2.3

1. 下载系统：网页下载，负责从万维网上自动下载各种类型的网页到网页库中，并且保持对万维网变化的同步；

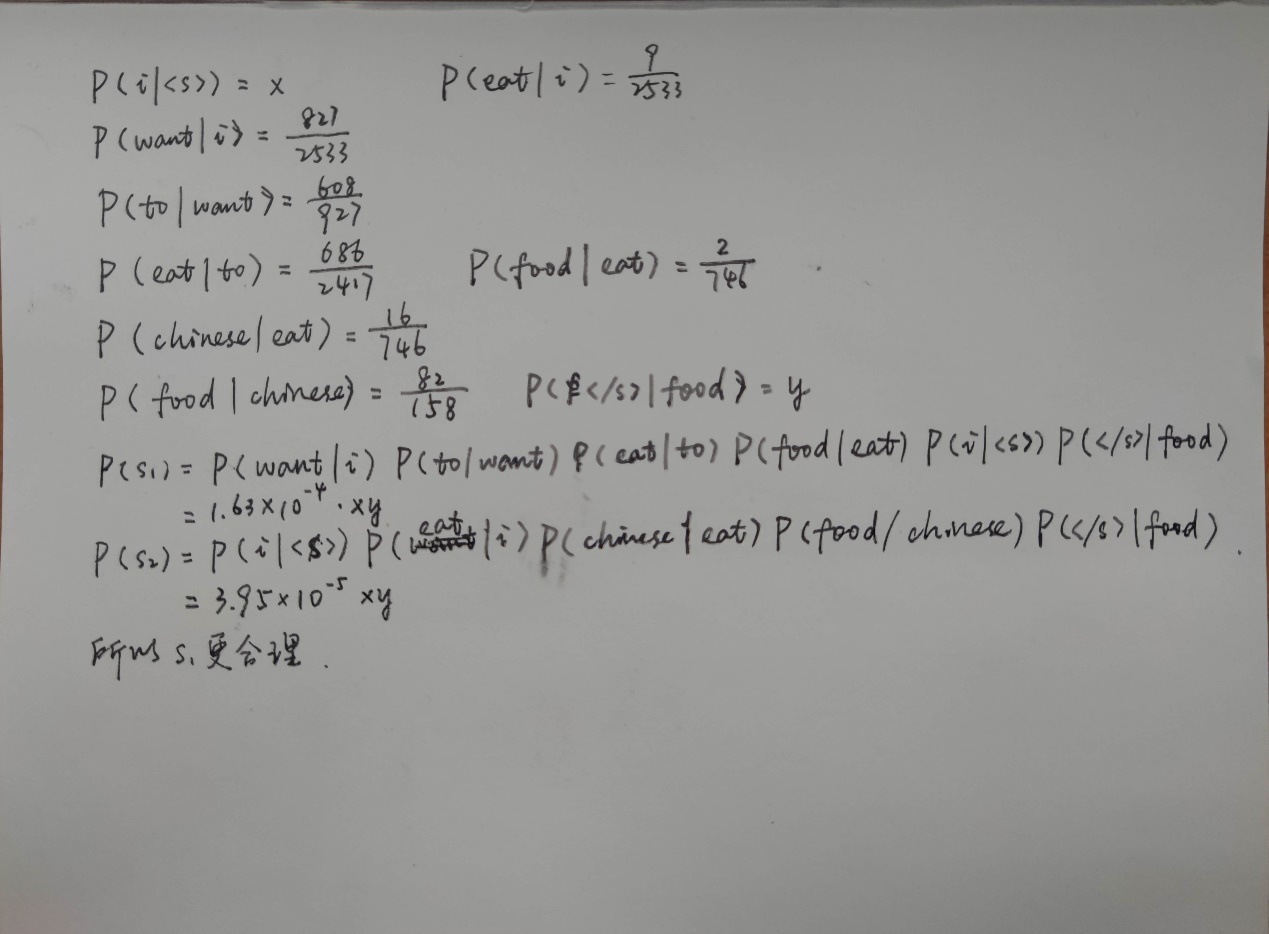
分析系统：信息抽取,负责抽取下载系统得到的网页数据，并进行PageRank(网页排序)和分词计算，发现信息；

