爬虫基类的程序调用结构

Start(url,option)

{

InitializtionData

{

ListPageRegAndParamInitialize

ContentPageRegAndParamInitialize

}

if (option.IsListUriTransform)

ListUriTransform

GetContentUrls

{

GetPageListHtmlData

{

if (Options.ListParamMethod == ParamTransMethod.Post)

GetPageDataByPost(PagePostParam)

else if (Options.ListParamMethod == ParamTransMethod.File)

GetPageDataByFile(PageFileRegex)

else if (Options.ListParamMethod == ParamTransMethod.Get)

GetPageDataByGet(PageGetRegex)

else//Dymanic

GetPageDataByDymanic(PageGetRegex, PageFileRegex, PagePostParam)

}

if (IsPageHasList)

GetContentHrefs

{

if (HasPagerInPage)

GetTotalPageCount(TotalPageCountRegex)

while (isKeepingCrawlerData)

GetCurrentConHrefs

{

HrefRegex

if (IsIgnoreUrl)

UrlToAbsUrl

if (IsKeepingCrawlerData)

}

GetPageListHtmlData

else

GetCurrentConHrefs

}

}

GetContentInfoAndSave

{

GetContentInfo

{

SwitchOptionAndRegex

if (Options.IsContentUriTransform)

ContentUriTransform

GetContentHtmlData

{

if (Options. ContentParamMethod == ParamTransMethod.Post)

GetPageDataByPost(ContentPostParam)

else if (Options. ContentParamMethod == ParamTransMethod.File)

GetPageDataByFile(ContentFileRegex)

else if (Options. ContentParamMethod == ParamTransMethod.Get)

GetPageDataByGet(ContentGetRegex)

else//Dymanic

GetPageDataByDymanic(ContentGetRegex,ContentPostParam,ContentFileRegex)

}

if (IsPageHasContent)

GetConTitle

{

TitleRegex

if (HasSubTitle)

SubTitleRegex

}

GetPublishDate{ PublishDateRegex }

GetPublishSource{ PublishSourceRegex }

GetContentAll

{

if (HasPagerInContent)

GetTotalPageCount{ TotalContentCountRegex }

for

GetContent{ ContentRegex }

GetContentHtmlData

else

GetContent

if (Options.IsTransContentImgUrl)

}

}

AddContentInfo

}

}

IsKeepingCrawlerData

{

if (Options.IsCrawlerByNum)

Options.CrawlerNum

Options.MaxCrawlerNum

else

Options.CrawlerDays

Options.MaxCrawlerDays

}