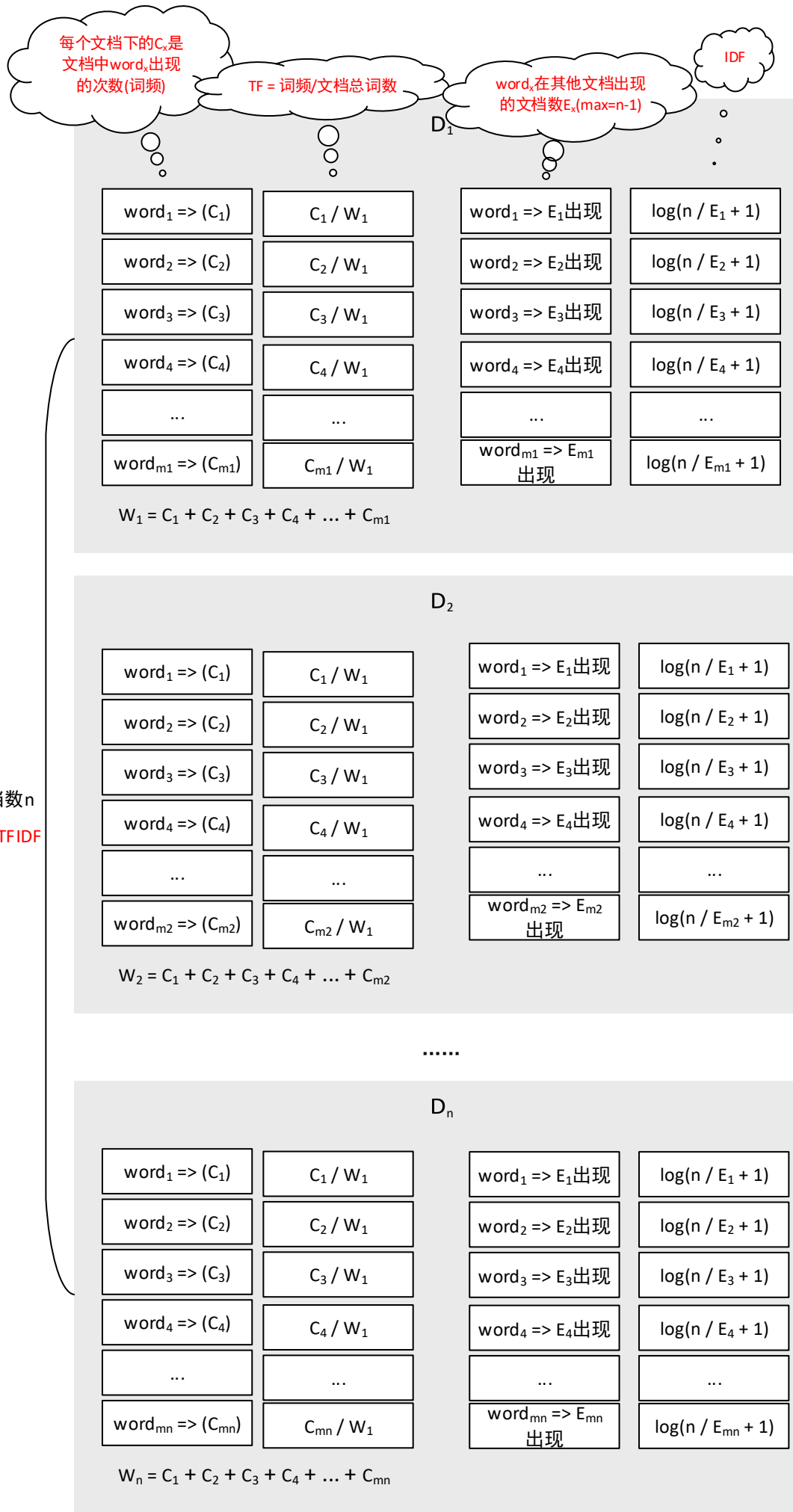


```
@author: natasha_yang

@e-mail: ityangna0402@163.com
```

我的想法一:
比如问题(比较短的文本)集合Q_1的label是A_1
可否将Q_1中所有的问题组成一个doc求TF-IDF的值
那如果要预测新的问题的时候要怎么办??

我的想法二:
可以将Q_1中所有的问题组成一个doc求TF-IDF的值
然后形成文档的词TF-IDF value的向量(词典顺序)
求主题的相似性

我的想法三:
在求IDF的时候可否考虑文档X中的word在非X的文档中出现的次数(感觉传统的IDF不能很好的体现主题的无关性)
比如说存在10个主题的文档,(机器)在类型1中出现了10000次,在其他9个类型文档中各出现一次(那么IDF=1)

我的想法四:
如果加上某个单词在当前X文档出现的次数/某个单词在所有文档中出现的次数作为权重,这个值越接近1/n就越不重要,但是这样的话就要注意输入集合的数据要平衡