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：201600301304	姓名：贾乘兴	班级：人工智能 16

实验题目：VSM to BM25

实验内容：

1. 对于一个检索问题，我们要检索与 query 相关的 text 时，存在一个 text 与 query 的 relevance，布尔检索是检索所有的 relevance 为 1 的结果。

2. 在之前的布尔检索中，我们所得到的最终结果是较为严格的匹配，同时，匹配的结果并没有严格的 ranking 次序，所以在本次实验中，我们将通过之前建立的 vsm 模型，对检索过程进行得分与 ranking。

3. 使用 vsm 建立文本的词向量，我们有 pivoted 和 bm25 两种计算公式

pivoted 计算如下：

$$f(q,d)=\sum_{w\in q\cap d} c(w,q)\frac{\ln\left(1+\ln\left(1+c(w,d)\right)\right)}{1-b+b\frac{|d|}{avdl}}\log\left(\frac{M+1}{df(w)}\right)$$

其中 b 为超参数，范围 0 到 1，avdl 是平均文档长度，df（w）为 w 的文档出现次数，c（w，q）为 w 在 q 中的分量，c（w，d）为 w 在 d 中的分量，M 为总的文档数，d 的模为文档 d 的长度

bm25 计算如下：

$$f(q,d)=\sum_{w\in q\cap d} c(w,q)\frac{(k+1)c(w,d)}{c(w,d)+k\left(1-b+b\frac{|d|}{avdl}\right)}\log\left(\frac{M+1}{df(w)}\right)$$

k 和 b 为超参数，b 用于惩罚较大的文本

4. 代码部分：

```
import json
import nltk
import re
import math

# 去除停用词
def cutstopwords(str):
    stopwords = {}.fromkeys([line.rstrip() for line in
open('estopwords.txt')])
    segs = str.replace('\n','').lower().split(' ')
    new_str = ''
    for seg in segs:
        if seg not in stopwords:
            new_str = new_str + " " +seg
    return new_str

# 去除标点
def cutsyms(str):
    new_str = re.sub('[,.\'\"\t\n*_+=?/|!@#$%^&*()`~<>:;\-\[\]]',"
",str)
    return new_str

# 词干提取
def stemming(str):
    s = nltk.stem.SnowballStemmer('english')
    segs = str.replace('\n', '').lower().split(' ')
    new_str = ''
    for seg in segs:
        new_str = new_str + " " + s.stem(seg)
    return new_str

k=5
b=0.5

path = 'tweets.txt'
file = open(path,'r',encoding='UTF-8',errors='ignore')
tweets = []
twords = {}
tdocnum = {}
tdocword = [0]*40000
counts = 0
for line in file:
    tweets.append(json.loads(line))
    tweets[counts]['text'] = tweets[counts]['text'].lower()
    tweets[counts]['text'] = cutsyms(tweets[counts]['text'])
    tweets[counts]['text'] = cutstopwords(tweets[counts]['text'])
    tweets[counts]['text'] = stemming(tweets[counts]['text'])
    if counts not in tdocnum.keys():
        tdocnum[counts] = {}
    for seg in tweets[counts]['text'].split(' '):
        tdocword[counts] = tdocword[counts] + 1
        if seg in tdocnum[counts].keys():
            length = len(twords[seg])
            tdocnum[counts][seg] = tdocnum[counts][seg] + 1
            if twords[seg][length-1]!=counts:
                twords[seg].append(counts)
                twords[seg][0] = twords[seg][0] + 1
        else:
            tdocnum[counts].update({seg:1})
            twords[seg] = []
            twords[seg].append(1)
            twords[seg].append(counts+1)
    counts = counts + 1
    print(counts)
file.close()

print(tweets[1])

sum = 0
avg = 0
for i in range(counts):
    sum = sum + tdocword[i]
avg = sum / counts

def qTest():
    # query
    path = '/Users/apple/Desktop/ir/hw3/topics.MB171-225.txt'
    file = open(path, 'r', encoding='UTF-8', errors='ignore')
    txt = file.read()
    file.close()
    txt.replace('\n', ' ')
    txtlist = txt.split(' ')
    numq=0;
    for i in range(len(txtlist)):
        if txtlist[i] == '</num>\n<query>':
            i = i + 1
            string = ""
            while txtlist[i] != '</query>\n<querytime>':
                string = string + txtlist[i] + " "
                i = i + 1

            numq = numq + 1
            string = string.lower()
            string = cutsyms(string)
            string = cutstopwords(string)
            string = stemming(string)
            print(string)

            [st1,st2]=rank(string,1000)

            for j in range(1000):
                outstr = str(numq + 170) + " " +
tweets[st1[j][0]]['tweetId']
                full_path = 'result1000bm25.txt'
                file = open(full_path, 'a+')
                file.write(outstr + "\n")
                file.close()

                outstr = str(numq + 170) + " " +
tweets[st2[j][0]]['tweetId']
                full_path = 'result1000pivo.txt'
                file = open(full_path, 'a+')
                file.write(outstr + "\n")
                file.close()

def rank(string,top):
    bm25 = {}
    pivoted = {}
    slist = string.lower().split(" ")
    for seg in slist:
        if seg != "":
            for w in range(counts):
                if w not in bm25.keys():
                    bm25[w] = 0
                    pivoted[w] = 0
                p = 0
                q = 0
                if seg in tdocnum[w].keys():
                    p = (math.log((counts + 1) / twords[seg][0])) * (k +
1) * (tdocnum[w][seg] / tdocword[w]) / ((tdocnum[w][seg] / tdocword[w])
+ k * (1 - b + b * tdocword[w] / avg))
                    q = (math.log((counts + 1) / twords[seg][0])) *
math.log(1 + math.log(1 + tdocnum[w][seg] / tdocword[w])) / (1 - b + b
* tdocword[w] / avg)
                bm25[w] = bm25[w] + p
                pivoted[w] = pivoted[w] + q

    sk1 = sorted(bm25.items(), key=lambda x: x[1], reverse=True)
    sk2 = sorted(pivoted.items(), key=lambda x: x[1], reverse=True)
    return sk1[0:top],sk2[0:top]

qTest()
```

