1.任务目标

给定一篇（些）论文，判断其是否为真论文，即非scigen生成的论文。

2.数据预处理

(一) 获取训练样本

正样本来源：由老师给定的三千多论文样本

负样本来源：于MIT scigen上爬取下来的论文，原打算爬三千多篇论文下来，由于有大量重复，最后只有一千六百多篇。

代码：crawelFalse.py

（二）转换pdf

用linux自带的pdftotext将pdf转化为txt

在pdf转化为txt已经自动进行了去除图和表格

（三）处理txt

去除标题和作者信息

去除Reference 信息

去除论文引用的标号 []

将其他括号，标点符号替换为空

所有单词转化为小写

数字归一化处理

代码：processTxt.py

3.判定

(一)词同现网络

1.词同现网络简介及网络构建

词同现网络：若两个词汇在同一单元（如邻接、段落）中共现，则认为它们存在关联关系。

词同现网络的思想是：对于每一个单词，对应词同现网络中的一个节点。若一个句子中的两个词之间间隔为n时存在词同现关系，即在网络中有边相连。

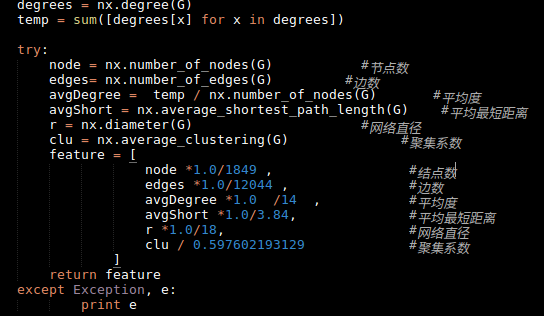
而大量的实践表明，n=2是比较合适的。两个词在句子中相邻出现是比较常见的。

如：zhe xue 间距为1，boy next door的boy 和door间距为2。

而如果n取过大，会引入大量的无关的词语，增加模型的复杂度。

在构建的词同现网络中，段和段之间相对是独立的，而一个段落里说明的内容是相关的。即一个段落中的两个相邻句子之间是有关联的。对于一个段落中相邻的两个句子，将其处理为第一个句子的最后一个词和第二个句子的第一个词之间的间距为2。

对于一篇论文的词同现网络，计算出它的：结点数、边数、平均度数、平均路径长度、网络直径、聚集系数，并且用这六个特征来表示这个网络。

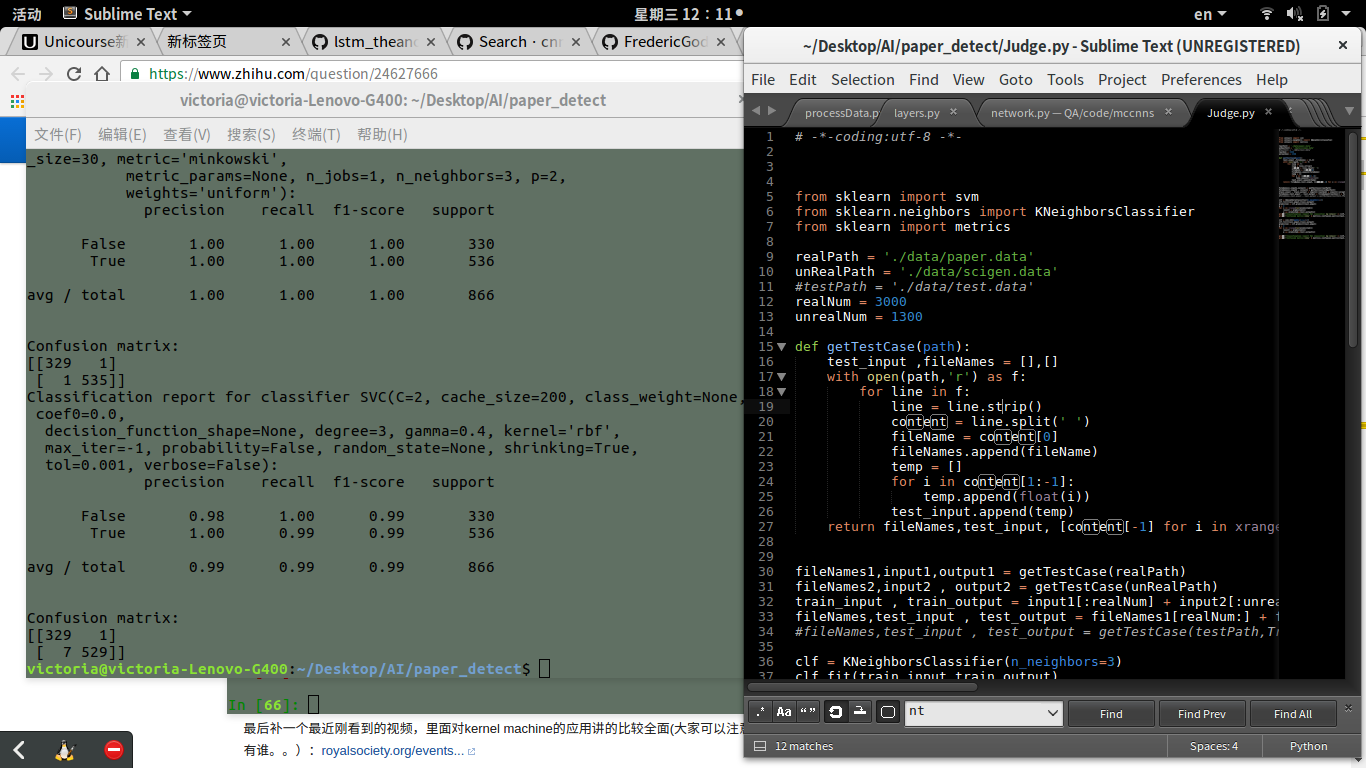


参考：基于频繁词集词共现网络的短文本聚类方法

基于词同现网络与支持向量机的科学论文甄别方法研究

2.分类器的选择

分别尝试了SVM和KNN进行训练和识别，结果如下：



1. 结论：

可以看出在训练集比测试集接近2：1的情况下准确率可以达到100%（只有几篇文章被判断错误），如果有大量的数据集，使得数据比列接近5：1则效果会更好。