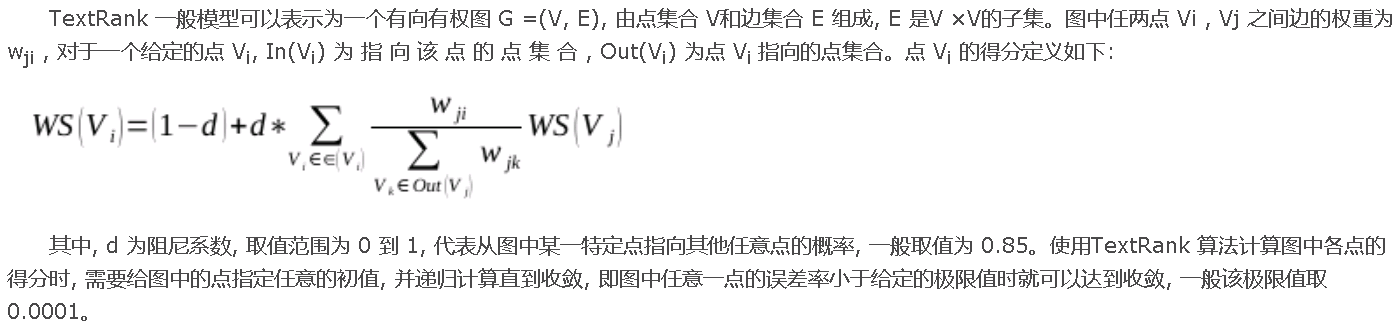
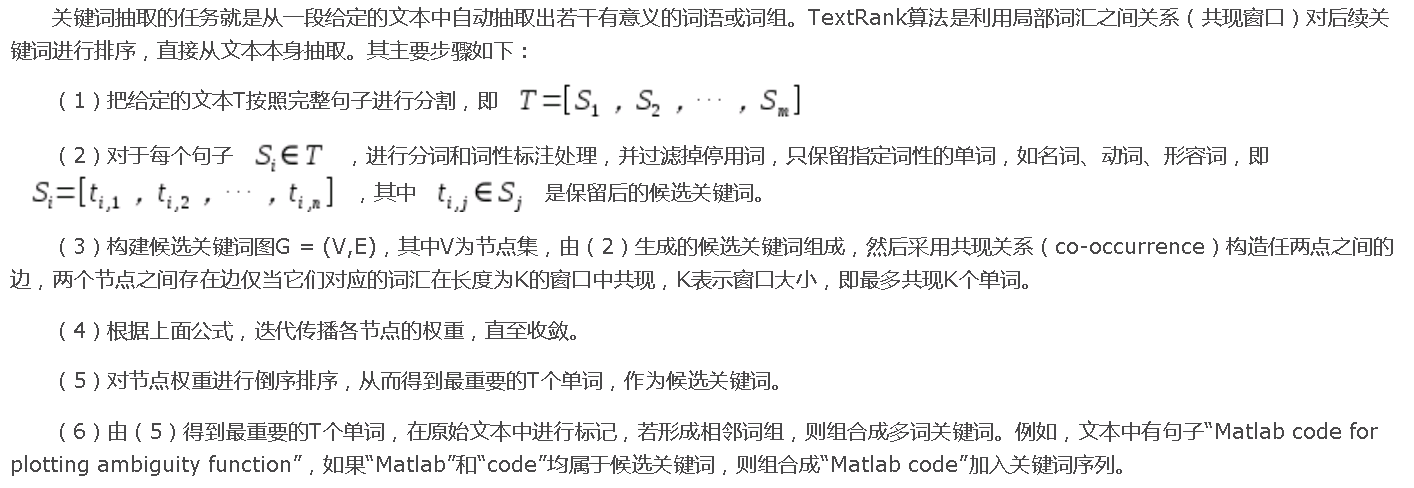
[TextRank](http://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf" \t "_blank)算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 [PageRank](http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf)算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同, TextRank不需要事先对多篇文档进行学习训练, 因其简洁有效而得到广泛应用。

* 如果一个单词出现在很多单词后面的话，那么说明这个单词比较重要
* 一个TextRank值很高的单词后面跟着的一个单词，那么这个单词的TextRank值会相应地因此而提高





