

# The First Experimental Report





















## 一、实验内容

---对给定的 20 组新闻数据进行预处理，并得到每个文本的 VSM 表示

---实现 KNN 分类器，测试其在给定数据集上的效果

## 二、数据集

20 个新闻数据文件夹

名称	修改日期	类型
 alt.atheism	2018/11/3 10:46	文件夹
 comp.graphics	2018/11/3 10:46	文件夹
 comp.os.ms-windows.misc	2018/11/3 10:46	文件夹
 comp.sys.ibm.pc.hardware	2018/11/3 10:46	文件夹
 comp.sys.mac.hardware	2018/11/3 10:46	文件夹
 comp.windows.x	2018/11/3 10:46	文件夹
 misc.forsale	2018/11/3 10:46	文件夹
 rec.autos	2018/11/3 10:46	文件夹
 rec.motorcycles	2018/11/3 10:46	文件夹
 rec.sport.baseball	2018/11/3 10:46	文件夹
 rec.sport.hockey	2018/11/3 10:46	文件夹
 sci.crypt	2018/11/3 10:46	文件夹
 sci.electronics	2018/11/3 10:46	文件夹
 sci.med	2018/11/3 10:46	文件夹
 sci.space	2018/11/3 10:46	文件夹
 soc.religion.christian	2018/11/3 10:46	文件夹
 talk.politics.guns	2018/11/3 10:46	文件夹
 talk.politics.mideast	2018/11/3 10:46	文件夹
 talk.politics.misc	2018/11/3 10:46	文件夹
 talk.religion.misc	2018/11/3 10:46	文件夹

每个文件夹里的数据文件

名称	日期时间	类型	大小
49960	1999/12/9 3:29	文件	12 KB
51060	1999/12/9 3:29	文件	32 KB
51119	1999/12/9 3:29	文件	4 KB
51120	1999/12/9 3:29	文件	2 KB
51121	1999/12/9 3:29	文件	1 KB
51122	1999/12/9 3:29	文件	5 KB
51123	1999/12/9 3:29	文件	1 KB
51124	1999/12/9 3:29	文件	2 KB
51125	1999/12/9 3:29	文件	3 KB
51126	1999/12/9 3:29	文件	1 KB
51127	1999/12/9 3:29	文件	1 KB
51128	1999/12/9 3:29	文件	1 KB
51130	1999/12/9 3:29	文件	2 KB
51131	1999/12/9 3:29	文件	3 KB
51132	1999/12/9 3:29	文件	3 KB
51133	1999/12/9 3:29	文件	2 KB
51134	1999/12/9 3:29	文件	2 KB
51135	1999/12/9 3:29	文件	2 KB
51136	1999/12/9 3:29	文件	3 KB
51139	1999/12/9 3:29	文件	2 KB

