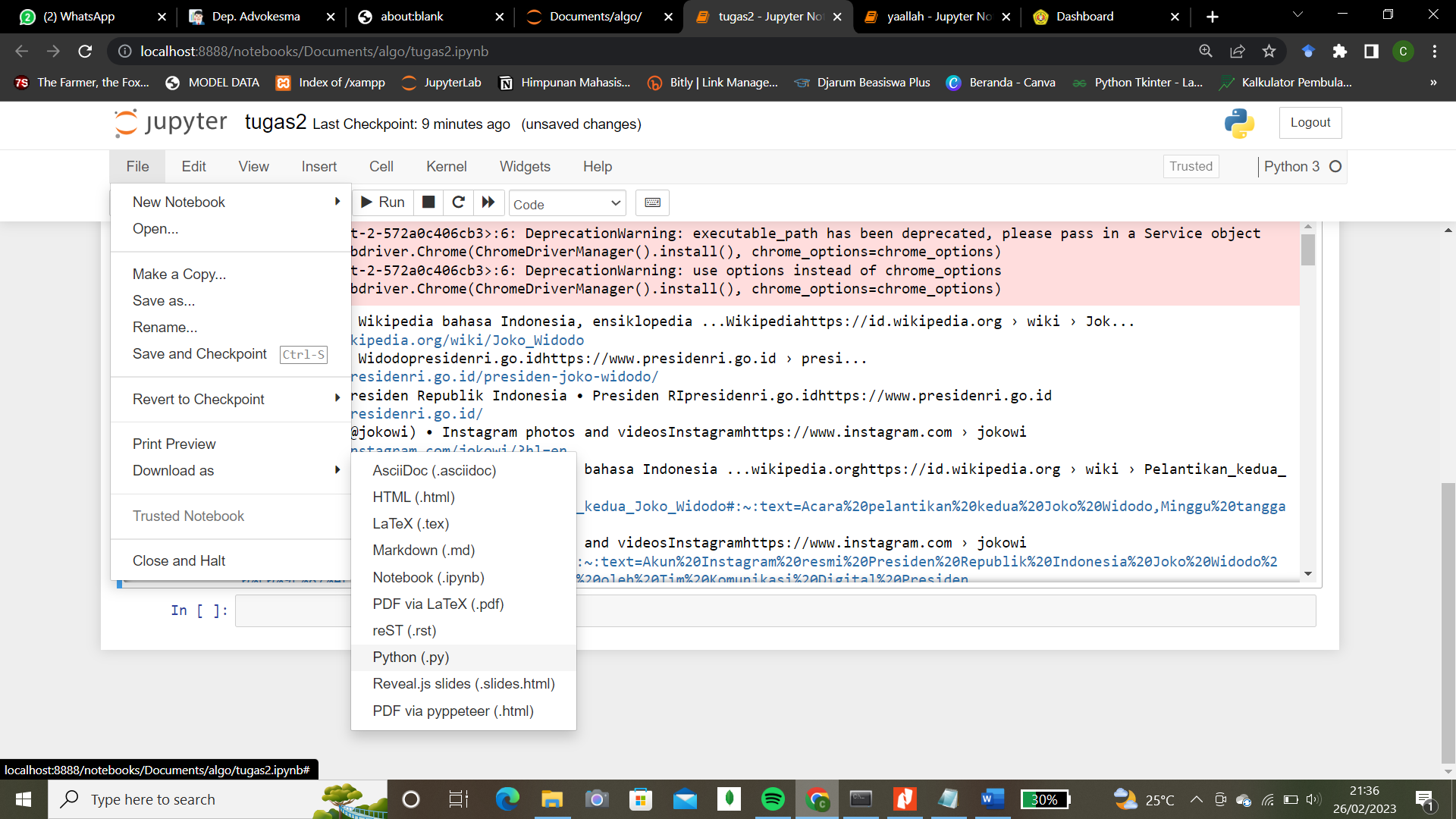
print(h.a.text)

print(h.a.get('href'))

