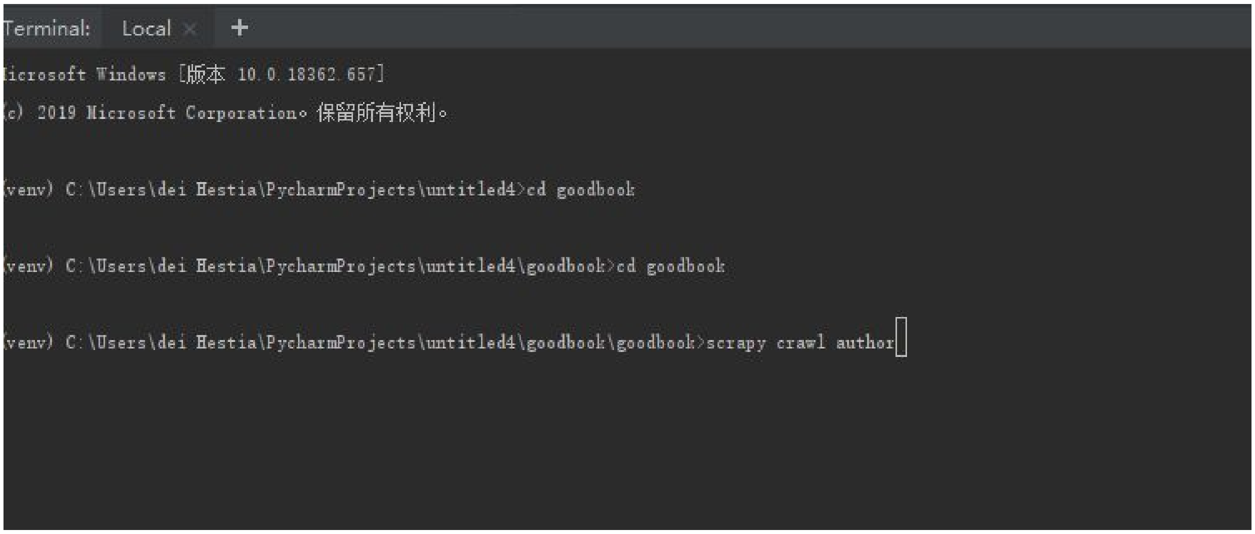
1. Usage

To crawl author information. Run scrapy crawl author. Eg

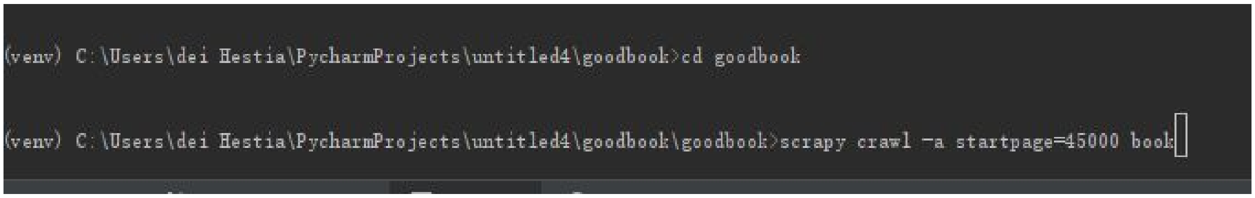


```
icrosoft Windows [版本 10.0.18362.657]
c) 2019 Microsoft Corporation。保留所有权利。

venv) C:\Users\dei Hestia\PycharmProjects\untitled4>cd goodbook

venv) C:\Users\dei Hestia\PycharmProjects\untitled4\goodbook>cd goodbook

venv) C:\Users\dei Hestia\PycharmProjects\untitled4\goodbook\goodbook>scrapy crawl author
```

And this program will automatically crawl 60 authors and store their information in database author.

2.
To crawl book information. You need to first give a index of book as startpoint.

It will search 220 books randomly after your start page.

