**1、TF-IDF算法分析**

在信息检索理论中，TF-IDF是Term Frequency - Inverse Document Frequency的简写。TF-IDF是一种数值统计，用于反映一个词对于语料中某篇文档的重要性。在信息检索和文本挖掘领域，它经常用于因子加权。

TF-IDF的主要思想就是：如果某个词在一篇文档中出现的频率高，也即TF高；并且在语料库中其他文档中很少出现，即DF的低，也即IDF高，则认为这个词具有很好的类别区分能力。

TF-IDF在实际中主要是将二者相乘，也即TF \* IDF，TF为词频（Term Frequency），表示词t在文档d中出现的频率；IDF为反文档频率（Inverse Document Frequency），表示语料库中包含词t的文档的数目的倒数。

**TF公式：**

TF计算公式为，

TF=count(t)count(di)TF=count(t)count(di)

式中，count(t)表示文档di中包含词t的个数；

count(di)表示文档di的词的总数；

**IDF公式：**

IDF计算公式为，

IDF=num(corpus)num(t)+1IDF=num(corpus)num(t)+1

式中，num(corpus)表示语料库corpus中文档的总数；

num(t)表示语料库corpus中包含t的文档的数目；

**应用到关键词抽取：**

1. 预处理，首先进行分词和词性标注，将满足指定词性的词作为候选词；

2. 分别计算每个词的TF-IDF值；

3. 根据每个词的TF-IDF值降序排列，并输出指定个数的词汇作为可能的关键词；

**2 .TextRank算法分析**

类似于PageRank的思想，将文本中的语法单元视作图中的节点，如果两个语法单元存在一定语法关系（例如共现），则这两个语法单元在图中就会有一条边相互连接，通过一定的迭代次数，最终不同的节点会有不同的权重，权重高的语法单元可以作为关键词。

节点的权重不仅依赖于它的入度结点，还依赖于这些入度结点的权重，入度结点越多，入度结点的权重越大，说明这个结点的权重越高；

TextRank迭代计算公式为，

WS(Vi)=(1−d)+d∗∑Vj∈In(Vi)wji∑Vk∈Out(Vj)wjk∗WS(Vj)WS(Vi)=(1−d)+d∗∑Vj∈In(Vi)wji∑Vk∈Out(Vj)wjk∗WS(Vj)

节点i的权重取决于节点i的邻居节点中i-j这条边的权重 / j的所有出度的边的权重 \* 节点j的权重，将这些邻居节点计算的权重相加，再乘上一定的阻尼系数，就是节点i的权重；

阻尼系数 d 一般取0.85；

**算法通用流程：**

1. 标识文本单元，并将其作为顶点加入到图中；

2. 标识文本单元之间的关系，使用这些关系作为图中顶点之间的边，边可以是有向或者无向，加权或者无权；

3. 基于上述公式，迭代直至收敛；

4. 按照顶点的分数降序排列；

* 1.本模型使用co-occurrence关系，如果两个顶点相应的语义单元共同出现在一个窗口中（窗口大小从2-10不等），那么就连接这两个顶点；
* 2.添加顶点到图中时，需要考虑语法过滤，例如只保留特定词性（如形容词和名词）的词；

**应用到关键短语抽取：**

1. 预处理，首先进行分词和词性标注，将单个word作为结点添加到图中；

2. 设置语法过滤器，将通过语法过滤器的词汇添加到图中；出现在一个窗口中的词汇之间相互形成一条边；

3. 基于上述公式，迭代直至收敛；一般迭代20-30次，迭代阈值设置为0.0001；

4. 根据顶点的分数降序排列，并输出指定个数的词汇作为可能的关键词；

5. 后处理，如果两个词汇在文本中前后连接，那么就将这两个词汇连接在一起，作为关键短语；