**基于深度学习的自然语言处理**

**第3次实验报告**

**一、实验名称**

使用并修改开源工具完成依存分析。

**二、实验日期**

2020.10.24 – 2020.11.06

**三、实验目的**

（1）熟悉依存分析

（2）熟悉开源工具的使用

**四、实验数据**

提供实验的新闻数据

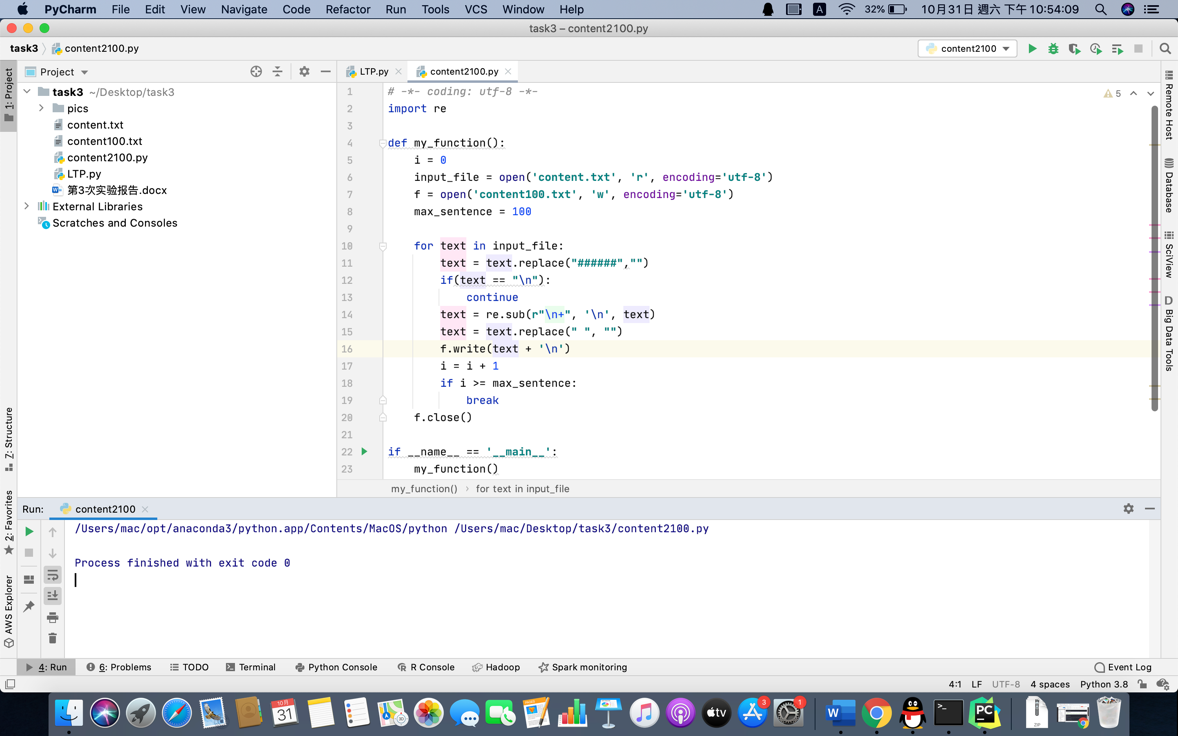
**五、实验步骤及结果分析**

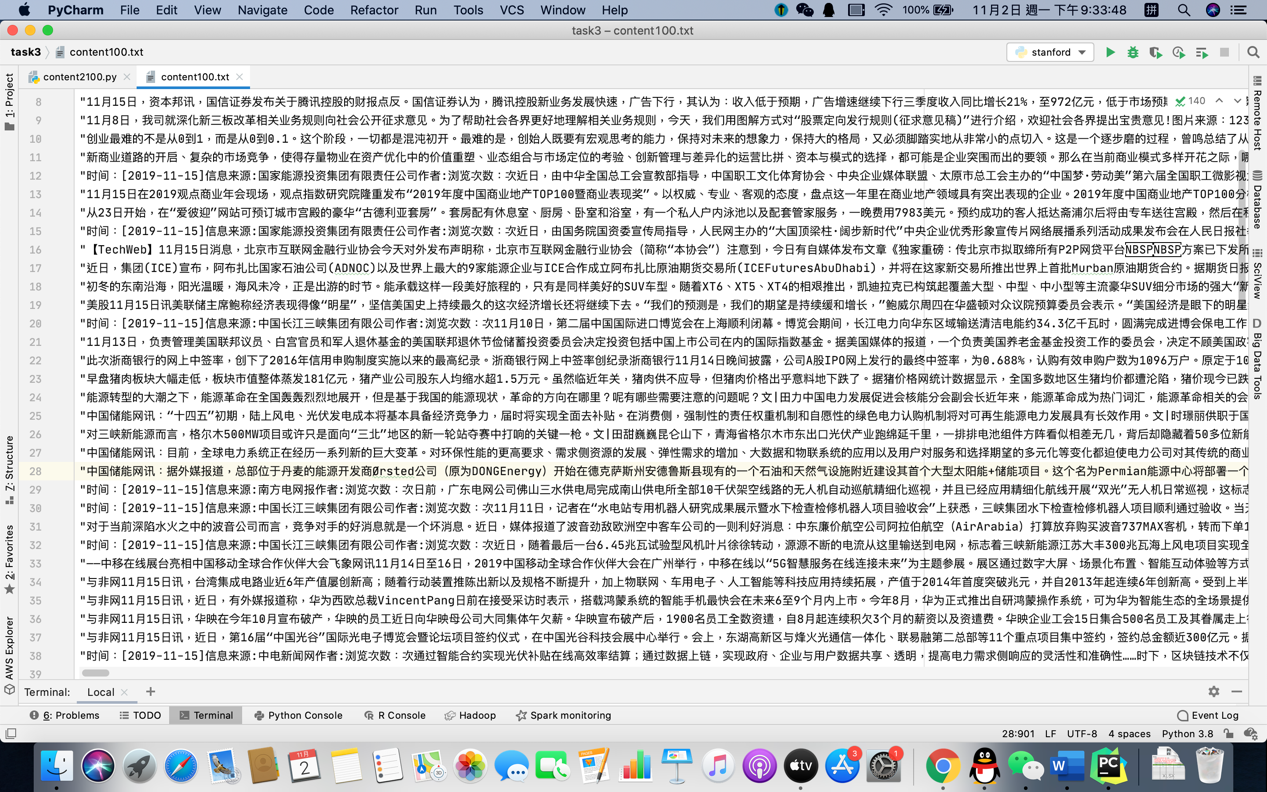
5.1 进入百度网盘下载语料content.txt

链接:https://pan.baidu.com/s/1QyVnxUTrJ2Dqj7FIpzuhgg 密码:lekc

5.2 对语料进行预处理：去掉一些无用的标点符号

这部分代码在content2100.py里，目的是从所有新闻中选出100条新闻用于进行依存分析。从content.txt中读取每一条新闻内容并去除作为分段符的######，多余的回车以及空格。若新闻较短或较长则不选用。然后将处理好的新闻保存到content100.txt中以备调用。程序及选择出的新闻如图所示。





5.3 使用LTP进行句法分析

5.3.1 参考链接 https://blog.csdn.net/asialee\_bird/article/details/102610588

