



自然语言处理

作业一：中文分词

数据科学与计算机学院 17大数据与人工智能

17341015 陈鸿峥

在本实验中我采用了三个分词工具对新华网¹的语料进行中文分词。其中三个工具说明如下：

- 北大分词(PKUSEG) [1]：北大基于之前发表的两篇ACL顶会论文整理出来的工具包，今年初才刚刚发布出来，宣称具有最高的准确率
- 结巴分词(jieba) [2]：号称要做最好的中文分词工具，也是最多人使用的开源中文分词工具包
- 清华分词(THULAC) [3]：清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包

我编写了一段Python脚本，从文本中读入语料，并进行分词并输出，程序如下。

```
import time
import pkuseg
import jieba
import thulac

with open("input.txt","r") as infile:
    intext = infile.read()
    # pkuseg
    print("Begin pkuseg...")
    seg = pkuseg.pkuseg()
    start = time.time()
    text = seg.cut(intext)
    end = time.time()
    res = "|".join(text)
    outfile = open("output_pkuseg.txt","w")
    outfile.write(res)
    print("Time (pkuseg): {:.4f}s".format(end-start),end="\n\n")

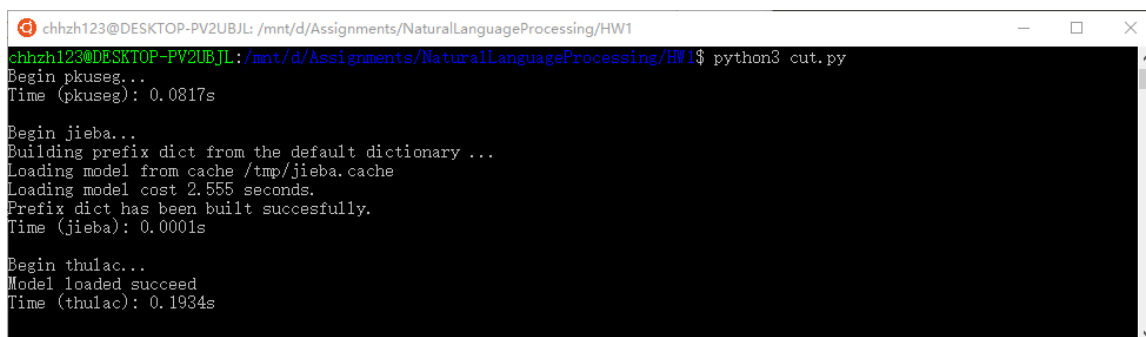
    # jieba
    print("Begin jieba...")
    start = time.time()
    text = jieba.cut(intext)
    end = time.time()
```

¹http://www.xinhuanet.com/politics/leaders/2019-09/21/c_1125023359.htm

```
res = "|".join(text)
outfile = open("output_jieba.txt", "w")
outfile.write(res)
print("Time (jieba): {:.4f}s".format(end-start), end="\n\n")

# thulac
print("Begin thulac...")
thu = thulac.thulac(seg_only=True)
start = time.time()
text = thu.cut(intext)
end = time.time()
text = [t[0] for t in text]
res = "|".join(text)
outfile = open("output_thulac.txt", "w")
outfile.write(res)
print("Time (thulac): {:.4f}s".format(end-start), end="\n\n")
```