数据文件里的内容

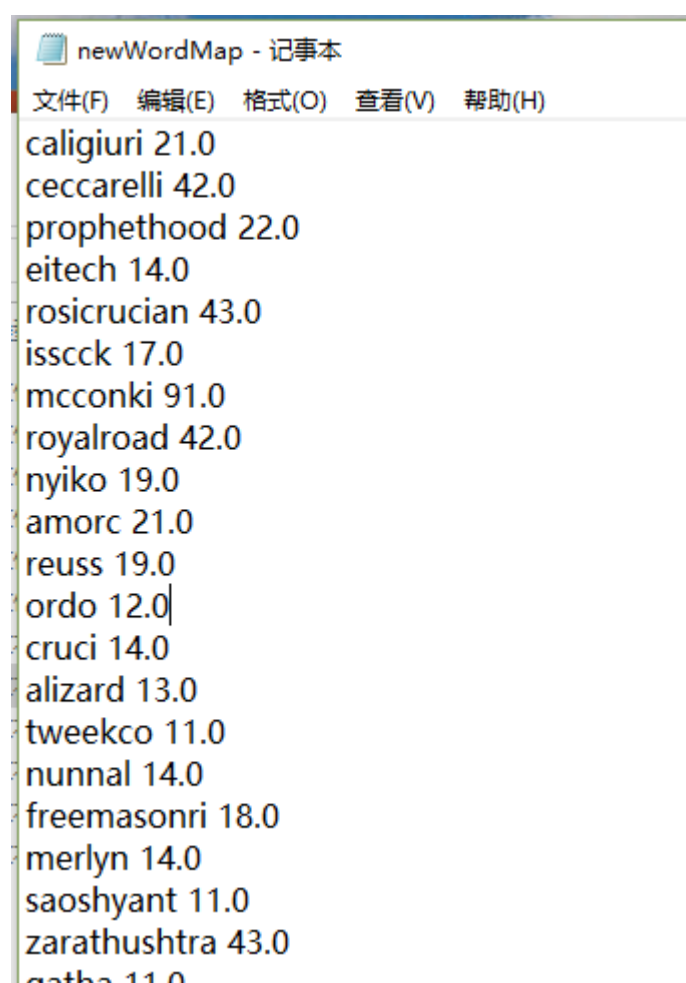
From: mathew <mathew@mantis.co.uk> Subject: Alt.Atheism FAQ: Atheist Resources  
Archive-name: atheism/resourcesAlt-atheism-archive-name: resources  
Last-modified: 11 December 1992Version: 1.0 Atheist Resources  
Addresses of Atheist Organizations USAFREEDOM  
FROM RELIGION FOUNDATIONDarwin fish bumper stickers and assorted other  
atheist paraphernalia areavailable from the Freedom From Religion Foundation in  
the US.Write to: FFRF, P.O. Box 750, Madison, WI 53701.Telephone: (608) 256-8900  
EVOLUTION DESIGNSEvolution Designs sell the "Darwin fish". It's a fish symbol, like  
the onesChristians stick on their cars, but with feet and the word "Darwin" written  
inside. The deluxe moulded 3D plastic fish is \$4.95 postpaid in the US.Write to:  
Evolution Designs, 7119 Laurel Canyon #4, North Hollywood, CA 91605.People  
in the San Francisco Bay area can get Darwin Fish from Lynn Gold --try mailing  
<figmo@netcom.com>. For net people who go to Lynn directly, theprice is \$4.95  
per fish.AMERICAN ATHEIST PRESSAAP publish various atheist books -- critiques of  
the Bible, lists ofBiblical contradictions, and so on. One such book is:"The Bible  
Handbook" by W.P. Ball and G.W. Foote. American Atheist Press.372 pp. ISBN 0-  
910309-26-4, 2nd edition, 1986. Bible contradictions,absurdities, atrocities,