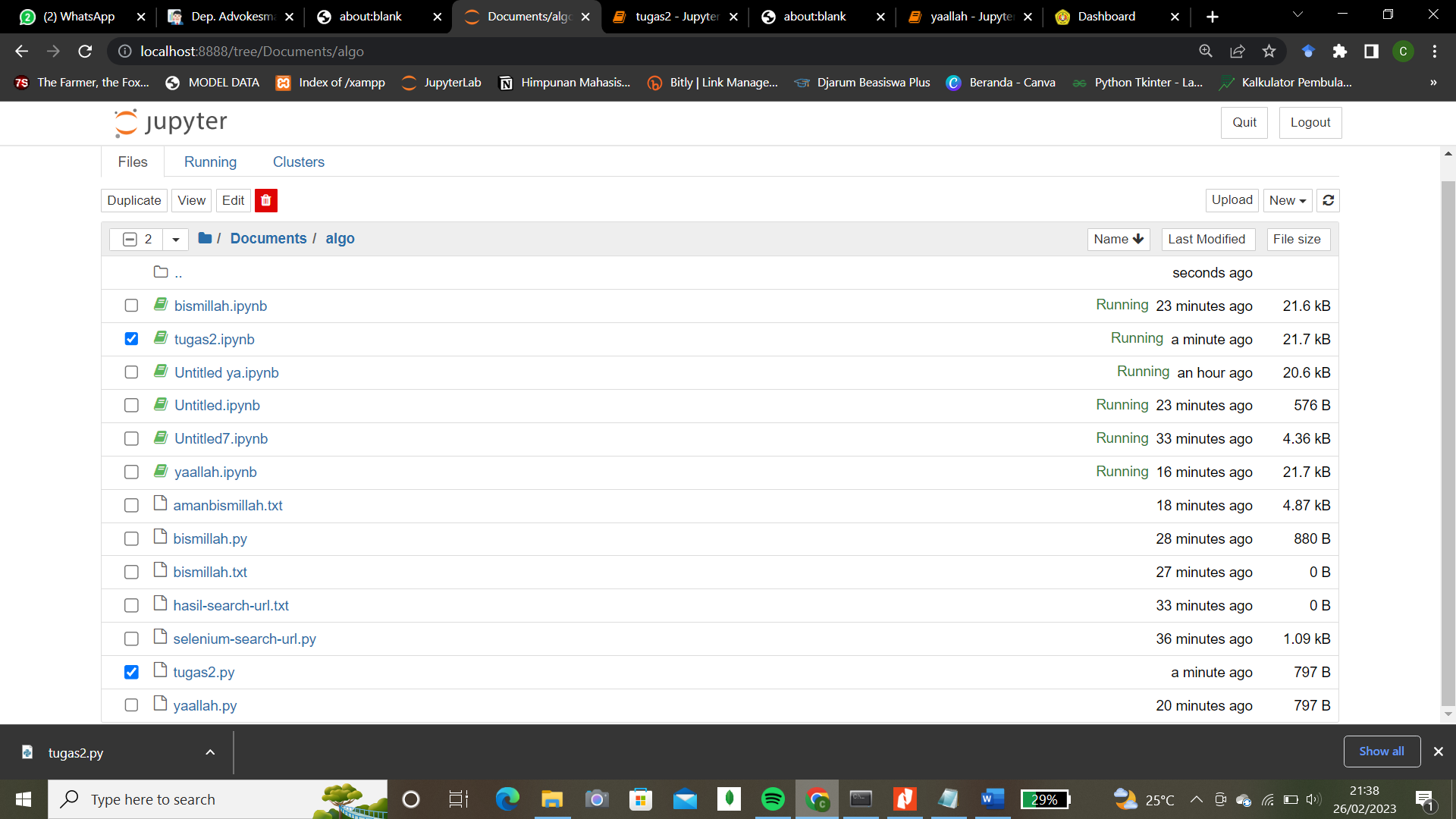


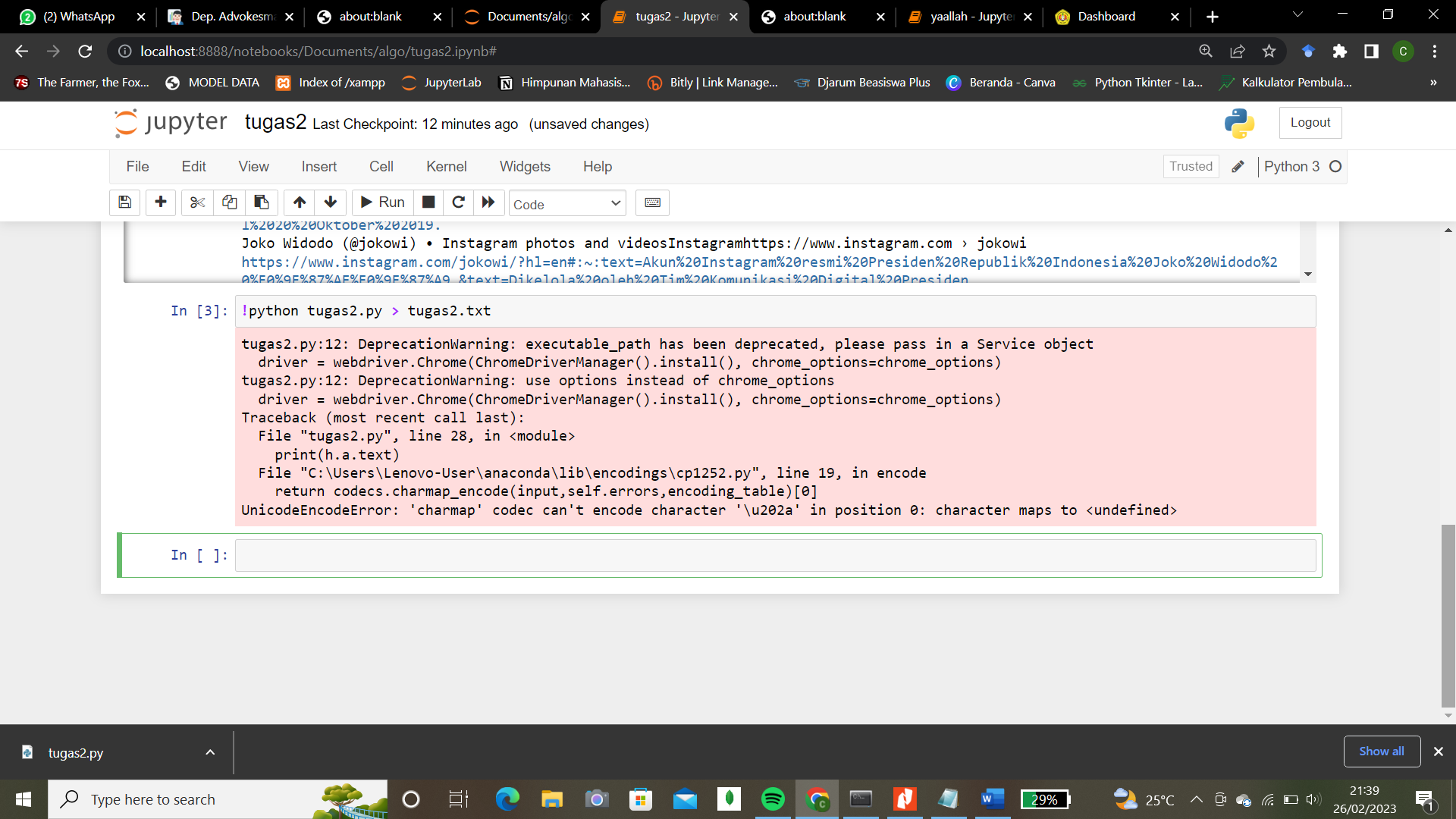
1. Setelah berhasil di run, hapus bagian install package dan download as .py