运行过程与时间如下图所示。



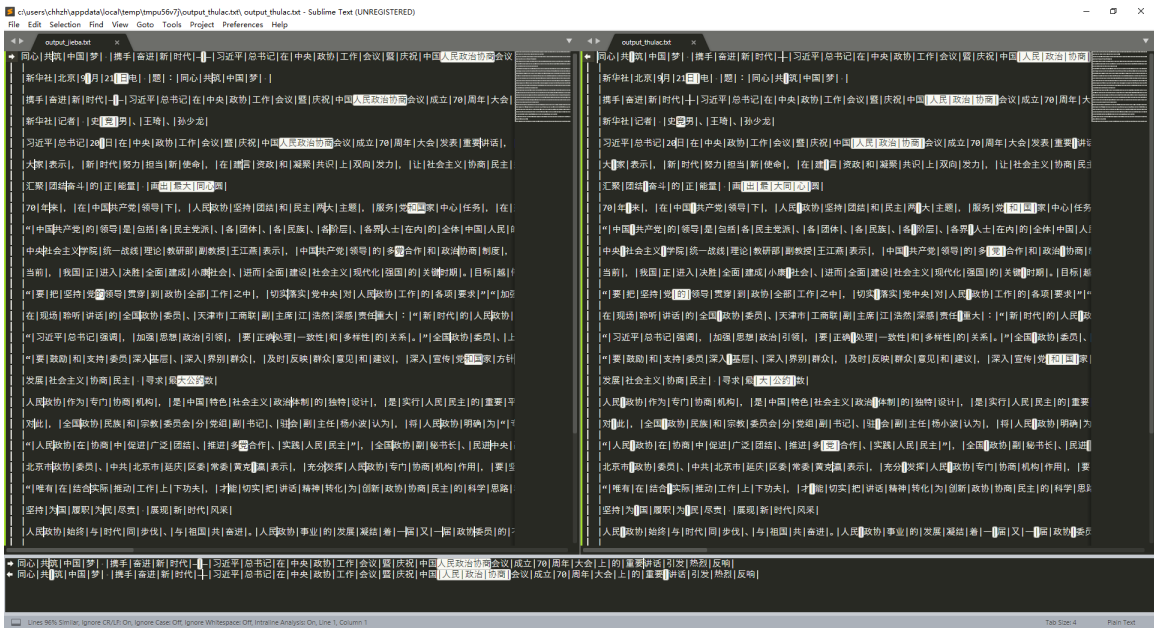
```
chhzh123@DESKTOP-PV2UBJL: /mnt/d/Assignments/NaturalLanguageProcessing/HW1$ python3 cut.py
Begin pkuseg...
Time (pkuseg): 0.0817s

Begin jieba...
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 2.555 seconds.
Prefix dict has been built successfully.
Time (jieba): 0.0001s

Begin thulac...
Model loaded succeed
Time (thulac): 0.1934s
```

可以看到结巴分词的速度最快，北大的工具包次之，清华的最慢。

我采用了Sublime Text的工具包Sublimerge来对比不同工具输出的分词，如下图所示。



通过分析，可以发现很多问题

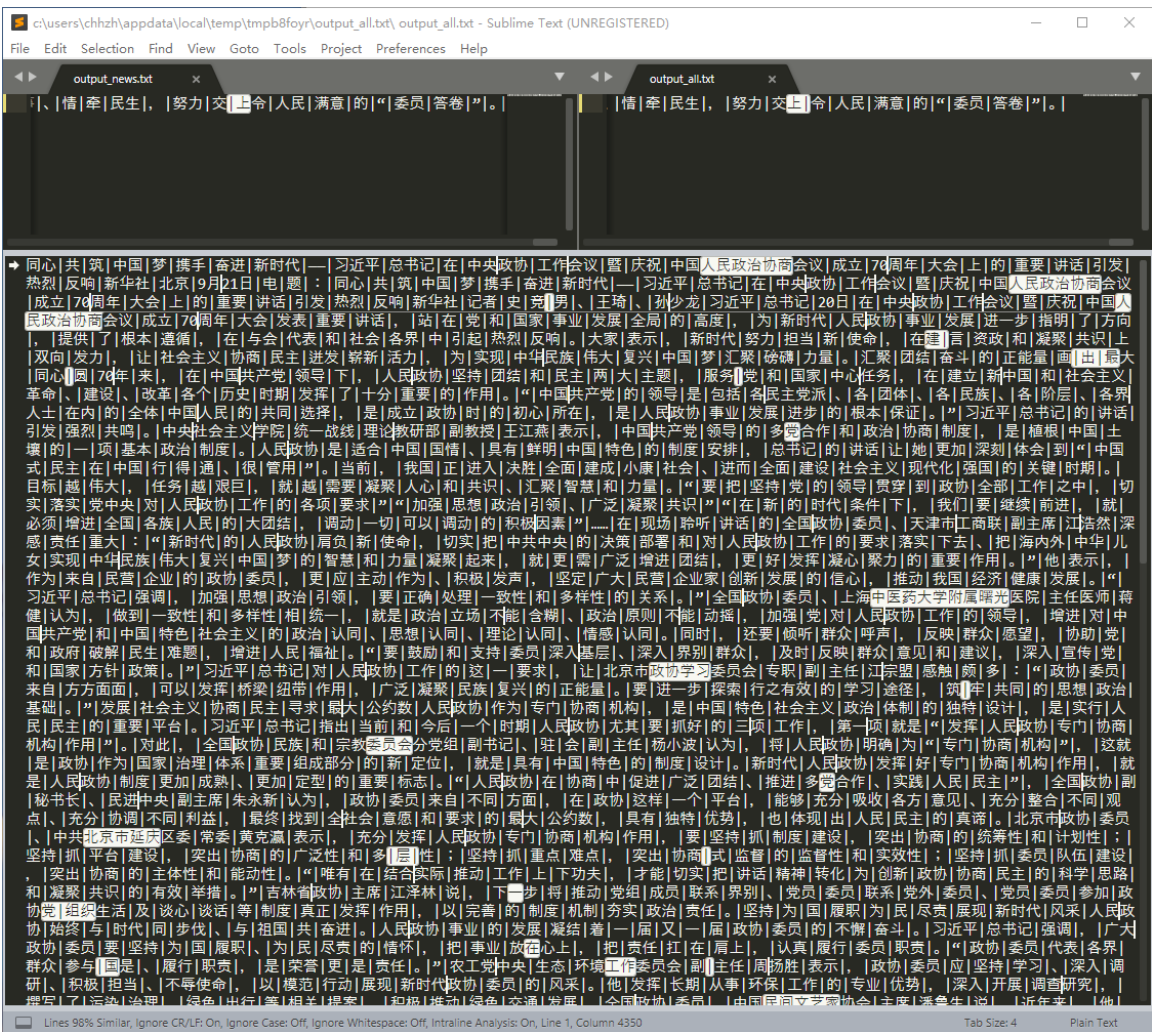
- 分词的粒度：清华的工具包可能没有专门针对新闻语料进行训练，因此会出现比较明显的问题，如下面第一条，清华工具包的断句明显是错的，它将大同连了起来，而没有辨别出同心圆。而且在可断可不断的点，清华工具包更倾向于断，如下表的后三条语料。从下面几个例子来看，结巴分词对于词语长度的掌握在几种工具中比较好。