5.3.2 LTP基础

LTP是由哈工大研发的语言技术平台，它提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等工作。它可以针对单一自然语言处理任务，生成统计机器学习的工具，或调用模型分析的编程接口，或使用基于云端的编程接口。还提供了系统可调用的，用于中文语言处理的模型文件。在这次实验中，主要用到的是LTP中提供的分句，分词，词性标注，依存分析等功能。

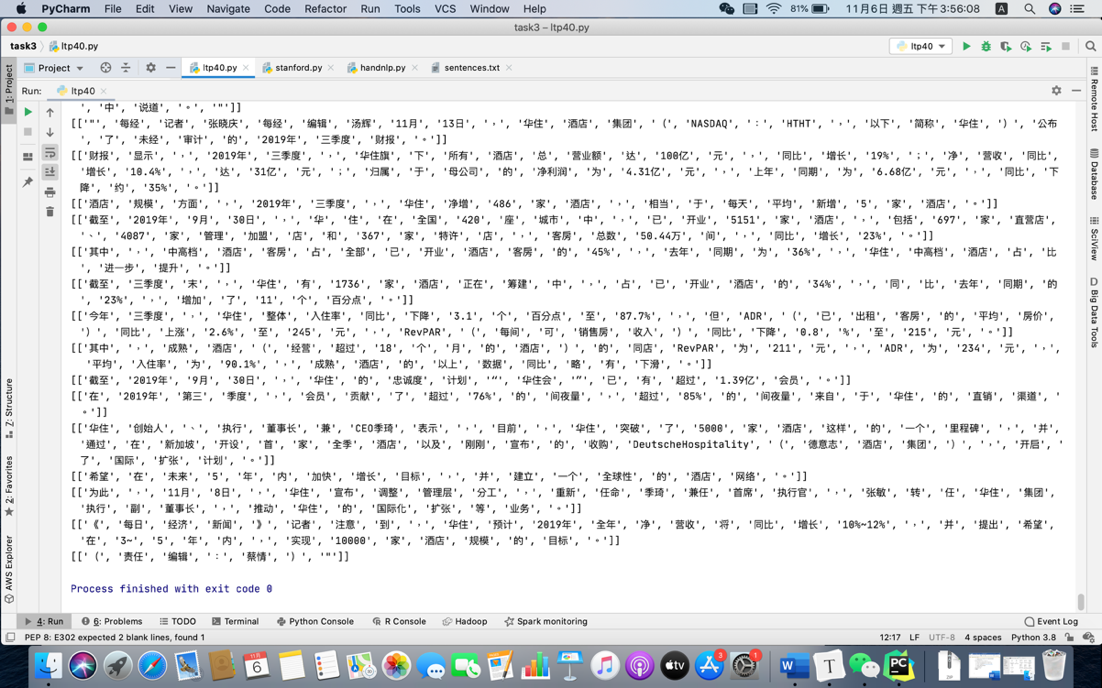
5.3.3 pyltp安装

目前，pyltp已经升级至新版本4.0.0，且原有的pyltp安装时出现错误，所以我选择使用pip安装最新版本的ltp。安装时如图所示。



5.3.4 pyltp实现句法分析

首先加载数据，将数据读入列表，然后先对原文进行分句，得出新闻中的每一句话。然后利用ltp中提供的分词，词性标注，依存分析等功能对每一句话进行分析。最后保存结果。运行结果如图所示。



5.3.5 结果分析

首先，ltp中自带的分句功能并不适用于商业新闻。商业新闻中含有大量数据，而ltp的分句功能会将数据中的小数点也当成句号，所以需要在分句完成后进一步处理。其分词功能表现较好，可以正确分出大多数中文单词，包括较难分出的人名和公司名。