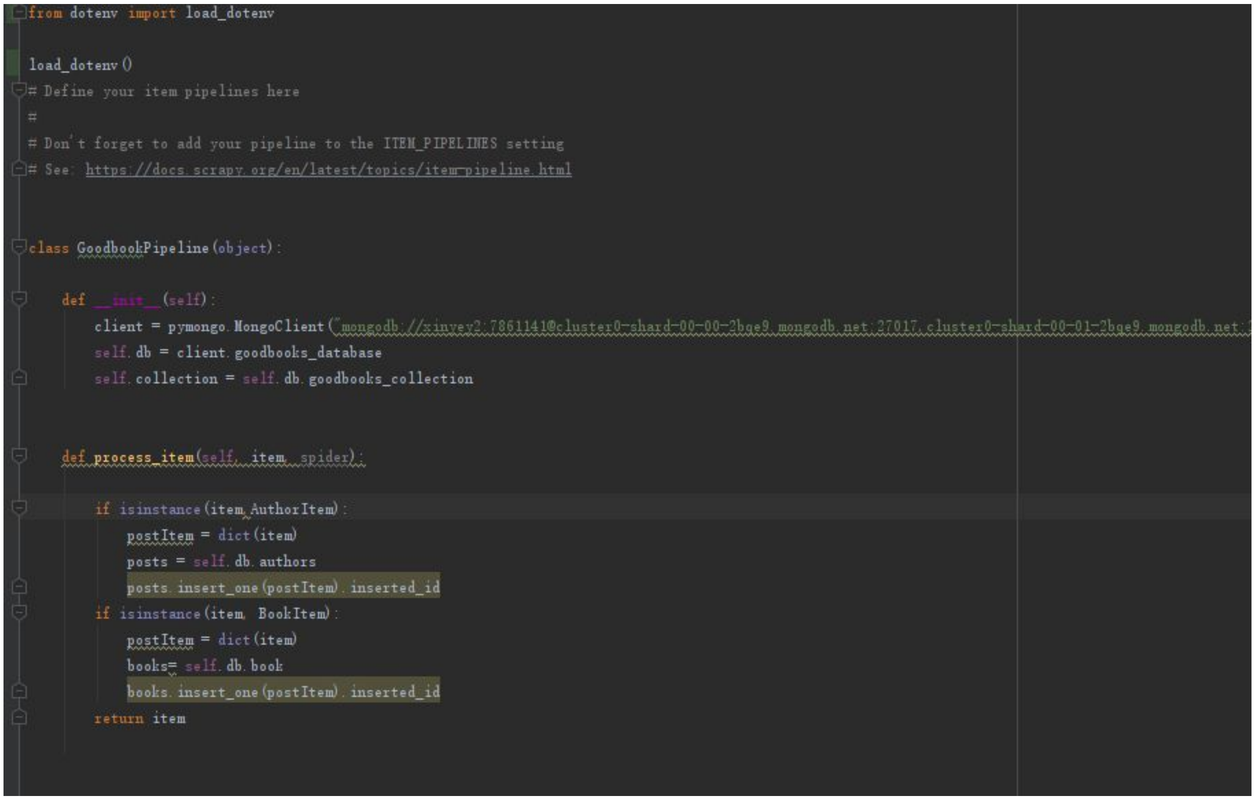


```
(venv) C:\Users\dei Hestia\PycharmProjects\untitled4\goodbook>cd goodbook

(venv) C:\Users\dei Hestia\PycharmProjects\untitled4\goodbook\goodbook>scrapy crawl -a startpage=45000 book
```

Here you can replace startpage to any number less than 99999 as you want.

After it finished, it will store all the data in json file at the AWS database books.

If you want to store data at other database, you can just modify the pipline function to do it

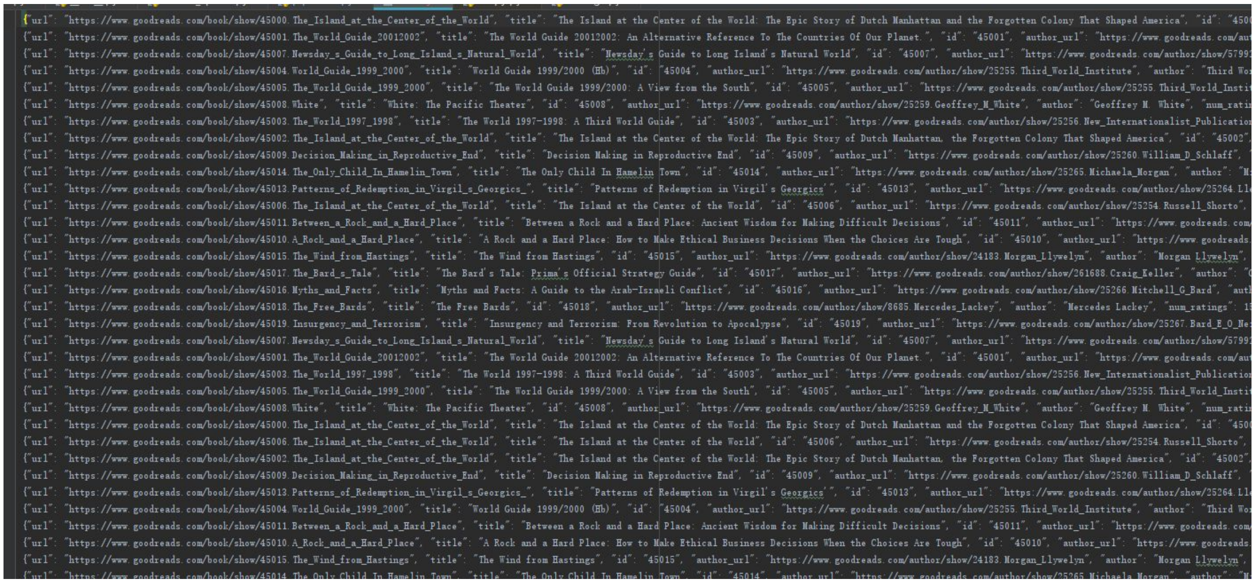


```
from dotenv import load_dotenv

load_dotenv()
# Define your item pipelines here
#
# Don't forget to add your pipeline to the ITEM_PIPELINES setting
# See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html


class GoodbookPipeline(object):

    def __init__(self):
        client = pymongo.MongoClient("mongodb://xinyey2:7861141@cluster0-shard-00-00-2bqe9.mongodb.net:27017,cluster0-shard-00-01-2bqe9.mongodb.net:
        self.db = client.goodbooks_database
        self.collection = self.db.goodbooks_collection


    def process_item(self, item, spider):

        if isinstance(item,AuthorItem):
            postItem = dict(item)
            posts = self.db.authors
            posts.insert_one(postItem).inserted_id
        if isinstance(item, BookItem):
            postItem = dict(item)
            books= self.db.book
            books.insert_one(postItem).inserted_id
        return item
```

