1. 从df\_qsw\_kin\_extracted中提取kinship的表达式，存为kinship.txt (just excel, without Python)
2. Kinship.txt 删去重复，得到kinship list: kinship\_final
3. 从quan\_song\_wen\_muzhi new version中提取墓志原文data，content\_id: content, qsw\_muzhi.xlsx, operated by Excel
4. Separate muzhi into list of sentences (separated by 。)
   1. Demo
   2. Sentence → subsentence

Extract kinship from subsentence (seperated by ;)

* 1. Extract officials
     1. 从df\_qsw\_kin\_extracted中提取official的表达式并预处理为officials\_renew4 list
     2. 添加officials列，从subsentence中提取包含的officials的信息
  2. Extract name
     1. 增加数据量，观察pattern
     2. Suppose that the string between kinship keyword and , is the first name
     3. Add last name by contendID\_subject.xlsx
  3. Add more kinship expressions, including children and female (2021.8.10)

# 五十三人/wsep/曾孫 含有wsep删掉

* 1. 手动补录kinship expression
  2. 手动补录officials

问题：

女子的姓氏分两种情况：妻子的姓氏从文中提取，女儿的姓氏与传主相同

已解决：

* Kinship list优先级排序，单字在后，多字在前
* Kinship还未收录女性、子女关键词，但可能与“祖”等单字重复
* 七代祖未收录在kinship list中 -> 已手动添加
* 烈祖未收录 -> 已手动添加