表 7‑5 items脚本模板

|  |
| --- |
| # -\*- coding: utf-8 -\*-  # Define here the models for your scraped items  #  # See documentation in:  # https://doc.scrapy.org/en/latest/topics/items.html  import scrapy  class TipdmspiderItem(scrapy.Item):  # define the fields for your item here like:  # name = scrapy.Field()  pass |

表 7‑6 TipDMSpider项目items脚本

|  |
| --- |
| class TipdmspiderItem(scrapy.Item):  # define the fields for your item here like:  # name = scrapy.Field()  title = scrapy.Field()  text = scrapy.Field()  time = scrapy.Field()  view\_count = scrapy.Field() |

表 7‑7 piplines脚本模板

|  |
| --- |
| # -\*- coding: utf-8 -\*-  # Define your item pipelines here  #  # Don't forget to add your pipeline to the ITEM\_PIPELINES setting  # See: https://doc.scrapy.org/en/latest/topics/item-pipeline.html  class TipdmspiderPipeline(object):  def process\_item(self, item, spider):  return item |

表 7‑12 TipDMSpider项目piplines脚本

|  |
| --- |
| # -\*- coding: utf-8 -\*-  # Define your item pipelines here  #  # Don't forget to add your pipeline to the ITEM\_PIPELINES setting  # See: https://doc.scrapy.org/en/latest/topics/item-pipeline.html  import pandas as pd  from sqlalchemy import create\_engine  class TipdmspiderPipeline(object):  def \_\_init\_\_(self):  self.engine = create\_engine('mysql+pymysql://root:335210@127.0.0.1:3306/tipdm')  def process\_item(self, item, spider):  data = pd.DataFrame(dict(item))  data.to\_sql('tipdm\_data',self.engine,if\_exists='append',index=False)  data.to\_csv('TipDM\_data.csv',mode='a+',index=False, sep='|', header=False) |

表 7‑14 genspider命令生成的spider脚本模板

|  |
| --- |
| # -\*- coding: utf-8 -\*-  import scrapy  class TipdmSpider(scrapy.Spider):  name = 'tipdm'  allowed\_domains = ['www.tipdm.com']  start\_urls = ['http://www.tipdm.com/']  def parse(self, response):  pass |

表 7‑15 获得网页数spider脚本

|  |
| --- |
| def parse(self, response):  # 网页解析  last\_page\_num = response.xpath("//div[@class='fpage']/div/a[last()]/text()").extract()  # 网址拼接  append\_urls = ['http://www.tipdm.com/tipdm/tddt/index\_%d.html'%i \  for i in range(2,int(last\_page\_num[0])+1)]  append\_urls.append('http://www.tipdm.com/tipdm/tddt')  # 回调  for url in append\_urls:  yield Request(url, callback=self.parse\_url, dont\_filter=True) |

表 7‑17 获得文章网址的方法

|  |
| --- |
| def parse\_url(self, response):  # 网页解析  urls = response.xpath("//div[@class='item clearfix']/div[1]/h1/a/@href").extract()  # 回调  for page\_url in urls:  text\_url = "http://www.tipdm.com"+page\_url  yield Request(text\_url, callback=self.parse\_text, dont\_filter=True) |

表 7‑19 解析文章网址的方法

|  |
| --- |
| def parse\_text(self,response):  item = TipdmspiderItem()  item['title'] = response.xpath("//div[@class='artTitle']/h1/text()").extract()  text = response.xpath("//div[@class='artCon']//p/text()").extract()  texts = " "  for strings in text:  texts = texts + strings + " \n"  item['text'] = [texts.strip()]  item['time'] = response.xpath("//span[@class='date']/text()").extract()  item['view\_count'] = response.xpath("//span[@class='view']/text()").extract()  yield item |

表 7‑21 TipDMSpider项目settings脚本追加设置

|  |
| --- |
| ROBOTSTXT\_OBEY = False  DOWNLOAD\_DELAY = 5  ITEM\_PIPELINES = {  'TipDMSpider.pipelines.TipdmspiderPipeline': 300,  }  HTTPCACHE\_ENABLED = True  HTTPCACHE\_DIR = 'E:/第7章/01-任务程序/tmp' |