5. 结果展示，map and ndcg

pivoted

query: 171 , AP: 0.9356617482020168

```
query: 172 ,AP: 0.28902352829720757
query: 173 ,AP: 0.48000337345679356
query: 174 ,AP: 0.5996708502522456
query: 175 ,AP: 0.38910505836575876
query: 176 ,AP: 0.7475601576484384
query: 177 ,AP: 0.3360503533818221
query: 178 ,AP: 0.4136904528448895
query: 179 ,AP: 0.4646854421828267
query: 180 ,AP: 0.17271157167530224
query: 181 ,AP: 0.9970760233918128
query: 182 ,AP: 0.19305019305019305
query: 183 ,AP: 0.425531914893617
query: 184 ,AP: 0.4861486064978063
query: 185 ,AP: 0.7584382689245911
query: 186 ,AP: 0.8146380434151332
query: 187 ,AP: 0.8543878308197999
query: 188 ,AP: 0.13129132788009404
query: 189 ,AP: 0.547520390537836
query: 190 ,AP: 0.5895706722607722
query: 191 ,AP: 0.4914901204502388
query: 192 ,AP: 0.6550790330718559
query: 193 ,AP: 0.4696833749185543
query: 194 ,AP: 0.8310246290545203
query: 195 ,AP: 0.24021521306458668
query: 196 ,AP: 0.6309794413689155
query: 197 ,AP: 0.7571194198289967
query: 198 ,AP: 0.472091130864241
query: 199 ,AP: 0.2375296912114014
query: 200 ,AP: 0.37652119412547275
query: 201 ,AP: 0.37037037037037035
query: 202 ,AP: 0.582944008433889
query: 203 ,AP: 0.07430297930183909
query: 204 ,AP: 0.8881906101318976
query: 205 ,AP: 0.48412481880693775
query: 206 ,AP: 0.5641746897699274
query: 207 ,AP: 0.5943919698393946
query: 208 ,AP: 0.303951367781155
query: 209 ,AP: 0.15766065819463979
query: 210 ,AP: 0.1899385830339053
query: 211 ,AP: 0.459578433578922
query: 212 ,AP: 0.551295687562442
query: 213 ,AP: 0.3124039411943941
query: 214 ,AP: 0.6740875669907214
query: 215 ,AP: 0.28332899055377503
```

```
query: 216 ,AP: 0.4961609400307482
query: 217 ,AP: 0.4569769872897383
query: 218 ,AP: 0.28969673597034107
query: 219 ,AP: 0.27143687617318446
query: 220 ,AP: 0.3257965265513425
query: 221 ,AP: 0.1988071570576541
query: 222 ,AP: 0.3155567232271073
query: 223 ,AP: 0.4482010315764862
query: 224 ,AP: 0.9400000000000001
query: 225 ,AP: 0.7027037606903075
MAP = 0.48588419036452474
query 171 , NDCG:  0.9224769955415072
query 172 , NDCG:  0.8644683690928974
query 173 , NDCG:  0.47691490372310086
query 174 , NDCG:  0.7189677902846582
query 175 , NDCG:  0.6720771828538592
query 176 , NDCG:  0.6865390003692599
query 177 , NDCG:  0.550244774174705
query 178 , NDCG:  0.7519136883699405
query 179 , NDCG:  0.619970351107371
query 180 , NDCG:  0.43425279753582346
query 181 , NDCG:  0.8177464231587351
query 182 , NDCG:  0.48699918945433257
query 183 , NDCG:  0.977700879065691
query 184 , NDCG:  0.6540467453713154
query 185 , NDCG:  0.7564575529860473
query 186 , NDCG:  0.7525754770493287
query 187 , NDCG:  0.739194759752893
query 188 , NDCG:  0.2916446495181594
query 189 , NDCG:  0.399878953319392
query 190 , NDCG:  0.631346130896955
query 191 , NDCG:  0.6690241200992628
query 192 , NDCG:  0.7644738038432504
query 193 , NDCG:  0.5031634881456607
query 194 , NDCG:  0.8187330464032292
query 195 , NDCG:  0.5836277080866716
query 196 , NDCG:  0.7428157406175602
query 197 , NDCG:  0.8284877508800681
query 198 , NDCG:  0.5309306822780093
query 199 , NDCG:  0.8957345912663868
query 200 , NDCG:  0.6567125424130342
query 201 , NDCG:  0.7290501920901167
query 202 , NDCG:  0.7423995604311945
query 203 , NDCG:  0.17295968083763788
query 204 , NDCG:  0.9026463715104112
query 205 , NDCG:  0.8677145329193771
query 206 , NDCG:  0.6421298087274195
query 207 , NDCG:  0.660025690039897 4
query 208 , NDCG:  0.44866567946447833
query 209 , NDCG:  0.5837835885802172
query 210 , NDCG:  0.4883026857892305
query 211 , NDCG:  0.3796586525712978
query 212 , NDCG:  0.8128884823345468
query 213 , NDCG:  0.8406430585905934