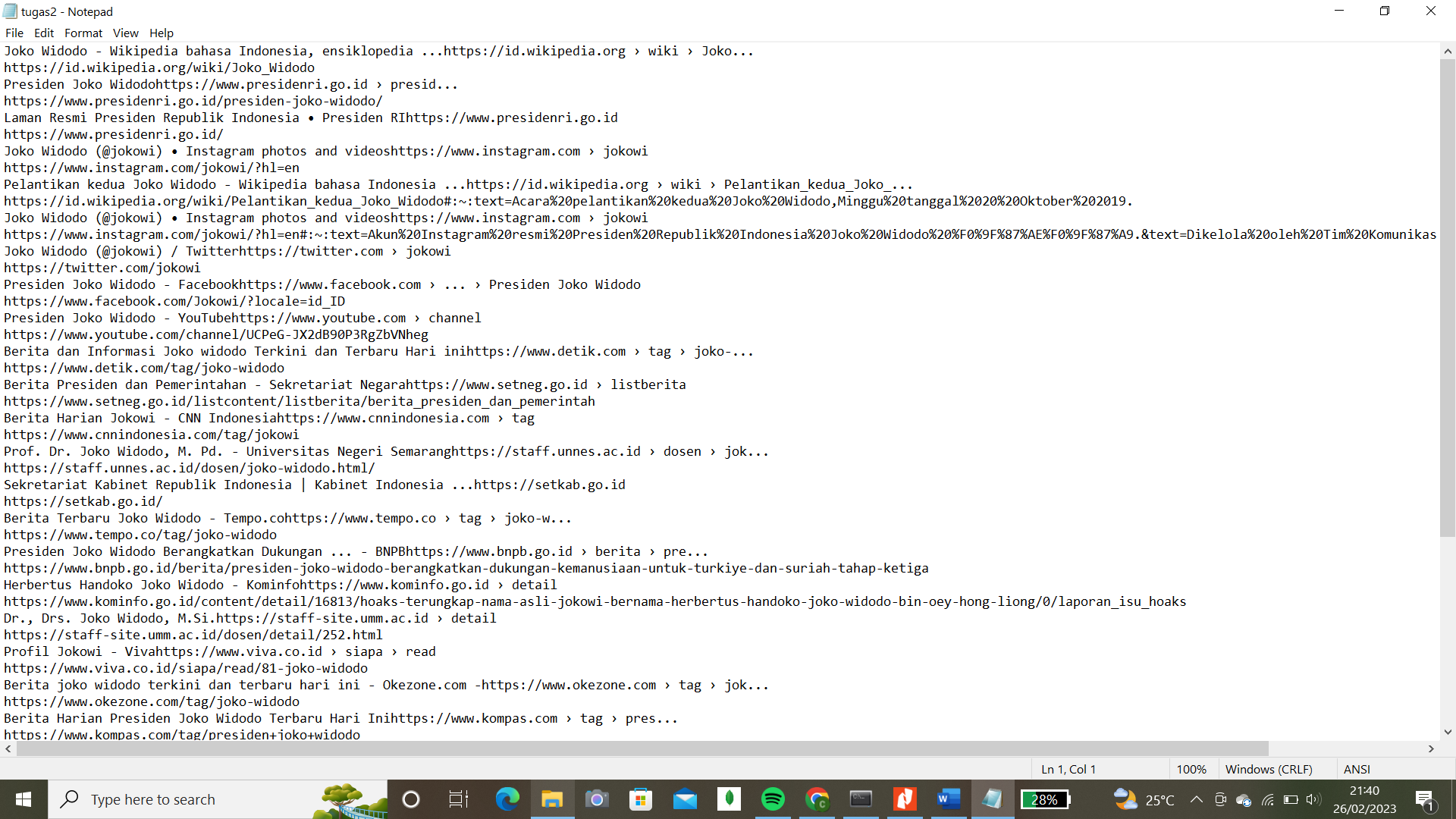


1. Tuliskan !python namafilemodule.py > namafiletxttujuan.txt, misal !python tugas2.py > tugas2.txt kemudian run. Pastikan file module dan file ipynb berada pada satu tempat