索引系统：负责将分析系统处理后的网页对象建立索引并入库；

查询系统：负责分析用户提交的查询请求，然后从索引库中检索出相关网页，返回查询结果。

1. (1) 统计；N；gram；频数；整体语料库；概率；乘积。

(2) ①各种搜索功能时根据用户输入产生的“联想”，即预测用户后续输入内容的功能可以通过N-gram模型计算文本库中每个gram出现的概率而实现；②对文章内容分类：通过N-gram模型获取文章“高频词”。

(3) 

2.4

1. Hit：表示索引词在文章中的位置（position）等信息；

Forward Index: 基于文档（doc）来查询关键字key；

Inverted Index：基于关键字key来查询文档（doc），关键词称为“索引词”。

1. 正向索引：优点：易维护，每新增一个doc只需要把每个key在该doc中 出现的次数和位置进行维护即可；缺点：搜索耗时太长

倒排索引：优点：搜索效率高；缺点：维护成本高，每新增一个doc，需要维护倒排索引

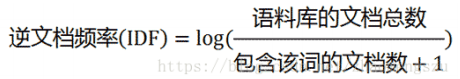


|  |  |  |  |
| --- | --- | --- | --- |
| 单词ID | 单词 | 文档频率 | 倒排列表（Doc ID; TF; <POS>） |
| 1 | The | 3 | (1;2;<1;5>), (2;1;<1>) |
| 2 | cat | 3 | (1;1;<2>), (2;1;<2>), (3;1;<5>) |
| 3 | sit | 1 | (1;1;<3>) |
| 4 | on | 1 | (1;1;<4>) |
| 5 | mat | 1 | (1;1;<6>) |
| 6 | catches | 1 | (2;1;<3>) |
| 7 | mice | 1 | (2;1;<4>) |
| 8 | Here | 1 | (3;1;<1>) |
| 9 | is | 1 | (3;1;<2>) |
| 10 | a | 1 | (3;1;<3>) |
| 11 | little | 1 | (3;1;<4>) |

2.5

1. 真实用户；检索代理；查询词；检索词。
2. “布尔模型”优点：易于实现，检索速度快。缺点：没有考虑文档和查询词的相关性问题没有区分查询词的权重问题，“效果”差；检索结果无法排序。
3. 词频：表示某关键词在文件中出现的次数，通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。

停用词：去掉没有用的词，例如‘的’，‘是’这样的词；

逆文档频率IDF：在词频的基础上，对每个词分配一个"重要性"权重，它是一个词语普遍重要性的度量；某一特定词语的IDF由总文件数目除以包含该词语之文件的数目，再将得到的商取以10为底的对数得到；

TF-IDF值：表示某个词对文章的重要性，"词频"（TF）和"逆文档频率"（IDF）相乘得到“TF-IDF” = TF × IDF，某词对文章重要性越高TF-IDF值就越大，排在最前面的几个词，就是这篇文章的关键词。



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 词表 | 水果 | 有 | 苹果 | 桃子 | 西瓜 | 菠萝 | 梨子 | 蔬菜 |
| 词频-内容1 | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 词频-内容2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 词频-内容3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 词频-内容4 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 词频-内容5 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 文档频率 | 3 | 3 | 3 | 1 | 2 | 2 | 2 | 1 |
| 逆文档频率 | 5/3 | 5/3 | 5/3 | 5 | 2.5 | 2.5 | 2.5 | 5 |
| TF-IDF简易计算 |  | | | | | | | |
| 内容1 | 10 | 5/3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 内容2 | 5/3 | 5/3 | 5/3 | 5 | 2.5 | 2.5 | 2.5 | 0 |
| 内容3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 内容4 | 5/3 | 0 | 5/3 | 0 | 0 | 0 | 2.5 | 0 |
| 内容5 | 0 | 5/3 | 5/3 | 0 | 2.5 | 2.5 | 0 | 0 |