code

```
def isChinese(ch):
    if 0x4e00 <= ord(ch) < 0x9fa6:
        return True
    return False

with open("红楼梦.txt", "r", encoding="utf-8", errors="ignore") as f:
    hl_str = f.read() #honglou
with open("水浒传.txt", "r", encoding="utf-8", errors="ignore") as f:
    sh_str = f.read() #shuihu

hl_str = "".join([x for x in hl_str if isChinese(x)])
sh_str = "".join([x for x in sh_str if isChinese(x)])

#第2题
ch2num_hl = {}
for i in hl_str:
    if i not in ch2num_hl:
        ch2num_hl[i] = 1
    else:
        ch2num_hl[i] += 1

sorted_ch_num_hl = list(sorted(ch2num_hl.items(), key=lambda x: x[1], reverse=True)
hl_len = len(hl_str)

f = open("2.txt", "w", encoding="utf-8")
for item in sorted_ch_num_hl[0:500]:  #top500
    frequency = item[1] / hl_len
    print(item[0] + ',' + '%.8f' % frequency, end="\n", file=f)
f.close()

ch2num_sh = {}
for i in sh_str:
    if i not in ch2num_sh:
        ch2num_sh[i] = 1
    else:
        ch2num_sh[i] += 1

sorted_ch_num_sh = list(sorted(ch2num_sh.items(), key=lambda x: x[1], reverse=True)
sh_len = len(sh_str)
secondfile = open("2.txt", "a", encoding="utf-8")
for item in sorted_ch_num_sh[0:500]:
    frequency = item[1] / sh_len
    print(item[0] + ',' + '%.8f' % frequency, end="\n", file=secondfile)
secondfile.close()


#第3题
hl_set = set(hl_str)
sh_set = set(sh_str)
thirdfile = open("3.txt", "w", encoding="utf-8")
print(str(len(hl_set)) + "," + str(len(sh_set)) + \
```

```
        "," + str(len(hl_set.difference(sh_set))) + "," + \
        str(len(sh_set.difference(hl_set))), file = thirdfile)
thirdfile.close()

#第4题
hl_only_set = hl_set.difference(sh_set)
sh_only_set = sh_set.difference(hl_set)
fourthfile = open("4.txt", "w", encoding="utf-8")
ch_num_hl_only = []
for i in hl_only_set:
    ch_num_hl_only.append((i, ch2num_hl[i]))
sorted_ch_num_hl_only = sorted(ch_num_hl_only, key=lambda x: x[1], reverse=True)
res = ""
for i in sorted_ch_num_hl_only:
    res += i[0] + ","
res = res[:-1] #去掉,
res += '\n' #换行

ch_num_sh_only = []
for i in sh_only_set:
    ch_num_sh_only.append((i, ch2num_sh[i]))
sorted_ch_num_sh_only = sorted(ch_num_sh_only, key=lambda x: x[1], reverse=True)
for i in sorted_ch_num_sh_only:
    res += i[0] + ","
fourthfile.write(res[:-1]) #去掉最后一个逗号.
fourthfile.close()

#第5题
union = sh_set.intersection(hl_set)
ch2fre_abs = {}
for i in union:
    ch2fre_abs[i] = abs(ch2num_sh[i] / sh_len - ch2num_hl[i] / hl_len)
sorted_ch_fre_abs = list(sorted(ch2fre_abs.items(), key=lambda x: x[1], reverse=Tru
fifthfile = open("5.txt", "w", encoding="utf-8")
res = ""
for item in sorted_ch_fre_abs[0:500]:
    res += item[0] + ","
fifthfile.write(res[:-1])
fifthfile.close()
```

总结

这次作业大部分同学都完成的不错。但仍有一部分同学忽略了细节，例如输出顺序反了、输出了多余的空格和逗号、输出了中文的逗号、频率多乘了100、输出了字的总数而不是种类数等等。