词性标注

班级：2016211303　姓名：黄若鹏 学号：2016212901

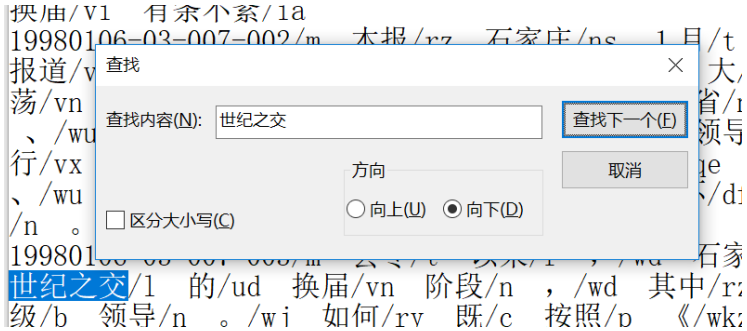
1.题目要求

Part-of-speech tagging: 30 points
-This data set contains one month of Chinese daily which are segment and POS tagged under Peking Univ. standard.
-Project idea:
 * Design a sequence learning method to predicate a POS tags for each word in sentence.
 * Use 80% data for model training and other 20% for testing (or 5-fold cross validation to test learner's performance. So it could be interesting to separate dataset.)

首先我想说这个数据集有太多坑了，导致训练时大部分都在处理数据，实在是耗费了大量时间
总是报各种由于数据集格式不规范导致的错误。数据量有这么大，实在是心累。
为了 debug 采用了 python 中的异常处理结构，这样一步一步 de 由于数据格式不规范的 bug。

