Assignment 4

Fall 2016
CS834 Introduction to Information Retrieval
Dr. Michael Nelson

Mathew Chaney

December 3, 2016

Contents

List of Figures

List of Tables

1 Question 8.3

1.1 Question

For one query in the CACM collection (provided at the book website), generate a ranking using Galago, and then calculate average precision, NDCG at 5 and 10, precision at 10, and the reciprocal rank by hand.

1.2 Approach

Galago version 3.10 was first downloaded from the Project Lemur Source Forge website, which can be found at the following URL: https://sourceforge.net/projects/lemur/files/lemur/galago-3.10/. The CACM document corpus was downloaded from the textbook's website, found here: http://www.search-engines-book.com/collections/. Galago was used to create an index of the CACM corpus and to run as a server to respond to queries on that index.

The getrel.py script (found in Listing 2) was created to issue queries to the Galago search server using the Python Requests library [1]. The HTML responses were then parsed using the Python Beautiful Soup library [2], where the CACM document identifiers were extracted for use in calculating the different evaluation scores for the Galago ranking.

The query used was from the CACM query set, number 10, and only the first 1000 retrieved documents were considered when calculating all scores for this experiment.

1.2.1 Initial Precision and Recall Calculations

Precision and Recall were calculated with the following equations:

$$Recall = \frac{\mid A \cap B \mid}{\mid A \mid}$$

$$Precision = \frac{\mid A \cap B \mid}{\mid B \mid}$$

In these equations, A is the relevant set of documents for the query, and B is the set of retrieved documents.

1.2.2 Calculating Precision at Specific Rankings

A list of precision values was created by calculating the cumulative precision at each document ranking with the set of retrieved documents up to that ranking.

1.2.3 Calculating Average Precision

Average precision was calculated by adding the precision at each retrieval ranking position for documents which are part of $A \cap B$, or the set of retrieved documents that are relevant, and then dividing by the size of that set to obtain the average. This can also be described as the area under the precision-recall curve, which can be expressed as the following summation:

$$AveP = \sum_{k=1}^{n} P(k)\Delta r(k)$$

where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k-1$ to k.

1.2.4 Calculating Normalized Discounted Cumulative Gain (NDCG)

First, discounted cumulative gain at rank p (DCG_p) was calculated with the following formula:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}$$

The ideal discounted cumulative gain at rank p ($IDCG_p$) is a simple series, expressed as:

$$IDCG_p = 1 + \sum_{i=2}^{p} \frac{1}{log_2 i}$$

Finally, normalized discounted cumulative gain at rank p ($NDCG_p$) is expressed as:

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

with rel_i being the relevancy for document i in the retrieval ranking. For this experiment, this value is either 0 or 1.

1.2.5 Calculating Reciprocal Rank

Reciprocal rank is defined as the reciprocal of the rank at which the first relevant document is found, so if the 3^{rd} document in the retrieval ranking list is the first relevant document, the reciprocal rank is $\frac{1}{3}$.

1.3 Results

After building the index, CACM query 10 was processed by the `getrel.py` script, the output of which can be found in Listing 1. This script calculates all the values shown in Table 1, which are all of the required values for the question.

```
[mchaney@mchaney-l getrel]$ python getrel.py -n 1000 -q 10
query 10
query: parallel languages  languages for parallel computation
precision: 0.027
recall: 0.771428571429
precision @10: 0.9
NDCG @5: 1.0
NDCG @10: 0.942709999032
avg precision: 0.697677898817
reciprocal rank: 1.0
```

Listing 1: Output from running the getrel.py script for queries 1 and 10 from the CACM collection.

Query #	Avg. Prec.	NDCG @5	NDCG @10	Prec. @10	Recip. Rank
10	0.697677898817	1.0	0.942709999032	0.9	1.0

Table 1: Calculations for CACM query 10 from top 1000 retrieved documents.

2 Question 8.4

2.1 Approach

Git er done

3 Appendix

3.1 Code listings

```
#!/usr/bin/python

import argparse
import re
import requests
import xmltodict
from math import log
from bs4 import BeautifulSoup
from pprint import pprint as pp


def parseargs():
    parser = argparse.ArgumentParser()
    parser.add_argument('-q', '--qnum', type=int, default=10, help='the query number to use')
    parser.add_argument('-n', type=int, default=10, help='the number of results to retrieve')
    return parser.parse_args()


args = parseargs()


def buildrel():
    rel = {}
    for line in open('cacm.rel').readlines():
        q, _, doc, _ = line.split()
        if q not in rel:
            rel[q] = []
        rel[q].append(int(doc.split('-')[1]))
    return rel


def buildqueries():
    with open('cacm.query.xml') as fd:
        return xmltodict.parse(fd.read())


REL = buildrel()
QUERIES = buildqueries()
RE = re.compile('/home/mchaney/workspace/edu/cs834-f16/assignments/assignment4/code/cacm/docs/CACM-([\d]+).html')
ID = {'id':'result'}
URL = 'http://0.0.0.0:{0}/search'
QUERY1 = 'what articles exist which deal with tss time sharing system an operating system for ibm computers'
PDICT = {'q': QUERY1, 'start': 0, 'n': args.n}

def query(qstr, port=54312):
    PDICT['q'] = qstr
    PDICT['n'] = args.n
    res = requests.get(URL.format(port), params=PDICT)
    if not res.ok:
        return None
    soup = BeautifulSoup(res.text, 'html.parser')
    return [int(RE.match(href.text).groups()[0]) for href in soup.select("#result a")]

# precision is the proportion of retrieved documents that are relevant
# recall is the proportion of relevant documents that are retrieved

def recall(rel, retr):
    relset = set(rel)
    retrset = set(retr)
    return float(len(relset.intersection(retrset))) / len(relset)


def precision(rel, retr):
    relset = set(rel)
    retrset = set(retr)
    return float(len(relset.intersection(retrset))) / len(retrset)

