实验报告

预处理与词频统计

首先是文本的预处理，在这一阶段我并没有使用 textblob 进行分词，而是参照网上的方法重新将文本打印了一遍，同时打印的时候每行只打印一个单词，程序中使用 createFiles(oripath,path)完成这一过程，在这一方法中调用了 createProcessFile 将单词分行打印，这样在后续读取时也可以达到分词的效果。

之后调用 lineProcess 方法对文本单词进行处理，使用 nltk 中现成的方法去停用词、提取词干，将非英文单词的部分去除，使用 countWords（path）进行词频统计。

计算 IDF-TF 值，建立向量空间

统计所有 Train 文档中出现的词干，多滤掉低频词，其他单词使用如下格式保存

Doc = {} # <word, (doc1,...,docN)>

IDF = {} # <word, IDF>

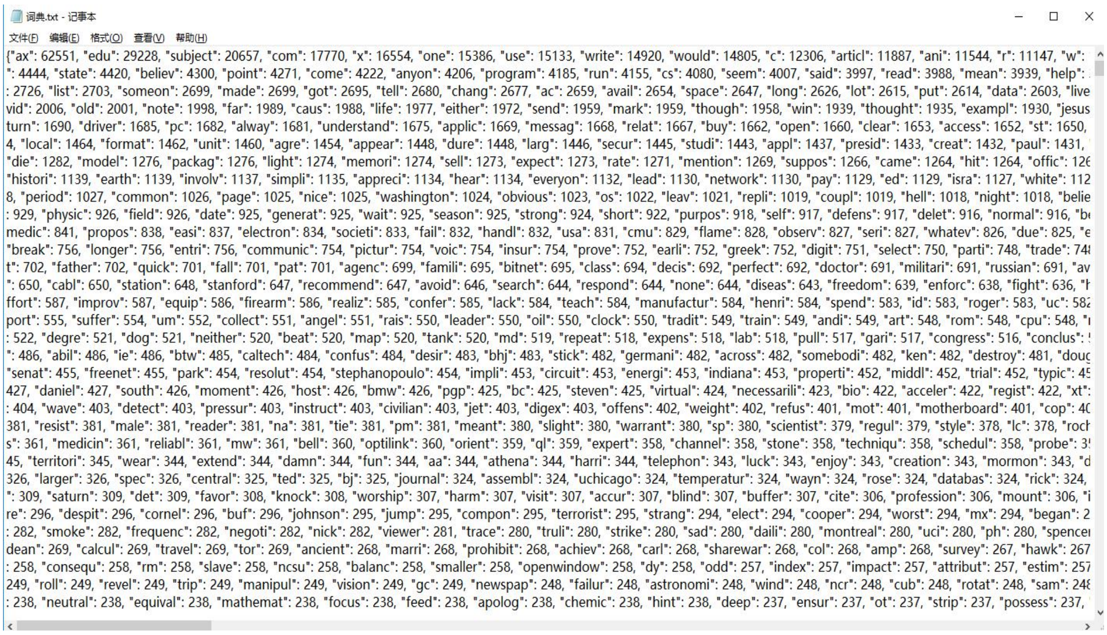
使用 CreateVSM(path)方法建立空间向量表示，每个训练数据都对应一个向量，向量形式以<category, document, (word1, para1), (word2, para2),...> 存入文件。

KNN 实现

将测试集中每个文档的{{word, TF}}与训练集<类_文件名，<word, TFIDF>>进行比较，通过余弦函数找到最相近的 20 个文档，通过 k 个文档所属类别计算出测试文档可能属于的类，返回与测试文档向量距离和最小的类。

