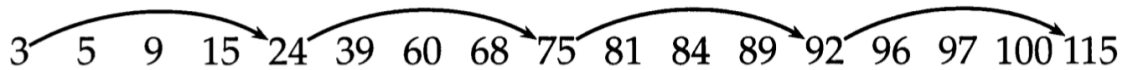
Web 信息处理与应用：课后作业 1
网页信息处理 + 网页索引部分（第 2-4 周课程）
请于 2022 年 10 月 11 日 23:59 前将作业电子版发送至课程邮箱：ustcweb2022@163.com

1 计算题

1.1 请推荐如下查询的处理次序。
(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)
其中，每个词项对应的倒排记录表的长度分别如下：

词项	倒排记录表长度
eyes	213 312
kaleidoscope	87 009
marmalade	107 913
skies	271 658
tangerine	46 653
trees	316 812

1.2 考虑利用如下带有跳表指针的倒排记录表



和一个中间结果表（如下所示，不存在跳表指针）进行合并操作。
3 5 89 95 97 99 100 101
采用基于跳表指针的倒排记录表合并算法，请问：
1) 跳表指针实际发生跳转的次数是多少？
2) 当两个表进行合并时，倒排记录之间的比较次数是多少？
3) 如果不使用跳表指针，那么倒排记录之间的比较次数是多少？

1.3 写出倒排记录表（777, 17743, 294068, 31251336）的可变字节编码。在可能的情况下对间距而不是文档 ID 编码。写出 8 位块的二元码（即可变长度编码）。

1.4 假设有三个城市，编号分别为 1、2、3。现在有一个商人在三个城市之间来回穿梭，已知三个城市作为起点的概率分别为（0.2，0.4，0.4）。同时，这个商人在城市之间旅行或同城停留的跳转概率如下表所示：

城市编号	->1	->2	->3
1	0.5	0.2	0.3

2	0.3	0.5	0.2
3	0.2	0.3	0.5

同时，还知道三座城市各自晴天/雨天的概率如下表所示：

城市编号	1	2	3
晴天概率	0.5	0.4	0.7
雨天概率	0.5	0.6	0.3

在某一次旅行中，商人连续三天观测到的天气状态是（晴天、雨天、晴天），请问，这三天内该名商人最有可能的旅行轨迹是什么？请给出计算过程。

2 问答题（言之有理即可）

2.1 针对海量数据爬取的任务，在商用环境下往往采用分布式爬虫，通过对 URL 的哈希结果来进行任务分配。然而，服务器陷阱、节点崩溃等原因将导致在实际运行中会出现节点的减少或新增。请思考在节点数量动态变更的情况下，采用何种策略可以保障负载均衡？

2.2 如何结合查询词项的分布细节，预设相对合理的跳表指针步长，或实现跳表指针步长的动态调节？

2.3 在信息检索系统中，如何同时使用位置索引（对倒排索引的位置信息扩展）和停用词表？潜在问题有哪些，如何解决？

2.4 机械分词的缺陷之一在于词汇频率无法对分词结果产生影响。请设计方案将词汇概率融入正/反向最大匹配分词，以提升分词效果。并考虑：在此种方案下，反向最大匹配分词的效果是否仍优于正向最大匹配分词，为什么？

2.5 Trie 树的缺陷在于“以空间换效率”，对于存储空间的压力较大。如何结合英文/中文的语言特点，适当放宽限定，以节约 Trie 树的存储空间？同时，请分析这一改进下对于查询效率的影响。