# 爬虫说明文档

# [OS Environment]

## 1. Ubuntu 14.04+

# [Develop Environment]

## 1. python 2.x (Ubuntu自带

## 2．scrapy, 爬虫框架

a. install

使用命令行安装：pip install Scrapy安装。

参考：<https://doc.scrapy.org/en/latest/intro/install.html>

## 3. python selenium api, 访问浏览器API的python接口

a.install

使用命令行安装: pip install selenium

b.webdriver selenium

需要安装对应的浏览器的驱动，并且复制到/usr/bin目录下，对应的浏览器需要更新到最新。

参考：http://selenium-python.readthedocs.io/

# [程序说明]

## 1.爬虫确认。

在root目录下，运行scrapy list可以看到所有的爬虫，目前可用的为jd。

## 2.爬虫运行

运行scrapy crawl jd –output=file.csv可执行爬虫，其结果保存在file.csv中。

## 3.爬虫修改

京东爬虫代码位于./spiders/jd.py内。

1. parse函数

用于抓取产品评论页面的url链接。形如：http://item.jd.com/3258432.html#comment

1. parse\_detail函数

抓起评论，目前该功能还未完成。