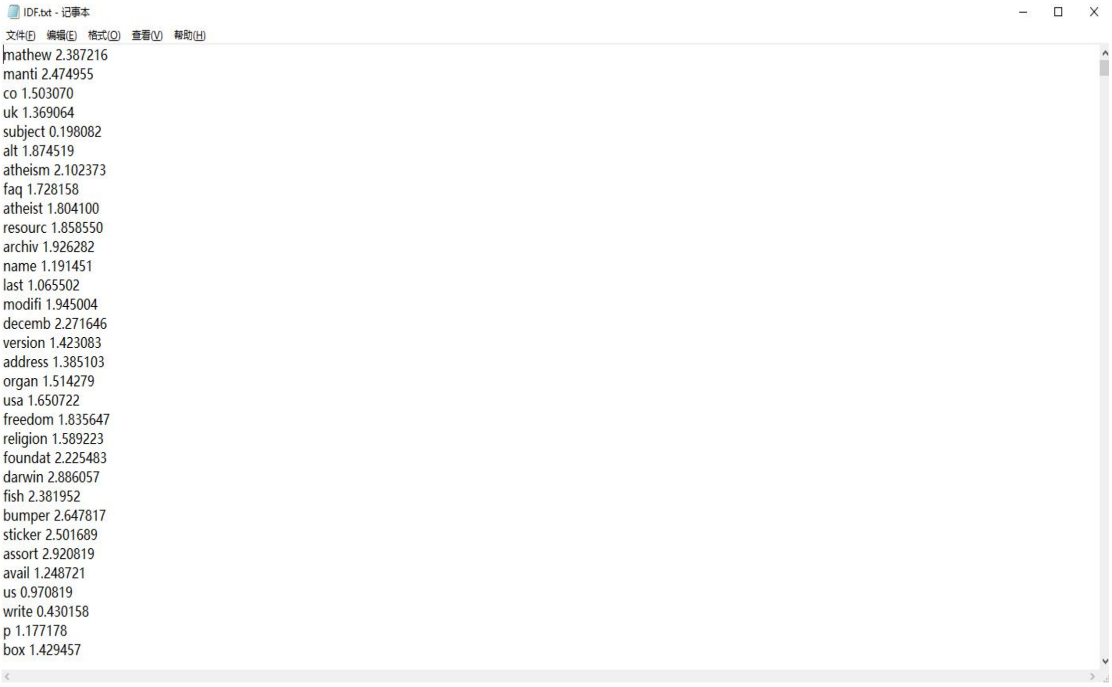
最终准确率在 0.78 左右

词频统计



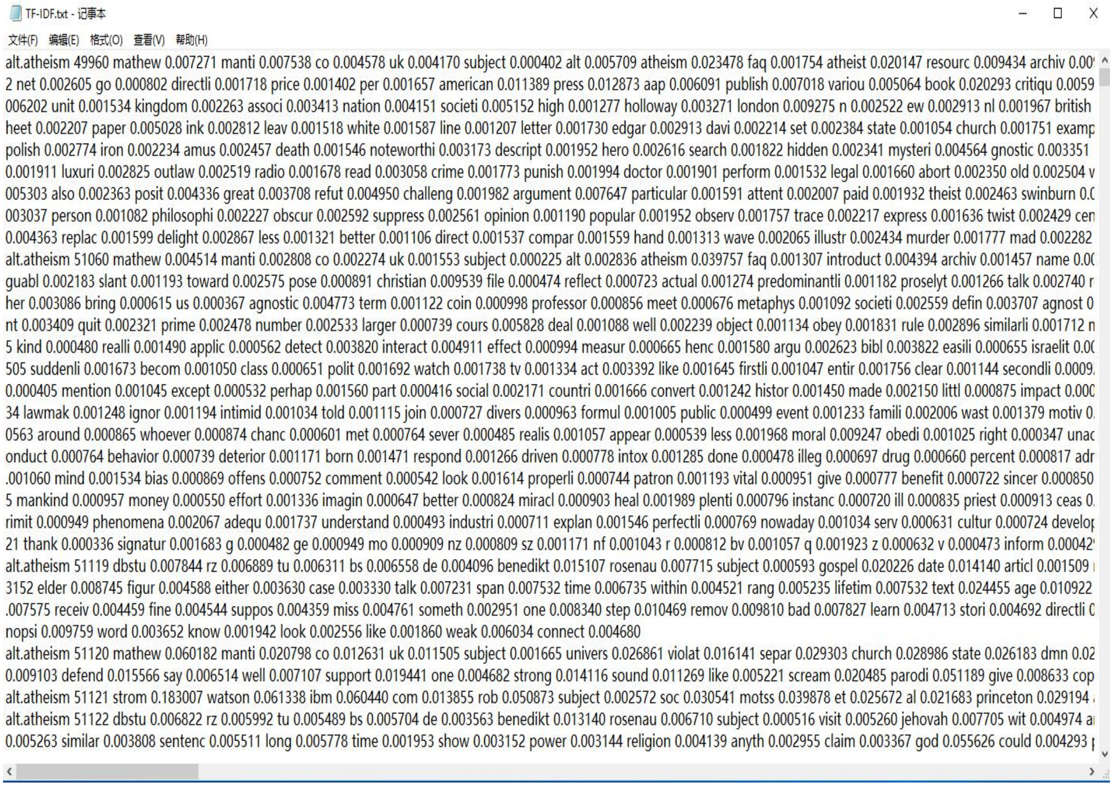
```
{"ax": 62551, "edu": 29228, "subject": 20657, "com": 17770, "x": 16554, "one": 15386, "use": 15133, "write": 14920, "would": 14805, "c": 12306, "articl": 11887, "ani": 11544, "r": 11147, "w":
": 4444, "state": 4420, "believ": 4300, "point": 4271, "come": 4222, "anyon": 4206, "program": 4185, "run": 4155, "cs": 4080, "seem": 4007, "said": 3997, "read": 3988, "mean": 3939, "help":
: 2726, "list": 2703, "someon": 2699, "made": 2699, "got": 2695, "tell": 2680, "chang": 2677, "ac": 2659, "avail": 2654, "space": 2647, "long": 2626, "lot": 2615, "put": 2614, "data": 2603, "live
vid": 2006, "old": 2001, "note": 1998, "far": 1989, "caus": 1988, "life": 1977, "either": 1972, "send": 1959, "mark": 1959, "though": 1958, "win": 1939, "thought": 1935, "exampl": 1930, "jesus
turn": 1690, "driver": 1685, "pc": 1682, "alway": 1681, "understand": 1675, "applic": 1669, "messag": 1668, "relat": 1667, "buy": 1662, "open": 1660, "clear": 1653, "access": 1652, "st": 1650,
4, "local": 1464, "format": 1462, "unit": 1460, "agre": 1454, "appear": 1448, "dure": 1448, "larg": 1446, "secur": 1445, "studi": 1443, "appl": 1437, "presid": 1433, "creat": 1432, "paul": 1431,
"die": 1282, "model": 1276, "packag": 1276, "light": 1274, "memori": 1274, "sell": 1273, "expect": 1273, "rate": 1271, "mention": 1269, "suppos": 1266, "came": 1264, "hit": 1264, "offic": 12
"histori": 1139, "earth": 1139, "involv": 1137, "simpli": 1135, "appreci": 1134, "hear": 1134, "everyon": 1132, "lead": 1130, "network": 1130, "pay": 1129, "ed": 1129, "isra": 1127, "white": 112
8, "period": 1027, "common": 1026, "page": 1025, "nice": 1025, "washington": 1024, "obvious": 1023, "os": 1022, "leav": 1021, "repli": 1019, "coupl": 1019, "hell": 1018, "night": 1018, "belie
: 929, "physic": 926, "field": 926, "date": 925, "generat": 925, "wait": 925, "season": 925, "strong": 924, "short": 922, "purpos": 918, "self": 917, "defens": 917, "delet": 916, "normal": 916, "b