query 214 , NDCG:  0.7783222740002366
query 215 , NDCG:  0.5159411204447049
query 216 , NDCG:  0.7528623182385635
query 217 , NDCG:  0.5789046552708403
query 218 , NDCG:  0.670153622484432
query 219 , NDCG:  0.3947052015863278
query 220 , NDCG:  0.38086672183526843
query 221 , NDCG:  0.6190203154051522
query 222 , NDCG:  0.40844270452949544
query 223 , NDCG:  0.4459743921546515
query 224 , NDCG:  0.9755435176818925
query 225 , NDCG:  0.5769574364929893
NDCG = 0.6466851336667289
Bm25
query: 171 ,AP: 0.9354639633918901
query: 172 ,AP: 0.28902352829720757
query: 173 ,AP: 0.47835502180844186
query: 174 ,AP: 0.5931893687707641
query: 175 ,AP: 0.38910505836575876
query: 176 ,AP: 0.7461624148676459
query: 177 ,AP: 0.33443766816480086
query: 178 ,AP: 0.4137784407966024
query: 179 ,AP: 0.4299416577205357
query: 180 ,AP: 0.17271157167530224
query: 181 ,AP: 0.9970760233918128
query: 182 ,AP: 0.19305019305019305
query: 183 ,AP: 0.425531914893617
query: 184 ,AP: 0.48018755234279964
query: 185 ,AP: 0.7548668403531625
query: 186 ,AP: 0.8146380434151332
query: 187 ,AP: 0.8229564407647003
query: 188 ,AP: 0.13070845170454543
query: 189 ,AP: 0.541568009585455
query: 190 ,AP: 0.5814051993440673
```

```
query: 191 ,AP: 0.4879487310409398
query: 192 ,AP: 0.6521358077148374
query: 193 ,AP: 0.4696833749185543
query: 194 ,AP: 0.8310246290545203
query: 195 ,AP: 0.23977525495535548
query: 196 ,AP: 0.62646257050671
query: 197 ,AP: 0.7469943413326292
query: 198 ,AP: 0.472091130864241
query: 199 ,AP: 0.2375296912114014
query: 200 ,AP: 0.37614013976672267
query: 201 ,AP: 0.37037037037037035
query: 202 ,AP: 0.5810065553429172
query: 203 ,AP: 0.07504174998741828
query: 204 ,AP: 0.8881906101318976
query: 205 ,AP: 0.48081915008265425
query: 206 ,AP: 0.5641746897699274
query: 207 ,AP: 0.5911903975948958
query: 208 ,AP: 0.303951367781155
query: 209 ,AP: 0.15748083656309378
query: 210 ,AP: 0.18748527502504175
query: 211 ,AP: 0.4525354605876438
query: 212 ,AP: 0.5506556274219794
query: 213 ,AP: 0.311215785832774
query: 214 ,AP: 0.6740875669907214
query: 215 ,AP: 0.28256237638846277
query: 216 ,AP: 0.4950833779301934
query: 217 ,AP: 0.45710379888394115
query: 218 ,AP: 0.28969673597034107
query: 219 ,AP: 0.2585361931441935
query: 220 ,AP: 0.3257965265513425
query: 221 ,AP: 0.1988071570576541
query: 222 ,AP: 0.3209691331820747
query: 223 ,AP: 0.43974893998391223
query: 224 ,AP: 0.9400000000000001
query: 225 ,AP: 0.6967050385923697
MAP = 0.4828574137315877
query 171 , NDCG:  0.9224142202898286
query 172 , NDCG:  0.8644011491077039
query 173 , NDCG:  0.47611191029208455
query 174 , NDCG:  0.7164363470771604
query 175 , NDCG:  0.6711055031420622
query 176 , NDCG:  0.6862692281563332
query 177 , NDCG:  0.5542935024376002
query 178 , NDCG:  0.746828297337751
query 179 , NDCG:  0.5485953607836835
query 180 , NDCG:  0.42841542789163045
query 181 , NDCG:  0.8177464231587351
query 182 , NDCG:  0.48700982401341825
query 183 , NDCG:  0.9775163252922493
query 184 , NDCG:  0.643849508735571
query 185 , NDCG:  0.7564575529860473
query 186 , NDCG:  0.7525754770493287
query 187 , NDCG:  0.7304691915956766
query 188 , NDCG:  0.29116830859977105
query 189 , NDCG:  0.39436148347305805
query 190 , NDCG:  0.6166660209719917
query 191 , NDCG:  0.6652588088089105
query 192 , NDCG:  0.762769544037964
query 193 , NDCG:  0.502900270136119
query 194 , NDCG:  0.8187330464032292
query 195 , NDCG:  0.5827484390552313
query 196 , NDCG:  0.7414111155909637
query 197 , NDCG:  0.8203056383545662
query 198 , NDCG:  0.5309306822780093
query 199 , NDCG:  0.8959380756834057
query 200 , NDCG:  0.6564423655177877
query 201 , NDCG:  0.7278194463398717
