1. 如何判断一个中文文档使用的字符集？

预先收集大量语料，建立一个通用语料库，并用不同编码进行表示，并分别建立统计语言模型。对于新加入的文档，可以截取一定长度的字符串，与各个统计语言模型比较，最大者对应编码集即为所用编码集。

1. html和xml两种表示的区别？

XML 不是 HTML 的替代，XML 和 HTML 为不同的目的而设计：

* HTML 被设计用来显示数据，其焦点是数据的外观。

HTML（超文本标记语言）是一种标记语言，用于创建网页和Web应用程序。通过HTML，我们可以创建自己的静态页面。它用于显示数据，而不是传输数据。

HTML是超文本和标记语言的组合。超文本定义了网页之间的链接。标记语言用于定义标记内的文本文档，该文档定义网页的结构。此语言用于注释（为计算机做笔记）文本，以便机器可以理解它并相应地操作文本。

* XML 被设计用来传输和存储数据，其焦点是数据的内容。

XML（可扩展标记语言）也用于创建网页和Web应用程序；它是动态的，因为它用于传输数据而不是用于显示数据。

XML的设计目标关注于Internet上的简单性、通用性和可用性。它是一种文本数据格式，通过Unicode对不同的人类语言提供了强大的支持。虽然XML的设计主要关注文档，但是这种语言广泛用于表示任意数据结构，比如web服务中使用的数据结构。

简而言之，HTML 旨在显示信息，而 XML 旨在传输信息。

因而，可以把XML看成是对 HTML 的补充：

XML 不会替代 HTML，理解这一点很重要。在大多数 Web 应用程序中，XML 用于传输数据，而 HTML 用于格式化并显示数据。

对 XML 最好的描述是：XML 是独立于软件和硬件的信息传输工具。

**从语法和标记上看：**

语法要求：

（1）在html中不区分大小写，在xml中严格区分。

（2） 在HTML中，有时不严格，如果上下文清楚地显示出段落或者列表键在何处结尾，那么你可以省略</p>或者</li>之类的结束标记。在XML中，是严格的树状结构，绝对不能省略掉结束标记。

（3） 在XML中，拥有单个标记而没有匹配的结束标记的元素必须用一个/ 字符作为结尾。这样分析器就知道不用查找结束标记了。

（4） 在XML中，属性值必须分装在引号中。在HTML中，引号是可用可不用的。

（5）在HTML中，可以拥有不带值的属性名。在XML中，所有的属性都必须带有相应的值。

（6） 在XML文档中，空白部分不会被解析器自动删除；但是html是过滤掉空格的。

标记：

（1）html使用固有的标记；而xml没有固有的标记。

（2）Html标签是预定义的；XML标签是免费的、自定义的、可扩展的。

1. 如何用概率模型实现“被访问次数”的估计？

随机选取一个页面P，对其赋予一个比较高的“访问次数”作为初始值；将访问次数乘上1/n（n为链出的数目），作为后继节点的访问次数，得到若干个非0新节点；对这些节点继续重复上述过程。。。

1. 为什么1,5,6三个网页的重要度变为0？

从网页的图结构看，因为2,3,4三个页面没有到1,5,6三个页面的链出，而2,3,4三个页面不仅内部链接形成了的环路，而且也有来自1,5,6三个页面（主要是1,6两个页面）的链入。

在我看来，通俗的说，迭代的过程就是一个“支持度传导”的过程——各个页面都将本页面的支持度（权值w）按条件概率不断传导给链出页面，又从链入页面接受支持度的过程。但由于上述的页面间的连接结构，导致2,3,4三个页面不仅能够“自给自足”，还能收到来自1,5,6三个页面的支持；相反1,5,6三个页面在传导的过程中“入不敷出”，相对于2,3,4三个页面而言偏向于“输出”的一方。在多次迭代后，1,5,6三个页面权值降为0，而2,3,4三个页面权值较高。

从计算上来看，B的第1,5,6行大多数为0，故相对更易变小；2,3,4行的非零项更多，容易增大。

解决这个问题，可以在迭代的过程中，在每次迭代过程中给每项人为的增添一个小的常数项，相当于输入“外源支持度”。