汽车之家口碑爬虫说明

# autohome\_app/items.py

brand\_id = scrapy.Field() ：品牌id

brand\_name = scrapy.Field() : 品牌名称

serialId = scrapy.Field() ：车系id

serialName = scrapy.Field() ：车系名称

AverageRating = scrapy.Field()：价格区间

ResultBad = scrapy.Field() ：差评标签

ResultGood = scrapy.Field() ：好评标签

Koubeiid = scrapy.Field() ：口碑id

goodContent = scrapy.Field() ：好评

badContent = scrapy.Field()：差评

# autohome\_app/middlewares.py

无需中间件

# autohome\_app/pipelines.py

初始化mongdb数据库，从settings.py导入mongodb数据库的信息

def \_\_init\_\_(self):

host = settings['MONGO\_HOST']

port = settings['MONGO\_PORT']

dbName = settings['MONGO\_DB']

client = pymongo.MongoClient(host=host, port=port)

tdb = client[dbName]

self.post = tdb[settings['MONGO\_COLL']]

以插入的方式插入到mongodb数据库中

def process\_item(self, item, spider):

article = dict(item)

self.post.insert(article)

return item

# autohome\_app/settings.py

mongodb数据库信息

MONGO\_HOST = "127.0.0.1" # 主机IP

MONGO\_PORT = 27017 # 端口号

MONGO\_DB = "koubei" # 库名

MONGO\_COLL = "autohome" # collection名

打开中间件，打开pipeline，设置多线程信息

# autohome\_app/spiders/koubei.py

通过下载的车系信息构造starturlList

function:parse

下载starturl下对应的内容通过json解析得到我想要的数据，

function:parse\_koubei

在新的页面里数据里同样通过json解析得到剩下的数据

function:get\_eid

根据eid翻页得到相应的口碑信息然后传到pipeline进行处理存进数据库

# 运行

cd到spiders目录下

scrapy crawl koubei