在词性标注部分，相应的词性含义如下图所示。该表摘自LTP官方文档的附录。

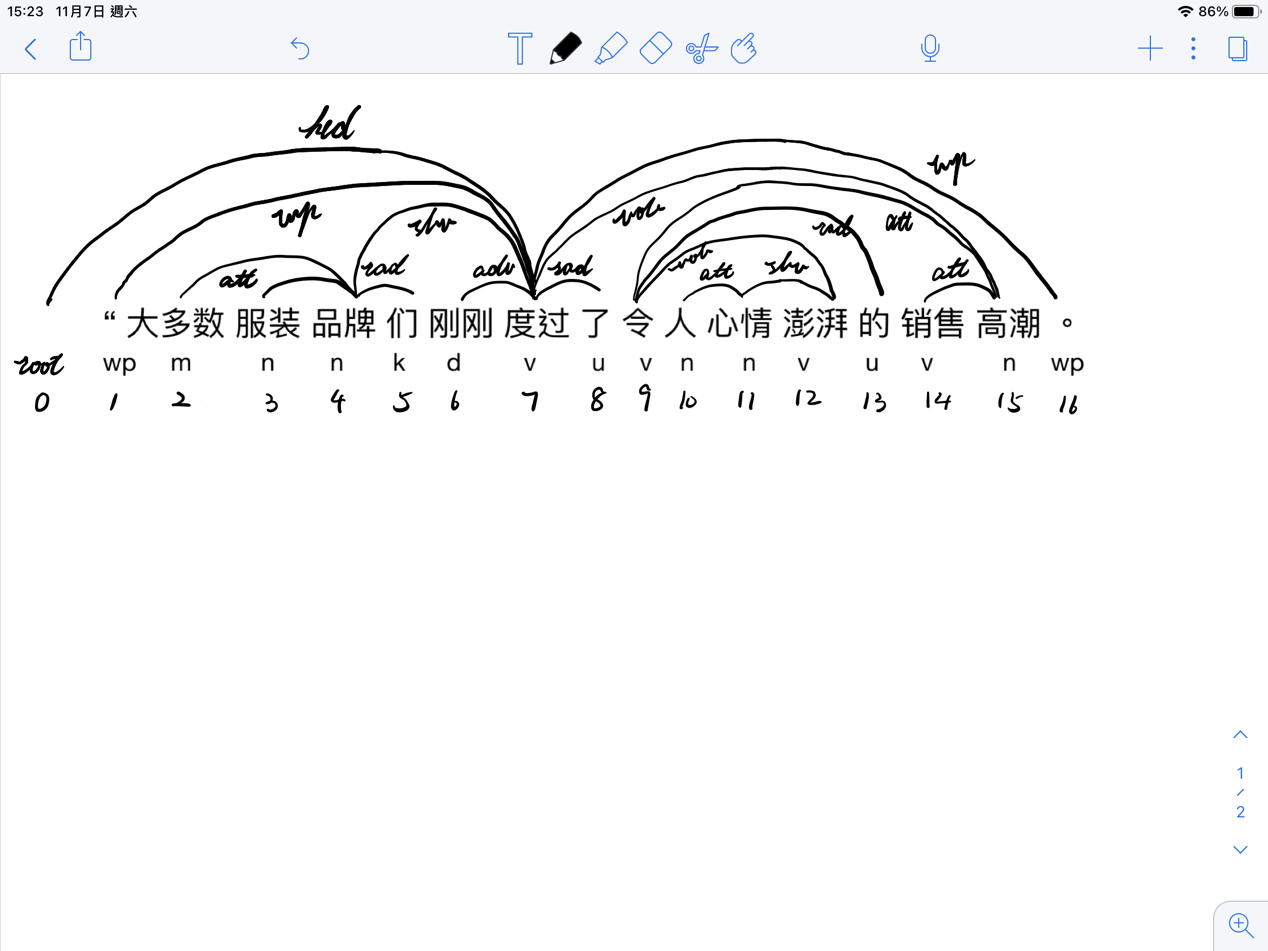


在依存句法分析部分，相应的含义如下图所示。该表摘自LTP官方文档的附录。

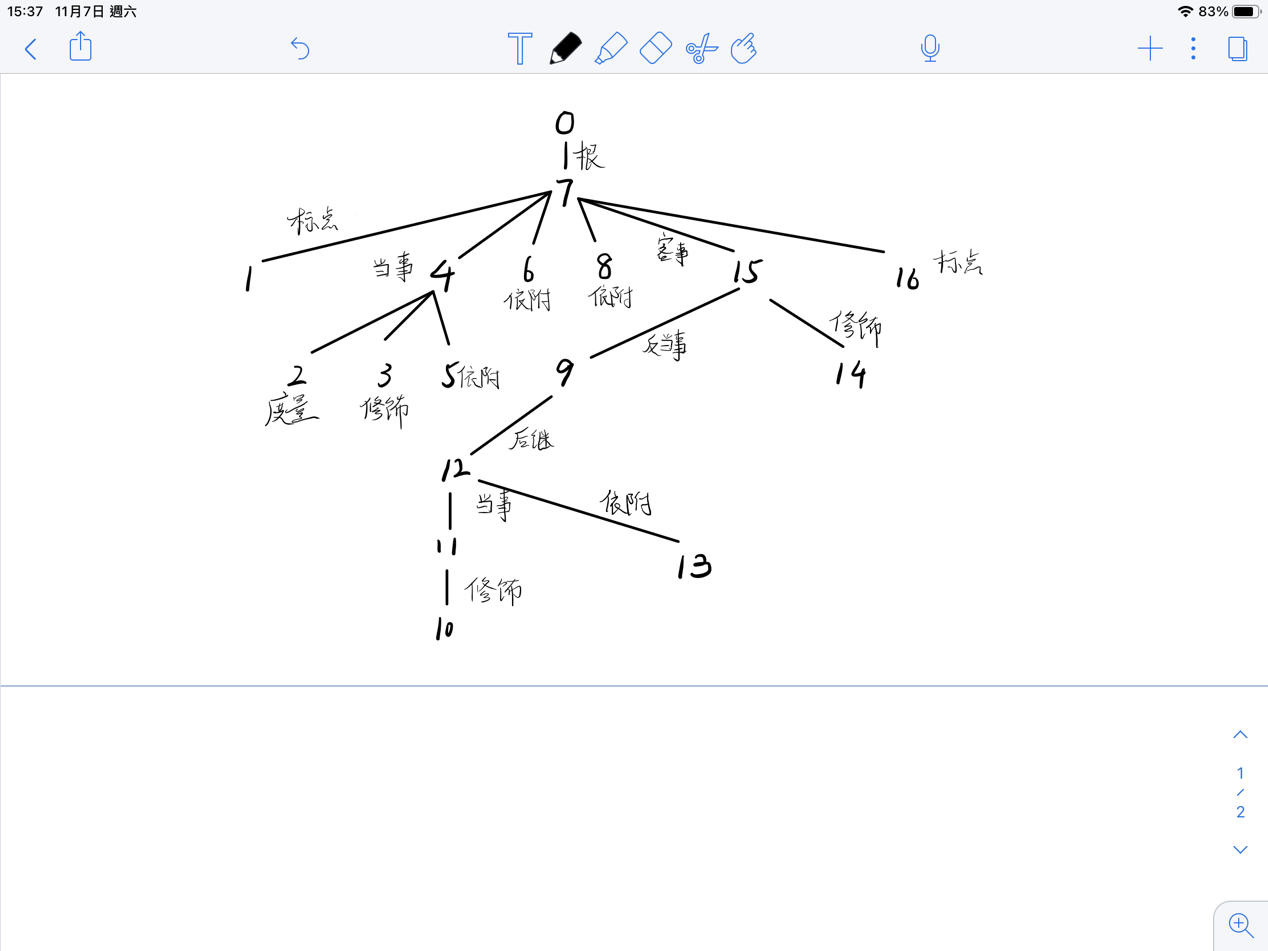


语义依存分析的图表过长，不再截图，请参见<http://ltp.ai/docs/appendix.html>。

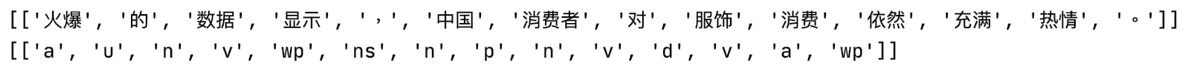
对于新闻的第一句话：大多数服装品牌们刚刚度过了令人心情澎湃的销售高潮。其词性标注和依存句法分析结果如下图所示：



其语义依存分析结果如下图所示：



由这两张图片并结合其他的分析结果可以看出， LTP的词性标注功能大部分情况下比较准确。它可以较好地识别出大部分名词、动词，而对于连词的识别并不好，经常将其判断为副词。有时面对具有多个词性也会出错，如在下面这句话中，热情显然被当作是名词比较合适，而这里将它当成了形容词。



在进行依存句法分析时，LTP可以很好的分出主谓，动宾，定中等关系。在树型结构构造过程中，ltp给出的结果较为准确，它可以正确分辨出令人心情澎湃的这一长短语。而在语义依存分析中，使用树和图的结构相差不大，一般都会给出树状的结构。该功能对于依附关系（标点、关系、依附）和周边角色（当事、客事、度量、修饰）分析较好，对语义结构关系的分析（反关系、先行、后继）相对来说差一些。