medic": 841, "propos": 838, "easi": 837, "electron": 834, "societi": 833, "fail": 832, "handl": 832, "usa": 831, "cmu": 829, "flame": 828, "observ": 827, "seri": 827, "whatev": 826, "due": 825, "
"break": 756, "longer": 756, "entri": 756, "communic": 754, "pictur": 754, "voic": 754, "insur": 754, "prove": 752, "earli": 752, "greek": 752, "digit": 751, "select": 750, "parti": 748, "trade": 74
t": 702, "father": 702, "quick": 701, "fall": 701, "pat": 701, "agenc": 699, "famili": 695, "bitnet": 695, "class": 694, "decis": 692, "perfect": 692, "doctor": 691, "militari": 691, "russian": 691, "av
": 650, "cabl": 650, "station": 648, "stanford": 647, "recommend": 647, "avoid": 646, "search": 644, "respond": 644, "none": 644, "diseas": 643, "freedom": 639, "enforc": 638, "fight": 636, "
ffort": 587, "improv": 587, "equip": 586, "firearm": 586, "realiz": 585, "confer": 585, "lack": 584, "teach": 584, "manufactur": 584, "henri": 584, "spend": 583, "id": 583, "roger": 583, "uc": 58
port": 555, "suffer": 554, "um": 552, "collect": 551, "angel": 551, "rais": 550, "leader": 550, "oil": 550, "clock": 550, "tradit": 549, "train": 549, "andi": 549, "art": 548, "rom": 548, "cpu": 548, "
: 522, "degre": 521, "dog": 521, "neither": 520, "beat": 520, "map": 520, "tank": 520, "md": 519, "repeat": 518, "expens": 518, "lab": 518, "pull": 517, "gari": 517, "congress": 516, "conclus":
": 486, "abil": 486, "ie": 486, "btw": 485, "caltech": 484, "confus": 484, "desir": 483, "bhj": 483, "stick": 482, "germani": 482, "across": 482, "somebodi": 482, "ken": 482, "destroy": 481, "doug
"senat": 455, "freenet": 455, "park": 454, "resolut": 454, "stephanopoulo": 454, "impli": 453, "circuit": 453, "energi": 453, "indiana": 453, "properti": 452, "middl": 452, "trial": 452, "typic": 4
427, "daniel": 427, "south": 426, "moment": 426, "host": 426, "bmw": 426, "pgp": 425, "bc": 425, "steven": 425, "virtual": 424, "necessarili": 423, "bio": 422, "acceler": 422, "regist": 422, "xt":
: 404, "wave": 403, "detect": 403, "pressur": 403, "instruct": 403, "civilian": 403, "jet": 403, "digex": 403, "offens": 402, "weight": 402, "refus": 401, "mot": 401, "motherboard": 401, "cop": 4