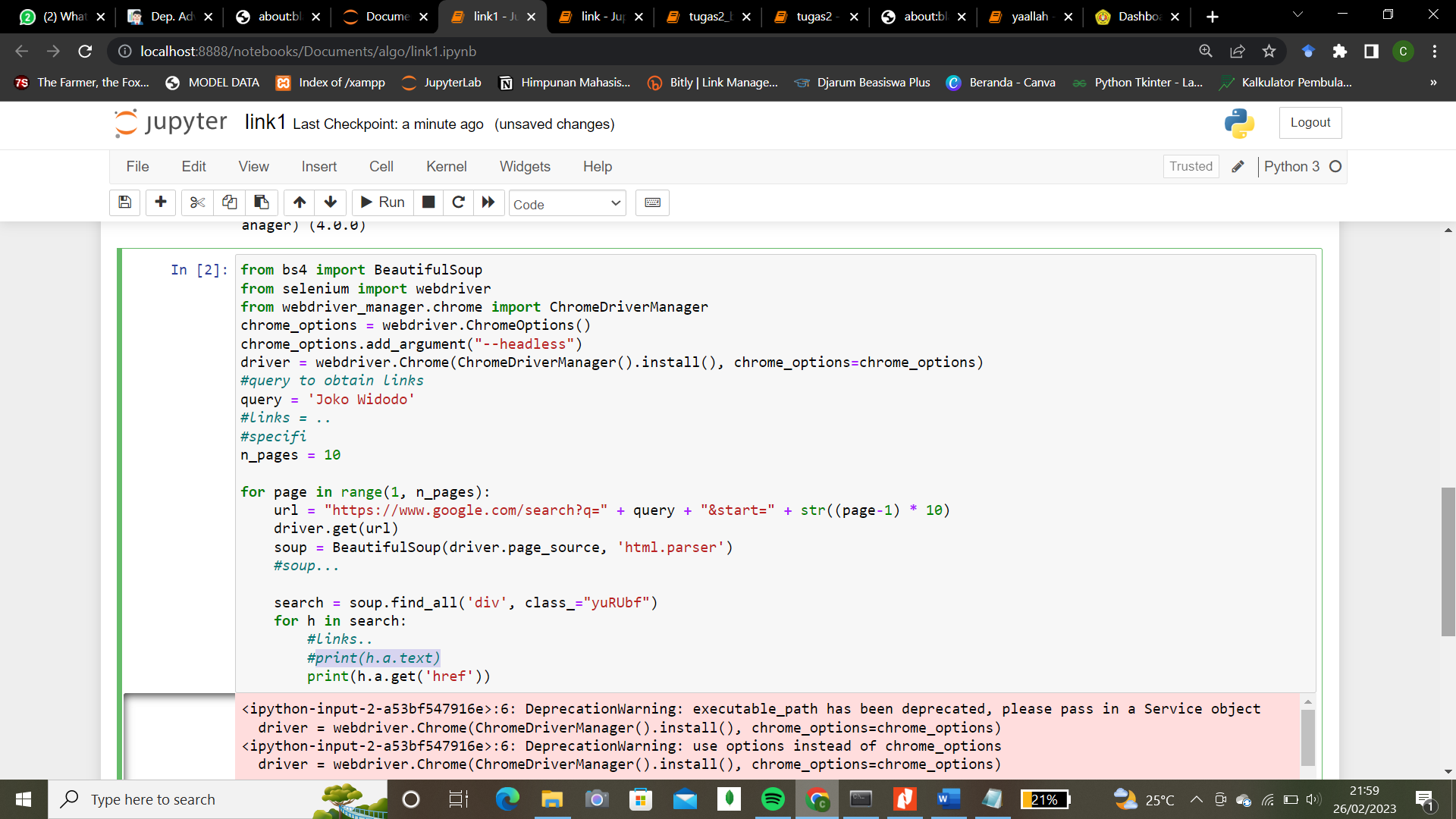




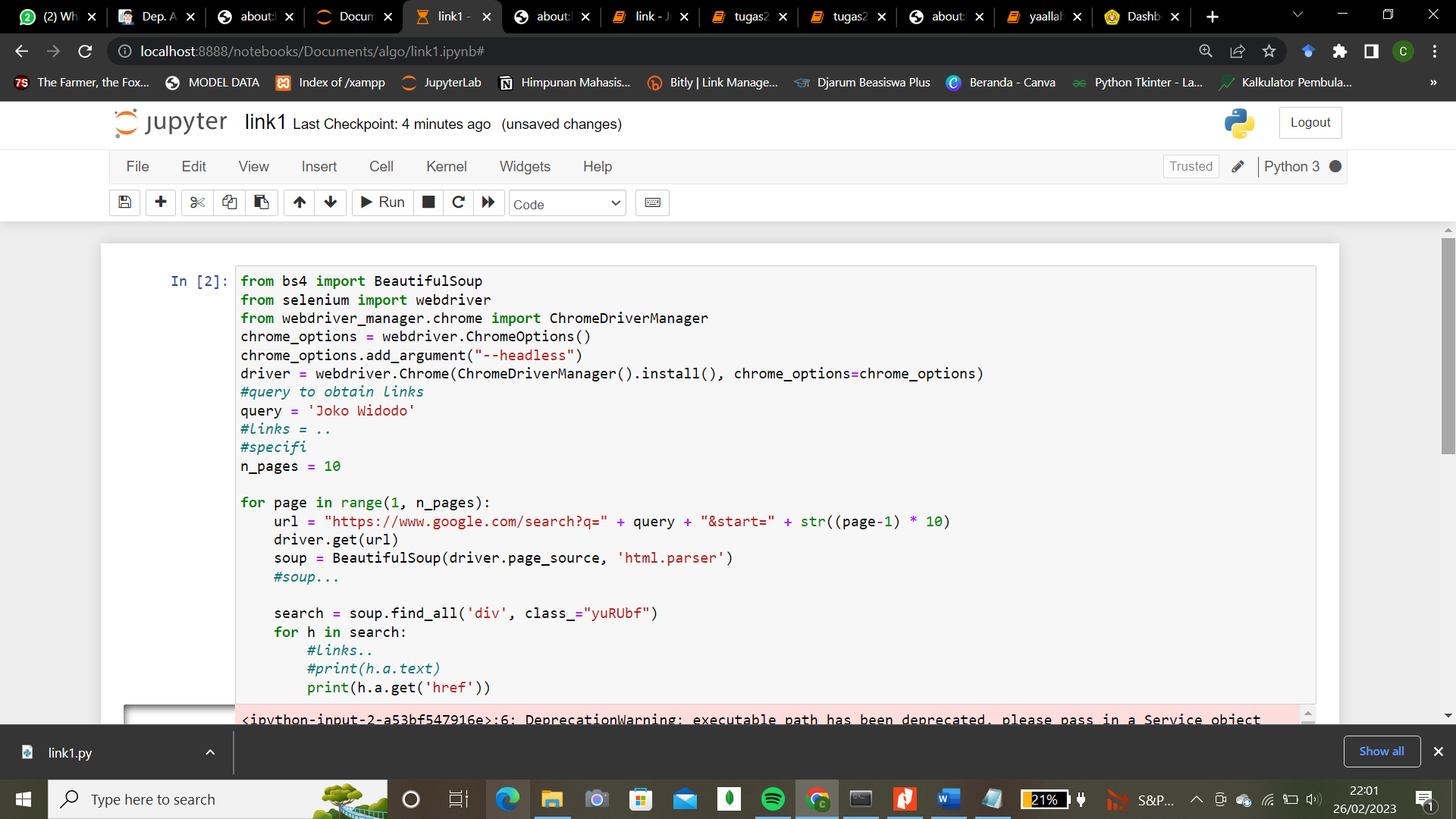
1. Cek pada file txt yang otomatis terbuat tadi, jika ada output berupa link scraping, maka proses berhasil