5.4 使用Stanford NLP进行句法分析

5.4.1 参考链接 https://blog.csdn.net/zkq\_1986/article/details/81583725

5.4.2 NLTK和StandfordNLP简介

NLTK是由宾夕法尼亚大学计算机和信息科学开发的，使用python语言实现的一种自然语言工具包，其收集了大量公开数据集、模型上提供了全面、易用的接口，涵盖了分词、词性标注(Part-Of-Speech tag, POS-tag)、命名实体识别(Named Entity Recognition, NER)、句法分析(Syntactic Parse)等各项 NLP 领域的功能。

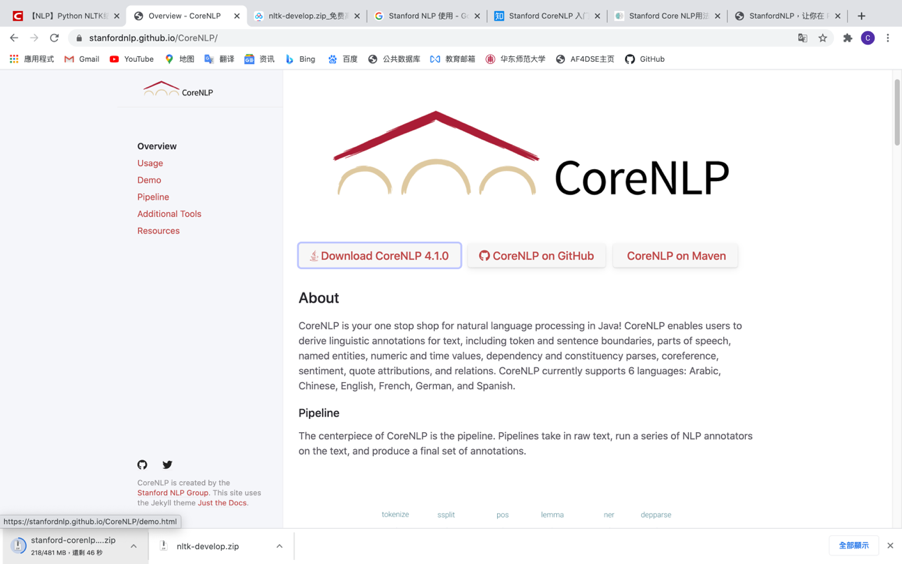
Stanford NLP是由斯坦福大学的 NLP 小组开源的使用 Java 实现的 NLP 工具包，同样对 NLP 领域的各个问题提供了解决办法。它集成了很多非常实用的功能，包括分词，词性标注，句法分析等等。这不是一个深度学习框架，而是一个已经训练好的模型，实际上可以类比为一个软件。在 2004 年 Steve Bird 在 NLTK 中加上了对 Stanford NLP 工具包的支持，通过调用外部的 jar 文件来使用 Stanford NLP 工具包的功能。

