舆情分析

文本情感分析又称作意见挖掘，倾向性分析等，简单而言是对具有感情色彩的主观性文本进行分析和处理以及归纳和总结的过程。互联网产生了大量的针对于任务、事件、产品等的评价，基于此，潜在的用户就可以浏览这些具有主管色彩的文本了解大众舆论的走向。

比如，酒店的评价。

具体总结：先针对训练文本进行分词处理，统计词频，通过词频-逆文档频率算法获得该词对样本语义的贡献，根据每一个词的贡献力度，构建有监督的分类学习模型，

训练后，把测试样本交给模型去预测分类结果。

1、名词解释

1.1、自然语言处理NLP(Nature Language Processing)

是计算机科学领域与人工智能领域的一个重要方向，研究能实现人与计算机之间用自然语言进行有效的通信。

1.2、词频TF(Term Frequency)

单词在矩阵中出现的次数除以句子所有的单词的总数称为词频，词频的大小可以反映该词的贡献程度。词频相对于单词出现的数量更能客观的评估某一个单词对于这个样本的语义的贡献程度，词频越高，对语义的贡献程度越大。

对词袋矩阵归一化就得到了词频。

1.3、文档频率DF(Document Frequency)

DF = 含有某一个单词的样本数/总的样本数

有些单词可能大部分的语句中都包含，比如我，这个，很，语气助词等等，对于判断样本的类别没有太大的作用，因此需要设计一种算法对这些单词对于语义的贡献程度降低，DF这个数越大，说明这个词越没用，这个数对于样本的语义贡献程度反相关。

1.4、逆文档频率IDF(Inverse Document Frequency)

如果一个词在许多文档中都出现(**IDF低**)，那么它可能不包含太多的信息，因为它普遍存在，相反，如果一个词只在少数文档中出现(**IDF高**)，那么它可能包含更多的信息，因为它具有区分不能文档的能力。逆文档频率的计算公式：

IDF=log(文档的总数/出现某单词的文档数+1)

这个计算过程可以有效的抑制那些在文档集中频繁出现的词，同时提升了那些在少数文档中出现的词的重要性。

这种算法在信息检索和文本挖掘中非常重要，因为它可以帮助确定哪些词对于区别文档更重要。

1.5、TF-IDF 词频-逆文档频率

是一种用于信息检索与数据挖掘的加权技术，其结合了一个词在文档中的词频和逆文档频率来计算某单词在文档集中的重要性。

2、环境、模块等

2.1 nltk模块(Nature Language Toolkit)自然语言处理共建包。

pip install nltk

2.2 jieba模块(结巴) 一款用于处理中文文本的分词工具。

pip install jieba

2.3 tokenizers(分词器)

NLP项目中，需要对文本进行编码，以便计算机识别，比如：I love you---🡪[12,45,56]，也就是一个字典进行一一对应，一些训练好的NLP模型中基本上都有自己的一套tokenizers工具，tokenizers的最终目的就是将输入的语句编码成数字。

安装好之后，nltk会自动的寻找tokenizers工具的词库punkt资源。具体做法，将词库放入Anaconda3的nltk\_data文件夹即可。