哈爾濱工業大學

自然语言处理

题　　目　作业一

专　　业　人工智能

学　　号　1190201303

班　　级　1903601

学　　生　王艺丹

指 导 教 师　杨沐昀

计算机科学与技术学院

2021 年 10 月

一、 请回溯汉字当初为什么无法在计算机内表示

最初计算机只在美国使用，八位的字节一共可以组合出 256(2^8)种不同的状态。他们把其中的编号从 0 开始的 32 种状态分别规定了特殊的用途，一直编到了第 127 号，这样计算机就可以用不同字节来存储英语的文字了。大家看到这样，都感觉很好，于是大家都把这个方案叫做 ANSI 的”Ascii”编码（American Standard Code for Information Interchange，美国信息互换标准代码）。当时世界上所有的计算机都用同样的 ASCII 方案来保存英文文字。

后来，世界各地的都开始使用计算机，但是很多国家用的不是英文，他们的字母里有许多是 ASCII 里没有的，为了可以在计算机保存他们的文字，他们决定采用 127 号之后的空位来表示这些新的字母、符号，还加入了很多画表格时需要用下到的横线、竖线、交叉等形状，一直把序号编到了最后一个状态 255。从 128 到 255 这一页的字符集被称“扩展字符集”。

等中国人们得到计算机时，已经没有可以利用的字节状态来表示汉字了。并且汉字是表意字符，一个字是一个方块图形；计算机对汉字信息处理需要对汉字本身进行编码，但汉字数量巨大，总数超过 6 万字，给汉字在计算机内部的表示、汉字的传输与交换、汉字的输入和输出等带来了一系列问题，过于复杂。

二、 请梳理支持汉字的字符编码方式。

2.1 ASCII 码

GB2312:一个小于 127 的字符的意义与原来相同，但两个大于 127 的字符连在一起时，就表示一个汉字，前面的一个字节（他称之为高字节）从 0xA1 用到 0xF7，后面一个字节（低字节）从 0xA1 到 0xFE，这样我们就可以组合出大约 7000 多个简体汉字了。在这些编码里，我们还把数学符号、罗马希腊的字母、日文的假名们都编进去了，连在 ASCII 里本来就有的数字、标点、字母都统统重新编了两个字节长的编码，这就是常说的“全角”字符，而原来在 127 号以下的那些就“半角”字符了。 中国人民看到这样很不错，于是就把这种汉字方案叫做“GB2312”。GB2312 是对 ASCII 的中文扩展

GBK:不再要求低字节一定是 127 号之后的内码，只要第一个字节是大于 127 就固定表示这是一个汉字的开始，不管后面跟的是不是扩展字符集里的内容。结果扩展之后的编码方案被称为 GBK 标准，GBK 包括了 GB2312 的所有内容，同时又增加了近 20000 个新的汉字（包括繁体字）和符号。

GB18030: 在 GBK 基础上再扩展，又加了几千个新的少数民族的字。

2.2 Unicode

从 unicode 开始，无论是半角的英文字母，还是全角的汉字，它们都是统一的“一个字符”，同时，也就是统一的“两个字节”。

UTF-8/UTF-16：UTF-8 就是每次 8 个位传输数据，而 UTF-16 就是每次 16 个位。UTF-8 就是在互联网上使用最广的一种 unicode 的实现方式，这是为传输而设计的编码，并使编码无国界，这样就可以显示全世界上所有文化的字符了。UTF-8 最大的一个特点，就是它是一种变长的编码方式。它可以使用 1~4 个字节表示一个符号，根据不同的符号而变化字节长度，当字符在 ASCII 码的范围时，就用一个字节表示，保留了 ASCII 字符一个字节的编码做为它的一部分，注意的是 unicode 一个中文字符占 2 个字节，而 UTF-8 一个中文字符占 3 个字节）。从 unicode 到 UTF-8 并不是直接的对应，而是要过一些算法和规则来转换

三、 谈谈对下述事情的看法

2001 年，中国工程院颁发了“二十世纪我国重大工程技术成就”评选结果，“汉字信息处理与印刷革命”当选第二项，比第一项“两弹一星”仅差一票。

答：1975 年，以王选为技术总负责人的科研团队开始从事我国“汉字信息处理系统工程”（简称“748 工程”）中“汉字精密照排”项目的研究。他带领团队研制的“汉字信息处理与激光照排系统”，攻克了汉字信息的数字化存储和输出等世界性难关，实现了原创性核心技术突破。为信息时代汉字和中华文化的传承与发展创造了条件。 这些成果的产业化和应用，开创了汉字印刷的一个崭新时代，引发了我国报业和印刷出版业“告别铅与火，迈入光与电”的技术革命，彻底改造了我国沿用上百年的铅字印刷技术，被誉为“汉字印刷的第二次发明”。实现了我国出版印刷行业“告别铅与火，迈入光和电”的技术革命，成为我国自主创新和用高新技术改造传统行业的典范，为激光照排行业创造了巨大的经济和社会效益。

我觉得这对于中国传统出版行业来说是一次历史性的变革，是伟大的里程碑，与两弹一星的领域不同，但都是对其领域内的一次重大突破，更是影响中国发展的重大工程技术成就，意义非凡。