5.4.3 安装配置过程中注意事项

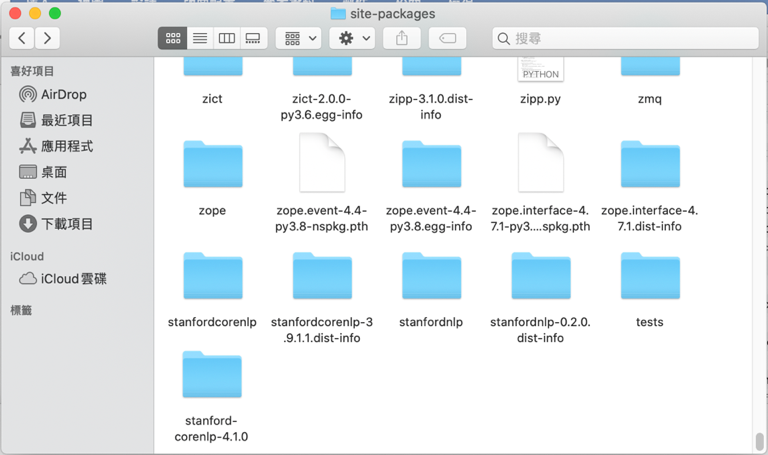
StanfordNLP的配置略复杂一些。首先我使用了pip install命令安装stanfordnlp和stanfordcorenlp库，安装时截图如图所示。



之后，我们需要从官网上下载CoreNLP工具，官网内容如图所示，点击下载按钮。

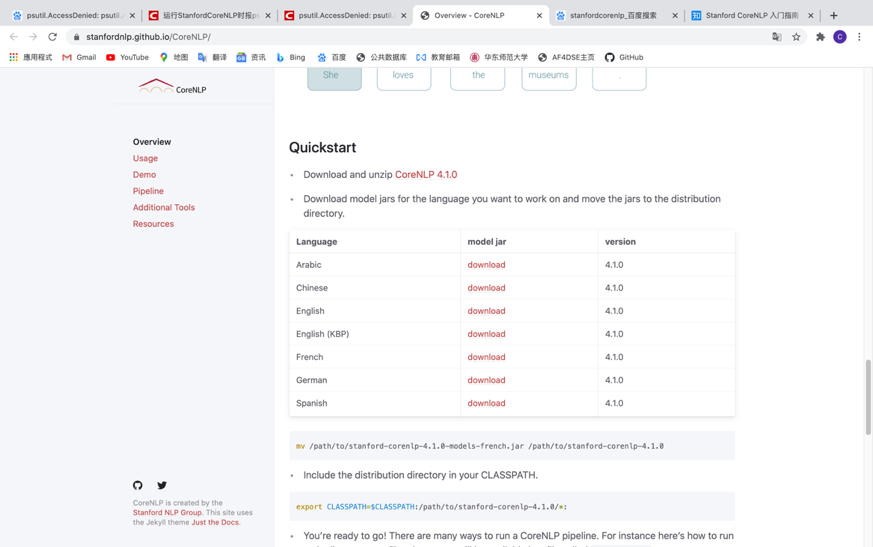


将下载的压缩包解压，放到python存放库的文件夹中，在我的电脑上即是/user/mac/opt/anaconda3/lib/python3.8/site-packages，完成之后文件夹如下图所示。

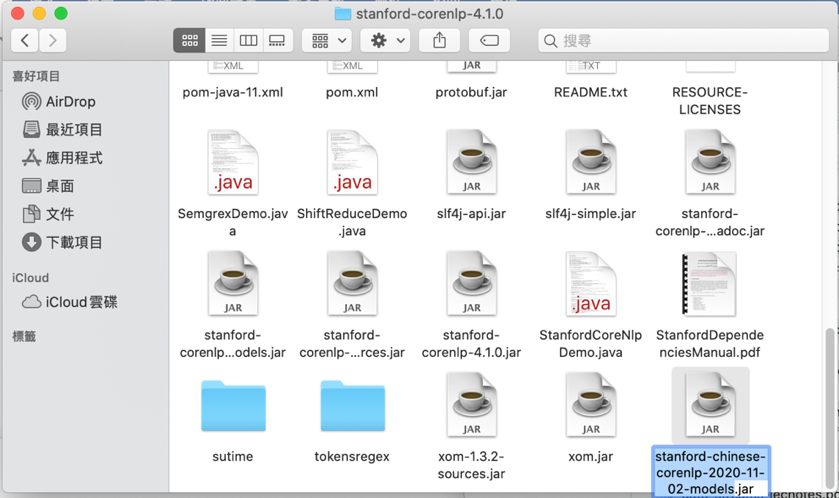


5.4.4 StandfordNLP必要工具包下载

在安装好库之后，我们还需要下载必要工具包，并且将其存储到正确的位置。我们先登陆官网，可以看到页面上有多个工具包，分别对应英文、中文、法文、德文、阿拉伯文等。这里需要下载中文对应的工具包。



在下载完工具包之后，需要将其放入stanfordcorenlp对应的文件夹下，其路径为/user/mac/opt/anaconda3/lib/python3.8/site-packages/stanford-corenlp-4.1.0，即上一步中解压缩并移动后的文件夹，如下图所示。然后将其重命名成正确的格式。这样在被调用时才能正常运行。至此，StanfordNLP就配置完成了。