Here, just replace the client to your client address, and it will automatically push data into database goodbooks_test/ author and goodbooks_test/ book depend on the item type it returns

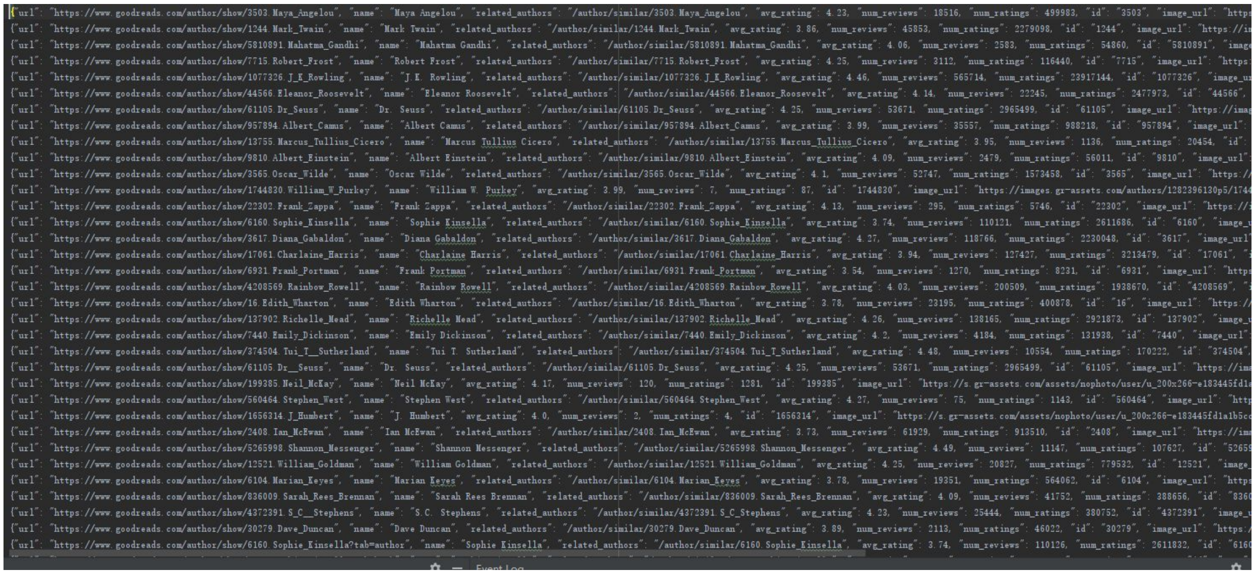
To test the function.

