TEST PROCESS:

شروع کار برنامه

1- وبسایت هایی را که قرار است داده های آنها را استخراج کند می خواند (از پایگاه داده، فایل یا هر منبع دیگر)

2- آماده سازی لینک ها جهت استخراج داده (تمیز کردن، بررسی کردن و اسپم نبودن لینک ها)

3- لینک ها به سه دسته تقسیم می شوند: دسته اول که نباید استخراج شوند که حذف می شوند. دسته دوم لینک های آماده به استخراج که مثلا محصول هستند. اینها مستقیم استخراج می شوند. دسته سوم لینکهایی هستند که کار بیشتر نیاز دارند. مثلا فایل robots.txt دارند که باید استخراج از طریق اون ها انجام بگیره

- در دسته دوم بررسی می شود که اگر لنیک آماده استخراج شدن باشد به صورت مستقیم اطلاعات موجود استخراج شود و در صورتی که استخراج موفقیت آمیر بود اطلاعات در پایگاه داده ذخیره شود

- اگر لینک ها آماده نبود ابتدا بررسی شود که لینک ها قابلیت استخراج را دارند یا خیر. اگر داشتند که همانند روش قبل عملیات استخراج روی آنها انجام گیرد

#################################################################

TEST STACK

1- db: sqlite3

2- chrome driver and requests

3- using futures.concurrent for pararelism

4- Run on-demand

##################################################################################################################################################

##################################################################################################################################################

##################################################################################################################################################

##################################################################################################################################################

##################################################################################################################################################

DEPLOY PROCESS:

شروع کار برنامه

1- وبسایت هایی را که قرار است داده های آنها را استخراج کند می خواند (از پایگاه داده، فایل یا هر منبع دیگر)

2- آماده سازی لینک ها جهت استخراج داده (تمیز کردن، بررسی کردن و اسپم نبودن لینک ها)

3- لینک ها به سه دسته تقسیم می شوند: دسته اول که نباید استخراج شوند که حذف می شوند. دسته دوم لینک های آماده به استخراج که مثلا محصول هستند. اینها مستقیم استخراج می شوند. دسته سوم لینکهایی هستند که کار بیشتر نیاز دارند. مثلا فایل robots.txt دارند که باید استخراج از طریق اون ها انجام بگیره

3-1- در دسته دوم بررسی می شود که اگر لنیک آماده استخراج شدن باشد به صورت مستقیم اطلاعات موجود استخراج شود و در صورتی که استخراج موفقیت آمیر بود اطلاعات در پایگاه داده ذخیره شود

3-2- اگر لینک ها آماده نبود ابتدا بررسی شود که لینک ها قابلیت استخراج را دارند یا خیر. اگر داشتند که همانند روش قبل عملیات استخراج روی آنها انجام گیرد

################################

DEPLOY STACK

1- db: mongodb

2- chrome driver and requests

3- using processing for pararelism

4- using crone job or apscheduler for scheduling

5- rabbitmq as message queue

هدف و تبلیغ پروژه:

در محیط رقابتی امروز دسترسی هر چه سریعتر و کاملتر و سازمان یافته تر به اطلاعات با استفاده از پیشرفته ترین و دقیقترین روش های ممکن نه تنها یک انتخاب اساسی، بلکه یک ضرورت می باشد. ربات استخراج گر و مدیریت داده این عمل را در سریعترین روش ممکن انجام می دهد