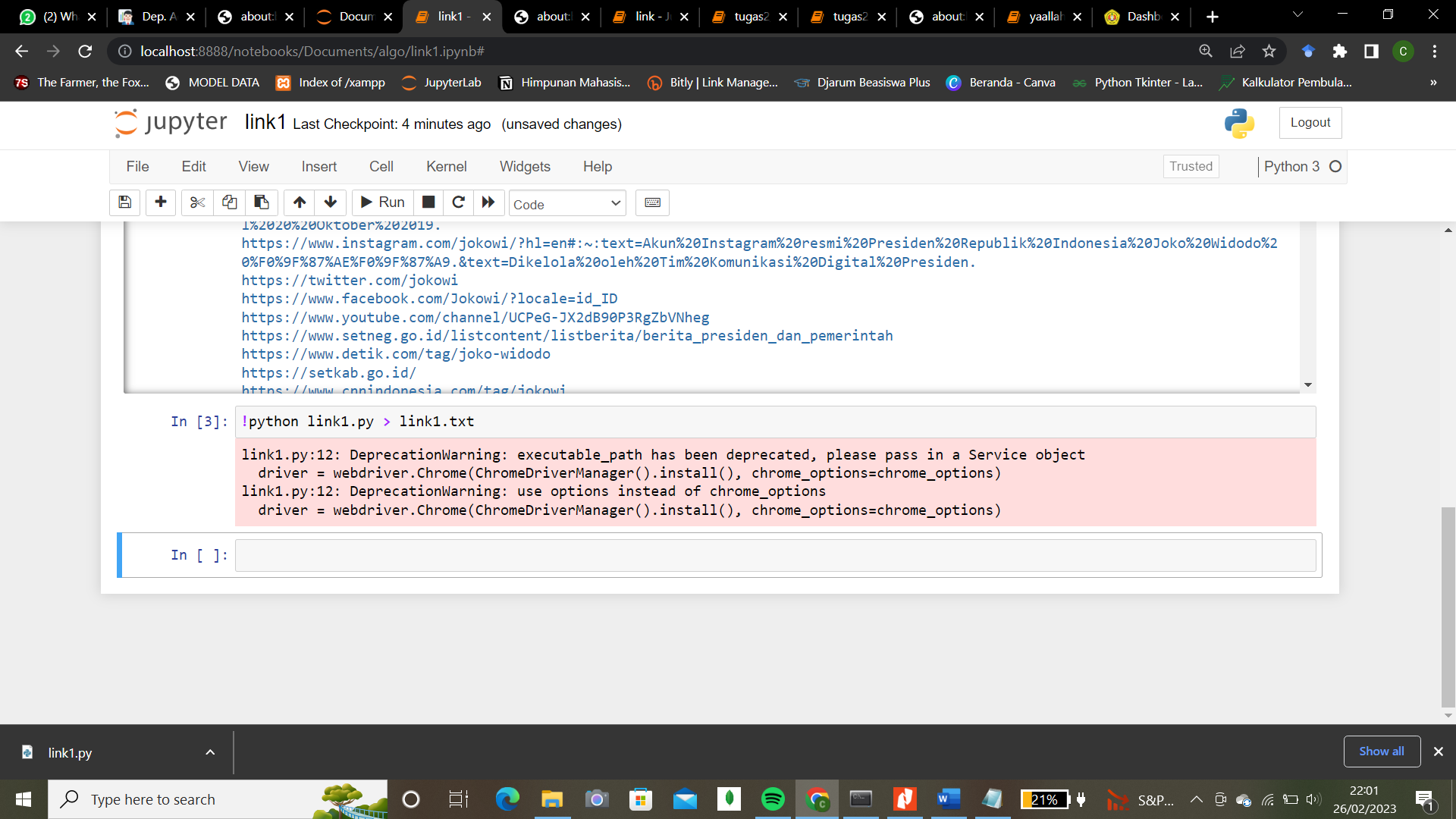


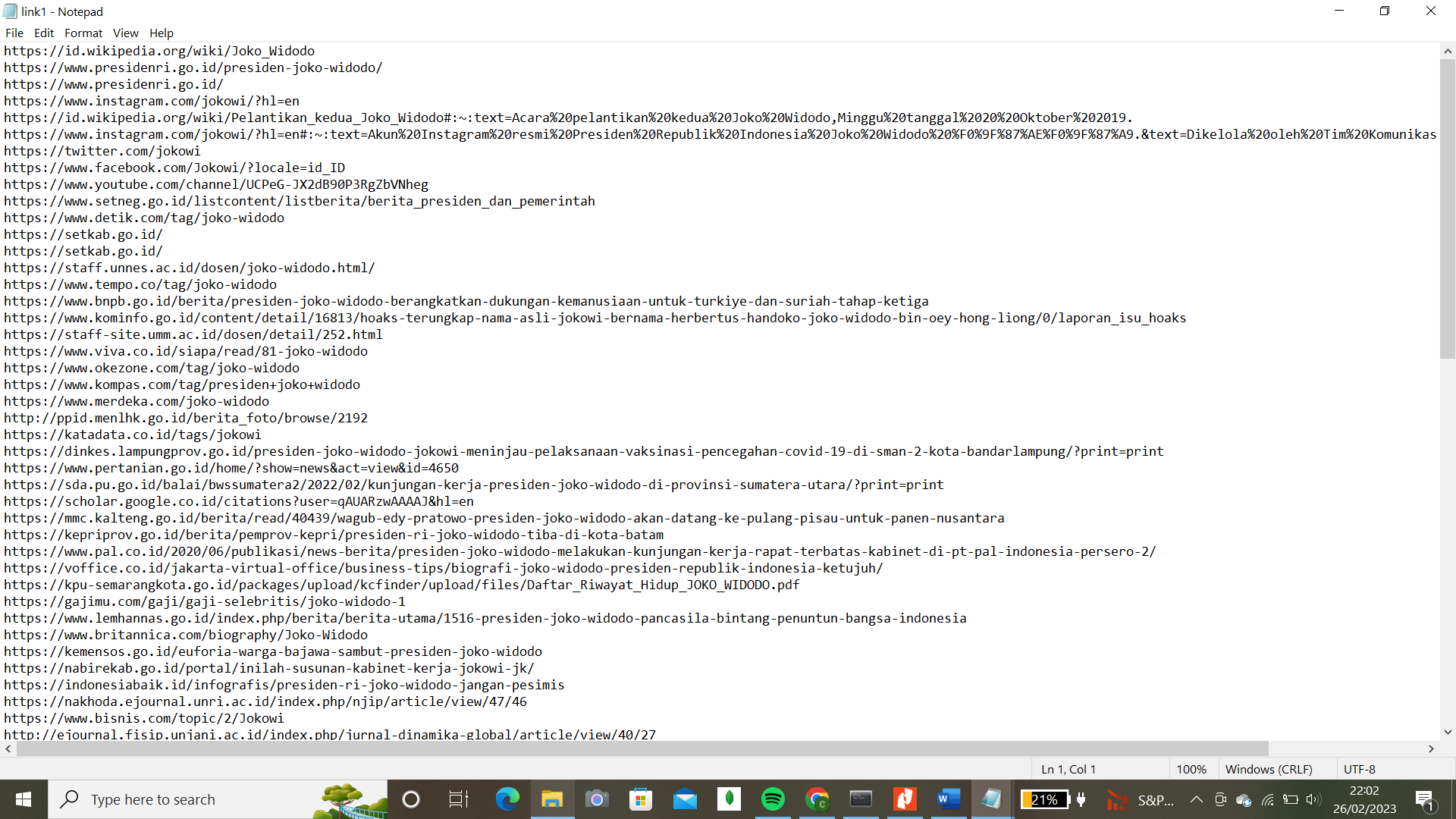
1. Jika ingin mengambil url-nya saja, script print(h.a.text) diberi # didepannya