You can run manully,
It will generate book.il json file for books



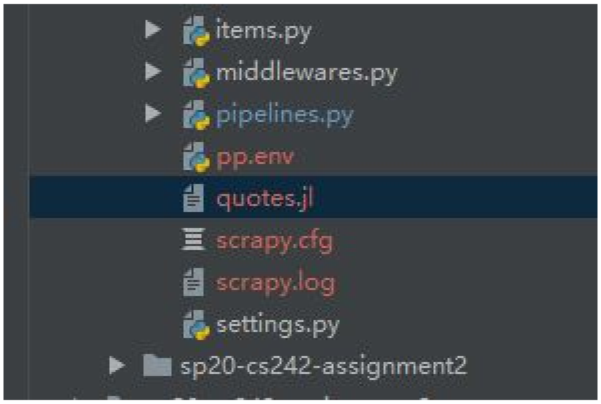
```
{"url": "https://www.goodreads.com/book/show/45000.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World: The Epic Story of Dutch Manhattan and the Forgotten Colony That Shaped America", "id": "4500
{"url": "https://www.goodreads.com/book/show/45001.The_World_Guide_20012002", "title": "The World Guide 20012002: An Alternative Reference To The Countries Of Our Planet.", "id": "45001", "author_url": "https://www.goodreads.com/aut
{"url": "https://www.goodreads.com/book/show/45007.Newsday_s_Guide_to_Long_Island_s_Natural_World", "title": "Newsday's Guide to Long Island's Natural World", "id": "45007", "author_url": "https://www.goodreads.com/author/show/5799
{"url": "https://www.goodreads.com/book/show/45004.World_Guide_1999_2000", "title": "World Guide 1999/2000 (Hb)", "id": "45004", "author_url": "https://www.goodreads.com/author/show/25255.Third_World_Institute", "author": "Third Wor
{"url": "https://www.goodreads.com/book/show/45005.The_World_Guide_1999_2000", "title": "The World Guide 1999/2000: A View from the South", "id": "45005", "author_url": "https://www.goodreads.com/author/show/25255.Third_World_Instit
{"url": "https://www.goodreads.com/book/show/45008.White", "title": "White: The Pacific Theater", "id": "45008", "author_url": "https://www.goodreads.com/author/show/25259.Geoffrey_M_White", "author": "Geoffrey M. White", "num_ratin
{"url": "https://www.goodreads.com/book/show/45003.The_World_1997_1998", "title": "The World 1997-1998: A Third World Guide", "id": "45003", "author_url": "https://www.goodreads.com/author/show/25256.New_Internationalist_Publicatio
{"url": "https://www.goodreads.com/book/show/45002.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World: The Epic Story of Dutch Manhattan, the Forgotten Colony That Shaped America", "id": "45002",
{"url": "https://www.goodreads.com/book/show/45009.Decision_Making_in_Reproductive_End", "title": "Decision Making in Reproductive End", "id": "45009", "author_url": "https://www.goodreads.com/author/show/25260.William_D_Schlaff", "
{"url": "https://www.goodreads.com/book/show/45014.The_Only_Child_In_Hamelin_Town", "title": "The Only Child In Hamelin Town", "id": "45014", "author_url": "https://www.goodreads.com/author/show/25265.Michaela_Morgan", "author": "Mi
{"url": "https://www.goodreads.com/book/show/45013.Patterns_of_Redemption_in_Virgil_s_Georgics_", "title": "Patterns of Redemption in Virgil's Georgics'", "id": "45013", "author_url": "https://www.goodreads.com/author/show/25264.Lle
{"url": "https://www.goodreads.com/book/show/45006.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World", "id": "45006", "author_url": "https://www.goodreads.com/author/show/25254.Russell_Shorto",
{"url": "https://www.goodreads.com/book/show/45011.Between_a_Rock_and_a_Hard_Place", "title": "Between a Rock and a Hard Place: Ancient Wisdom for Making Difficult Decisions", "id": "45011", "author_url": "https://www.goodreads.com/
{"url": "https://www.goodreads.com/book/show/45010.A_Rock_and_a_Hard_Place", "title": "A Rock and a Hard Place: How to Make Ethical Business Decisions When the Choices Are Tough", "id": "45010", "author_url": "https://www.goodreads.
{"url": "https://www.goodreads.com/book/show/45015.The_Wind_from_Hastings", "title": "The Wind from Hastings", "id": "45015", "author_url": "https://www.goodreads.com/author/show/24183.Morgan_Llywelyn", "author": "Morgan Llywelyn",
{"url": "https://www.goodreads.com/book/show/45017.The_Bard_s_Tale", "title": "The Bard's Tale: Prima's Official Strategy Guide", "id": "45017", "author_url": "https://www.goodreads.com/author/show/261688.Craig_Keller", "author": "C
{"url": "https://www.goodreads.com/book/show/45016.Myths_and_Facts", "title": "Myths and Facts: A Guide to the Arab-Israeli Conflict", "id": "45016", "author_url": "https://www.goodreads.com/author/show/25266.Mitchell_G_Bard", "auth
