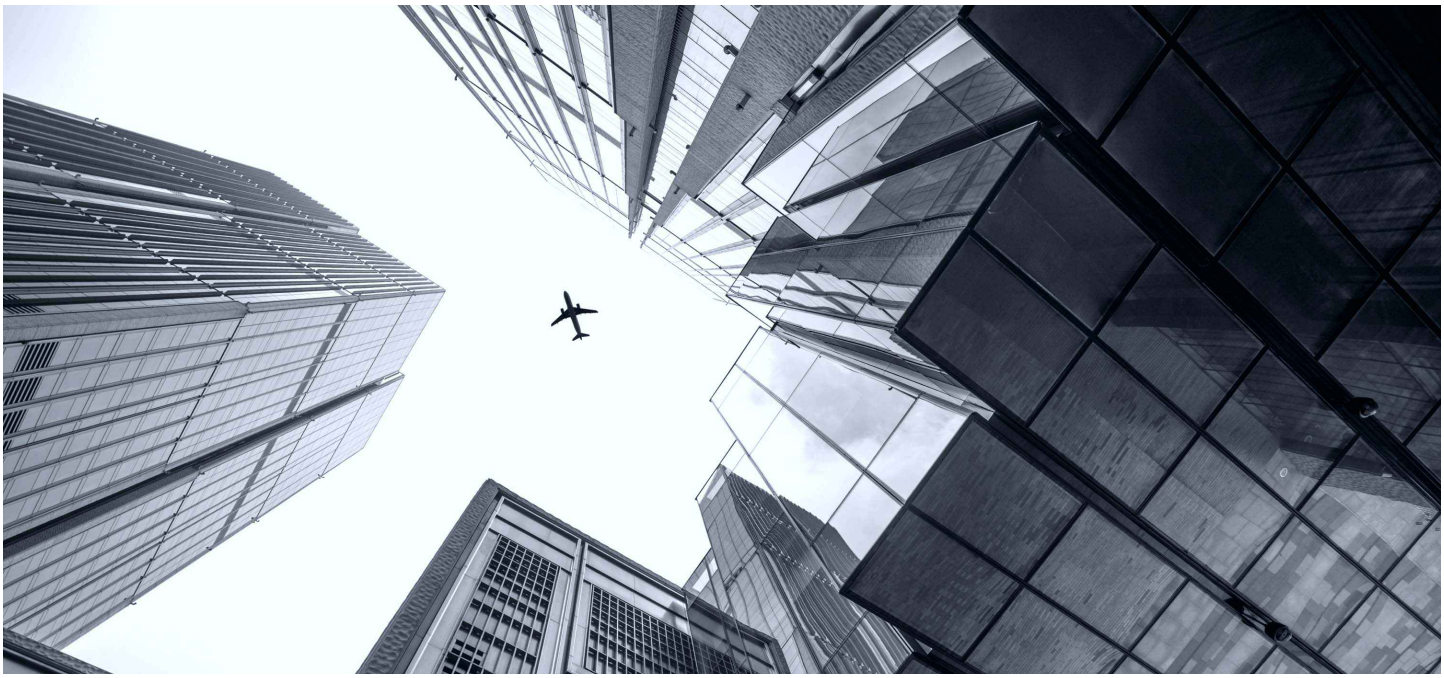

TFIDF

TFIDF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。其核心思想是，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。计算上，TFIDF由词频TF（term frequency）和逆文档频率IDF（inverse document frequency）两部分相乘得到。

关键词：关键词提取，文本特征，文本间相似度

一、计算

词 i 对文档 j 的权重为：

$$tfidf_{ij} = tf_{ij} \times idf_{ij} \tag{1}$$

$$tf_{ij} = \frac{\text{i在j中出现的次数}}{\text{j的总词数}} \tag{2}$$

$$idf_{ij} = \lg \frac{\text{语料库的总文档数}}{\text{包含}i\text{的文档数}} \tag{3}$$

边界情况：公式3的分母在为0的时候，要加1做平滑。

二、思考

TF的值域是[0,1]，随着词频增大，TF取最大值1；

IDF的值域是[0, logN]，随着逆文档概率增大，IDF取最小值0；

实战技巧：停用词需要过滤，否则会误抬高相似度；

优点：

- 冷启动，实现简单；
- 可以作为特征；

缺点：

- 依赖高质量的语料。如果语料库是同类语料库，那么重点词会被掩盖；
- 忽略了位置信息以及词与词之间的相互关系。文章标题或者首段的重要性应该更高；
- 终究是匹配的方法，无法搞定语义问题；

改进方法

IDF改成IWF

$$IWF_{ij} = \frac{i在语料库所有文档中的词频}{i在j中的词频}$$

小改进，以上的几个缺点仍然没有解决。

三、应用

1. 关键词提取

可以和TextRank、LDA等主题模型进行对比。

2. 文档间的相似度

通过TFIDF，可以给每个文档输出一个one-hot向量，每个分量上的值为n或者0，这样便可以计算两个文档的相似。