

QINMIN HU

2018年5月29日

LAB 5 检索模型

一. 实验目的

~ 在unigram假设下进行检索

二. 实验步骤

1. 用户提取query、倒排索引字典、获取全部的文档名和文档长度

```
def getDocLength(dirname,filename):
    print('getdocument')
    if filename != '.DS_Store':
        xmldoc = ET.parse(dirname + '/' + filename)
        for doc in xmldoc.findall('DOC'):
            doc_list=[]
            for s1 in doc.findall('TEXT'):

                if s1.text==None:
                    doc_list.append("")
                else:
                    for word in s1.text.split():
                        while '_' in word or '"' in word or '.' in
word or '-' in word or '!' in word or ';' in word or ',' in word or '$' in word or '/' in word or
'/' in word or '{' in word or '}' in word or '*' in word or '#' in word or '^' in word or '|' in
word or '~' in word or '=' in word or '\' in word or '+' in word or ':' in word or '?' in word:
                            word = word.replace("_",
""").replace("'", """).replace(".", """).replace("-", """).replace("!", """).replace("; ",
""").replace(", ", """).replace("$", """).replace("//", """).replace('/',
""").replace('{', """).replace('}', """).replace('*', """).replace('^', """).replace("| ",
""").replace('~', """).replace('=', """).replace("\", """).replace('+', """).replace(':', """).replace('?',
""")

                            if re.findall(r'\d+', word) or word ==

""":

                            continue
                        doc_list.append(word)
                    for s2 in doc.findall('DOCNO'):
```

```

        docno=s2.text.strip()
        doc_filename[docno]=len(doc_list)

def loadDocument(dirname):
    print('loaddocument')
    for parent,dirnames,filenames in os.walk(dirname):
        for e in filenames:
            print(e)
            getDocLength(dirname,e)

def loadDic():
    print('loadaddic')
    f_dic=open("inverted_index.json",encoding='utf-8')
    word_list=json.load(f_dic)
    for word in word_list:
        word_dic[word[0]]=word[1]

def loadQuery():
    print('loadquery')
    xmldoc = ET.parse('topics_new.xml')
    for doc in xmldoc.findall('top'):
        test=""
        for s1 in doc.findall('num'):
            query_id.append(s1.text.replace("Number:", "").strip())
        for s2 in doc.findall('desc'):
            test=test+' '+s2.text.replace("Descript :", "").strip()
        for s3 in doc.findall('narr'):
            test=test+' '+s3.text.replace("Narr :", "").strip()
        for s4 in doc.findall('title'):
            test=test+' '+s4.text.replace("Topic :", "").strip()
        while '_' in test or '"' in test or '.' in test or '-' in test or '!' in test or ';' in test or ',' in test or '$' in test or '/' in test or '/' in test or '{' in test or '}' in test or '*' in test or '#' in test or '^' in test or '|' in test or '~' in test or '=' in test or '\' in test or '+' in test or ':' in test or '?' in test:
            test = test.replace("_", "").replace('"', "").replace(".", "").replace('-', "").replace('!', "").replace("; ", "").replace(", ", "").replace("$", "").replace("//", "").replace('/', "").replace('{', "").replace('}', "").replace('*', "").replace('^', "").replace("| ", "").replace('~', "").replace('=', "").replace('\', "").replace('+', "").replace(':', "").replace('?', "")
            if re.findall(r'\d+', test) or test == "":
                continue
        query[s1.text.replace("Number:", "").strip()]=test.split()

```

2. 根据查询在unigram假设下进行检索，打分

$$P(D|Q) \propto P(D) \prod_{t \in Q} [\alpha P_{MLE}(t|M_D) + (1-\alpha) P_{MLE}(t|M_C)]$$

$$P_{MLE}(t|M_D) = (tf(t,D))/L_D$$

$$P_{MLE}(t|M_C) = (cf_t)/T$$

平滑方式：Add-one、线性插值、Dirichlet...

10152130138_丁婉宁_unigram.py 中：

```
def getUnigram(title, words):
    docDic={}
    for doc, doc_len in doc_filename.items():
        p=1
        for word in words:
            if doc in word_dic[word]:
                pwd1=word_dic[word][doc]*1.0/doc_len
            else:
                pwd1=1
            wDoc=0
            for d, freq in word_dic[word].items():
                wDoc=wDoc+freq
            p=p*(0.9*pwd1+(1-0.9)*(wDoc/len(doc_filename)))
        docDic[doc]=p
    score[title]=docDic
```

10152130138_丁婉宁_unigram_dic.py 中：

```
def getUnigram(title, words, T):
    print('getunigram')
    docDic={}
    for doc, doc_len in doc_filename.items():
        p=1
        for word in words:
            if word in word_dic.keys():
                if doc in word_dic[word]:
                    pwd1=word_dic[word][doc]*1.0/doc_len
                else:
                    pwd1=0
            else:
                pwd1=0
            p=p*(0.9*pwd1+(1-0.9)*(word_freq[word]/T))
        docDic[doc]=p
    score[title]=docDic
```