{"url": "https://www.goodreads.com/book/show/45018.The_Free_Bards", "title": "The Free Bards", "id": "45018", "author_url": "https://www.goodreads.com/author/show/8685.Mercedes_Lackey", "author": "Mercedes Lackey", "num_ratings": 15
{"url": "https://www.goodreads.com/book/show/45019.Insurgency_and_Terrorism", "title": "Insurgency and Terrorism: From Revolution to Apocalypse", "id": "45019", "author_url": "https://www.goodreads.com/author/show/25267.Bard_E_O_Ne
{"url": "https://www.goodreads.com/book/show/45007.Newsday_s_Guide_to_Long_Island_s_Natural_World", "title": "Newsday's Guide to Long Island's Natural World", "id": "45007", "author_url": "https://www.goodreads.com/author/show/5799
{"url": "https://www.goodreads.com/book/show/45001.The_World_Guide_20012002", "title": "The World Guide 20012002: An Alternative Reference To The Countries Of Our Planet.", "id": "45001", "author_url": "https://www.goodreads.com/aut
{"url": "https://www.goodreads.com/book/show/45003.The_World_1997_1998", "title": "The World 1997-1998: A Third World Guide", "id": "45003", "author_url": "https://www.goodreads.com/author/show/25256.New_Internationalist_Publicatio
{"url": "https://www.goodreads.com/book/show/45005.The_World_Guide_1999_2000", "title": "The World Guide 1999/2000: A View from the South", "id": "45005", "author_url": "https://www.goodreads.com/author/show/25255.Third_World_Instit
{"url": "https://www.goodreads.com/book/show/45008.White", "title": "White: The Pacific Theater", "id": "45008", "author_url": "https://www.goodreads.com/author/show/25259.Geoffrey_M_White", "author": "Geoffrey M. White", "num_ratin
{"url": "https://www.goodreads.com/book/show/45000.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World: The Epic Story of Dutch Manhattan and the Forgotten Colony That Shaped America", "id": "4500
{"url": "https://www.goodreads.com/book/show/45006.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World", "id": "45006", "author_url": "https://www.goodreads.com/author/show/25254.Russell_Shorto",
{"url": "https://www.goodreads.com/book/show/45002.The_Island_at_the_Center_of_the_World", "title": "The Island at the Center of the World: The Epic Story of Dutch Manhattan, the Forgotten Colony That Shaped America", "id": "45002",
{"url": "https://www.goodreads.com/book/show/45009.Decision_Making_in_Reproductive_End", "title": "Decision Making in Reproductive End", "id": "45009", "author_url": "https://www.goodreads.com/author/show/25260.William_D_Schlaff", "
{"url": "https://www.goodreads.com/book/show/45013.Patterns_of_Redemption_in_Virgil_s_Georgics_", "title": "Patterns of Redemption in Virgil's Georgics'", "id": "45013", "author_url": "https://www.goodreads.com/author/show/25264.Lle
{"url": "https://www.goodreads.com/book/show/45004.World_Guide_1999_2000", "title": "World Guide 1999/2000 (Hb)", "id": "45004", "author_url": "https://www.goodreads.com/author/show/25255.Third_World_Institute", "author": "Third Wor
{"url": "https://www.goodreads.com/book/show/45011.Between_a_Rock_and_a_Hard_Place", "title": "Between a Rock and a Hard Place: Ancient Wisdom for Making Difficult Decisions", "id": "45011", "author_url": "https://www.goodreads.com/
{"url": "https://www.goodreads.com/book/show/45010.A_Rock_and_a_Hard_Place", "title": "A Rock and a Hard Place: How to Make Ethical Business Decisions When the Choices Are Tough", "id": "45010", "author_url": "https://www.goodreads.
{"url": "https://www.goodreads.com/book/show/45015.The_Wind_from_Hastings", "title": "The Wind from Hastings", "id": "45015", "author_url": "https://www.goodreads.com/author/show/24183.Morgan_Llywelyn", "author": "Morgan Llywelyn",
{"url": "https://www.goodreads.com/book/show/45014.The_Only_Child_In_Hamelin_Town", "title": "The Only Child In Hamelin Town", "id": "45014", "author_url": "https://www.goodreads.com/author/show/25265.Michaela_Morgan", "author": "Mi
```

