下面这个小工具包含了 判断 unicode 是否是汉字，数字，英文，或者其他字符。 全角符号转半角符号。 unicode 字符串归一化等工作。 还有一个能处理多音字的汉字转拼音的程序，还在整理中。

```
  #!/usr/bin/env python
# -*- coding:GBK -*-

"""汉字处理的工具:
判断 unicode 是否是汉字，数字，英文，或者其他字符。
全角符号转半角符号。"""

def is_chinese(uchar):
                """判断一个 unicode 是否是汉字"""
                if uchar >= u'\u4e00' and uchar<=u'\u9fa5':
                                return True
                else:
                                return False

def is_number(uchar):
                """判断一个 unicode 是否是数字"""
                if uchar >= u'\u0030' and uchar<=u'\u0039':
                                return True
                else:
                                return False

def is_alphabet(uchar):
                """判断一个 unicode 是否是英文字母"""
                if (uchar >= u'\u0041' and uchar<=u'\u005a') or (uchar >= u'\u0061'
and uchar<=u'\u007a'):
                                return True
                else:
                                return False

def is_other(uchar):
```

```
        """判断是否非汉字，数字和英文字符"""
        if not (is_chinese(uchar) or is_number(uchar) or
is_alphabet(uchar)):
                return True
        else:
                return False

def B2Q(uchar):
        """半角转全角"""
        inside_code=ord(uchar)
        if inside_code<0x0020 or inside_code>0x7e:          #不是半角字
符就返回原来的字符
                return uchar
        if inside_code==0x0020:#除了空格其他的全角半角的公式为:半角=全角
-0xfee0
                inside_code=0x3000
        else:
                inside_code+=0xfee0
        return unichr(inside_code)

def Q2B(uchar):
        """全角转半角"""
        inside_code=ord(uchar)
        if inside_code==0x3000:
                inside_code=0x0020
        else:
                inside_code-=0xfee0
        if inside_code<0x0020 or inside_code>0x7e:          #转完之后不
是半角字符返回原来的字符
                return uchar
        return unichr(inside_code)
```

```
def stringQ2B(ustring):
        """把字符串全角转半角"""
        return "".join([Q2B(uchar) for uchar in ustring])

def uniform(ustring):
        """格式化字符串，完成全角转半角，大写转小写的工作"""
        return stringQ2B(ustring).lower()

def string2List(ustring):
        """将 ustring 按照中文，字母，数字分开"""
        retList=[]
        utmp=[]
        for uchar in ustring:
                if is_other(uchar):
                        if len(utmp)==0:
                                continue
                        else:
                                retList.append("
".join(utmp))
                                utmp=[]
                else:
                        utmp.append(uchar)
        if len(utmp)!=0:
                retList.append("".join(utmp))
        return retList

if __name__=="__main__":
        #test Q2B and B2Q
        for i in range(0x0020,0x007F):
                print Q2B(B2Q(unichr(i))),B2Q(unichr(i))
```

```
#test uniform
ustring=u'中国 人名 a 高频 A'
ustring=uniform(ustring)
ret=string2List(ustring)
```