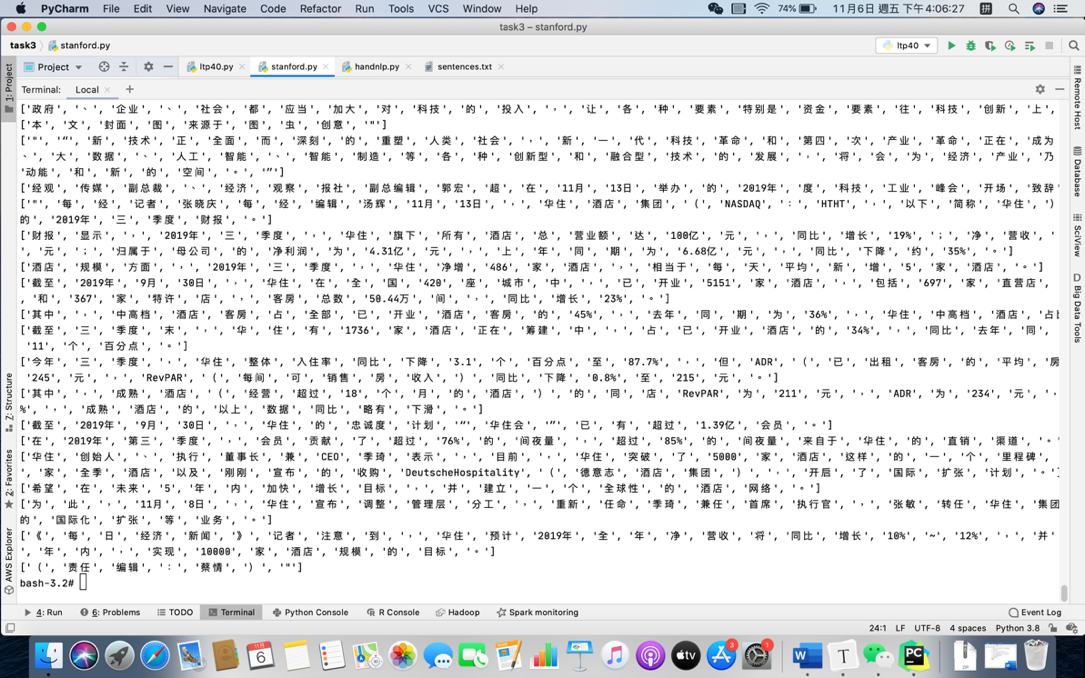


5.4.5 StandfordNLP相关核心操作

StanfordNLP主要的功能包括分句，分词，词性标注，命名实体识别，句法成分分析，依存句法分析等功能。某些功能只对特定的语言有效，如分句无法对中文使用。此外，在运行调用了stanfordcorenlp库的程序时，需要在终端中输入sudo -s指令切换到管理员账户，并输入python stanford.py以运行程序。

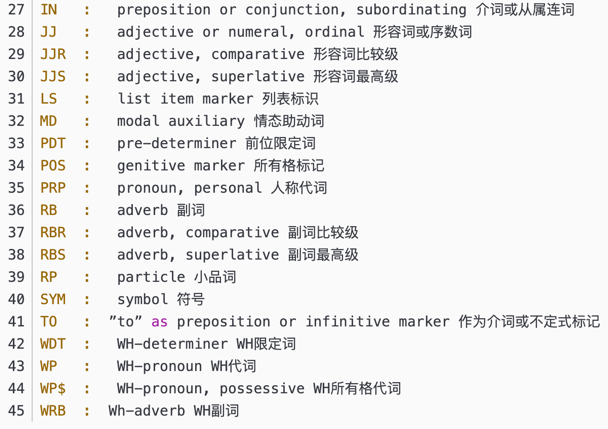
5.4.6 StanfordDependencyParser中文依存句法分析

首先我们加载好用上一段程序中的ltp分好的句子。然后对其中的每个句子进行分词，词性标注，命名实体识别，句法成分分析，依存句法分析等功能。然后输出并保存，程序的运行结果如下图所示。

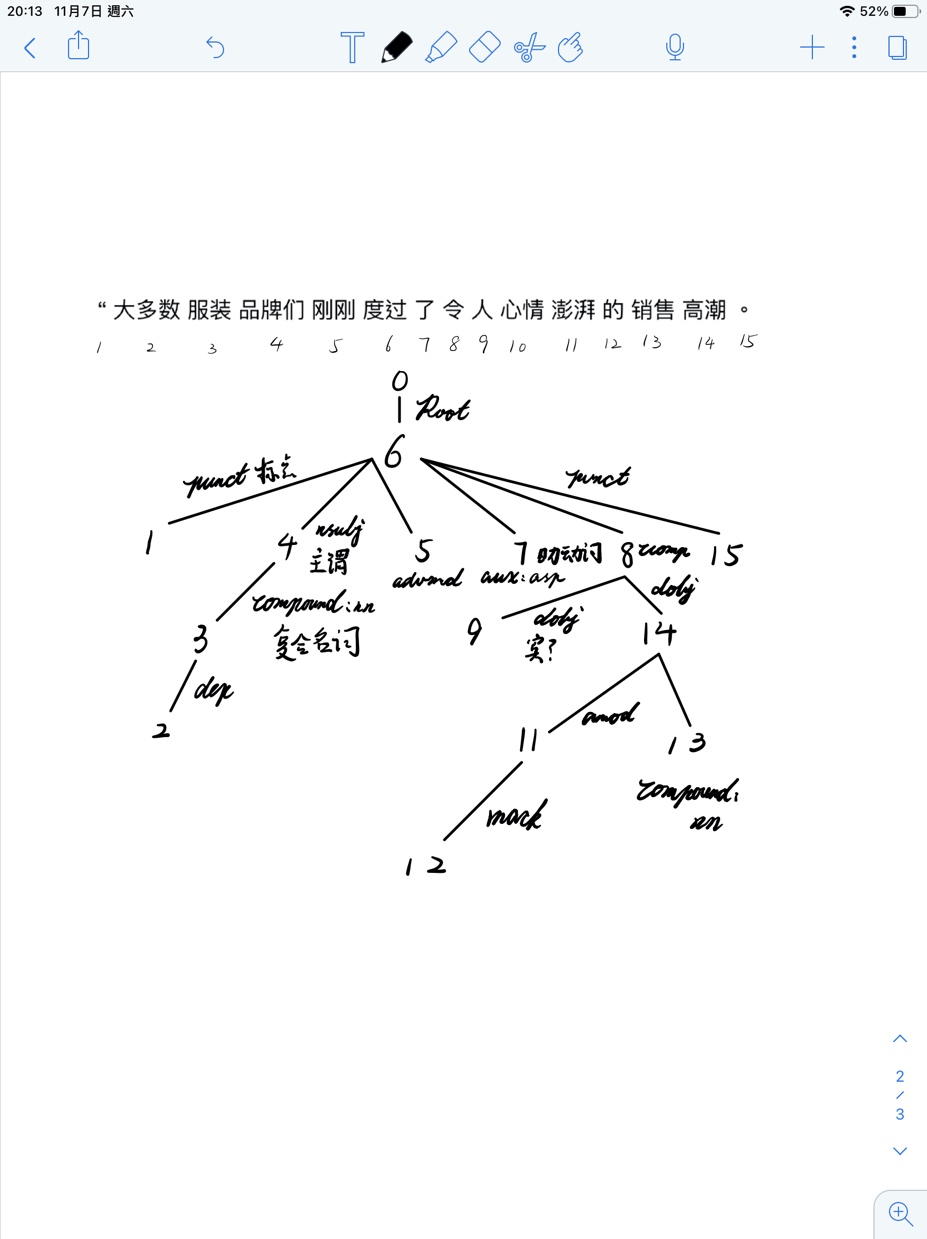
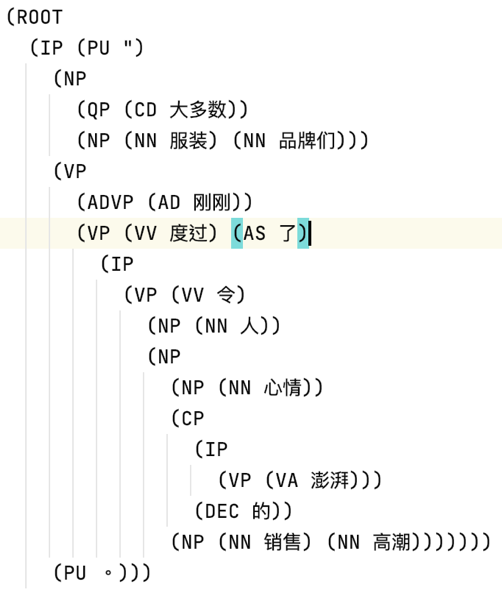


