课下作业 1：汉字的计算机表示

1. 请回溯汉字当初为什么无法在计算机内表示

首先需要明确的是所有的数据在存储和运算的时候都是使用二进制数来表示的，这些数据包括字母、数字以及一些常见的字符，而具体使用哪些二进制数表示哪些符号就需要制定标准，由于最早的计算机字符表示的标准是由美国的有关标准化组织出台的 ASCII 编码来表示的，由于汉字并不能使用字母表示，因此汉字当初无法在计算机内表示。

2. 请梳理支持汉字的字符编码方式

由于 ASCII 中除了常用字符、数字就只有英文字母，这些显然是无法满足全世界所有国家使用的，因此每个国家都开始编写自己国家的编码，统称为 MBCS（Muilti-Bytes Charecter Set）。在这个阶段，中国在 1980 年发布了 **GBK2312**，这个编码用区位码（94 个区，每区 94 个字符）的方式可以支持 7000 多个汉字，它所收录的汉字已经覆盖中国大陆 99.75%的使用频率，基本可以满足汉字计算机的需要了。接下来 1995 年中国发布了 **GBK1.0**，可以支持 2 万多个汉字，并在 2000 年发布了 **GB18030**，可以支持 2 万 7 千多个汉字。

在 MBCS 阶段，由于各个国家各自对各自的语言进行编码，因此存在两个国家的编码中一个二进制表示可能在两种编码中表示两种意思，为了进行全世界的编码统一，1990 年开始研发，1994 年正式发布的 Unicode 编码方式同意了世界上的各种文字和符号。**Unicode**（统一码、万国码、单一码）是一种在计算机上使用的字符编码。Unicode 是为了解决传统的字符编码方案的局限而产生的，它为每种语言中的每个字符设定了统一并且唯一的二进制编码，规定所有的字符和符号最少由 16 位来表示（2 个字节）。

而由于 Unicode 编码对于编码长度有所浪费，因此在 1992 年推出了可变长度的编码形式 **UTF-8**，在这种编码中，ASCII 码中的内容用 1 个字节保存、欧洲的字符用 2 个字节保存，东亚的字符用 3 个字节保存。

3. 请谈谈你对这件事的看法

汉字信息处理与印刷革命主要进行的就是汉字信息的数字化存储以及激光照排进行印刷的技术革新。在我看来这一研究的意义就在于当前我们处在信息化的时代，因此在交互过程中不可避免的必须使用计算机，因此如果汉字无法在计算机中进行数字化存储那么就意味着所有中国使用的计算机必须使用英文作为交流，在计算机上进行的一切交流、操作都需要使用英文作为载体，同时汉字印刷技术同理，如果不从铅字印刷产生突破，那么就意味着要么坚持一本书印刷时间需要一年，或者接受英文作为印刷物的核心语言。汉字信息处理与印刷革命的主要意义就在于在信息时代汉字和中华文化的传承与发展创造了条件，这一点对于我国来说意义相当重大。