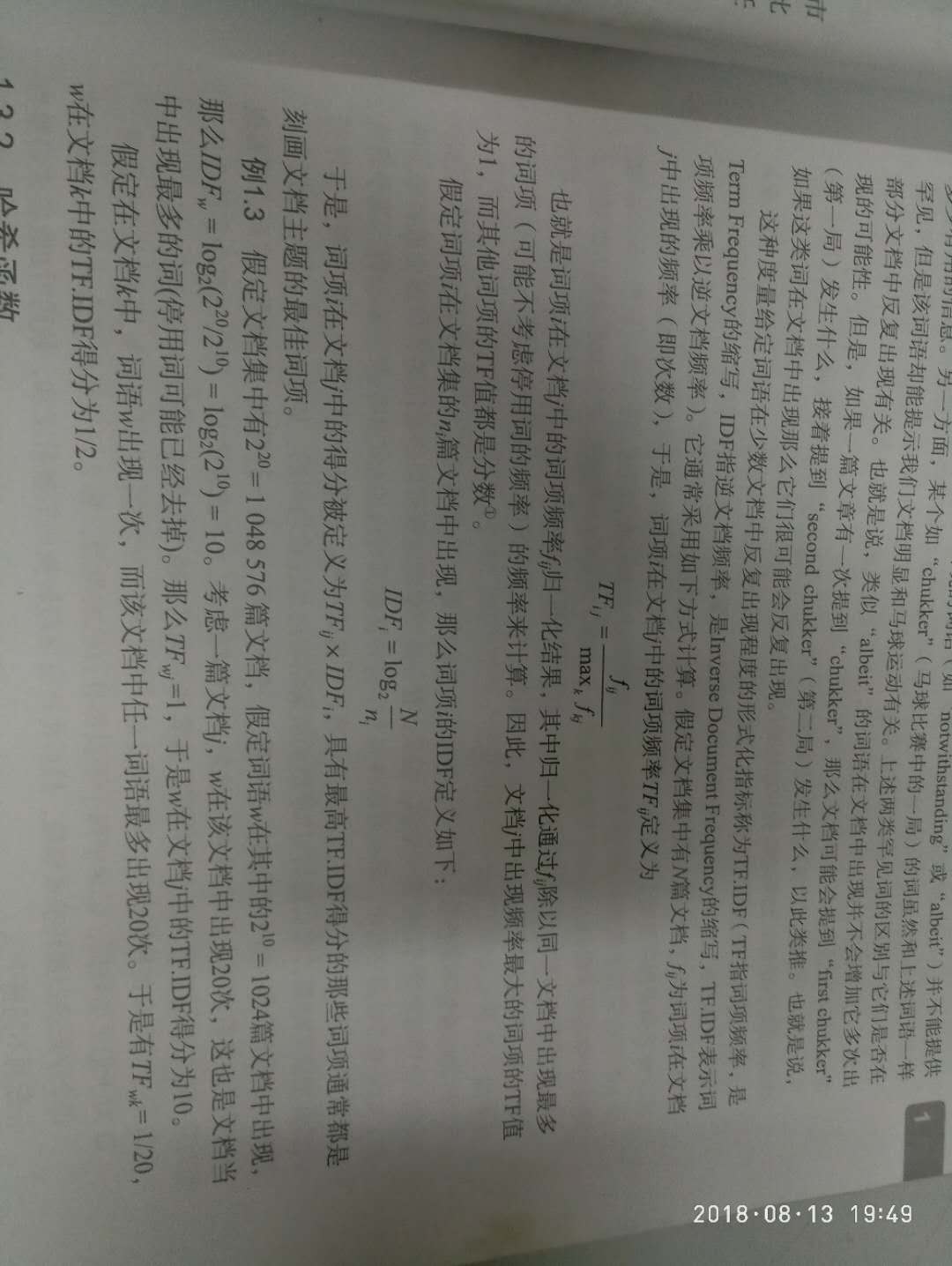
数据挖掘：data mining

数据汇总形式：PageRank，聚类

度量给定词语在少数文档中反复出现程度的形式化指标成为TF.IDF（TF指词项频率，Term Frequency，IDF是逆文档频率，是Inverse Document Frequency的缩写，TF.IDF表示词项频率乘以逆文档频率）

哈希函数：Hash，一般翻译做"散列"，也有直接音译为"哈希"的，就是把任意长度的输入（又叫做预映射， pre-image），通过散列算法，变换成固定长度的输出，该输出就是散列值。这种转换是一种压缩映射，也就是，散列值的空间通常远小于输入的空间，不同的输入可能会散列成相同的输出，而不可能从散列值来唯一的确定输入值。简单的说就是一种将任意长度的消息压缩到某一固定长度的消息摘要的函数。

哈希表是一种简单地索引构建方法

磁盘组织成块结构，每个块是操作系统用于在内存和磁盘之间传输数据的最小单元

幂定律（power law），两个变量在对数空间下呈现出线性关系

马太效应：强者越强，弱者越弱，是社会学家和经济学家们的常用术语

邦弗朗尼原理：在考察数据是，如果将某些对象视为数据的有趣特征，而这些对象中的许多都可能会在随机数据中出现，那么这些显著的特征就不可依赖，对于那些实际中并不充分罕见的特征来说，上述观察结果限制了从这些数据特征中进行挖掘的能力

分布式文件系统（Distributed File System）

文件块（chunk）

分布式文件系统只在文件巨大、更新很少的情况下才更有意义

关系代数运算：

1. 选择（Selection）
2. 投影（Projection）
3. 并（Union）、交（Intersection）及差（Difference）
4. 自然连接（Natural Join）
5. 分组和聚合（Grouping and Aggregation）

Jaccard相似度，集合S与T的交集与S和T的并集的比值，可以记为SIM（S，T）