And generate quotes.il for Author information.



```
{"url": "https://www.goodreads.com/author/show/3503.Maya_Angelou", "name": "Maya Angelou", "related_authors": "/author/similar/3503.Maya_Angelou", "avg_rating": 4.23, "num_reviews": 18516, "num_ratings": 499983, "id": "3503", "image_url": "https
{"url": "https://www.goodreads.com/author/show/1244.Mark_Twain", "name": "Mark Twain", "related_authors": "/author/similar/1244.Mark_Twain", "avg_rating": 3.86, "num_reviews": 45853, "num_ratings": 2279098, "id": "1244", "image_url": "https://i
{"url": "https://www.goodreads.com/author/show/5810891.Mahatma_Gandhi", "name": "Mahatma Gandhi", "related_authors": "/author/similar/5810891.Mahatma_Gandhi", "avg_rating": 4.06, "num_reviews": 2583, "num_ratings": 54860, "id": "5810891", "image
{"url": "https://www.goodreads.com/author/show/7715.Robert_Frost", "name": "Robert Frost", "related_authors": "/author/similar/7715.Robert_Frost", "avg_rating": 4.25, "num_reviews": 3112, "num_ratings": 116440, "id": "7715", "image_url": "https:
{"url": "https://www.goodreads.com/author/show/1077326.J_K_Rowling", "name": "J.K. Rowling", "related_authors": "/author/similar/1077326.J_K_Rowling", "avg_rating": 4.46, "num_reviews": 565714, "num_ratings": 23917144, "id": "1077326", "image_u
{"url": "https://www.goodreads.com/author/show/44566.Eleanor_Roosevelt", "name": "Eleanor Roosevelt", "related_authors": "/author/similar/44566.Eleanor_Roosevelt", "avg_rating": 4.14, "num_reviews": 22245, "num_ratings": 2477973, "id": "44566",
{"url": "https://www.goodreads.com/author/show/61105.Dr_Seuss", "name": "Dr. Seuss", "related_authors": "/author/similar/61105.Dr_Seuss", "avg_rating": 4.25, "num_reviews": 53671, "num_ratings": 2965499, "id": "61105", "image_url": "https://imag
{"url": "https://www.goodreads.com/author/show/957894.Albert_Camus", "name": "Albert Camus", "related_authors": "/author/similar/957894.Albert_Camus", "avg_rating": 3.99, "num_reviews": 35557, "num_ratings": 988218, "id": "957894", "image_url":
{"url": "https://www.goodreads.com/author/show/13755.Marcus_Tullius_Cicero", "name": "Marcus Tullius Cicero", "related_authors": "/author/similar/13755.Marcus_Tullius_Cicero", "avg_rating": 3.95, "num_reviews": 1136, "num_ratings": 20454, "id":
{"url": "https://www.goodreads.com/author/show/9810.Albert_Einstein", "name": "Albert Einstein", "related_authors": "/author/similar/9810.Albert_Einstein", "avg_rating": 4.09, "num_reviews": 2479, "num_ratings": 56011, "id": "9810", "image_url":
{"url": "https://www.goodreads.com/author/show/3565.Oscar_Wilde", "name": "Oscar Wilde", "related_authors": "/author/similar/3565.Oscar_Wilde", "avg_rating": 4.1, "num_reviews": 52747, "num_ratings": 1573458, "id": "3565", "image_url": "https://
{"url": "https://www.goodreads.com/author/show/1744830.William_W_Purkey", "name": "William W. Purkey", "avg_rating": 3.99, "num_reviews": 7, "num_ratings": 87, "id": "1744830", "image_url": "https://images.gr-assets.com/authors/1282396130p5/1744
{"url": "https://www.goodreads.com/author/show/22302.Frank_Zappa", "name": "Frank Zappa", "related_authors": "/author/similar/22302.Frank_Zappa", "avg_rating": 4.13, "num_reviews": 295, "num_ratings": 5746, "id": "22302", "image_url": "https://i
{"url": "https://www.goodreads.com/author/show/6160.Sophie_Kinsella", "name": "Sophie Kinsella", "related_authors": "/author/similar/6160.Sophie_Kinsella", "avg_rating": 3.74, "num_reviews": 110121, "num_ratings": 2611686, "id": "6160", "image_u
{"url": "https://www.goodreads.com/author/show/3617.Diana_Gabaldon", "name": "Diana Gabaldon", "related_authors": "/author/similar/3617.Diana_Gabaldon", "avg_rating": 4.27, "num_reviews": 118766, "num_ratings": 2230048, "id": "3617", "image_url"
{"url": "https://www.goodreads.com/author/show/17061.Charlaine_Harris", "name": "Charlaine Harris", "related_authors": "/author/similar/17061.Charlaine_Harris", "avg_rating": 3.94, "num_reviews": 127427, "num_ratings": 3213479, "id": "17061", "i
{"url": "https://www.goodreads.com/author/show/6931.Frank_Portman", "name": "Frank Portman", "related_authors": "/author/similar/6931.Frank_Portman", "avg_rating": 3.54, "num_reviews": 1270, "num_ratings": 8231, "id": "6931", "image_url": "https