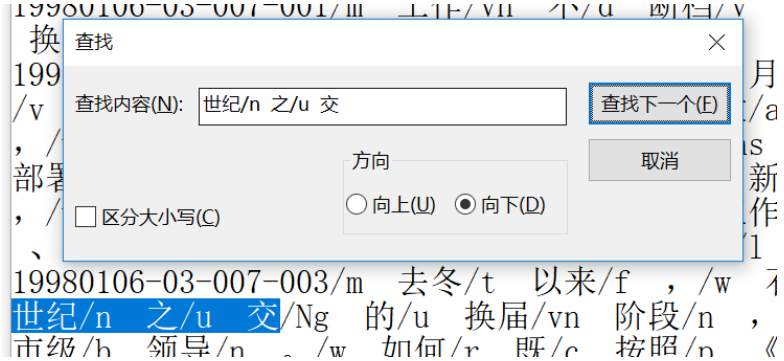
2.处理数据

2.1.这个语料好像还有不同版本，有些细节，很不同



上图中世纪之交是一个词

而请看下图

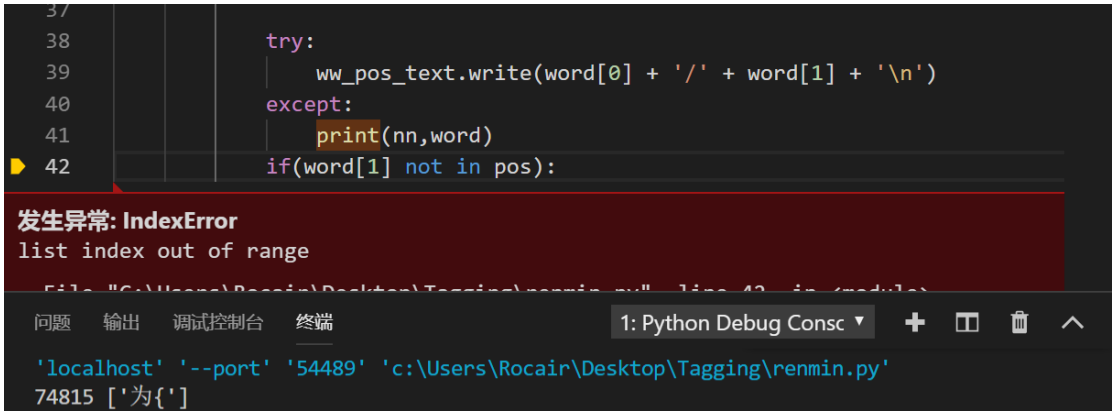


世纪之交却拆成了三个词。

这个问题如果统一语料，不算问题，但是在实验中，我交叉使用了不同的数据，使得，训练过程中，大量词无法正确匹配！因此只能自己编写处理数据的程序。

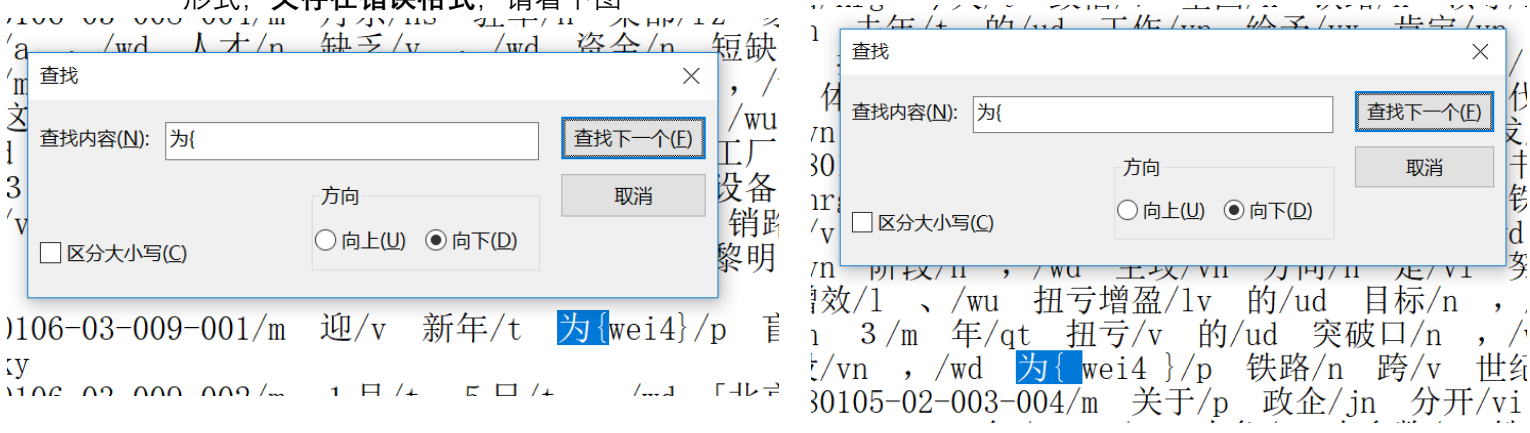
2.2.数据格式不规范

我本以为处理同一数据集就不会遇到啥奇葩问题了，看来我还是太单纯了。



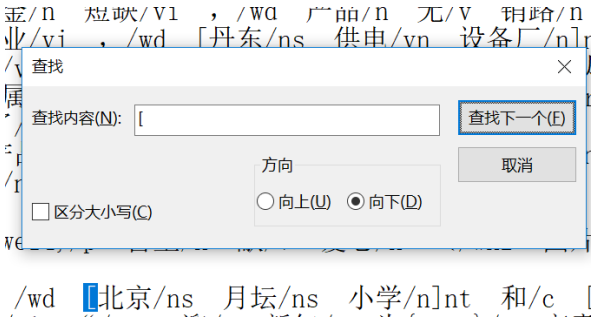
上图是我进行处理数据，然后按相应的格式输出到另一个文件中。方法是先对数据按空格 split 然后在按“/”符号 split 将词和词性分割开，但是却遇到上述问题，一开始，是花费了很长时间检查代码，是不是哪里错了，可是代码逻辑很简单，应该不会出错，面对这种 bug 只能采用异常处理，try except 方式了（C++中是 try catch），打印结果。

后来去检查数据集，奇葩问题出现，在同一个文档中，对于一个词的描述既存在正确形式，**又存在错误格式**，请看下图



虽然差异不大，左图是正确表达格式，而右图是错误表达格式，“{”后面不应该出现空格，这样会导致 split 函数操作后结果不正确。虽说对于人类来说很好分别，但是对于代码来判断，却要考虑大量逻辑！