381, "resist": 381, "male": 381, "reader": 381, "na": 381, "tie": 381, "pm": 381, "meant": 380, "slight": 380, "warrant": 380, "sp": 380, "scientist": 379, "regul": 379, "style": 378, "lc": 378, "roc
s": 361, "medicin": 361, "reliabl": 361, "mw": 361, "bell": 360, "optilink": 360, "orient": 359, "ql": 359, "expert": 358, "channel": 358, "stone": 358, "techniqu": 358, "schedul": 358, "probe": 3
45, "territori": 345, "wear": 344, "extend": 344, "damn": 344, "fun": 344, "aa": 344, "athena": 344, "harri": 344, "telephon": 343, "luck": 343, "enjoy": 343, "creation": 343, "mormon": 343, "d
326, "larger": 326, "spec": 326, "central": 325, "ted": 325, "bj": 325, "journal": 324, "assembl": 324, "uchicago": 324, "temperatur": 324, "wayn": 324, "rose": 324, "databas": 324, "rick": 324,
": 309, "saturn": 309, "det": 309, "favor": 308, "knock": 308, "worship": 307, "harm": 307, "visit": 307, "accur": 307, "blind": 307, "buffer": 307, "cite": 306, "profession": 306, "mount": 306, "i
re": 296, "despit": 296, "cornel": 296, "buf": 296, "johnson": 295, "jump": 295, "compon": 295, "terrorist": 295, "strang": 294, "elect": 294, "cooper": 294, "worst": 294, "mx": 294, "began": 2
: 282, "smoke": 282, "frequenc": 282, "negoti": 282, "nick": 282, "viewer": 281, "trace": 280, "truli": 280, "strike": 280, "sad": 280, "daili": 280, "montreal": 280, "uci": 280, "ph": 280, "spence
dean": 269, "calcul": 269, "travel": 269, "tor": 269, "ancient": 268, "marri": 268, "prohibit": 268, "achiev": 268, "carl": 268, "sharewar": 268, "col": 268, "amp": 268, "survey": 267, "hawk": 267
: 258, "consequ": 258, "rm": 258, "slave": 258, "ncsu": 258, "balanc": 258, "smaller": 258, "openwindow": 258, "dy": 258, "odd": 257, "index": 257, "impact": 257, "attribut": 257, "estim": 257
249, "roll": 249, "revel": 249, "trip": 249, "manipul": 249, "vision": 249, "gc": 249, "newspap": 248, "failur": 248, "astronomi": 248, "wind": 248, "ncr": 248, "cub": 248, "rotat": 248, "sam": 24
: 238, "neutral": 238, "equival": 238, "mathemat": 238, "focus": 238, "feed": 238, "apolog": 238, "chemic": 238, "hint": 238, "deep": 237, "ensur": 237, "ot": 237, "strip": 237, "possess": 237,
```