query 202 , NDCG:  0.7419593933787226
query 203 , NDCG:  0.17406760331172222
query 204 , NDCG:  0.9026463715104112
query 205 , NDCG:  0.856411035473557
query 206 , NDCG:  0.6209823968342716
query 207 , NDCG:  0.6592021839373902
query 208 , NDCG:  0.44870812480593786
query 209 , NDCG:  0.5805221790961462
query 210 , NDCG:  0.48741835705033165
query 211 , NDCG:  0.3796586525712978
query 212 , NDCG:  0.8126720581809838
query 213 , NDCG:  0.8392975238828411
query 214 , NDCG:  0.7780658570864358
query 215 , NDCG:  0.5154950856305499
query 216 , NDCG:  0.747041905433647
query 217 , NDCG:  0.5772204047094308
query 218 , NDCG:  0.670153622484432
query 219 , NDCG:  0.39001571427974346
query 220 , NDCG:  0.38086672183526843
query 221 , NDCG:  0.6181957277375177
query 222 , NDCG:  0.40979103523504845
query 223 , NDCG:  0.44109883055961846
query 224 , NDCG:  0.9755435176818925
query 225 , NDCG:  0.5749278152485202
NDCG = 0.6430892838646084

对于每个query，展示top5结果
ron weasley birthday
([(17011, 16.382447406471105), (16938, 13.179496131538759),
(11128, 10.92163160431407), (17023, 10.915177394116876), (10314,
9.236911787838896)], [(17011, 10.68243012349938), (16938,
8.573381926387846), (11128, 7.121620082332919), (17023,
7.092316850936076), (10314, 5.998949099721806)])
 merging air american
([(12917, 3.638392464705625), (12270, 3.2508537786195233),
(12450, 3.2508537786195233), (11526, 2.9348689847975264), (10562,
2.6651858101015464)], [(12917, 2.364105616978692), (12270,
2.113421443093096), (12450, 2.113421443093096), (11526,
1.9088106999228374), (10562, 1.73418107868624)])
 muscle pain statins
([(4872, 1.812574586532745), (21399, 1.812574586532745), (20179,
1.6254268893097616), (15954, 1.3325929050507732), (21453,
1.3325929050507732)], [(4872, 1.1779112598614037), (21399,
1.1779112598614037), (20179, 1.056710721546548), (15954,
0.86709053934312), (21453, 0.86709053934312)])
 hubble oldest star
([(19063, 5.460815802157035), (19258, 3.99777871515232), (19411,
3.99777871515232), (19762, 3.268895724362861), (19388,
3.0813987695953324)], [(19063, 3.5608100411664596), (19258,
2.60127161802936), (19411, 2.60127161802936), (19762,
2.100812753029735), (19388, 2.007715632430522)])
 commentary naming storm nemo
([(11058, 7.032578463221057), (11099, 7.032578463221057), (10588,
5.951282746979509), (10647, 5.951282746979509), (10662,
5.951282746979509)], [(11058, 4.569534047803563), (11099,
4.569534047803563), (10588, 3.8650842507975813), (10647,
3.8650842507975813), (10662, 3.8650842507975813)])
 book club members
([(11003, 4.620976871486273), (14064, 4.620976871486273), (19002,
4.620976871486273), (21307, 4.620976871486273), (10749,
3.5573293287789873)], [(11003, 3.0011595956007175), (14064,
3.0011595956007175), (19002, 3.0011595956007175), (21307,
3.0011595956007175), (10749, 2.28496895240748)])
 boko haram kidnapped french tourists
([(18729, 5.298884056444082), (23491, 5.298884056444082), (15964,
4.461905753939547), (16026, 4.461905753939547), (14895,
4.4023034771962894)], [(18729, 3.4410218663254373), (23491,
3.4410218663254373), (15964, 2.9058940942695886), (16026,
2.9058940942695886), (14895, 2.8632160498842563)])
 tiger woods regains title
([(19778, 9.76287115453756), (19786, 8.786330754359172), (23437,
7.27678492941125), (9334, 6.157941191892597), (19834,
6.157941191892597)], [(19778, 6.224715117329302), (19786,
5.71558795092523), (23437, 4.728211233957384), (9334,
3.999299399814537), (19834, 3.999299399814537)])
 care iditarod dogs
([(22277, 4.876280667929285), (19148, 3.2508537786195233),
(19165, 3.2508537786195233), (19894, 2.8479031859859534), (22069,
2.649442028222041)], [(22277, 3.1701321646396443), (19148,