2.3 复合词问题

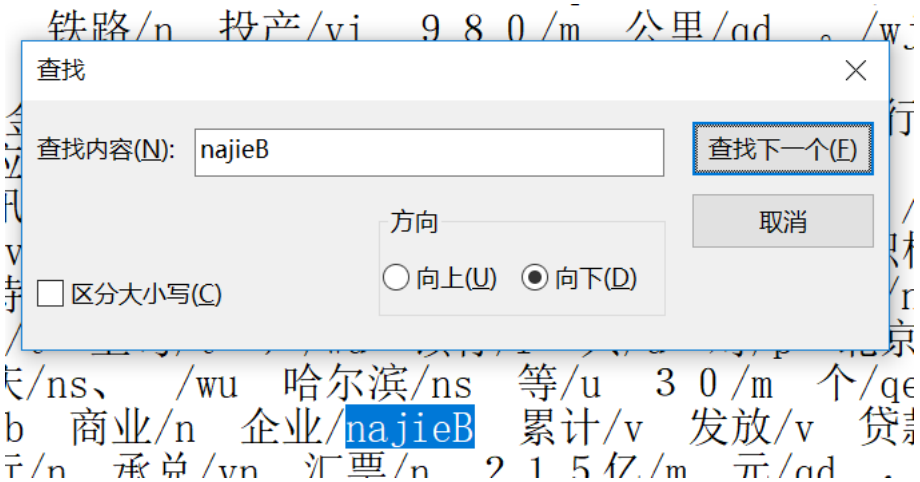


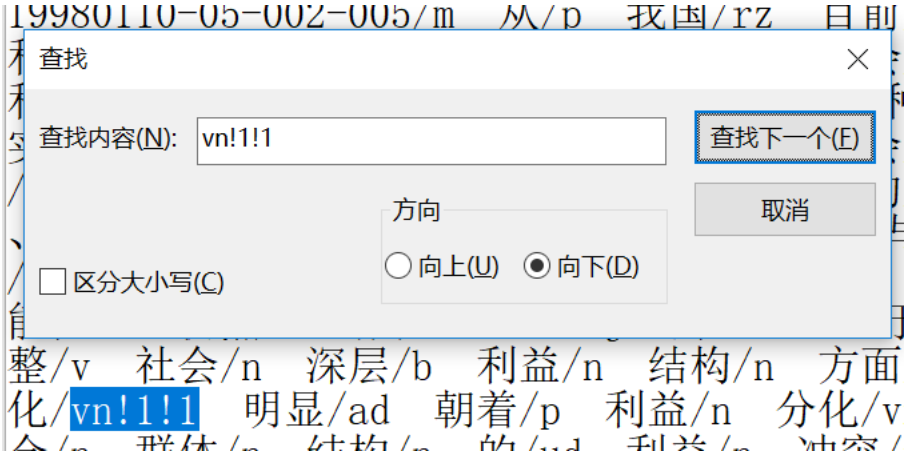
就是还存在这种复合词的情况，这里确实很麻烦。

这里最合理的处理方法应该是，根据复合词，建立一个新词，但是这里，为了简便，不生成新的词汇，只处理这个复合词成为不同的小词，并分别处理

```
        word = tmp[i].split('/')
        index = word[1].find("]", 0) #处理复合词
        if(index!=-1):
            word[1]=word[1][:index]
        word[0]=word[0].replace('[', '')#处理复合词
```

2.4.其他各种奇葩问题





不知道这些词性是如何创造出来的

总之，做其他题时，可能面对这些奇葩问题，无法感知，因此也就略过了，而做这道题如果就用提供的数据集，要解决一堆这样的问题，恰逢期末考试周，本作业难以做的完美，望见谅。

3.运行环境

系统：windows10

工具：VsCode

4.输入输出

输入 :只有一个 1998-01-105-带音.txt 文件，其余文件均有，DataProcessing.py 程序生成

输出：词性标注正确率

5.概要

词性标注（part-of-speech tagging）,又称为词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或者其他词性的过程。

我通过上网查询，发现主要有以下几种统计方法：
（1）基于最大熵的词性标注

（2）基于统计最大概率输出词性

（3）基于 HMM 的词性标注

我是通过 HMM 隐马尔科夫模型来进行词性标注。观测序列即为分词后的语句，隐藏序列即为经过标注后的词性标注序列。起始概率 发射概率和转移概率和分词中的含义大同小异，可以通过大规模语料统计得到。观测序列到隐藏序列的计算可以通过 viterbi 算法，利用统计得到的起始概率 发射概率和转移概率来得到。得到隐藏序列后，就完成了词性标注过程。

