基础部分：TextRank

概述

阅读论文【TextRank Bringing Order into Texts】，实现文中的textRank的keyword extraction算法，并在给定数据集上实验。

提供内容：

1.dataSet文件夹下提供了推荐的数据集(Krapivin2009)，这是一个有人工标注关键词(keyphrase)的数据集。

txt文件本身比较复杂，不必全部使用，数据使用时选取自己需要的部分使用即可（比如只使用title、abstract和introduction部分）。

一共大致2000篇，建议大家全部使用来测评自己的算法。

2.提供了一个简单的范例测评文件evaluation.py

要求：

实现textRank算法的完整流程，即：

（1） 对于给定的文本进行读取，分词，统一大小写，去除停用词，筛选名词形容词等预处理操作。

（2） 构建候选关键词图G=(V,E)，其中V为节点集，由候选关键词组成，使用某种度量来构造点之间的边。

（例如共现关系：两个节点之间仅当它们对应的词汇在长度为K的窗口中共现则存在边，K表示窗口大小即最多共现K个词汇）

（3） 根据公式迭代计算各节点的权重，直至收敛；

（4） 对节点权重进行倒序排列，得到排名前TopN个词汇作为文本关键词并输出。也可以将始终相邻的关键词连接组成关键词组。

其中(2)(3)步要求自己实现，其他部分允许调用第三方包（推荐nltk，pke，github等...）

进阶部分

以下部分任选1~2项完成。

【注意！】请在报告中声明自己所做的部分

1.使用textRank方法完成sentence extraction，为文章生成自动摘要。

2.参考另一篇论文【TopicRank Graph-Based Topic Ranking for Keyphrase Extraction】，简单复现topicRank的方法。

- 先验reference keyphrase candidates提取的部分允许调包

- 不要求完全一致，有类似思路的实现即可

3.做一个对比测评，使用更多的指标和算法，例如

算法	precision	recall	F-measure	...(其他方面的测评指标)
textRank（window size=5）				
textRank（window size=10）				
singleRank				
……(其他keyphrase extraction算法)				

【textRank以外的算法允许调包】

4.基于自己建立的text graph应用一个社区发现类的算法，探索并输出不同的聚集社区。

5.其他自己感兴趣的拓展模块。（自己的各类思考发现创新都是可以的）

报告提交

书写作业报告

内容格式不限。

可以写自己的探索历程，工作内容，创新思考，心得体会，遇到的困难与解决方法……

记得标记哪些是自己的工作，哪些是调用外部包。

总之不管你做了什么都可以写到报告里，有什么亮点和创新也请高光描述，充实的报告意味着高的评价！~

实现代码

代码语言建议C、C++、python，除核心算法部分外大部分均允许调用开源代码。

【希望你的代码有较高的可读性，添加必要的注释】

【将自己的代码和报告打包成zip或7z上传教学网。】

ps：如果速度太慢也可发邮件至liukaibo_0223@163.com或者联系助教刘恺博，支持物理交作业（U盘）。

【！！但请同时在教学网上提交一份报告。】