三、实验过程

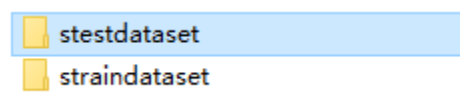
(1) 先对数据集进行预处理

去停用词，词干分析，去除非字母字符

- (2) 对预处理后的数据集进行分割，20%作为测试集，80%作为训练集
- (3) 针对训练集构造词典，过滤掉出现次数大于 10 的词



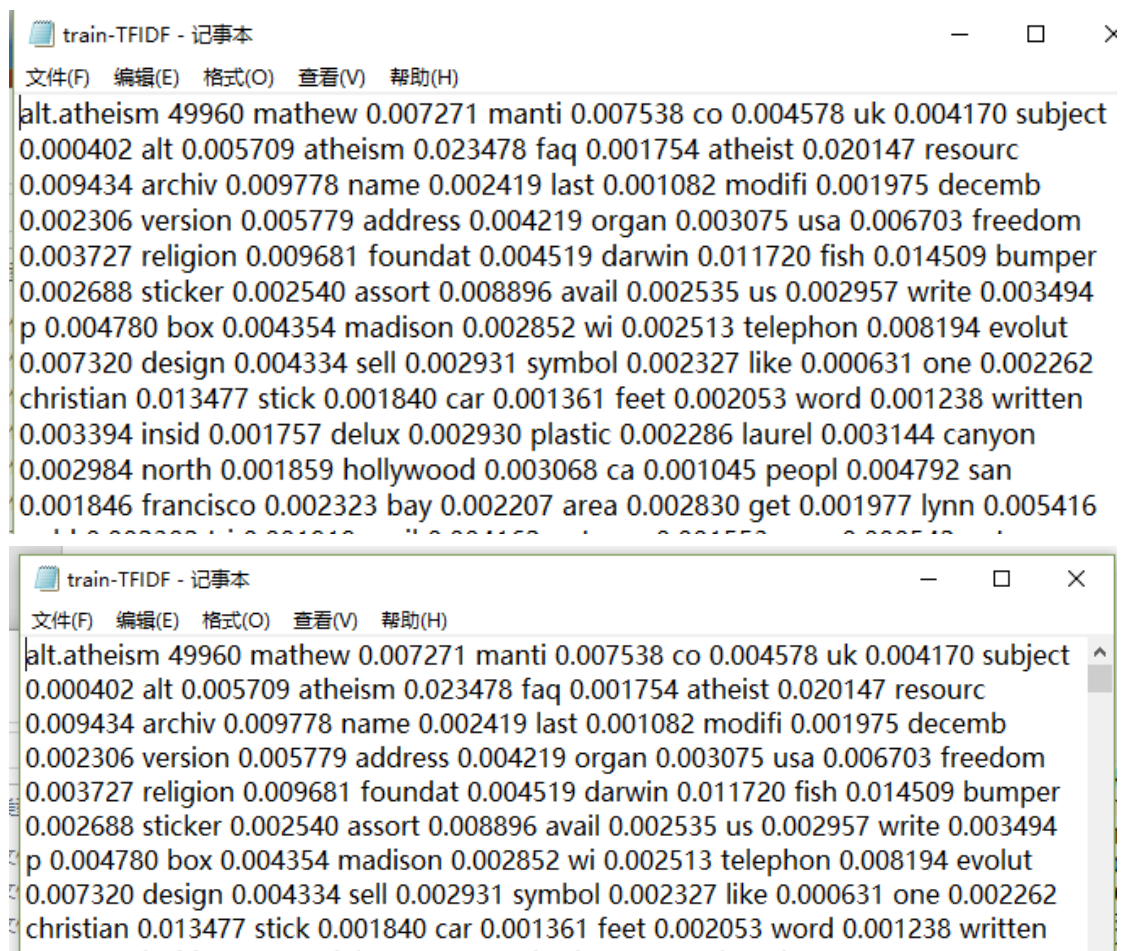
- (4) 根据构造的词典，分别针对测试集与训练集进行特征词选取，生成 20 个文件夹，每个文件夹的每篇文档存放本篇数据文件的特征词



- (5) 分别计算训练集与测试集的 TF-IDF 值
- 先计算 IDF，生成 test-IDF 和 train-IDF 的 txt 文件



再计算 TF 与 IDF 的成绩



#### (6) 实现 KNN

根据 TF-IDF 的值计算距离，距离越近代表相似度越大，根据距离进行分类。

统计分类正确的次数，计算准确率。

### 四、实验结果

```
KNN x
↑ talk.religion.misc soc.religion.christian rightCount: 3016
↓ this is 3767 round
≡ talk.religion.misc talk.religion.misc rightCount: 3017
⇅ rightCount 3017,count: 3767,acc: 0.800903
🖨️
🗑️ Process finished with exit code 0
```

## 五、实验心得

本次实验是用 python 写的，由于之前没有接触过 python，所以写起来比较有难度，要一边学习 python，一边学习 knn 是如何实现的。

编写过程中也遇到了很多困难，比如如何构造词典，如何划分数据集等，通过查找资源，询问同学等，困难得以解决，自己也有所收获。由于对 python 不太了解，其中一些语法规则，模块函数等，在使用时总会出错，多亏同学的帮助，找出错误，加以纠正。

通过本次实验，真正的理解了 knn 算法的思想，也对 python 这个语言有了大致了解，确实是个很人性化的语言，需要我进一步去学习。

实验是集中在最后一周做的，时间比较紧张，这也是个教训。以后再做事情最好不要拖延。