2.113421443093096), (19165, 2.113421443093096), (19894,
1.8564001957290088), (22069, 1.7205109331627186)])
 sherlock elementary bbc cbs
([(8130, 15.171379474964327), (9536, 15.171379474964327), (11049,
15.171379474964327), (12623, 15.171379474964327), (13669,
15.171379474964327)], [(8130, 10.437009743251545), (9536,
10.437009743251545), (11049, 10.437009743251545), (12623,
10.437009743251545), (13669, 10.437009743251545)])
 costa concordia shipwreck
([(18695, 3.2508537786195233), (16337, 2.6651858101015464),
(15776, 2.2309528769697735), (16028, 2.2309528769697735), (17115,
2.2309528769697735)], [(18695, 2.113421443093096), (16337,
1.73418107868624), (15776, 1.4529470471347943), (16028,
1.4529470471347943), (17115, 1.4529470471347943)])
 chinua achebe death
([(22314, 14.150655601531675), (22366, 14.150655601531675),
(22731, 14.150655601531675), (22998, 14.150655601531675), (22261,
10.92163160431407)], [(22314, 9.273681608526097), (22366,
9.273681608526097), (22731, 9.273681608526097), (22998,
9.273681608526097), (22261, 7.121620082332919)])
 evernote hacked
([(17872, 5.460815802157035), (17362, 5.298884056444082), (17432,
5.298884056444082), (18095, 3.62514917306549), (18118,
3.62514917306549)], [(17872, 3.5608100411664596), (17362,
3.4410218663254373), (17432, 3.4410218663254373), (18095,
2.355822519722807), (18118, 2.355822519722807)])
 election hugo chavez successor
([(18565, 4.620976871486273), (18778, 4.393165377179586), (20045,
3.99777871515232), (23530, 3.74567896117469), (18570,
3.62514917306549)], [(18565, 3.0011595956007175), (18778,
2.857793975462615), (20045, 2.60127161802936), (23530,
2.4077973507133548), (18570, 2.3558225197228073)])
 national zoo panda  insemination
([(23987, 4.461905753939547), (23956, 3.3464293154546603),
(23963, 3.0813987695953324), (23960, 2.8479031859859534), (112,
2.649442028222041)], [(23987, 2.905894094269588), (23956,
2.1794205707021916), (23963, 2.007715632430522), (23960,
1.8564001957290088), (112, 1.7205109331627186)])
 dorner  truck  compensation
([(11221, 1.8624619608947692), (21123, 1.760646709003587),
(21180, 1.760646709003587), (21194, 1.4287078065629173), (21127,
1.1846668769189337)], [(11221, 1.2023363521043098), (21123,
1.1481721196799375), (21180, 1.1481721196799375), (21194,
0.9328599540233485), (21127, 0.77439674567645)])
 pope washed muslims feet
([(23752, 6.157941191892597), (23787, 3.62514917306549), (23724,
2.9348689847975264), (23791, 2.9348689847975264), (23757,
2.6651858101015464)], [(23752, 3.999299399814537), (23787,
2.3558225197228073), (23724, 1.9088106999228374), (23791,
1.9088106999228374), (23757, 1.73418107868624)])
 bombing police headquarters  kirkuk
([(9117, 3.62514917306549), (8488, 2.649442028222041), (11491,
2.649442028222041), (8386, 2.4328269470356636), (8339,
2.054265846396888)], [(9117, 2.3558225197228073), (8488,
1.7205109331627186), (11491, 1.7205109331627186), (8386,
1.5837046325024189), (8339, 1.3384770882870145)])
 injuries pets
([(9808, 1.368941271771643), (1440, 1.3325929050507732), (8152,
1.0406064067872873), (25877, 1.0239769358577462), (12941,
0.9581924490975584)], [(9808, 0.8903463018823515), (1440,
0.86709053934312), (8152, 0.6777131070669506), (25877,
0.6577438100718219), (12941, 0.624319701141017)])
 organized crime  sports doping  australia
([(10523, 4.620976871486273), (10457, 3.2508537786195233),
(10526, 2.9348689847975264), (10417, 2.4328269470356636), (10506,
2.4328269470356636)], [(10523, 3.0011595956007175), (10457,
2.113421443093096), (10526, 1.9088106999228374), (10417,
1.5837046325024189), (10506, 1.5837046325024189)])
 irish laundries apology
([(14457, 3.99777871515232), (14442, 3.3464293154546603), (9883,