数据处理
写程序时，我首先对提供的数据进行了清洗和整理，把时间信息给去掉，得到总训练语料之后，我又建立了一个所有词的集合（所有出现的词，但不能重复），和一个所有词性的的集合，这里都是为了 HMM 模型建立那个大表做准备。

建立模型
使用 HMM 隐马尔可夫模型，进行词性标注。
训练集是这个数据的 80%，这里我也查了 5-fold cross validation 的概念，这里由于程序训练比较耗时，因此在程序中是将整个文件读进去，为了方便直接取前 80%为训练集，后 20%为测试集，也可以自定义训练集和测试集

输入输出
这次实验使用的就是老师提供的新闻报纸语料
输入训练集是标注好词性的文件的前 80%
测试集是没有标注词性的文件后 20%
输出为，标注好词性的测试集。
检测性能时 可以通过模型输出与元数据集做比较，统计标注正确性。

HMM 的模型表示

HMM 由**隐含状态 S**、**可观测状态 O**、**初始状态概率矩阵 pi**、**隐含状态概率转移矩阵 A**、**可观测值转移矩阵 B**（混淆矩阵）组成。

pi 和 A 决定了状态序列，B 决定了观测序列，因此，HMM 可以由三元符号表示：

$$\lambda = (A, B, \pi)$$

HMM 的两个性质：

1. 齐次假设：

$$P(i_t | i_{t-1}, o_{t-1}, i_{t-2}, o_{t-2} \cdots i_1, o_1) = P(i_t | i_{t-1})$$

2.观测独立性假设：

$$P(o_t | i_T, o_T, i_{T-1}, o_{T-1} \cdots i_1, o_1) = P(o_t | i_t)$$

viterbi 用于词性标注

词性标注问题映射到隐马模型可以表述为：模型中状态(词性)的数目为词性符号的个数 N；从每个状态可能输出的不同符号(单词)的数目为词汇的个数 M。假设在统计意义上每个词性的概率分布只与上一个词的词性有关(即词性的二元语法)，而每个单词的概率分布只与其词性相关。那么，我们就可以通过对已分词并做了词性标注的训练语料进行统计，统计出 HMM 的参数 λ，当然这就是上述学习问题。

