山东大学 计算机科学与技术 学院

信息检索与数据挖掘 课程实验报告

学号：201600130077	姓名： 于海洋	班级： 16 人工智能

实验题目：预处理文本数据集，并得到每个文本的 VSM

实验内容：

1、利用从网上找的 stopword 集，建立一个 list 存放，用于筛选

```
    ff_stop = open("stopword", "rb")
    for stop_line in ff_stop:
        str_stop = str(stop_line).strip()
        for sword in re.findall(letters, str_stop):
            f_stop.append(sword)
```

2、计算 df(t)——出现单词 t 的文章数
同时统计 c(t,d)——在文章 d 中 t 出现的次数

```
def get_df(ffile):

    tdic = {}
    ff = open(ffile, "rb")
    for line in ff:
        str_line = str(line).strip().lower()
        for word in re.findall(letters, str_line):
            if word not in f_stop:
                if word in tdic:
                    tdic[word] += 1
                else:
                    tdic[word] = 1
    for word in tdic:
        if word in DF:
            DF[word] += 1
        else:
            DF[word] = 1
    return tdic
```

3、统计文章数和 TF(t, d) = 1 + log(c(t, d)) ,if c(t, d)>0

```
def get_tf():
    num_doc = 0
    root_1dir = r"20news-18828"
    for pack in os.listdir(root_1dir):
        root_dir = root_1dir + "\\" + pack
        for file in os.listdir(root_dir):
            file_name = root_dir + "\\" + file
            num_doc += 1
            cword = get_df(file_name)
            for k in cword:
                if cword[k] > 0:
                    TF[k, pack + "\\" + file] = 1.0 + math.log(cword[k])
                else:
                    TF[k, pack + "\\" + file] = 0
    return num_doc
```

4、统计 IDF(t)=log(N/df(t))

```
def get_idf():
    N = get_tf()
    for word in DF:
        IDF[word] = math.log(N / DF[word])
```

5、计算 w(t, d) = TF(t, d) * IDF(t)

```
def get_w():
    for k1, k2 in TF:
        w[k1, k2] = TF[k1, k2] * IDF[k1]
```

实验过程中遇到和解决的问题：

（记录实验过程中遇到的问题，以及解决过程和实验结果。可以适当配以关键代码辅助说明，但不要大段贴代码。）

1、高低频词的影响，起先统计 words bag 的时候考虑了频率的影响，就先去掉了，但后来速度方面发现没有太大影响，就都统计进去了

2、tokenization 和 stop word 都做了，但是 stemming 还没有实现

结论分析与体会：

词频的舍去很难把握，词量太大，但是还是把词库做好了，准备为下面垃圾邮件识别做准备