# مرحله اول)

در این مرحله نیاز است که ابتدا با استفاده از اسکریپت [**DataBase.sql]** دیتابیس مورد نظر را ساخته.

در جدول jobinja companies لینک های شرکت ها را نگهداری میکنیم در جدول JobAds

اطلاعات آگهی هارا نگهداری میکنیم.

سپس وارد فایل **configs.py** شده و اطلاعات پایگاه داده sql server خود را وارد کنید

توجه کنید تمام متغیر ها باید string باشند

# مرحله دوم)

برای استفاده از کرالر جابینجا باید ابتدا از کلاس CrawlCompanies یک instance ساخته و در ورودی لینک جابینجا و تعداد صفحاتی که برای کرال نیاز دارید وارد کنید ( اخرین صفحه جابینجا 10000 است )

obj = CrawlCompanies(

'https://jobinja.ir/companies',

Put\_number\_of\_pages\_here(int))

print(obj.run\_threads())

درین مرحله لینک های مورد نیاز شرکت ها کرال میشود

# مرحله سوم)

obj = CrawlPage(10)  
print(obj.run\_threads())

برای استفاده ازین کلاس نیاز به ورودی داریم که به معنای تعداد شرکت هایی است که نیاز دارید توجه کنید که قبل ان حتما باید از پر بودن جدول jobinja companies اطمینان حاصل کنید

به جای 10 هر تعداد شرکتی که در jobinja companies وجود دارد میتونید بگذارید

بعد از اجرای کد برای تکرار نشدن اگهی ها این تعداد شرکت از جدول jobinja companies پاک میشود

# مرحله چهارم )

obj = CrawlJobsJobVision(تعداد آگهی ها,  
 min\_range = عدد شروع,  
 max\_range =عدد پایان )

print(obj.run\_threads())

استراتژی ما درین کرالرو کمی تفاوت دارد به دلیل اینکه جابویژن از یک مکانیزم دفاعی برای کرال نکردن دیتا استفاده میکرد ما با توجه به شناسایی الگوی url ساختن این سایت برای اگهی ها به الگویی رسیدیم

در این الگو اگهی ها یک عدد دارند ولی رنج مشخصی ندارند , باتوجه به این موضوع ما از سمت کاربر یک رنج شروع و پایان گرفته و یک تعداد اگهی , کرالر یک رنج در این بازه میسازد و به تعداد عدد ورودی برای تعداد اگهی ها ما به صورت رندوم یک عدد از بین این بازه بر میداریم و یک url تولید میشود

الگوی URL آگهی ها:

هر آگهی در جاب ویژن دارای یک شناسه عددی منحصر به فرد است.

رنج این شناسه ها مشخص نیست و می تواند از هر عددی شروع شود.

URL آگهی ها بر اساس الگوی مشخصی با استفاده از این شناسه ها و پارامترهای اضافی مانند نوع شغل و موقعیت مکانی ساخته می شوند.

استراتژی کرال:

دریافت اطلاعات از کاربر:

رنج شروع و پایان شناسه آگهی ها

تعداد آگهی های مورد نظر

محدود کردن رنج:

اعمال محدودیت بر رنج ورودی کاربر با توجه به ظرفیت جاب ویژن

تولید URL رندوم:

انتخاب تصادفی تعدادی شناسه از بین رنج محدود شده

اضافه کردن پارامترهای مورد نظر به شناسه ها

ساخت URL آگهی ها با استفاده از الگوی شناخته شده

استخراج اطلاعات:

کرال صفحات آگهی های منتخب

استخراج اطلاعات آگهی ها مانند عنوان، شرح وظایف، حقوق و مزایا

مزایای این استراتژی:

عدم کرال تکراری: انتخاب تصادفی آگهی ها از تکرار کرال جلوگیری می کند.

قابلیت تنظیم: کاربر می تواند پارامترهای مورد نظر خود را برای کرال مشخص کند.

حفظ کارایی: محدود کردن رنج و استفاده از الگوی URL از بار اضافی روی سرور جاب ویژن جلوگیری می کند.

معایب این استراتژی:

عدم کرال تمام آگهی ها: انتخاب تصادفی آگهی ها ممکن است منجر به عدم کرال تمام آگهی ها شود.

احتمال کرال آگهی های منقضی شده: وجود آگهی های منقضی شده در بین آگهی های کرال شده

نیاز به بررسی: بررسی صفحات کرال شده برای صحت اطلاعات و عدم وجود آگهی های تکراری یا منقضی شده