评测:

test_10152130138_Unigram_0.5.txt:

test1.txt		
runid	all	10152130138_dingwanning_unigram
num_q	all	50
num_ret	all	25000
num_rel	all	9805
num_rel_ret	all	1641
map	all	0.0417
gm_map	all	0.0177
Rprec	all	0.1167
bpref	all	0.0996
recip_rank	all	0.2018
iprec at recall 0.00	all	0.3033
iprec at recall 0.10	all	0.1731
iprec at recall 0.20	all	0.1156
iprec at recall 0.30	all	0.0580
iprec at recall 0.40	all	0.0008
iprec at recall 0.50	all	0.0000
iprec at recall 0.60	all	0.0000
iprec at recall 0.70	all	0.0000
iprec at recall 0.80	all	0.0000
iprec at recall 0.90	all	0.0000
iprec at recall 1.00	all	0.0000
P_5	all	0.1080
P_10	all	0.1320
P_15	all	0.1507
P_20	all	0.1440
P_30	all	0.1533
P_100	all	0.1448
P_200	all	0.1195
P_500	all	0.0656
P_1000	all	0.0328

test_10152130138_Unigram_500.txt:

test1.txt		
runid	all	10152130138_dingwanning_unigram
num_q	all	50
num_ret	all	50000
num_rel	all	9805
num_rel_ret	all	1667
map	all	0.0202
gm_map	all	0.0070
Rprec	all	0.0664
bpref	all	0.0764
recip_rank	all	0.1060
iprec at recall 0.00	all	0.1813
iprec at recall 0.10	all	0.0866
iprec at recall 0.20	all	0.0571
iprec at recall 0.30	all	0.0173
iprec at recall 0.40	all	0.0044
iprec at recall 0.50	all	0.0005
iprec at recall 0.60	all	0.0000
iprec at recall 0.70	all	0.0000
iprec at recall 0.80	all	0.0000
iprec at recall 0.90	all	0.0000
iprec at recall 1.00	all	0.0000
P_5	all	0.0400
P_10	all	0.0340
P_15	all	0.0493
P_20	all	0.0510
P_30	all	0.0520
P_100	all	0.0790
P_200	all	0.0802
P_500	all	0.0549
P_1000	all	0.0333

test_10152130138_Unigram.txt

test1.txt		
runid	all	10152130138_dingwanning_unigram
num_q	all	50
num_ret	all	25000
num_rel	all	9805
num_rel_ret	all	1286
map	all	0.0252
gm_map	all	0.0082
Rprec	all	0.0869
bpref	all	0.0772
recip_rank	all	0.1388
iprec at recall 0.00	all	0.2462
iprec at recall 0.10	all	0.1110
iprec at recall 0.20	all	0.0617
iprec at recall 0.30	all	0.0115
iprec at recall 0.40	all	0.0009
iprec at recall 0.50	all	0.0000
iprec at recall 0.60	all	0.0000
iprec at recall 0.70	all	0.0000
iprec at recall 0.80	all	0.0000
iprec at recall 0.90	all	0.0000
iprec at recall 1.00	all	0.0000
P_5	all	0.0640
P_10	all	0.0740
P_15	all	0.0960
P_20	all	0.1020
P_30	all	0.1147
P_100	all	0.1224
P_200	all	0.0957
P_500	all	0.0514
P_1000	all	0.0257

三. 实验结果

10152130138_Unigram_0.5.res:

10152130138_Unigram_0.5.res			UNREGISTERED	
10152130138_Unigram_0.5.res			10152130138_Unigram_500.res	10152130138_Unigram.res
1	151	0 WSJ870428-0145 0 1.3008190732032897e-125	10152130138_dingwanning_unigram	
2	151	0 AP890406-0100 1 5.896700272629206e-126	10152130138_dingwanning_unigram	
3	151	0 AP890523-0138 2 3.553181629073012e-126	10152130138_dingwanning_unigram	
4	151	0 AP890125-0145 3 7.531226362386954e-127	10152130138_dingwanning_unigram	
5	151	0 AP890913-0046 4 1.9074778899955163e-127	10152130138_dingwanning_unigram	
6	151	0 AP891030-0018 5 8.144039778966308e-128	10152130138_dingwanning_unigram	
7	151	0 AP890105-0262 6 1.4636779391863454e-128	10152130138_dingwanning_unigram	
8	151	0 AP890408-0117 7 1.0173370520313563e-128	10152130138_dingwanning_unigram	
9	151	0 AP890408-0070 8 5.776391485679636e-129	10152130138_dingwanning_unigram	
10	151	0 AP880815-0076 9 4.268883393873505e-129	10152130138_dingwanning_unigram	
11	151	0 AP881004-0197 10 2.349420486325124e-129	10152130138_dingwanning_unigram	
12	151	0 AP890116-0061 11 2.2997460463828658e-129	10152130138_dingwanning_unigram	
13	151	0 AP890105-0166 12 1.5516801742835816e-129	10152130138_dingwanning_unigram	
14	151	0 AP890411-0092 13 1.43831556843307e-129	10152130138_dingwanning_unigram	
15	151	0 AP880608-0313 14 1.4145921081424862e-129	10152130138_dingwanning_unigram	
16	151	0 AP881005-0001 15 1.2120340590503986e-129	10152130138_dingwanning_unigram	
17	151	0 AP881108-0087 16 8.0572678265047e-130	10152130138_dingwanning_unigram	
18	151	0 AP890309-0047 17 4.245453886030364e-130	10152130138_dingwanning_unigram	
19	151	0 AP881002-0015 18 2.9974456071351656e-130	10152130138_dingwanning_unigram	
20	151	0 AP890912-0107 19 1.6730402843477924e-130	10152130138_dingwanning_unigram	