IDF



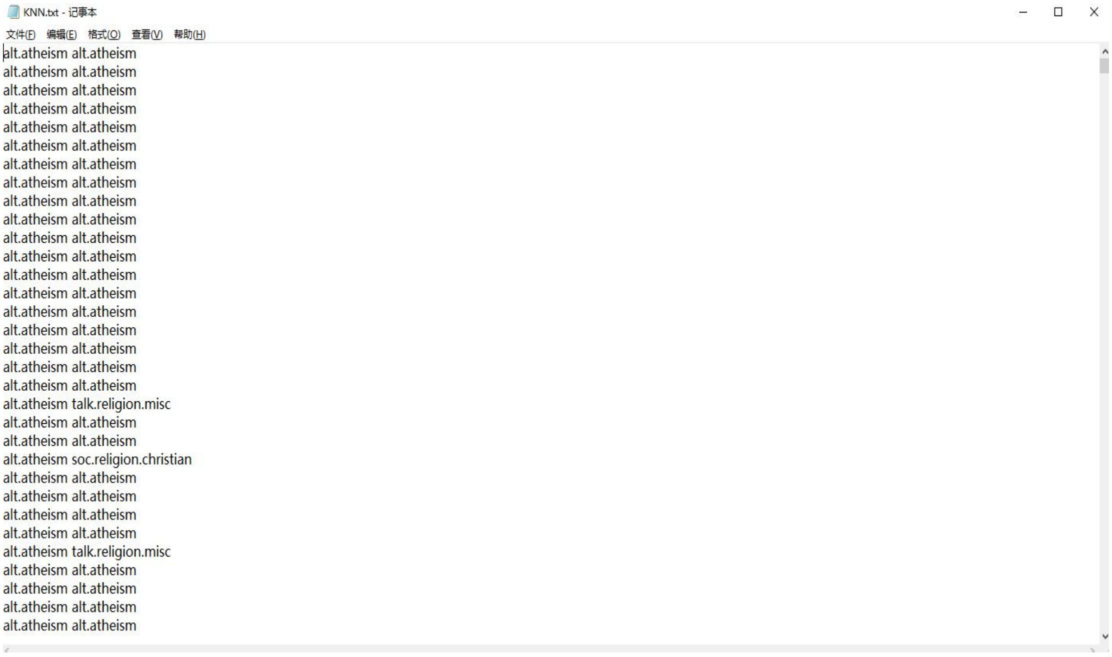
```
mathew 2.387216
manti 2.474955
co 1.503070
uk 1.369064
subject 0.198082
alt 1.874519
atheism 2.102373
faq 1.728158
atheist 1.804100
resourc 1.858550
archiv 1.926282
name 1.191451
last 1.065502
modifi 1.945004
decemb 2.271646
version 1.423083
address 1.385103
organ 1.514279
usa 1.650722
freedom 1.835647
religion 1.589223
foundat 2.225483
darwin 2.886057
fish 2.381952
bumper 2.647817
sticker 2.501689
assort 2.920819
avail 1.248721
us 0.970819
write 0.430158
p 1.177178
box 1.429457
```

