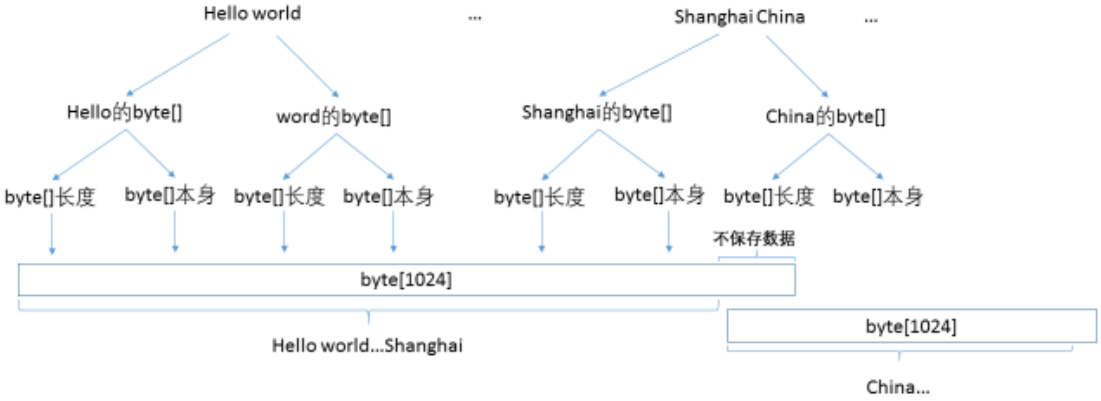
编程作业 1: 编程基础 (总分 20 分)

[1] 数据结构化存储和查询：10分

1、给定文本文件（sample.txt）按照标点符号、空格、换行符等特殊分隔符提取单词（不计数字），并对所提取单词按照alphabet进行排序（小写a-z，然后A-Z），对每个单词转化为字节数组byte[]。

2、依次将每个单词对应的字节数组byte[]拷贝到一个单位长度为1024的字节数组byte[1024]，在进行copy的过程中，若copy第i个单词的字节数组byte[]超出byte[1024]范围，那么当前字节数组byte[1024]则只保存截止第(i-1)个单词的字节数组byte[]，然后将该字节数组byte[1024]写入文件（文件名为sort.dat）；然后重新初始化字节数组byte[1024]，将第i个单词的字节数组byte[]拷贝到字节数组byte[1024]。示意图如下所示，确保将单词的字节数组以固定长度1024的结构形式进行保存。



3、读取test.txt中的所有words，以random access方式访问sort.dat文件，查询test.txt中每一个word在sort.dat中的下一个排序单词，并写到按行日志文件(out.log)中。

4、在以上步骤中，统计步骤2的存储时间和步骤3的查询时间，追加到out.log最后2行

测试要求：助教会给定10组sample.txt和test.txt文件，学生要求输出日志文件out.log，并画曲线图，x轴为每组文件的编号，y轴为(存储/查询)时间.

[2] 多客户端查询：10分

1、在服务器端启动Server Socket服务；客户端通过Socket接口连接服务器，将客户端本地test.txt文件中words发送至服务器；服务器获取客户端发送的words，按照[1]作业的要求将结果返回至客户端。

2、在满足上述功能的情况，通过多线程服务优化客户端获取访问结果的时间，具体如下，设定服务器端启动m个线程(自1－5)、服务器端启动n个线程(自1－5)，计算客户端完成获取结果的整体时间，要求画一个3维曲线图，x轴为m的值，y轴为n的值，z轴为时间。

作业提交

1、Deadline：**2019/10/20 23：59** PM，代码实现：Java （推荐）或者 Python。

2、内容：代码+文档（含作图）：hw1-[学号]-zip

3、目录结构：以每个任务q1/q2的source代码、binary代码、doc文档、in输入数据、out输出结果，来组织目录结构。根据该目录结构，助教的测试代码会首先将助教提供的10组sample.txt依次copy到in子目录，然后从binary子目录的main程序（该命名为hard code）载入可执行代码，该main程序要求从in子目录读取数据（包括sample.txt和test.txt），然

后将结果输出至out子目录, 助教的测试代码然后对比out子目录的输出结果与正确结果, 最终根据对比结果自动打分。