**使用TextRank提取关键短语**

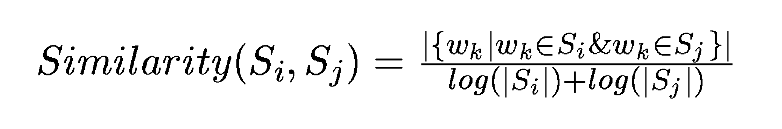
参照“使用TextRank提取关键词”提取出若干关键词。若原文本中存在若干个关键词相邻的情况，那么这些关键词可以构成一个关键短语。

例如，在一篇介绍“支持向量机”的文章中，可以找到三个关键词支持、向量、机，通过关键短语提取，可以得到支持向量机。  
使用TextRank提取摘要

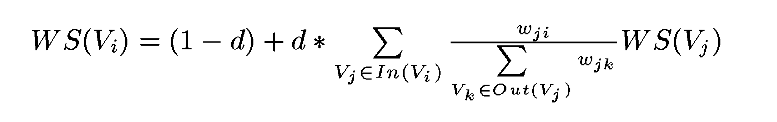
将每个句子看成图中的一个节点，若两个句子之间有相似性，认为对应的两个节点之间有一个无向有权边，权值是相似度。

通过pagerank算法计算得到的重要性最高的若干句子可以当作摘要。

论文中使用下面的公式计算两个句子Si和Sj的相似度：

[](https://www.letiantian.me/content/images/2014/12/textrank-01.png)

分子是在两个句子中都出现的单词的数量。|Si|是句子i的单词数。

由于是有权图，PageRank公式略做修改：  
[](https://www.letiantian.me/content/images/2014/12/textrank-02.png)

**实现TextRank**

基于Python 2.7的TextRank针对中文文本的库TextRank4ZH。位于：

<https://github.com/letiantian/TextRank4ZH>

下面是一个例子：

*# -\*- encoding:utf-8 -\*-*

**import** codecs

**from** textrank4zh **import** TextRank4Keyword, TextRank4Sentence

text = codecs.open('./text/01.txt', 'r', 'utf-8').read()

tr4w = TextRank4Keyword(stop\_words\_file='./stopword.data') *# 导入停止词*

*# 使用词性过滤，文本小写，窗口为2*

tr4w.train(text=text, speech\_tag\_filter=**True**, lower=**True**, window=2)

**print** '关键词：'

*# 20个关键词且每个的长度最小为1*

**print** '/'.join(tr4w.get\_keywords(20, word\_min\_len=1))

**print** '关键短语：'

*# 20个关键词去构造短语，短语在原文本中出现次数最少为2*

**print** '/'.join(tr4w.get\_keyphrases(keywords\_num=20, min\_occur\_num= 2))

tr4s = TextRank4Sentence(stop\_words\_file='./stopword.data')

*# 使用词性过滤，文本小写，使用words\_all\_filters生成句子之间的相似性*

tr4s.train(text=text, speech\_tag\_filter=**True**, lower=**True**, source = 'all\_filters')

**print** '摘要：'

**print** '\n'.join(tr4s.get\_key\_sentences(num=3)) *# 重要性最高的三个句子*

运行结果如下：

关键词：

媒体/高圆圆/微/宾客/赵又廷/答谢/谢娜/现身/记者/新人/北京/博/展示/捧场/礼物/张杰/当晚/戴/酒店/外套

关键短语：

微博

摘要：

中新网北京12月1日电(记者 张曦) 30日晚，高圆圆和赵又廷在京举行答谢宴，诸多明星现身捧场，其中包括张杰(微博)、谢娜(微博)夫妇、何炅(微博)、蔡康永(微博)、徐克、张凯丽、黄轩(微博)等

高圆圆身穿粉色外套，看到大批记者在场露出娇羞神色，赵又廷则戴着鸭舌帽，十分淡定，两人快步走进电梯，未接受媒体采访

记者了解到，出席高圆圆、赵又廷答谢宴的宾客近百人，其中不少都是女方的高中同学

另外， [jieba分词](https://github.com/fxsjy/jieba" \t "_blank)提供的基于TextRank的关键词提取工具。 [snownlp](https://github.com/isnowfy/snownlp" \t "_blank)也实现了关键词提取和摘要生成。

**下载：**

<https://github.com/letiantian/TextRank4ZH>