BOW&TFIDF

1.BOW(词袋法)

词袋法是一种文本向量的表示方法，它忽略了词汇的顺序、文本语义，通过统计文本中出现的单词数量，来简单地构造文本向量。词袋法构造的文本向量是高维稀疏的向量（维度由词典的单词数量决定）。

例如，对于文本：

```
undefined
content= ['this is the first document',
        'this is the second second document',
        'and the third one',
        'is this the first document']
```

通过词袋法构建文本向量可以表示为（元组中第一个数字表示单词对应在词典中的序号，第二个数字表示单词在文本中出现的频数）：

```
content= [[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)],
        [(0, 1), (1, 1), (2, 1), (5, 2), (4, 1)],
        [(6, 1), (2, 1), (7, 1), (8, 1)],
        [(1, 1), (0, 1), (2, 1), (3, 1), (4, 1)]]
```

词典为：

```
{'this': 0, 'is': 1, 'the': 2, 'first': 3, 'document': 4, 'second': 5, 'and': 6, 'third': 7, 'one': 8})
```

2.TF-IDF

TF表示词条在文本中出现的频率，这个数字通常会被归一化(一般是词频除以文本总词数), 以**防止它偏向长的文本**（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否）。

TF计算公式为：$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$,其中i表示文本中的某个单词i，j表示所有文档中的某个文本，$n_{i,j}$表示编号为i的单词在第j个文档中的频数，$\sum_k n_{k,j}$ 表示第j个文档的所有单词个数。

以上述例子为例，第一个文本'this is the first document'中'this'的tf值可以计算为：$tf = \frac{1}{5} = 0.2$

IDF表示关键词的普遍程度(它是一个表示某单词在文档全局状况的指标)。如果包含该词条文档越少， IDF越大，则说明该词条具有很好的类别区分能力。某一特定词语的IDF，可以由**总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。**

IDF计算公式:$IDF_i = log\frac{|D|}{1+|j:t_i \in d_j}$,$|D|$表示文档所有文本的个数，$1+|j:t_i \in d_j$表示文档中包含单词i的所有文本个数加1。

以BOW中的例子为例，第一个文本'this is the first document'中'this'的idf值可以计算为：$idf = log\frac{4}{3+1}$

TF-IDF为:TF*IDF，以BOW中的例子为例，计算出的TF-IDF为:

```
content= [[(2, 0.0), (1, 0.0), (3, -0.44628710262841945), (4, 0.05753641449035617), (0, 0.0)],
        [(2, 0.0), (1, 0.0), (3, -0.37190591885701 62), (5, 0.2310490601866 4842), (0, 0.0)],
        [(7, 0.17328679513998632), (3, -0.05578588782855243), (6, 0.17328679513998632), (8, 0.17328679513998632)],
        [(1, 0.0), (2, 0.0), (3, -0.44628710262841945), (4, 0.05753641449035617), (0, 0.0)]]
```

词典为：

```
{'first': 0,'is': 1,'document': 2,'this': 3,'the': 4,'second': 5,'one': 6 'and': 7, 'third': 8}
```