然后可以根据已知的词语，通过 viterbi 算法，求出每个词语对应的词性，即完成词性标注。

6.部分数据

6.1.词性表

v
n
ud
a
wp
t
wkz
m
qe
wky
nt
wu
nrf
nrg
wd
wj
Vg
k
wm
p

vi
f
rr
dc
ui
vn
ns
c
rz
s
wt
vl
df
qt
d
ad
wyz
jn
wyy
lv
ul

qv
mq
j
r
vx
an
vu
u
b
l
jb
i
ia
in
vq
iv
wf
q
vd
ue

nr
Tg
uz
qd
y
nx
ry
qz
ld
lb
qc
ww
uo
jv
h
Ag
ib
qb
o
e
ws
Dg
qr
rzw

……

6.2.所有词汇的“集合”，所有词不能重复

迈向充满希望的新世纪——一九九八年新年讲话（附图片1张）

中共中央总书记、国家主席江泽民

一九九七年十二月三十一日１２月３１日，发表１９９８年《》。

新华社记者 光红 摄

同胞们 朋友们 女士们 先生们

中央人民广播电台 中国国际电视台

电台 电视台

向全国各族各界同胞，香港特别行政区、澳门特别行政区同胞，台湾同胞，海外侨胞，世界各国朋友，以及国际友人致以诚挚问候

１９９７年是发展历史上非常重要很不平凡的一年。

我们决心继承邓小平同志遗志，继续把建设有中国特色社会主义事业

。。。。。。。。

6.3.训练集

迈向/v 充满/v 希望/n 的/ud 新/a 世纪/n ——/wp 一九九八年/t 新年/t 讲话/n （/wkz 附/v 图片/n 1/m 张/qe ）/wky 中共中央/nt 总书记/n 、/wu 国家/n 主席/n 江/nrf 泽民/nrg （/wkz 一九九七年/t 十二月/t 三十一日/t ）/wky 12月/t 31日/t ，/wd 中共中央/nt 总书记/n 、/wu 国家/n 主席/n 江/nrf 泽民/nrg 发表/v 1998年/t 新年/t 讲话/n 《/wkz 迈向/v 充满/v 希望/n 的/ud 新/a 世纪/n 》/wky 。/wj （/wkz 新华社/nt 记者/n 兰/nrf 红光/nrg 摄/Vg ）/wky 同胞/n 们/k 、/wu 朋友/n 们/k 、/wu 女士/n 们/k 、/wu 先生/n 们/k ：/wm 在/p 1998年/t 来临/vi 之际/f ，/wd 我/rr 十分/dc 高兴/a 地/ui 通过/p 中央/n 人民/n 广播/vn 电台/n 、/wu 中国/ns 国际/n 广播/vn 电台/n 和/c 中央/n 电视台/n ，/wd 向/p 全国/n 各族/rz 人民/n ，/wd 向/p 香港/ns 特别/a 行政区/n 同胞/n 、/wu 澳门/ns 和/c 台湾/ns 同胞/n 、/wu 海外/s 侨胞/n ，/wd 向/p 世界/n 各国/rz 的/ud 朋友/n 们/k ，/wd 致以/v 诚挚/a 的/ud 问候/vn 和/c 良好/a 的/ud 祝愿/vn ！/wt 1997年/t ，/wd 是/vl 中国/ns 发展/vn 历史/n 上/f 非常/dc 重要/a 的/ud 很/dc 不/df 平凡/a 的/ud 一/m 年/qt 。/wj 中国/ns 人民/n 决心/d 继承/v 邓/nrf 小平/nrg 同志/n 的/ud 遗志/n ，/wd 继续/v 把/p 建设/v 有/v 中国/ns 特色/n 社会主义/n 事业/n 推向/v 前进/vi 。/wj 中国/ns 政府/n 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/wd 并/c 按照/p “/wyz 一国两制/jn ”/wyy 、/wu “/wyz 港人治港/lv ”/wyy 、/wu 高度/d 自治/vi 的/ud 方针/n 保持/v 香港/ns 的/ud 繁荣/vn 稳定/vn 。/wj 中国/ns 共产党/n 成功/a 地/ui 召开/v 了/ul 第十五/m 次/qv 全国/n 代表大会/n ，/wd 高举/v 邓小平理论/n 伟大/a 旗帜/n ，/wd 总结/v 百年/mq 历史/n ，/wd 展望/v 新/a 的/ud 世纪/n ，/wd 制定/v 了/ul 中国/ns 跨/v 世纪/n 发展/v 的

