Web Information Management: Project #1

Due on April 23, 2015

Professor Fang TTh 12:10

Rick Sullivan

Problem 1

Test the performance of retrieval algorithm "RawTF" with two types of text data.

1. Results from RawTF without further processing are on page 5. Results were, as expected, unexceptional. Stemming and removig stopwords gave the results on page 5.

   Using stemming and removing stopwords improved our results across the board. We improve interpolated recall-precision averages, average precision, and R-precision.

   These first stemmed results are using the Porter stemming algorithm. For future reference, anytime I mention stemming in this report, assume I am using the Porter algorithm.
   Using the Krovetz stemming algorithm instead seems to further improve our results, which can be seen on page 6.

2. The previously mentioned results using Porter stemming also had their stopwords removed. To test the remaining cases around stopwords, I tested RawTF with removing stopwords, and Porter stemming without removing stopwords, which can be found on pages 7 and 5, respectively.

Figure 1: Results from RawTF and stemmed RawTF including or removing stopwords.

	RawTF		RawTF with Porter stemming	
	Avg precision	R-precision	Avg precision	R-precision
Stopwords included	0.0449	0.0712	0.0444	0.0546
Stopwords removed	0.1495	0.1732	0.1224	0.1372

Removing stopwords improves our query results in all tested cases.

Problem 2

Implement three different retrieval algorithms and evaluate their performance.

All algorithms were tested using Porter stemming and stopword removal. Raw results can be found in the Appendix from page 8 and on.

```
// compute the weight of a matched term
double computeRawTFIDFWeight(int docID,
                             int termID,
                             int docTermFreq,
                             double qryTermWeight,
                             Index *ind)
{
  return docTermFreq*log2(ind->docCount()/ind->docCount(termID))*qryTermWeight
      ;
}
```

Figure 2: RawTFIDF implementation.

```
// compute the weight of a matched term
double computeLogTFIDFWeight(int docID,
                             int termID,
                             int docTermFreq,
                             double qryTermWeight,
                             Index *ind)
{
  return (log2(docTermFreq) + 1)*log2(ind->docCount()/ind->docCount(termID))*
      qryTermWeight;
}
```

Figure 3: LogTFIDF implementation.

```
// compute the weight of a matched term
double computeOkapiWeight(int docID,
                          int termID,
                          int docTermFreq,
                          double qryTermWeight,
                          Index *ind)
{
  /* Ugly, but removes necessity of short-lived variables */
  return (docTermFreq/(docTermFreq + 0.5 + 1.5*(ind->docLength(docID)/ind->
      docLengthAvg())))*log2((ind->docCount()-ind->docCount(termID) + 0.5)/(
      ind->docCount(termID) + 0.5))*((8 + qryTermWeight)/(7 + qryTermWeight));
}
```

Figure 4: Okapi implementation.

Problem 2 continued on next page...

```
double computeCustomWeight(int docID,
                           int termID,
                           int docTermFreq,
                           double qryTermWeight,
                           Index *ind)
{
  return pow(qryTermWeight,2)*docTermFreq/ind->docCount(termID);
}

// compute the adjusted score
double computeCustomAdjustedScore(double origScore, // the score from the
    accumulator
                                  int docID, // doc ID
                                  Index *ind) // index
{
  return origScore/ind->docLength(docID);
}
```

Figure 5: Custom algorithm implementation.

Figure 6: Performance results from various retrival algorithms.

	Avg precision	R-precision
RawTF	0.1224	0.1372
RawTFIDF	0.1491	0.1557
LogTFIDF	0.1770	0.1958
Okapi	0.1694	0.1701
Custom	0.1698	0.1713

LogTFIDF seems to edge out the other algorithms for our test cases. LogTFIDF is popular in practice, and these results seem to validate its popularity.

My custom algorithm is aimed at increasing the weight of important query terms, as well as inflating the score of short documents. While these tests do not illustrate whether shorter documents are actually favored, my custom algorithm seems to keep up with other IDF algorithms, which it is somewhat similar to.

Appendix

Part 1 results

RawTF

```
Queryid (Num):       30
Total number of documents over all queries
    Retrieved:     3000
    Relevant:       442
    Rel_ret:        108
Interpolated Recall - Precision Averages:
    at 0.00       0.1760
    at 0.10       0.1180
    at 0.20       0.0844
    at 0.30       0.0539
    at 0.40       0.0396
    at 0.50       0.0349
    at 0.60       0.0234
    at 0.70       0.0072
    at 0.80       0.0072
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.0449
Precision:
  At    5 docs:   0.0733
  At   10 docs:   0.0833
  At   15 docs:   0.0689
  At   20 docs:   0.0633
  At   30 docs:   0.0611
  At  100 docs:   0.0360
  At  200 docs:   0.0180
  At  500 docs:   0.0072
  At 1000 docs:   0.0036
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.0712
```