TF-IDF



```
alt.atheism 49960 mathew 0.007271 manti 0.007538 co 0.004578 uk 0.004170 subject 0.000402 alt 0.005709 atheism 0.023478 faq 0.001754 atheist 0.020147 resourc 0.009434 archiv 0.00
2 net 0.002605 go 0.000802 directli 0.001718 price 0.001402 per 0.001657 american 0.011389 press 0.012873 aap 0.006091 publish 0.007018 variou 0.005064 book 0.020293 critiqu 0.0059
006202 unit 0.001534 kingdom 0.002263 associ 0.003413 nation 0.004151 societi 0.005152 high 0.001277 holloway 0.003271 london 0.009275 n 0.002522 ew 0.002913 nl 0.001967 british
heet 0.002207 paper 0.005028 ink 0.002812 leav 0.001518 white 0.001587 line 0.001207 letter 0.001730 edgar 0.002913 davi 0.002214 set 0.002384 state 0.001054 church 0.001751 examp
polish 0.002774 iron 0.002234 amus 0.002457 death 0.001546 noteworthi 0.003173 descript 0.001952 hero 0.002616 search 0.001822 hidden 0.002341 mysteri 0.004564 gnostic 0.003351
0.001911 luxuri 0.002825 outlaw 0.002519 radio 0.001678 read 0.003058 crime 0.001773 punish 0.001994 doctor 0.001901 perform 0.001532 legal 0.001660 abort 0.002350 old 0.002504 v
005303 also 0.002363 posit 0.004336 great 0.003708 refut 0.004950 challeng 0.001982 argument 0.007647 particular 0.001591 attent 0.002007 paid 0.001932 theist 0.002463 swinburn 0.0
003037 person 0.001082 philosophi 0.002227 obscur 0.002592 suppress 0.002561 opinion 0.001190 popular 0.001952 observ 0.001757 trace 0.002217 express 0.001636 twist 0.002429 cen
0.004363 replac 0.001599 delight 0.002867 less 0.001321 better 0.001106 direct 0.001537 compar 0.001559 hand 0.001313 wave 0.002065 illustr 0.002434 murder 0.001777 mad 0.002282
alt.atheism 51060 mathew 0.004514 manti 0.002808 co 0.002274 uk 0.001553 subject 0.000225 alt 0.002836 atheism 0.039757 faq 0.001307 introduct 0.004394 archiv 0.001457 name 0.0
guabl 0.002183 slant 0.001193 toward 0.002575 pose 0.000891 christian 0.009539 file 0.000474 reflect 0.000723 actual 0.001274 predominantli 0.001182 proselyt 0.001266 talk 0.002740 r
her 0.003086 bring 0.000615 us 0.000367 agnostic 0.004773 term 0.001122 coin 0.000998 professor 0.000856 meet 0.000676 metaphys 0.001092 societi 0.002559 defin 0.003707 agnost 0
nt 0.003409 quit 0.002321 prime 0.002478 number 0.002533 larger 0.000739 cours 0.005828 deal 0.001088 well 0.002239 object 0.001134 obey 0.001831 rule 0.002896 similarli 0.001712 n
5 kind 0.000480 realli 0.001490 applic 0.000562 detect 0.003820 interact 0.004911 effect 0.000994 measur 0.000665 henc 0.001580 argu 0.002623 bibl 0.003822 easili 0.000655 israelit 0.0
505 suddenli 0.001673 becom 0.001050 class 0.000651 polit 0.001692 watch 0.001738 tv 0.001334 act 0.003392 like 0.001645 firstli 0.001047 entir 0.001756 clear 0.001144 secondli 0.0009
0.000405 mention 0.001045 except 0.000532 perhap 0.001560 part 0.000416 social 0.002171 countri 0.001666 convert 0.001242 histor 0.001450 made 0.002150 littl 0.000875 impact 0.000
34 lawmak 0.001248 ignor 0.001194 intimid 0.001034 told 0.001115 join 0.000727 divers 0.000963 formul 0.001005 public 0.000499 event 0.001233 famili 0.002006 wast 0.001379 motiv 0.