3.0789705959462985), (9857, 2.9348689847975264), (9975,
2.9348689847975264)], [(14457, 2.60127161802936), (14442,
2.1794205707021916), (9883, 1.9996496999072686), (9857,
1.9088106999228374), (9975, 1.9088106999228374)])
```

```
 whooping cough epidemic
([(16883, 3.62514917306549), (19850, 3.62514917306549), (10410,
3.2508537786195233), (10590, 3.2508537786195233), (15988,
3.2508537786195233)], [(16883, 2.3558225197228073), (19850,
2.3558225197228073), (10410, 2.113421443093096), (10590,
2.113421443093096), (15988, 2.113421443093096)])
 bulgarian protesters immolate
([(14695, 3.0789705959462985), (19700, 2.9348689847975264),
(14627, 2.3409158552611555), (6325, 1.812574586532745), (14683,
1.812574586532745)], [(14695, 1.9996496999072686), (19700,
1.9088106999228374), (14627, 1.4931166361562707), (6325,
1.1779112598614037), (14683, 1.1779112598614037)])
 cherry blossom washington
([(3151, 2.3104884357431366), (20634, 2.2309528769697735),
(21923, 2.054265846396888), (18853, 1.760646709003587), (9406,
1.527678340781292)], [(3151, 1.5005797978003588), (20634,
1.4529470471347943), (21923, 1.3384770882870145), (18853,
1.1481721196799375), (9406, 0.9970806987338444)])
 argo wins oscar
([(15617, 8.786330754359172), (15570, 6.61076382466176), (15581,
6.157941191892597), (15594, 6.157941191892597), (15536,
5.460815802157035)], [(15617, 5.71558795092523), (15570,
4.240145378744836), (15581, 3.999299399814537), (15594,
3.999299399814537), (15536, 3.5608100411664596)])
 fines google street view
([(20566, 4.4023034771962894), (20647, 4.4023034771962894),
(20674, 3.99777871515232), (20703, 3.99777871515232), (20736,
3.99777871515232)], [(20566, 2.8632160498842563), (20647,
2.8632160498842563), (20674, 2.60127161802936), (20703,
2.60127161802936), (20736, 2.60127161802936)])
 mad men season 6
([(18060, 6.931465307229409), (20330, 6.501707557239047), (20216,
5.869737969595053), (20805, 5.869737969595053), (20097,
5.330371620203093)], [(18060, 4.501739393401076), (20330,
4.226842886186192), (20216, 3.8176213998456747), (20805,
3.8176213998456747), (20097, 3.46836215737248)])
 hostess bought apollo
([(18334, 2.3104884357431366), (11772, 2.037088279409492),
(10987, 1.812574586532745), (20577, 1.6377296992916117), (20681,
1.4674344923987632)], [(18334, 1.5005797978003588), (11772,
1.3233578664143462), (10987, 1.1779112598614037), (20577,
1.0684670279521713), (20681, 0.9544053499614187)])
 ed koch death
([(8167, 13.675764336742828), (8356, 13.675764336742828), (8960,
13.675764336742828), (8250, 8.491464435632631), (8418,
8.491464435632631)], [(8167, 8.962459958283594), (8356,
8.962459958283594), (8960, 8.962459958283594), (8250,
5.523774732692926), (8418, 5.523774732692926)])
 uk passes marriage bill
([(10231, 10.915177394116876), (9911, 4.620976871486273), (9972,
4.620976871486273), (10006, 4.620976871486273), (13945,
4.393165377179586)], [(10231, 7.092316850936076), (9911,
3.0011595956007175), (9972, 3.0011595956007175), (10006,
3.0011595956007175), (13945, 2.857793975462615)])
 higgs boson discovery
([(21048, 7.27678492941125), (21596, 7.27678492941125), (20962,
6.111264838228477), (21273, 6.111264838228477), (21072,
4.620976871486273)], [(21048, 4.728211233957384), (21596,
4.728211233957384), (20962, 3.9700735992430385), (21273,
3.9700735992430385), (21072, 3.0011595956007175)])
 boko haram  amnesty opposition
([(18729, 7.948326084666123), (23491, 7.948326084666123), (18685,
6.111264838228477), (20360, 5.437723759598235), (21366,
4.876280667929285)], [(18729, 5.161532799488156), (23491,
5.161532799488156), (18685, 3.9700735992430385), (20360,