这里，是整理数据过后的结果，但是绝对这个整理过后的数据依然不完美，肯定存在一堆标记有问题的地方。

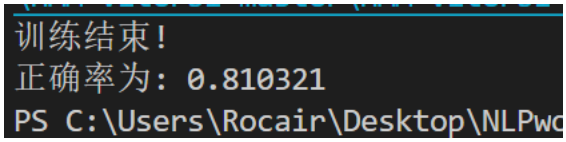
6.4.测试集

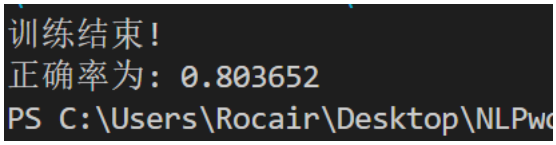
交通 银行 坚持 改革 ， 加强 管理 ， 努力 防范 金融 风险 ， 1997年 继续 以 良好 的 经营 业绩 令 同行 们 刮目相看 。 去年 该行 全年 完成 利润 47.84亿 元 ， 剔除 政策性 因素 ， 按 同 口径 计算 ， 利润 为 63.59亿 元 ， 比 上年 增长 18.57% ， 名列 银行 同业 盈利 前茅 。 1997年 交通 银行 认真 贯彻 国家 经济 金融 政策 ， 自觉 服从 金融 监管 ， 全面 推进 内部 管理 制度 ， 明显 加快 结构 调整 步伐 ， 各项 业务 发展 迅速 。 1997年 各项 人民币 存款 余额 2700亿 元 ， 比 1996年 增长 20% ； 各项 人民币 贷款 余额 1800 多 亿 元 ， 比 上年 增长 18% 。 全行 完成 80亿 元 凭证式 国债 承销 计划 ， 超额 完成 420亿 元 现金 回笼 计划 。 本报 讯 著名 家用电器 生产商 瑞典 伊莱克斯 公司 于 一月 十四日 在 海口 向 来自 全国 的 六百 多 名 代理商 宣布 ： 推出 九八 冰箱 新 系列 『 新静界 』 系列 。 其 最 突出 的 特点 是 低 噪音 设计 。 （ 马 艳 ） 本报 讯 深圳 先科 电子 有限公司 努力 提高 产品 质量 ， 该 公司 生产 的 VCD 视盘机 最近 通过 电子 工业部 优等品 检测 。 为{wei4} 促进 销售 ， 先科 公司 最近 推出 一 元 行动 ， 凡 购买 先科 产品 的 消费者 ， 只要 再 加 一 元 钱 ， 即 可 得到 麦克风 一 支 或 卡拉OK 碟 一 套 。 （ 唐 纯 ） 本报 讯 辽河 油田 在 稳定 石油 、 天然气 生产 的 同时 ， 以 国家 利益 为重 ， 克服 资金 紧张 等 困难 因素 ， 及时 申报 缴纳 税款 ， 去年 共 实现 增值税 、 消费税 十三点七亿 元 ， 比 上年 增加 一点六八亿 元 。 （ 夏 才源 李 兴华 ） 本报 讯 一九九七年 ， 青岛市 国税局 深入 推行 税收 征管 改革 ， 完善 构建 了 税收收入 全 过程 监控 体系 ， 确保 了 税收 足额 、 均衡 入库 。 全年 实现 工商 各税 七十六点九六亿 元 ， 比 去年 同期 增长 百分之二十三点四一 ， 增收 十五点六亿 元 。 （ 张 梦谦 ） 本报 讯 茂名 石化 三十万 吨 乙烯 装置 投 料 试车 第 一 年 成功 地 生产 三十万 吨 乙烯 ， 实现 了 一九九七 年度 产量 达标 ， 而且 投产 后 装置

测试集是不含任何标注，词语和符号文件。

7.结果分析

通过多次测试，在抽取的测试集上，平均准确度为 0.80622





8.部分代码

```
#计算概率矩阵
line = 0

while(True):
    text = fin.readline()
    if(text == '\n'):
        continue
    if(text == ""):#结束
        break
    tmp = text.split()#按空格分割
    len_temp = len(tmp)#语料长度
    line += 1
    for i in range(0, len_temp):
        word = tmp[i].split('/')#按/分割
        pre = tmp[i-1].split('/')#前一个词按/分割]
        #print(word[1])
        fre[word[1]] += 1#统计该词出现频率
        if(i == 1):
            pi[word[1]] += 1#先验概率
        elif(i > 0):
            A[pre[1]][word[1]] += 1#状态转移矩阵概率
        B[word[1]][word[0]] += 1 #观察概率矩阵
        B1[word[1]][word[0]] += 1
fin.close()
```

```
for i in pos:

    for j in pos:#避免概率 0
        A[i][j] += 1
        # print(A[i][j])
    for j in ww:
        B[i][j] += 1
        # print(B[i][j])

for i in pos:
    pi[i] = pi[i]*1.0/line
    for j in pos:
        A[i][j] = A[i][j]*1.0/(fre[i])
    for j in ww:
        B[i][j] = B[i][j]*1.0/(fre[i])
        # print(B[i][j])
print("训练结束！")
```

```
#维特比算法
    max_end = ""
    for i in range(0, num-1):
        max = 0

        for each in res[text[i]]:
            for each1 in res[text[i+1]]:
                if(A[each][each1] > max):
                    max = A[each][each1]
                    max_res[i] = each
                    if(i == num-2):
                        max_end = each1

    max_res[num-1] = max_end
```

```
#结果分析
tag_result = open('tag_result.txt','r')#标注结果
reference = open('ww_pos.txt', 'r')#训练集

a=tag_result.read().split()[start:end]
b=reference.read().split()[start:end]
correctnum=0,num=0
for t1,t2 in zip(a,b):
    num+=1
    if t1.split('/')[1]==t2.split('/')[1]:
```

```
        correctnum+=1

print("正确率为:",correctnum/num)
```