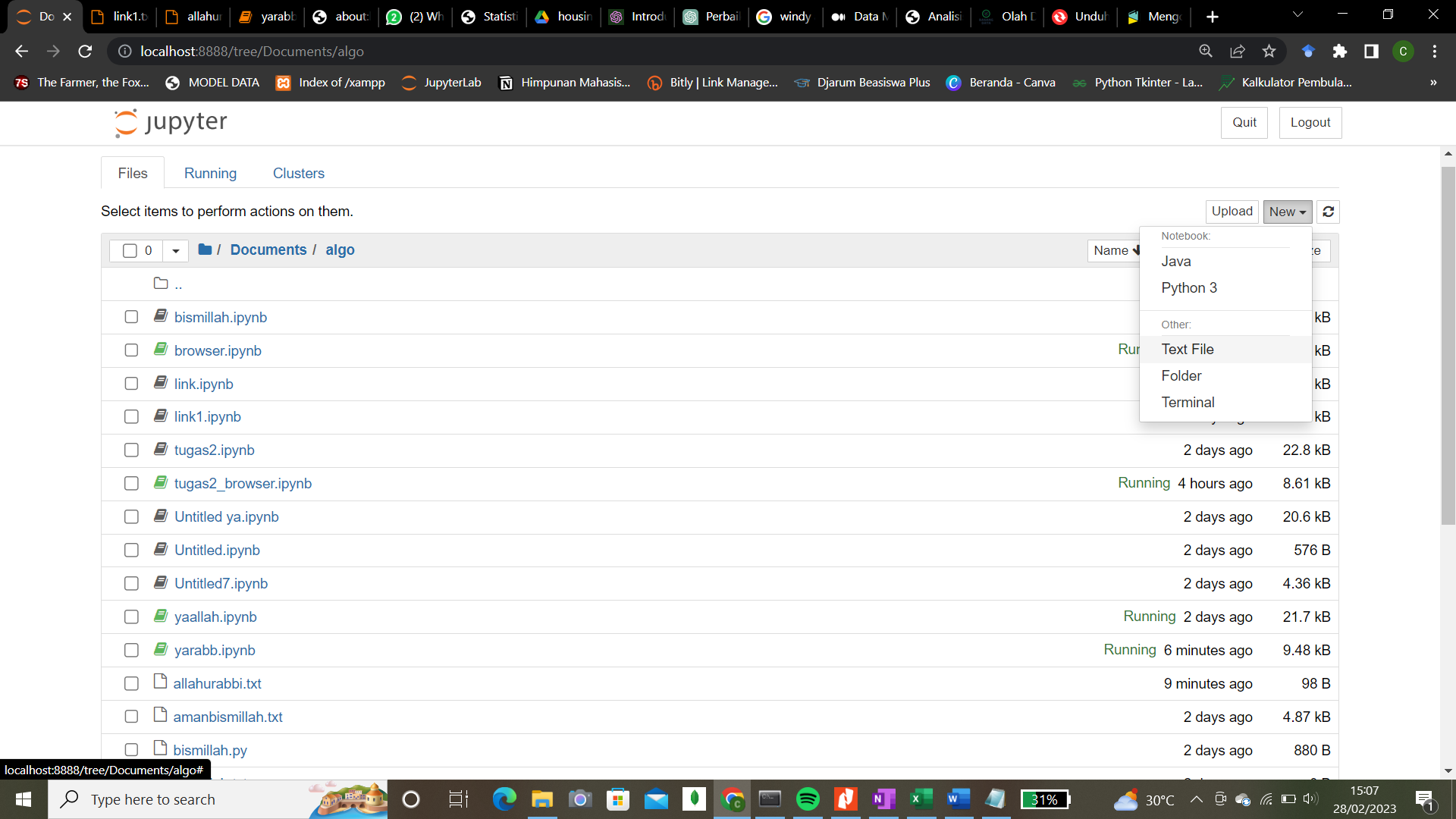
1. Ulangi Langkah 3 sampai 5

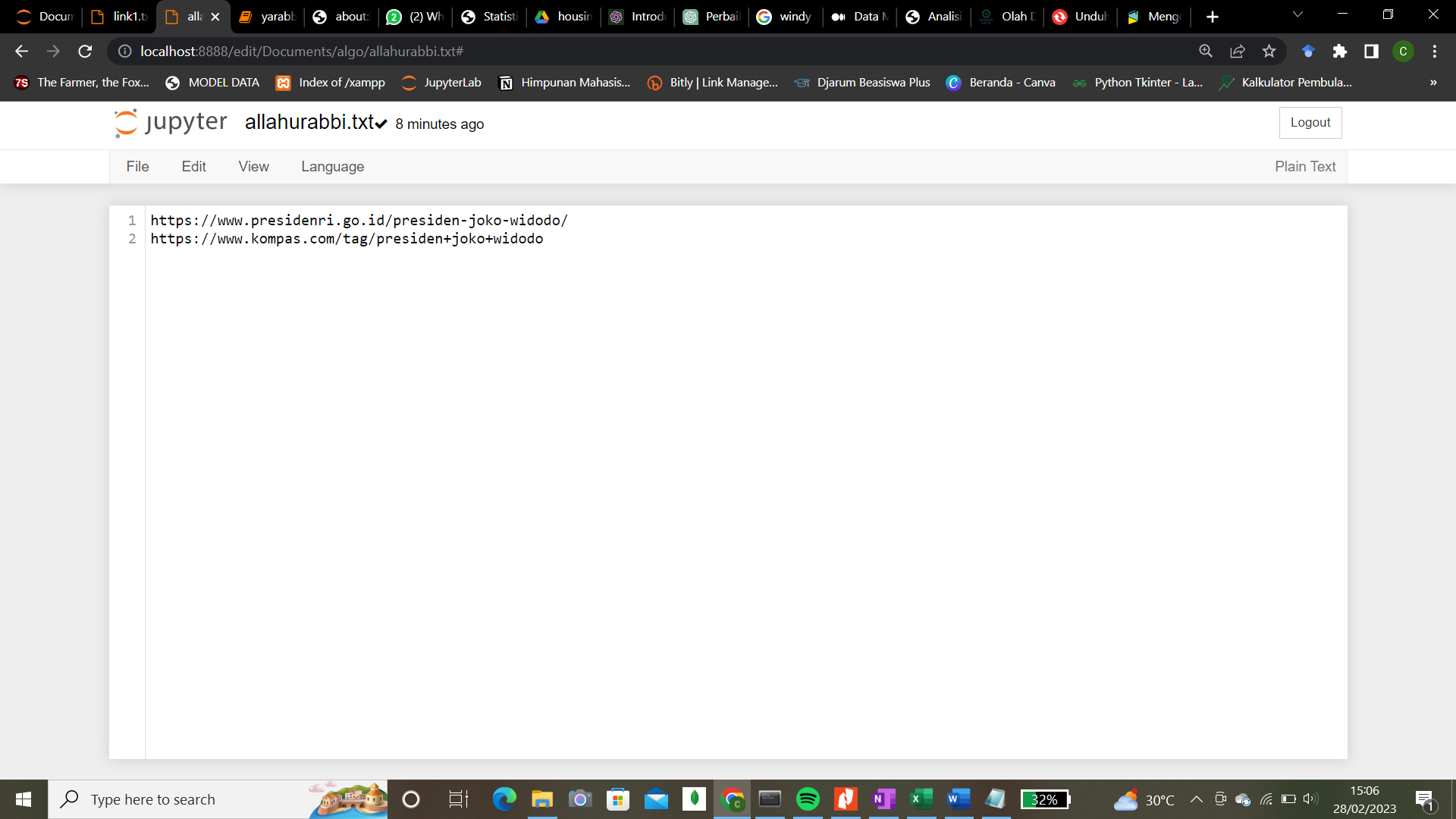






1. Buat file text baru pada folder tempat semua file tersimpan





1. Ambil 2 link dari file.txt yang hanya berupa link tadi, kemudian save
2. Buat file baru berisikan script berikut

Script :

from selenium import webdriver

import sys, getopt

import argparse

from webdriver\_manager.chrome import ChromeDriverManager

from selenium.webdriver.chrome.options import Options

from selenium.webdriver.chrome.service import Service

from selenium.webdriver.common.by import By

options = Options()

options.add\_argument("--headless")

driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()),options=options)

full\_text=[]

def parse\_args():

parser = argparse.ArgumentParser()

parser.add\_argument('-i', '--infile', default='', help='input filename')

parser.add\_argument('-o', '--outfile', default='', help='output filename')

return parser.parse\_args()

def main():

args = parse\_args()

outfile = args.outfile

infile = args.infile

with open(infile) as f:

content = f.read().splitlines()

f.close()

f = open(outfile,"w")

for u in content:

driver.get(u)

elems = driver.find\_element(By.TAG\_NAME, 'body').text.encode("utf-8")

full\_text.append(elems)

print(full\_text)

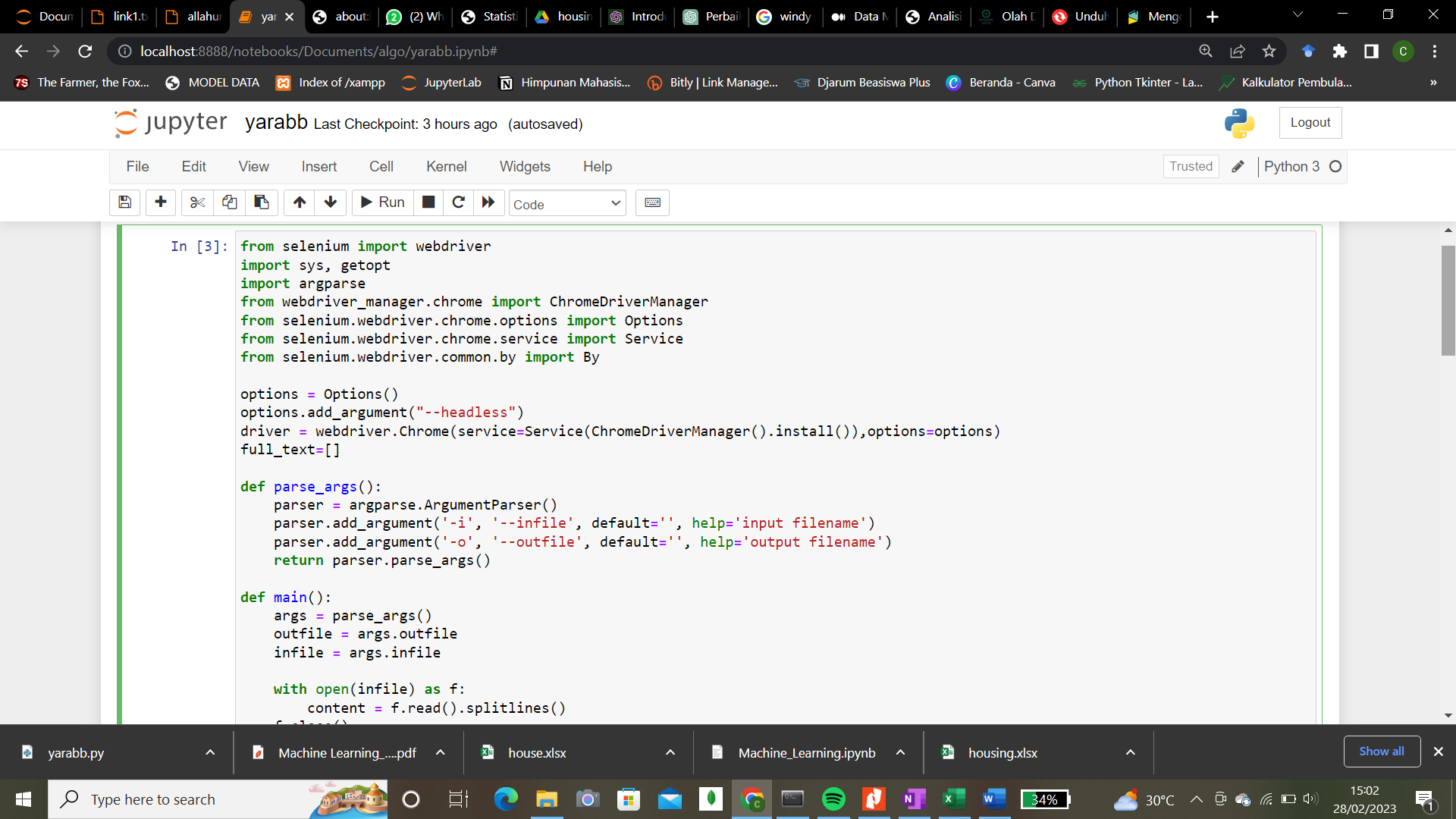
f.write(str(full\_text))