0563 around 0.000865 whoever 0.000874 chanc 0.000601 met 0.000764 sever 0.000485 realis 0.001057 appear 0.000539 less 0.001968 moral 0.009247 obedi 0.001025 right 0.000347 unac
onduct 0.000764 behavior 0.000739 deterior 0.001171 born 0.001471 respond 0.001266 driven 0.000778 intox 0.001285 done 0.000478 illeg 0.000697 drug 0.000660 percent 0.000817 adr
.001060 mind 0.001534 bias 0.000869 offens 0.000752 comment 0.000542 look 0.001614 properli 0.000744 patron 0.001193 vital 0.000951 give 0.000777 benefit 0.000722 sincer 0.000850
5 mankind 0.000957 money 0.000550 effort 0.001336 imagin 0.000647 better 0.000824 miracl 0.000903 heal 0.001989 plenti 0.000796 instanc 0.000720 ill 0.000835 priest 0.000913 ceas 0.
rimit 0.000949 phenomena 0.002067 adequ 0.001737 understand 0.000493 industri 0.000711 explan 0.001546 perfectli 0.000769 nowaday 0.001034 serv 0.000631 cultur 0.000724 develop
21 thank 0.000336 signatur 0.001683 g 0.000482 ge 0.000949 mo 0.000909 nz 0.000809 sz 0.001171 nf 0.001043 r 0.000812 bv 0.001057 q 0.001923 z 0.000632 v 0.000473 inform 0.0004
alt.atheism 51119 dbstu 0.007844 rz 0.006889 tu 0.006311 bs 0.006558 de 0.004096 benedikt 0.015107 rosenau 0.007715 subject 0.000593 gospel 0.020226 date 0.014140 articl 0.001509
3152 elder 0.008745 figur 0.004588 either 0.003630 case 0.003330 talk 0.007231 span 0.007532 time 0.006735 within 0.004521 rang 0.005235 lifetim 0.007532 text 0.024455 age 0.010922
.007575 receiv 0.004459 fine 0.004544 suppos 0.004359 miss 0.004761 someth 0.002951 one 0.008340 step 0.010469 remov 0.009810 bad 0.007827 learn 0.004713 stori 0.004692 directli 0
nopsi 0.009759 word 0.003652 know 0.001942 look 0.002556 like 0.001860 weak 0.006034 connect 0.004680
alt.atheism 51120 mathew 0.060182 manti 0.020798 co 0.012631 uk 0.011505 subject 0.001665 univers 0.026861 violat 0.016141 separ 0.029303 church 0.028986 state 0.026183 dmn 0.02
0.009103 defend 0.015566 say 0.006514 well 0.007107 support 0.019441 one 0.004682 strong 0.014116 sound 0.011269 like 0.005221 scream 0.020485 parodi 0.051189 give 0.008633 cop
alt.atheism 51121 strom 0.183007 watson 0.061338 ibm 0.060440 com 0.013855 rob 0.050873 subject 0.002572 soc 0.030541 motss 0.039878 et 0.025672 al 0.021683 princeton 0.029194
alt.atheism 51122 dbstu 0.006822 rz 0.005992 tu 0.005489 bs 0.005704 de 0.003563 benedikt 0.013140 rosenau 0.006710 subject 0.000516 visit 0.005260 jehovah 0.007705 wit 0.004974 a
0.005263 similar 0.003808 sentenc 0.005511 long 0.005778 time 0.001953 show 0.003152 power 0.003144 religion 0.004139 anyth 0.002955 claim 0.003367 god 0.055626 could 0.004293
```

Knn



```
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism talk.religion.misc
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism soc.religion.christian
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism talk.religion.misc
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
alt.atheism alt.atheism
```