{"url": "https://www.goodreads.com/author/show/4208569.Rainbow_Rowell", "name": "Rainbow Rowell", "related_authors": "/author/similar/4208569.Rainbow_Rowell", "avg_rating": 4.03, "num_reviews": 200509, "num_ratings": 1938670, "id": "4208569", "i
{"url": "https://www.goodreads.com/author/show/16.Edith_Wharton", "name": "Edith Wharton", "related_authors": "/author/similar/16.Edith_Wharton", "avg_rating": 3.78, "num_reviews": 23195, "num_ratings": 400878, "id": "16", "image_url": "https://
{"url": "https://www.goodreads.com/author/show/137902.Richelle_Mead", "name": "Richelle Mead", "related_authors": "/author/similar/137902.Richelle_Mead", "avg_rating": 4.26, "num_reviews": 138165, "num_ratings": 2921873, "id": "137902", "image_u
{"url": "https://www.goodreads.com/author/show/7440.Emily_Dickinson", "name": "Emily Dickinson", "related_authors": "/author/similar/7440.Emily_Dickinson", "avg_rating": 4.2, "num_reviews": 4184, "num_ratings": 131938, "id": "7440", "image_url":
{"url": "https://www.goodreads.com/author/show/374504.Tui_T__Sutherland", "name": "Tui T. Sutherland", "related_authors": "/author/similar/374504.Tui_T_Sutherland", "avg_rating": 4.48, "num_reviews": 10554, "num_ratings": 170222, "id": "374504",
{"url": "https://www.goodreads.com/author/show/61105.Dr__Seuss", "name": "Dr. Seuss", "related_authors": "/author/similar/61105.Dr_Seuss", "avg_rating": 4.25, "num_reviews": 53671, "num_ratings": 2965499, "id": "61105", "image_url": "https://ima
{"url": "https://www.goodreads.com/author/show/199385.Neil_McKay", "name": "Neil McKay", "avg_rating": 4.17, "num_reviews": 120, "num_ratings": 1281, "id": "199385", "image_url": "https://s.gr-assets.com/assets/nophoto/user/u_200x266-e183445fd1a
{"url": "https://www.goodreads.com/author/show/560464.Stephen_West", "name": "Stephen West", "related_authors": "/author/similar/560464.Stephen_West", "avg_rating": 4.27, "num_reviews": 75, "num_ratings": 1143, "id": "560464", "image_url": "http
{"url": "https://www.goodreads.com/author/show/1656314.J_Humbert", "name": "J. Humbert", "avg_rating": 4.0, "num_reviews": 2, "num_ratings": 4, "id": "1656314", "image_url": "https://s.gr-assets.com/assets/nophoto/user/u_200x266-e183445fd1a1b5cc
{"url": "https://www.goodreads.com/author/show/2408.Ian_McEwan", "name": "Ian McEwan", "related_authors": "/author/similar/2408.Ian_McEwan", "avg_rating": 3.73, "num_reviews": 61929, "num_ratings": 913510, "id": "2408", "image_url": "https://ima
{"url": "https://www.goodreads.com/author/show/5265998.Shannon_Messenger", "name": "Shannon Messenger", "related_authors": "/author/similar/5265998.Shannon_Messenger", "avg_rating": 4.49, "num_reviews": 11147, "num_ratings": 107627, "id": "52659
{"url": "https://www.goodreads.com/author/show/12521.William_Goldman", "name": "William Goldman", "related_authors": "/author/similar/12521.William_Goldman", "avg_rating": 4.25, "num_reviews": 20827, "num_ratings": 779532, "id": "12521", "image_
{"url": "https://www.goodreads.com/author/show/6104.Marian_Keyes", "name": "Marian Keyes", "related_authors": "/author/similar/6104.Marian_Keyes", "avg_rating": 3.78, "num_reviews": 19351, "num_ratings": 564062, "id": "6104", "image_url": "https
{"url": "https://www.goodreads.com/author/show/836009.Sarah_Rees_Brennan", "name": "Sarah Rees Brennan", "related_authors": "/author/similar/836009.Sarah_Rees_Brennan", "avg_rating": 4.09, "num_reviews": 41752, "num_ratings": 388656, "id": "8360
{"url": "https://www.goodreads.com/author/show/4372391.S_C__Stephens", "name": "S.C. Stephens", "related_authors": "/author/similar/4372391.S_C_Stephens", "avg_rating": 4.23, "num_reviews": 25444, "num_ratings": 380752, "id": "4372391", "image_u
{"url": "https://www.goodreads.com/author/show/30279.Dave_Duncan", "name": "Dave Duncan", "related_authors": "/author/similar/30279.Dave_Duncan", "avg_rating": 3.89, "num_reviews": 2113, "num_ratings": 46022, "id": "30279", "image_url": "https:/
{"url": "https://www.goodreads.com/author/show/6160.Sophie_Kinsella?tab=author", "name": "Sophie Kinsella", "related_authors": "/author/similar/6160.Sophie_Kinsella", "avg_rating": 3.74, "num_reviews": 110126, "num_ratings": 2611832, "id": "6160
```

