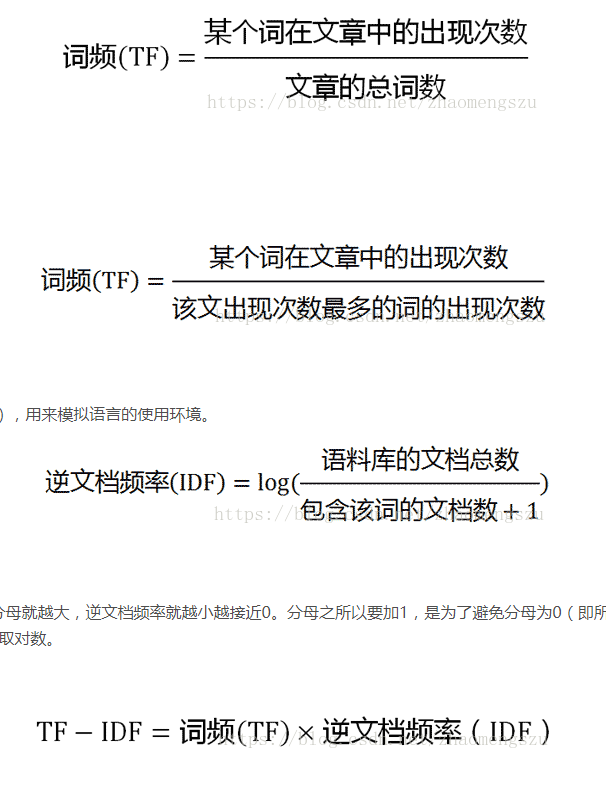


取10个主题的标题各1000条，每组都进行分词和词频统计

将10组中的词总和在一起成为一个语料库，库中的词频只计算该词在10组中出现的组数，全都出现过词频就为10，只出现在一组中，词频就为1，并且删除其中的停用词，如“下面”这种（删除停用词这步还没做，因为还有很多没有意义的词并不在停用词库中）



利用tfidf的公式计算每个词的权值，可以得出每个主题中的权值排行

（这一步有问题，因为科技一次明显在人工智能中频率很高，但是因为他在其他9组中都出现了，所以idf值=0，会忽略科技在人工智能主题中的作用）

计算标题权值：输入标题并且进行分词，将每个词的权值相加并且除以标题长度

去除该分类中权重的最后20%的标题用以过滤（因为每组中的标题错误率都很高），再重新进行权值计算

计划的改良方案：只针对一类主题，如”人工智能”，分类为人工智能和非人工智能，用类似卡方检验的公式，可以让科技这种在一类中出现频率高，一类中出现频率低的词脱颖而出