Task: 调研开源网络爬虫，并且说明各自的特点、局限性以及相互之间的区别.

Python:

主要有Scrapy和Pyspider.

Scrapy

特点：模块化，基于中间件机制，扩展性强，易开发。扩展可支持分布式。

支持抓取和抽取，数据抽取结构化抽取支持xpath和css提取网页数据。

局限性：不支持JS页面抓取。单机多线程实现，不支持分布式，默认中间过程的网页不保存，只保存抽取结果。

Pysider

特点：强大的爬虫管理，有任务监控UI。适用于页面监控与更新。支持JS页面抓取。支持分布式。

局限性：无插件。

Java:

主要有Nutch,Crawler4j,WebCollector,WebMagic.

Nutch

特点：基于Hadoop任务的分布式处理模型，总体分为抓取和搜索两个部分。存储层剥离，支持存储HBase, Cassandra, MySql等数据库；

基于插件式设计，扩展和定制比较方便；支持网页解析和索引，可以搭建通用的搜索引擎。

局限性：开发复杂,上手难度高，不支持JS页面抓取。

Crawler4j

特点：用于简单数据采集

局限性：无插件，不支持分布式，不支持JS页面抓取。

WebCollector

特点：易开发，支持JS页面抓取，

局限性：不支持分布式。