3.533733779584211), (21366, 3.1701321646396443)])
 eastern australia floods
([(8492, 6.111264838228477), (485, 3.62514917306549), (659,
2.9348689847975264), (88, 2.6651858101015464), (450,
2.6651858101015464)], [(8492, 3.9700735992430385), (485,
2.3558225197228073), (659, 1.9088106999228374), (88,
1.73418107868624), (450, 1.73418107868624)])
 sotomayor  prosecutor  racial comments
([(15913, 3.3464293154546603), (16063, 3.233287261871984),
(16053, 2.9348689847975264), (16052, 2.8479031859859534), (15968,
2.6651858101015464)], [(15913, 2.1794205707021916), (16063,
2.0894182185844077), (16053, 1.9088106999228374), (16052,
1.8564001957290088), (15968, 1.73418107868624)])
 port football riot  death sentences
([(19443, 6.0820673675891594), (19501, 5.869737969595053),
(19417, 5.330371620203093), (19447, 4.876280667929285), (19450,
4.461905753939547)], [(19443, 3.9592615812560474), (19501,
3.8176213998456747), (19417, 3.46836215737248), (19447,
3.1701321646396443), (19450, 2.9058940942695886)])
 yarn bombing
([(11491, 5.298884056444082), (19966, 3.62514917306549), (21120,
3.62514917306549), (21138, 3.0789705959462985), (21620,
3.0789705959462985)], [(11491, 3.4410218663254373), (19966,
2.3558225197228073), (21120, 2.3558225197228073), (21138,
1.9996496999072686), (21620, 1.9996496999072686)])
 david cameron  apology  amritsar
([(14745, 4.074176558818984), (14660, 3.62514917306549), (14623,
3.0813987695953324), (14629, 3.0813987695953324), (14657,
3.0813987695953324)], [(14745, 2.6467157328286923), (14660,
2.3558225197228073), (14623, 2.007715632430522), (14629,
2.007715632430522), (14657, 2.007715632430522)])
 olympics drops wrestling
([(11717, 7.032578463221057), (11834, 5.121055281411086), (12107,
5.121055281411086), (11760, 4.767183103403362), (11813,
4.767183103403362)], [(11717, 4.569534047803563), (11834,
3.3255423244384534), (12107, 3.3255423244384534), (11760,
3.0992064485170845), (11813, 3.0992064485170845)])
 chelyabinsk meteor damage
([(13945, 4.098299058453045), (15324, 3.9374485574123788),
(12901, 3.1417562140935997), (13293, 3.1417562140935997), (13294,
3.1417562140935997)], [(13945, 2.665980757230311), (15324,
2.5578928388729163), (12901, 2.0424957269705364), (13293,
2.0424957269705364), (13294, 2.0424957269705364)])
 arrest craig wilson drive shooting d c
([(8918, 5.094289501738195), (21306, 4.682788858288293), (6141,
4.098299058453045), (4061, 3.638392464705625), (19992,
3.220497738135868)], [(8918, 3.321810855299859), (21306,
3.052466880145864), (6141, 2.665980757230311), (4061,
2.364105616978692), (19992, 2.101076660135735)])
 downton abbey  lady mary  beau
([(17739, 7.27678492941125), (14486, 3.62514917306549), (14628,
3.62514917306549), (17239, 3.62514917306549), (14214,
2.9348689847975264)], [(17739, 4.728211233957384), (14486,
2.3558225197228073), (14628, 2.3558225197228073), (17239,
2.3558225197228073), (14214, 1.9088106999228374)])
 kate middleton maternity wear
([(15753, 4.074176558818984), (15892, 4.074176558818984), (21544,
4.074176558818984), (14549, 3.6492404205534954), (16721,
3.6492404205534954)], [(15753, 2.6467157328286923), (15892,
2.6467157328286923), (21544, 2.6467157328286923), (14549,
2.375556948753628), (16721, 2.375556948753628)])
 embassy ankara bombed
([(887, 4.393165377179586), (8379, 3.62514917306549), (8233,
3.2508537786195233), (8307, 3.2508537786195233), (8729,
3.2508537786195233)], [(887, 2.857793975462615), (8379,
2.3558225197228073), (8233, 2.113421443093096), (8307,
2.113421443093096), (8729, 2.113421443093096)])
```

```
 math common core
([(23280, 4.620976871486273), (14951, 3.8878689645077653),
(12038, 3.62514917306549), (21756, 3.62514917306549), (23211,
