TextRank算法 18051429 俞皓

[TextRank](http://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf" \t "https://www.cnblogs.com/clover-siyecao/p/_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" \t "https://www.cnblogs.com/clover-siyecao/p/_blank)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。

TextRank 一般模型可以表示为一个有向有权图 G =(V, E), 由点集合 V和边集合 E 组成, E 是V ×V的子集。图中任两点 Vi , Vj 之间边的权重为 wji , 对于一个给定的点 Vi, In(Vi) 为 指 向 该 点 的 点 集 合 , Out(Vi) 为点 Vi指向的点集合。

关键词抽取的任务就是从一段给定的文本中自动抽取出若干有意义的词语或词组。TextRank算法是利用局部词汇之间关系（共现窗口）对后续关键词进行排序，直接从文本本身抽取。其主要步骤如下：

（1）把给定的文本T按照完整句子进行分割,即T=S1,S2...Sm。

（2）对于每个句子Si属于T,进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的单词，如名词、动词、形容词，即Si=ti1,ti2,ti3...tin，其中tij属于Sj是保留后的候选关键词。

（3）构建候选关键词图G = (V,E)，其中V为节点集，由（2）生成的候选关键词组成，然后采用共现关系（co-occurrence）构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为K的窗口中共现，K表示窗口大小，即最多共现K个单词。

（4）根据上面公式，迭代传播各节点的权重，直至收敛。

（5）对节点权重进行倒序排序，从而得到最重要的T个单词，作为候选关键词。

（6）由（5）得到最重要的T个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。例如，文本中有句子“Matlab code for plotting ambiguity function”，如果“Matlab”和“code”均属于候选关键词，则组合成“Matlab code”加入关键词序列。