#print(full\_text)

driver.close()

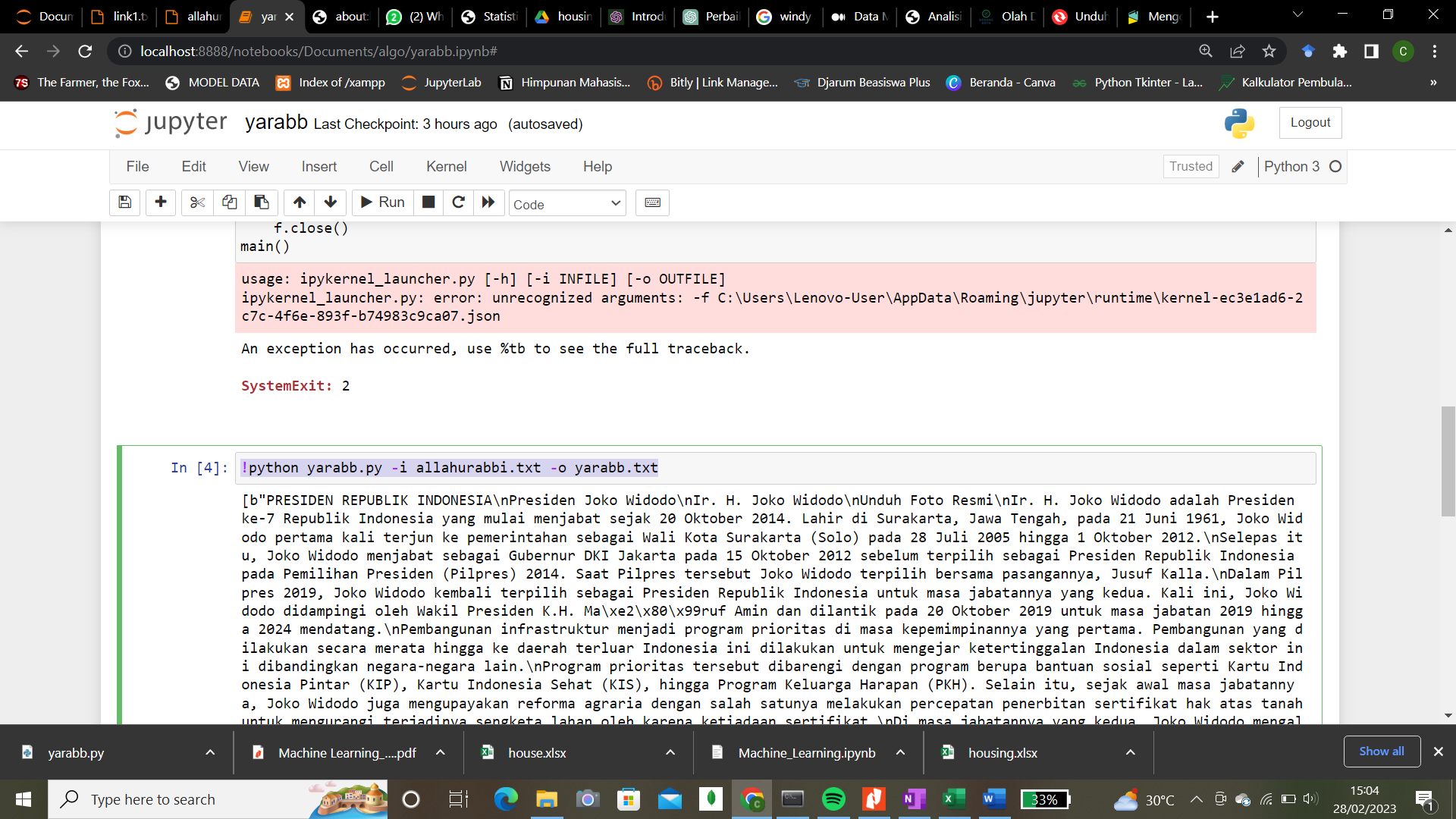
f.close()