10152130138_Unigram_500.res:

Line	DocID	Freq	Score	Term
1	WSJ870428-0145	0	2.966294800934832e-130	10152130138_dingwanning_unigram
2	AP890406-0100	1	1.8993693485223924e-133	10152130138_dingwanning_unigram
3	AP891030-0018	2	4.477836188587675e-136	10152130138_dingwanning_unigram
4	AP890523-0138	3	3.2655917724449605e-136	10152130138_dingwanning_unigram
5	AP890913-0046	4	2.6520367703044315e-136	10152130138_dingwanning_unigram
6	AP890105-0262	5	3.381877664960443e-137	10152130138_dingwanning_unigram
7	AP880815-0076	6	1.746231922084771e-138	10152130138_dingwanning_unigram
8	AP890408-0117	7	1.2715904138688501e-138	10152130138_dingwanning_unigram
9	AP890125-0145	8	5.5990525653093035e-139	10152130138_dingwanning_unigram
10	AP880814-0095	9	2.4318603847752114e-139	10152130138_dingwanning_unigram
11	AP881004-0197	10	1.1885734940888844e-139	10152130138_dingwanning_unigram
12	AP890912-0107	11	8.06115651039704e-140	10152130138_dingwanning_unigram
13	AP881108-0087	12	5.770263941300024e-140	10152130138_dingwanning_unigram
14	WSJ870205-0131	13	3.2140311559944274e-140	10152130138_dingwanning_unigram
15	AP881005-0001	14	1.6794300171573255e-141	10152130138_dingwanning_unigram
16	AP890116-0061	15	1.322688797059853e-141	10152130138_dingwanning_unigram
17	AP890309-0047	16	1.2605834900396723e-141	10152130138_dingwanning_unigram
18	AP890502-0187	17	5.9002353905420734e-142	10152130138_dingwanning_unigram
19	AP880726-0195	18	4.8635372134695955e-142	10152130138_dingwanning_unigram
20	AP880608-0313	19	3.7480416733933035e-142	10152130138_dingwanning_unigram

10152130138_Unigram.res:

Line	DocID	Freq	Score	Term
1	WSJ870428-0145	0	2.966294800934832e-130	10152130138_dingwanning_unigram
2	AP890406-0100	1	1.8993693485223924e-133	10152130138_dingwanning_unigram
3	AP891030-0018	2	4.477836188587675e-136	10152130138_dingwanning_unigram
4	AP890523-0138	3	3.2655917724449605e-136	10152130138_dingwanning_unigram
5	AP890913-0046	4	2.6520367703044315e-136	10152130138_dingwanning_unigram
6	AP890105-0262	5	3.381877664960443e-137	10152130138_dingwanning_unigram
7	AP880815-0076	6	1.746231922084771e-138	10152130138_dingwanning_unigram
8	AP890408-0117	7	1.2715904138688501e-138	10152130138_dingwanning_unigram
9	AP890125-0145	8	5.5990525653093035e-139	10152130138_dingwanning_unigram
10	AP880814-0095	9	2.4318603847752114e-139	10152130138_dingwanning_unigram
11	AP881004-0197	10	1.1885734940888844e-139	10152130138_dingwanning_unigram
12	AP890912-0107	11	8.06115651039704e-140	10152130138_dingwanning_unigram
13	AP881108-0087	12	5.770263941300024e-140	10152130138_dingwanning_unigram
14	WSJ870205-0131	13	3.2140311559944274e-140	10152130138_dingwanning_unigram
15	AP881005-0001	14	1.6794300171573255e-141	10152130138_dingwanning_unigram
16	AP890116-0061	15	1.322688797059853e-141	10152130138_dingwanning_unigram
17	AP890309-0047	16	1.2605834900396723e-141	10152130138_dingwanning_unigram
18	AP890502-0187	17	5.9002353905420734e-142	10152130138_dingwanning_unigram
19	AP880726-0195	18	4.8635372134695955e-142	10152130138_dingwanning_unigram
20	AP880608-0313	19	3.7480416733933035e-142	10152130138_dingwanning_unigram