Robots协议：爬虫协议，网络爬虫排除标准，告诉爬虫那些目录可以抓取，那些不可以抓取，一般放在网站根目录下。

如：

<https://www.jianshu.com/robots.txt>

例：程序crawler\_test/robots\_test.py

# Copyright(C) 2018 刘珅珅  
# Environment: python 3.6.4  
# Date: 2018.9.8  
# robots协议测试  
from urllib.robotparser import RobotFileParser  
url = 'http://www.jianshu.com/robots.txt'  
rp = RobotFileParser(url)  
rp.read()  
# 在scrapy中，如果希望爬取robots.txt禁止的页面，需要禁用robots  
print(rp.can\_fetch('\*', 'http://www.jianshu.com/p/b67554025d7d')) # False，不允许爬取