1. 内容提取核心算法
   1. 分析HTML页面内容
   2. 将页面转换成为符合XML的标记的格式
   3. 使用中文匹配方式 得到一段段连续的中文字.出现的数据集合.
   4. 对这个中文数据集合进行排列,将与上一个表达式相隔的位置记录下来
   5. 然后将这些最近的中文进行组合,中间用正则表达式来联接.
   6. 填充距离，位置记数．
   7. 遍历有效内容集合，对 元素 距离 位置记数 进行分析 如果两个 相近的原素间 space 值过大则可能是两个不同的原素了
   8. 对有效结果进行分组，当space超过一个设定的经验值时将结果分到不同的组.
   9. 当分组完成后,提取元素最多的那个组做为有效内容
   10. 然后将这些最近的中文进行组合,中间用正则表达式来联接.
   11. 当距离最近的几个联接完成后,从原始文档中提取这部分内容.
   12. 对提取的这部分内容进行分析,得出得到包含这些内容的最小的一个元素标记．
   13. 从这个原素标记中提取 文章的核心内容
2. 列表分析
   1. 对一某一站点上的某一种类的数据进行分析.可以得到一个页面上所包含某类型链接最多的数据.这类数据就是当前页面的重点.
3. 内容路径分析