```

```
68
69 def run(rel, retr, func):
70     rr = []
71     for i in range(1, len(retr)+1):
72         rr.append(func(rel, retr[:i]))
73     return rr
74
75
76 def avg(rel, retr, func):
77     prun = run(rel, retr, precision)
78     res = []
79     for i in range(len(retr)):
80         if retr[i] in rel:
81             res.append(prun[i])
82     return float(sum(res))/len(res)
83
84
85 def getrel(rel, retr, i):
86     return 1 if retr[i] in rel else 0
87
88
89 def DCG(rel, retr, p):
90     sum = 0
91     for i in range(2, p+1):
92         sum += float(getrel(rel, retr, i-1)) / log(i, 2)
93     return getrel(rel, retr, 0) + sum
94
95 def IDCG(p):
96     sum = 0
97     for i in range(2, p+1):
98         sum += 1 / log(i, 2)
99     return 1 + sum
100
101
102 def NDCG(rel, retr, p):
103     dcg = DCG(rel, retr, p)
104     idcg = IDCG(p)
105     return dcg / idcg
106
107
108 def reciprank(rel, retr):
109     for i in range(1, len(retr)+1):
110         if retr[i-1] in rel:
111             return 1.0 / i
112     return 0.0
113
114
115 def getquery(qnum):
116     return QUERIES['parameters']['query'][qnum-1]['text']
117
118
119 def process(qnum):
120     qstr = getquery(qnum)
121     retr = query(qstr)
122     rel = REL[str(qnum)]
123     prun = run(rel, retr, precision)
124     prec = precision(rel, retr)
125     rec = recall(rel, retr)
126     avgprec = avg(rel, retr, precision)
127     ndcg5 = NDCG(rel, retr, 5)
128     ndcg10 = NDCG(rel, retr, 10)
129     recip = reciprank(rel, retr)
130     return qnum, qstr, retr, rel, prun, prec, rec, ndcg5, ndcg10, avgprec, recip
131
132
133 def printresults(qnum, qstr, retr, rel, prun, prec, rec, ndcg5, ndcg10, avgprec, recip):
134     print 'query {0}'.format(qnum)
135     print 'query: {0}'.format(qstr)
136     if args.n == 10:
137         print 'relevant: {0}'.format(rel)
138         print 'retrieved: {0}'.format(retr)
139         print 'p-run: {0}'.format(prun)
140     print 'precision: {0}'.format(prec)
141     print 'recall: {0}'.format(rec)
142     print 'precision @10: {0}'.format(prun[9])
143     print 'NDCG @5: {0}'.format(ndcg5)
144     print 'NDCG @10: {0}'.format(ndcg10)
```

```
    print 'avg precision: {0}'.format(avgprec)
    print 'reciprocal rank: {0}'.format(recip)


TABLE = """\\begin{{table}}[h!]
\\centering
\\begin{{tabular}}{{ | c | c | c | c | c | c | }}
\\hline
Query \# & Avg. Prec. & NDCG @5 & NDCG @10 & Prec. @10 & Recip. Rank \\\\
\\hline
{0} & {1} & {2} & {3} & {4} & {5} \\\\
\\hline
\\end{{tabular}}
\\caption{{Calculations for CACM query {6} from top {7} retrieved documents.}}
\\label{{tab:query20}}
\\end{{table}}
"""

def printtab(qnum, qstr, retr, rel, prun, prec, rec, ndcg5, ndcg10, avgprec, recip):
    fname = 'query{0}.tab'.format(qnum)
    with open(fname, 'w') as fd:
        fd.write(TABLE.format(qnum, avgprec, ndcg5, ndcg10, prun[9], recip, qnum, args.n))


results = process(args.qnum)
printresults(*results)
printtab(*results)
```

Listing 2: getrel.py

4 References

[1] Kenneth Reitz. Requests: HTTP for Humans. Available at http://docs.python-requests.org/en/master/. Accessed: 2016/09/20.

[2] Leonard Richardson. Beautiful Soup. Available at: https://www.crummy.com/software/beautifulsoup/. Accessed: 2016/09/20.