5.4.7 结果分析

参考网站<https://blog.csdn.net/u010454729/article/details/88358932>可得句法成分分析的符号解释如后图，参照网站<https://nlp.stanford.edu/nlp/javadoc/javanlp-3.5.0/edu/stanford/nlp/trees/international/pennchinese/ChineseGrammaticalRelations.html>可得依存句法分析的符号解释。



同样以上一段中的句子为例。句法成分分析和依存句法分析结果如图所示。



可以看出stanfordnlp能给出大致正确的结果，但是其中也有些小瑕疵。句法成分分析过于依赖从句（IP）的方式界定句子成分，像澎湃一词应该不会在正确语法中当成一整个句子。而依存句法分析中将人和高潮均当成了宾语，这是不太合理的。总的来说，该库对于中文的依存分析略逊一筹。

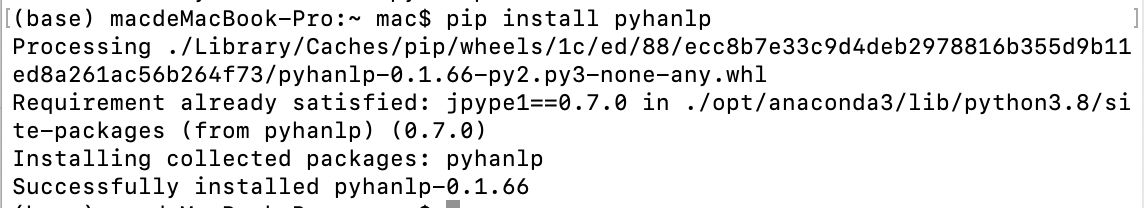
5.5 使用pyhanlp进行句法分析

5.5.1 参考链接 <https://github.com/yangby888/handnlp>

由清华大学开发的pyhanlp是面向生产环境的多语种自然语言处理工具包，基于 TensorFlow 2.x，目标是普及落地最前沿的NLP技术。HanLP具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。