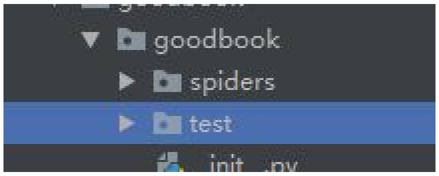
All the error information will be stored at logg.txt in the root source to view.



```
    ▶ items.py
    ▶ middlewares.py
    ▶ pipelines.py
      pp.env
      quotes.jl
      scrapy.cfg
      scrapy.log
      settings.py
  ▶ sp20-cs242-assignment2
```

Like scrapy.log

You can also run the unittest test case to test the functionality of author spider and book spider inside the test fold.

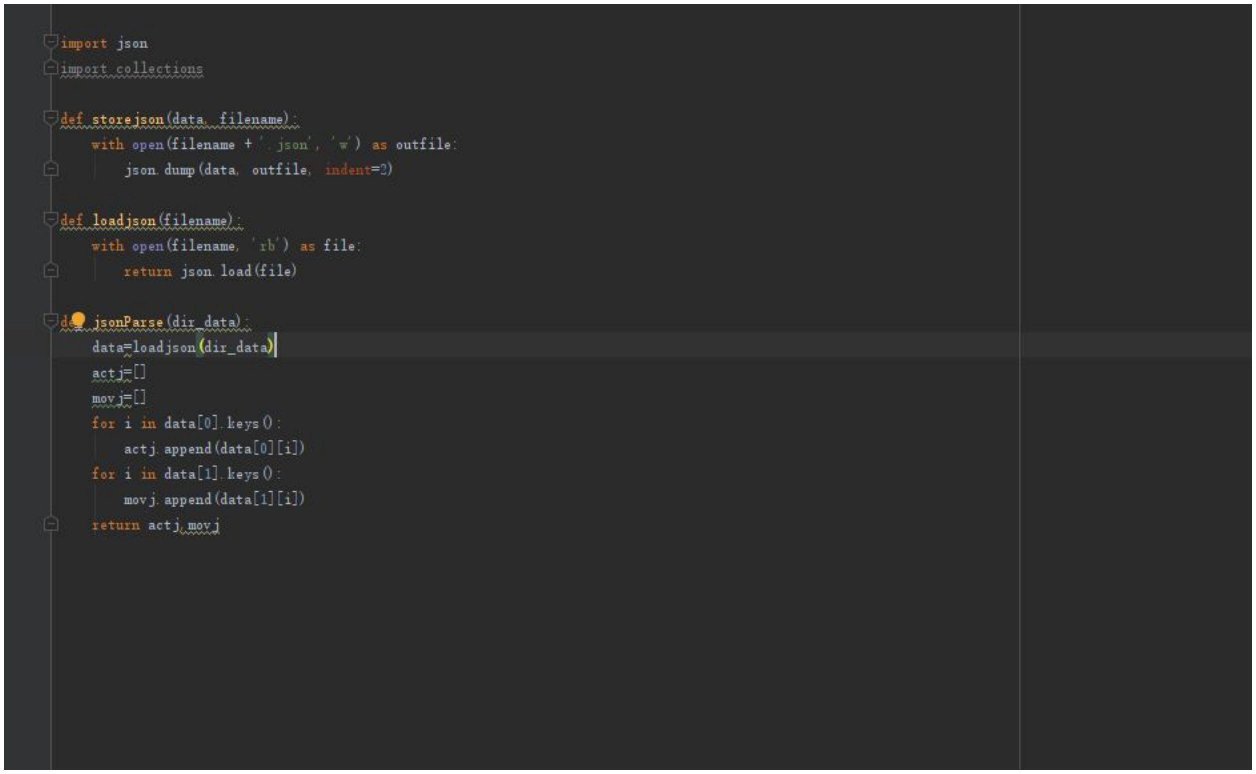


It contains basic assert configuration to see if the programs are running smothly

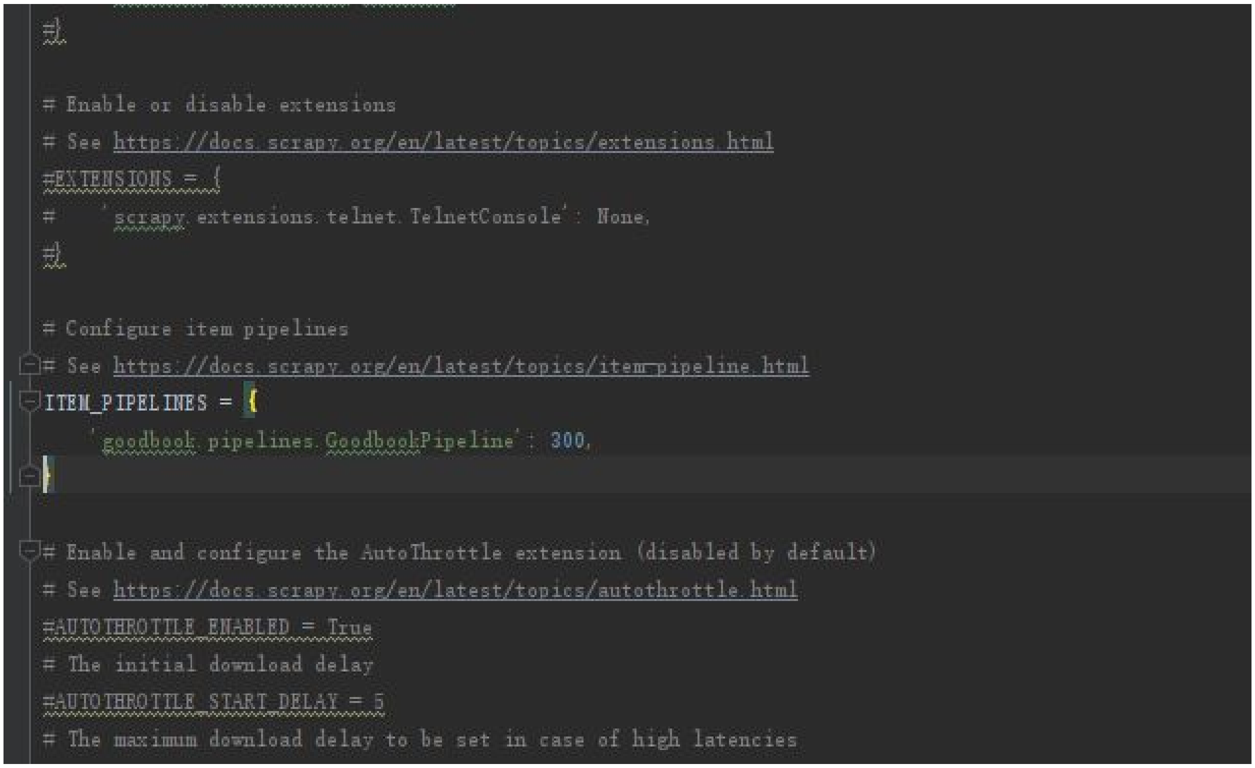
Finally, if you want to speed up the search or modify your result.

In helper.py, you can import json file or write to json files.

In the setting.py, you can set your limit, delay time, and whether or not to use pipline



```
import json
import collections

def storejson(data, filename):
    with open(filename + '.json', 'w') as outfile:
        json.dump(data, outfile, indent=2)

def loadjson(filename):
    with open(filename, 'rb') as file:
        return json.load(file)

def jsonParse(dir_data):
    data=loadjson(dir_data)
    actj=[]
    movj=[]
    for i in data[0].keys():
        actj.append(data[0][i])
    for i in data[1].keys():
        movj.append(data[1][i])
    return actj,movj
```



```
#}

# Enable or disable extensions
# See https://docs.scrapy.org/en/latest/topics/extensions.html
#EXTENSIONS = {
#    'scrapy.extensions.telnet.TelnetConsole': None,
#}

# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
    'goodbook.pipelines.GoodbookPipeline': 300,
}

# Enable and configure the AutoThrottle extension (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/autothrottle.html
#AUTOTHROTTLE_ENABLED = True
# The initial download delay
#AUTOTHROTTLE_START_DELAY = 5
# The maximum download delay to be set in case of high latencies
```