2.1.1爬虫的定义:

网络爬虫(Web Crawler)又被称为Ants, Automatic Indexer, Spiders, Robots

或者Wanderers，就是抓取特定网站网页的HTML数据。网络爬虫从一个存放URL

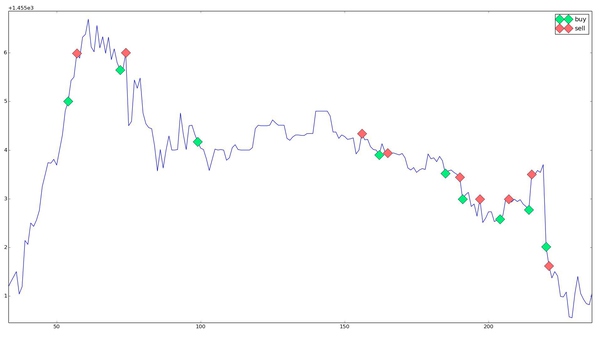
的集合开始进行爬取，利用HTTP协议连续地从不同的服务器中获取Web页面,首先从队列中获取一个URL并下载此网页，提取该网页中的其它URL并放入队列中。此过程将重复直至关闭的计算机程序。现在很多网站，尤其是搜索引擎都将网络爬虫作为一种获取、提供互联网最新数据的手段川。

2.1.2爬虫的作用与分类:

爬虫能用来干什么.如今每天互联网上的流量足足有10亿GB左右，不可能毫无选择的全部爬取下来，这是要根据我们自己的需求来有选择的爬取相对应的数据.那爬虫可以用来干什么呢？我们根据实际需要将爬虫的目标分为三类.

第一类是通过搜索引擎的网络爬虫来充实搜索引擎的索引列表.这部分需求所需要的数据是各种网页的地址，标题，主题字等等.这方面比较知名的有google爬虫，baidu爬虫，Yahoo爬虫等等。可以大致分为 批量型爬虫（Batch Crawler），增量型爬虫（Incremental Crawler），垂直型爬虫(Focused Crawter）.

第二类建立自己的数据仓库，然后用各种机器学习模型来得出一些寻常无法得出的预测和分析.我们可以通过这些数据集来进行预测.比如说股票，比特币和各种风险交易。



只不过要达到这种水平，那么需要爬取到特别优质的数据集.而比特币市场比起股票来说是一个很简单的市场，但是这也能证明爬虫在该方面的运用.

注：该MIT论文的链接是[https://arxiv.org/pdf/1410.1231v1.pdf](https://arxiv.org/pdf/1410.1231v1.pdf%29.这篇论文以比特币市场为例，用爬虫爬取的各种高频交易的信息，通过机器学习建立的人工智能成功的预测比特币短期价格的涨跌，然后做高频交易.下面就是一张MIT只用了3天的training)

第三类为各种论文以及文章提供有力的数据支撑。

我们一般在报告中看到的有关于各行各业的数据分析图，那都是通过抓取特定行业的专业数据，然后用各种数据分析也好，自己使用第三方可视化JavaScript库也好，最终得出的这些表图.同样用例子来说明.比例如，我们要找出全国气温最低的地方并直观的表示出来，那么我们仅仅需要爬取一下全国的气象信息，然后通过类似于D3.js这种可视化JavaScript库就可以得到一张柱状图.如下图所示，这样，全国最低气温就能直观的显示出来。同时也可以用过各种气象论文中论点的有力支撑.