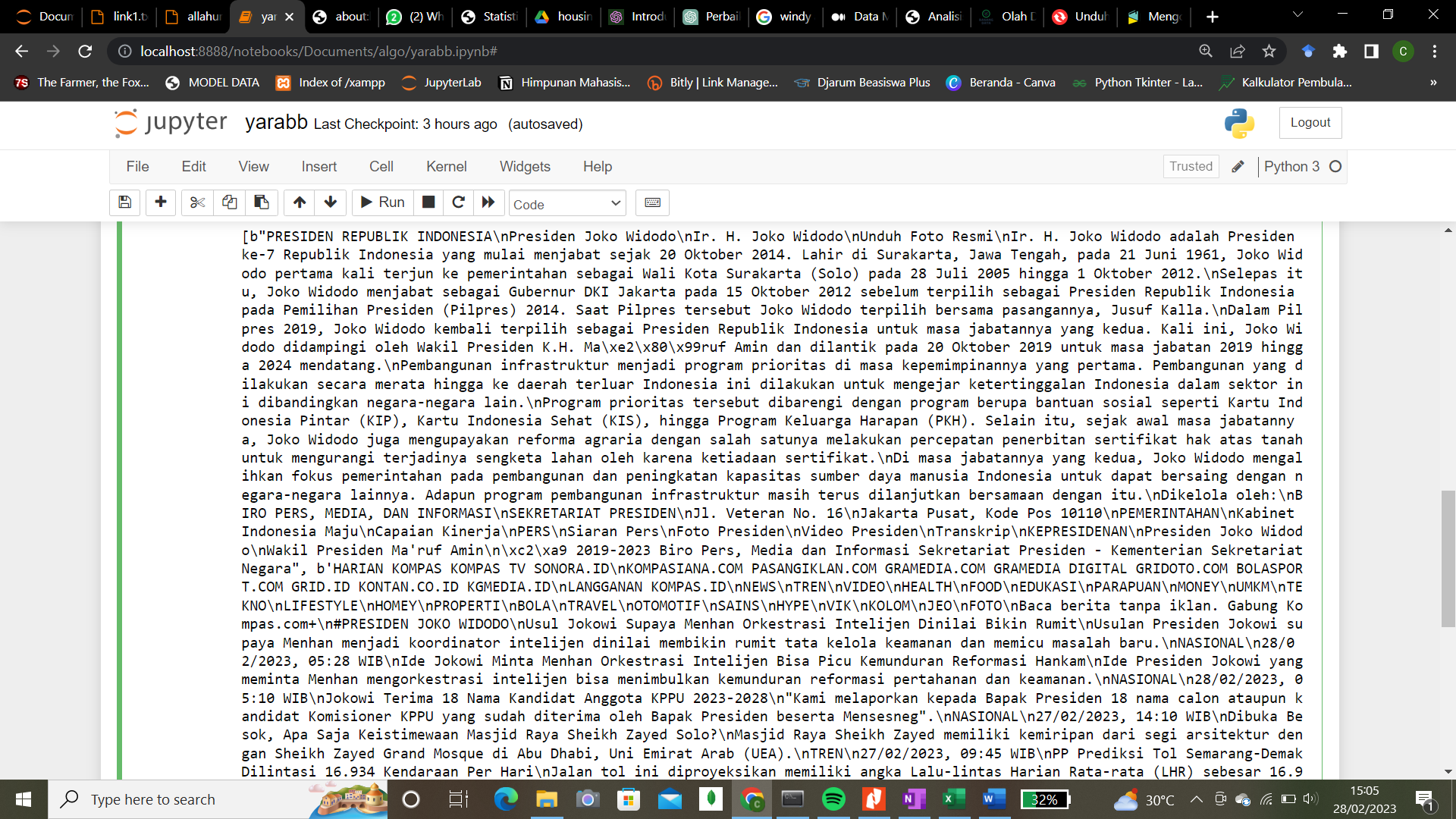
main()



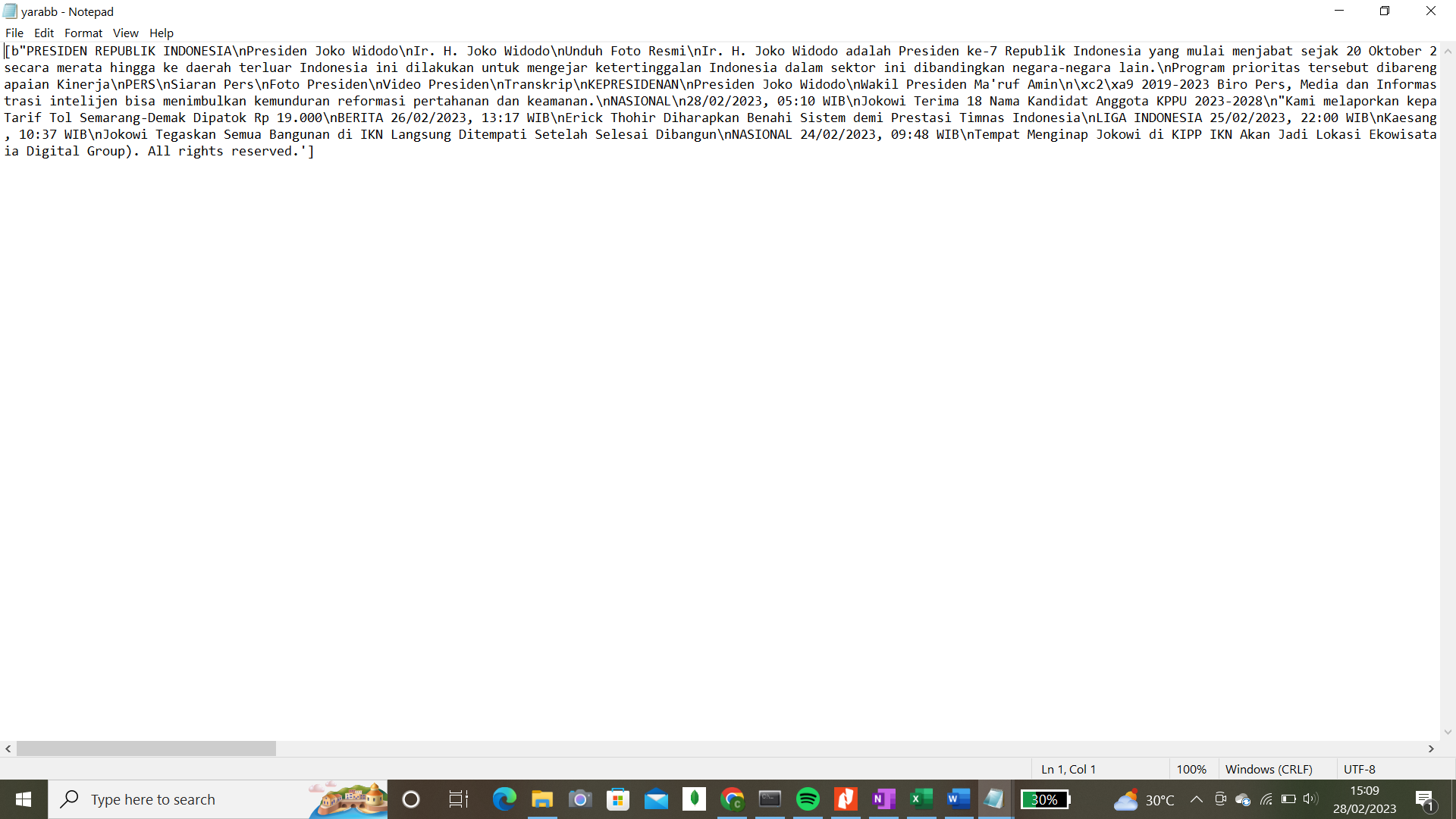
Download dalam bentuk .py

1. Pada file script tadi dalam bentuk ipynb, tuliskan !python namafilemodulebrowser.py -i filetxtbaru.txt -o filetxttujuan.txt seperti contoh punya saya !python yarabb.py -i allahurabbi.txt -o yarabb.txt. Tunggu sampai proses selesai dan hasil akan keluar





1. Hasil tadi akan muncul di file txt tujuan



1. Untuk merapikan text, bisa menggunakan script berikut

Script :

import nltk

from nltk.tokenize import sent\_tokenize

import re

with open('dikitlagi.txt', 'r') as file:

text = file.read() #membuka file dan menyimpan dalam text

tokens = nltk.tokenize.sent\_tokenize(text) #tokenisasi text

array = text.split() #memecah text menjadi array

with open('terakhir.txt', 'w') as file:

file.write(text) #mencetak hasil

print(tokens, array)