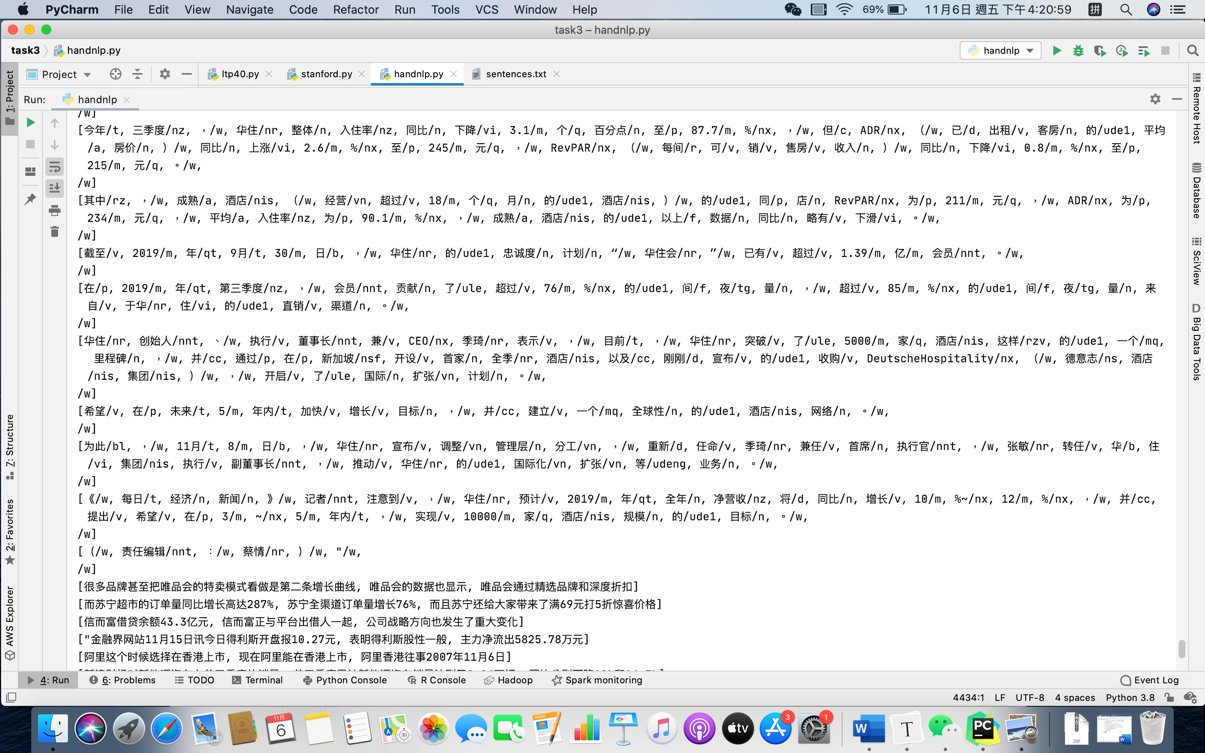
5.5.2 安装

在命令行中输入pip install pyhanlp命令安装pyhanlp库，安装过程如下图所示。



5.5.3 依存句法分析

首先读入新闻中的句子，然后对每个句子输出其分词结果和依存句法分析的结果。所用的handnlp还可以对一整段的文字进行关键词提取和摘要。程序运行结果如下图所示。

****

5.5.4 结果分析

同样使用第一个句子为例，其依存分析结果如下图。



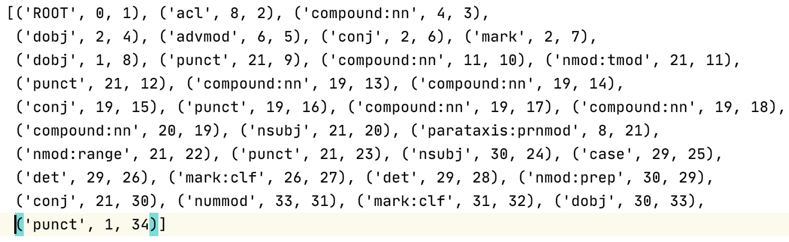
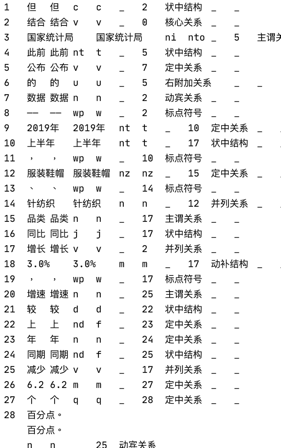
可以看到该依存分析结果比较合理，不但分出了较为简单的定中，主谓，动宾等关系，还分出了比较难以判断的兼语、补语等关系。其树状结构与stanfordnlp给出的结果大致一致，并不能很好区分复杂结构的长短语。还有一个小瑕疵是句末的标点并没有被正确分出。

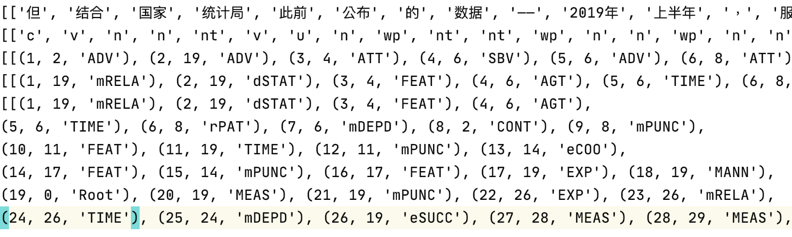
此外，我还使用了handnlp的关键字和摘要功能，均表现不佳。关键字中经常会出现百分号或分词错误时的结果。而摘要大部分情况下不符合语法，让人匪夷所思。

5.6 三种方法的分析与比较

综合三种进行依存分析的方法，我认为在中文这方面，ltp最为准确，handnlp次之，stanfordnlp表现较差。三种方法对于主谓、动宾、定中、状中等结构都可以较好地识别，而对于比较复杂的结构，ltp可以给出比较准确的树型结构。三种方法的依存分析树在大部分情况下均不会发生交错。在ltp中使用依存分析图中也很少不同于依存分析树。核心成分左右两边的成分一般不会跨越核心成分，满足依存分析的规律。

对于核心词的选择大致上是基本一致的，但也有例外。如：但结合国家统计局此前公布的数据——2019年上半年，服装鞋帽、针纺织品类同比增长3.0%，增速较上年同期减少6.2个百分点。在这句话中，handnlp和stanfordnlp给出的核心词是第一个分句中的结合一词，而ltp给出了更为合理的核心词：增长。如图所示，上方为stanfordnlp，下方为ltp，右侧为handnlp。





对于stanfordnlp来说，中文的分析可能比较难，其无法很好识别中文中的常用短语，如换句话说等。并且其在句法分析中使用了太多的从句，中文中其实更多的是一个个的分句。我猜测这可能是因为斯坦福大学对于中文不如国内学校了解导致的。对于stanfordnlp的分析结果仅适用于其中文nlp分析，不能推广至其它语种。

此外，本次实验中使用的主要是商业新闻，其具有专业性强，文学性差，专业名词较多，形容词较少，几乎没有描写性文字等特征。并且出现了大量人名、公司名、阿拉伯数字及字母简称，用词专业而非口语化。这些特征使其和日常对话，文学作品等类型的语料有很大的差别，故在实验过程中有许多特殊情况也许无法被测试到。如果要进一步测试，还需要使用其它语料进行分析。

**六、实验感想**

通过编写这次实验的代码，我学会了使用ltp，stanfordnlp，handnlp这三个重要的开源工具包。在分析实验结果的过程中，我加深了对依存句法分析的印象。这一方面的知识不仅需要计算机和大数据的知识，还需要结合语言学的部分内容。最后，在分析和撰写这次实验报告的过程中，受限于我匮乏的语言知识、低下的语文功底和对方言的长期使用，对于某些句子成分和关系的分析不准确，还望指正。