3.3464293154546603)], [(23280, 3.0011595956007175), (14951,
2.496050989026187), (12038, 2.355822519722807З), (21756,
2.3558225197228073), (23211, 2.1794205707021916)])
 snow blower problems
([(10848, 8.680247355816668), (11121, 7.27678492941125), (10851,
6.157941191892597), (10868, 6.157941191892597), (11164,
6.157941191892597)], [(10848, 5.576465635337361), (11121,
4.728211233957384), (10851, 3.999299399814537), (10868,
3.999299399814537), (11164, 3.999299399814537)])
 type ii diabetes research
([(21086, 4.30381025656917), (17607, 3.1243596636900826), (8677,
2.7841867342976574), (19197, 2.6678131204093436), (20725,
2.581883349360876)], [(21086, 2.799157001805189), (17607,
2.0085680241201214), (8677, 1.814866749659858), (19197,
1.7169738513718924), (20725, 1.6837259075556958)])
 pope candidates
([(23752, 3.0789705959462985), (11504, 2.9348689847975264),
(13901, 2.6651858101015464), (11493, 2.2309528769697735), (20524,
2.2309528769697735)], [(23752, 1.9996496999072686), (11504,
1.9088106999228374), (13901, 1.73418107868624), (11493,
1.4529470471347943), (20524, 1.4529470471347943)])
 sinkhole rescues
([(18304, 2.649442028222041), (18269, 1.9883455117731628),
(16909, 1.812574586532745), (17009, 1.812574586532745), (17206,
1.812574586532745)], [(18304, 1.7205109331627186), (18269,
1.2697822084890888), (16909, 1.1779112598614037), (17009,
1.1779112598614037), (17206, 1.1779112598614037)])
 russian meteorite conspiracy
([(13888, 4.876280667929285), (13908, 4.876280667929285), (13785,
2.4328269470356636), (13425, 2.3409158552611555), (13934,
2.3104884357431366)], [(13888, 3.1701321646396443), (13908,
3.1701321646396443), (13785, 1.5837046325024189), (13934,
1.5005797978003588), (13989, 1.4956210481007666)])
 shahbag protest
([(14742, 2.054265846396888), (14874, 2.054265846396888), (6045,
1.812574586532745), (14829, 1.812574586532745), (10776,
1.760646709003587)], [(14742, 1.3384770882870145), (14874,
1.3384770882870145), (6045, 1.1779112598614037), (14829,
1.1779112598614037), (10776, 1.1481721196799375)])
 hiv baby cured
([(17975, 13.179496131538759), (18126, 10.915177394116876),
(17729, 9.236911787838896), (17951, 9.236911787838896), (18227,
9.236911787838896)], [(17975, 8.573381926387846), (18126,
7.092316850936076), (17729, 5.998949099721806), (17951,
5.998949099721806), (18227, 5.998949099721806)])
  oz  great powerful  opens
([(18652, 3.99777871515232), (14029, 3.6492404205534954), (20686,
3.6492404205534954), (19827, 3.638392464705625), (20273,
3.638392464705625)], [(18652, 2.60127161802936), (14029,
2.375556948753628), (20686, 2.375556948753628), (19827,
2.364105616978692), (20273, 2.364105616978692)])
 dog leash
([(14238, 4.465898434638259), (9217, 3.3941859985154315), (8813,
3.1417562140935997), (8928, 3.1417562140935997), (22069,
2.471613253189045)], [(14238, 2.900441684267073), (9217,
2.205428430824871), (8813, 2.0424957269705364), (8928,
2.0424957269705364), (22069, 1.6050313912757348)])
 dark pool trading
([(23784, 4.4023034771962894), (21704, 2.054265846396888),
(16190, 1.9806713870750008), (21721, 1.4992038761402333), (9270,
1.4674344923987632)], [(23784, 2.8632160498842563), (21704,
1.3384770882870145), (16190, 1.2779506074643607), (21721,
0.9810218772848289), (9270, 0.9544053499614187)])
 barbara walters  chicken pox
([(8937, 6.501707557239047), (18181, 5.298884056444082), (16164,
3.521293418007174), (18286, 3.2508537786195233), (10334,
2.6651858101015464)], [(8937, 4.226842886186192), (18181,
3.4410218663254373), (16164, 2.296344239359875), (18286,
2.113421443093096), (10334, 1.73418107868624)])
```

结论分析与体会：通过本次实验，实现了简单的有 ranking 的检索，同时将之前的知识进行整合与回顾，对知识体系结构有了更好的了解