**工业机理模型的多标签学习模型使用说明**

1. 模型用途

根据模型各部件性能及其制造流程等因素，对模型添加标签或关键词。输入模型各部件性能、所需原材料及其制造流程等因素，输出标签或关键词。

1. 数据说明

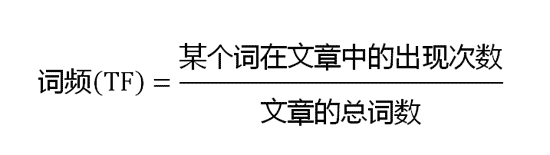
输入数据需要一个文本文件，该文本文件对模各部件性能及其制造流程进行详细描述。

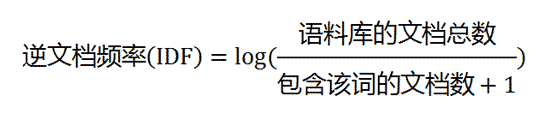
输出文件中为输入文本文件对应的关键词。

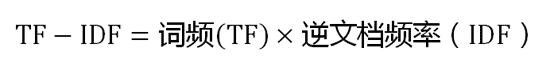
1. 算法原理

工业机理模型的多标签学习模型提供两种方法生成关于模型的关键词。第一种是TF-IDF算法， TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。 TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF的计算过程如下：







计算出文档的每个词的TF-IDF值，然后按降序排列，取排在最前面的几个词

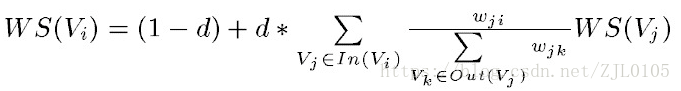
另一种方法是Textrank方法，TextRank 用于关键词提取的算法如下：

（1）把给定的文本 T 按照完整句子进行分割，即:

（2）对于每个句子，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，其中 ti,j 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 TextRank 的公式，



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。

本模型使用了jieba工具包中提供的TF-IDF方法和TextRank方法，对输入文本进行关键词提取，通过参数控制关键词算法的选择。

1. 使用方法

具体输入命令：

**python model.py --test\_dir待抽取关键词文本 --tags\_dir 结果保存地址 --topK 返回的关键词个数 --关键词算法**

例如：

**python model.py --test\_dir ./sample.txt --tags\_dir ./sample\_keywords.txt --topK 20**

**--method tf-idf**

分析结果会存放在**结果保存地址**中