原始语料	PKUSEG	Jieba	THULAC
画出最大同心圆	画出 最 大 同心圆	画出 最大 同心圆	画 出 最 大 同 心 圆
来自不同方面	来自 不同 方面	来自 不同 方面	来 自 不 同 方面
中国民间文艺家协会	中国 民间 文艺家 协会	中国民间文艺家协会	中国 民间 文艺 家 协会
最大公约数	最 大 公约数	最大公约数	最 大 公 约 数

- 对于人名等专有名词的辨别：这也是我这次实验中比较关注的点，会发现这些分词工具对于人名的辨别还是比较差的，经常需要附加的标点符号或标志词（如“说”）来进行辨别。而在这一点上，结巴分词就做得比较糟糕了，经常辨别不出人名，如下面的例子，常常将人名中的名字部分分开。而且三个分词软件都不能很好辨别江浩然这一人名，因为在大部分语料中，浩然是自成一词，而这些模型都没有足够的具体情况具体分析的能力，故都会选择断开。

原始语料	PKUSEG	Jieba	THULAC
史竞男	史 竞男	史 竞 男	史竞男
黄克瀛	黄克瀛	黄克 瀛	黄克瀛
江浩然	江 浩然	江 浩然	江 浩然

- 不同语料训练出来的模型：这里主要看PKUSEG提供的不同预训练模型。



可以看到，用新闻语料预训练出来的模型分词粒度会比较大，能够很好地辨别专有名词，而混杂语料训练出来的模型则倾向于细粒度的划分。具体的例子包括中国人民政治协商会议、中国民间文艺家协会、上海中医药大学附属曙光医院等，这些新闻语料预训练模型都能将其当成一个整体看待。

所有的程序及输出文件都附在附件中，可以进行查看。

参考文献

- [1] Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, Xu Sun. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. Arxiv. 2019. <https://github.com/lancopku/PKUSEg-python>
- [2] 结巴分词, <https://github.com/fxsjy/jieba>
- [3] 孙茂松, 陈新雄, 张开旭, 郭志芑, 刘知远. THULAC: 一个高效的中文词法分析工具包. 2016. <https://github.com/thunlp/THULAC-Python>