RawTF with Porter stemming and stopwords removed

```
Queryid (Num):       30
Total number of documents over all queries
    Retrieved:     3000
    Relevant:       442
    Rel_ret:        168
Interpolated Recall - Precision Averages:
    at 0.00       0.3271
    at 0.10       0.2264
    at 0.20       0.2061
```

```
    at 0.30       0.1779
    at 0.40       0.1501
    at 0.50       0.1224
    at 0.60       0.1085
    at 0.70       0.0724
    at 0.80       0.0593
    at 0.90       0.0096
    at 1.00       0.0049
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1224
Precision:
  At    5 docs:   0.1600
  At   10 docs:   0.1467
  At   15 docs:   0.1378
  At   20 docs:   0.1300
  At   30 docs:   0.1122
  At  100 docs:   0.0560
  At  200 docs:   0.0280
  At  500 docs:   0.0112
  At 1000 docs:   0.0056
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1372
```

RawTF with Krovetz stemming and stopwords removed

```
Queryid (Num):      30
Total number of documents over all queries
    Retrieved:    3000
    Relevant:      442
    Rel_ret:       226
Interpolated Recall - Precision Averages:
    at 0.00       0.5020
    at 0.10       0.3610
    at 0.20       0.2922
    at 0.30       0.2585
    at 0.40       0.2144
    at 0.50       0.1634
    at 0.60       0.1363
    at 0.70       0.0654
    at 0.80       0.0415
    at 0.90       0.0138
    at 1.00       0.0090
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1707
Precision:
  At    5 docs:   0.2333
  At   10 docs:   0.1967
  At   15 docs:   0.1867
  At   20 docs:   0.1633
```

```
  At   30 docs:   0.1456
  At  100 docs:   0.0753
  At  200 docs:   0.0377
  At  500 docs:   0.0151
  At 1000 docs:   0.0075
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.2007
```

RawTF with Porter stemming (stopwords included)

```
Queryid (Num):      30
Total number of documents over all queries
    Retrieved:    3000
    Relevant:      442
    Rel_ret:        78
Interpolated Recall - Precision Averages:
    at 0.00       0.1590
    at 0.10       0.0827
    at 0.20       0.0746
    at 0.30       0.0561
    at 0.40       0.0540
    at 0.50       0.0414
    at 0.60       0.0308
    at 0.70       0.0121
    at 0.80       0.0070
    at 0.90       0.0048
    at 1.00       0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.0444
Precision:
  At    5 docs:   0.0600
  At   10 docs:   0.0667
  At   15 docs:   0.0600
  At   20 docs:   0.0550
  At   30 docs:   0.0511
  At  100 docs:   0.0260
  At  200 docs:   0.0130
  At  500 docs:   0.0052
  At 1000 docs:   0.0026
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.0546
```

RawTF with stopwords removed)

```
Queryid (Num):      30
Total number of documents over all queries
    Retrieved:    3000
    Relevant:      442
    Rel_ret:       193
```

```
Interpolated Recall - Precision Averages:
    at 0.00       0.4982
    at 0.10       0.3773
    at 0.20       0.2842
    at 0.30       0.2057
    at 0.40       0.1428
    at 0.50       0.1234
    at 0.60       0.0883
    at 0.70       0.0474
    at 0.80       0.0391
    at 0.90       0.0153
    at 1.00       0.0153
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1495
Precision:
  At    5 docs:   0.2067
  At   10 docs:   0.1700
  At   15 docs:   0.1444
  At   20 docs:   0.1283
  At   30 docs:   0.1189
  At  100 docs:   0.0643
  At  200 docs:   0.0322
  At  500 docs:   0.0129
  At 1000 docs:   0.0064
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1732
```

Part 2 results

RawTFIDF

```
Queryid (Num):       30
Total number of documents over all queries
    Retrieved:     3000
    Relevant:       442
    Rel_ret:        176
Interpolated Recall - Precision Averages:
    at 0.00       0.3798
    at 0.10       0.2905
    at 0.20       0.2752
    at 0.30       0.2241
    at 0.40       0.1851
    at 0.50       0.1433
    at 0.60       0.1294
    at 0.70       0.0751
    at 0.80       0.0586
    at 0.90       0.0227
    at 1.00       0.0046
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1491
Precision:
  At    5 docs:   0.1867
  At   10 docs:   0.1600
  At   15 docs:   0.1600
  At   20 docs:   0.1517
  At   30 docs:   0.1233
  At  100 docs:   0.0587
  At  200 docs:   0.0293
  At  500 docs:   0.0117
  At 1000 docs:   0.0059
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1557
```

LogTFIDF

```
Queryid (Num):       30
Total number of documents over all queries
    Retrieved:     3000
    Relevant:       442
    Rel_ret:        175
Interpolated Recall - Precision Averages:
    at 0.00       0.4110
    at 0.10       0.3626
    at 0.20       0.3425
    at 0.30       0.2737
    at 0.40       0.2156
```

```
    at 0.50       0.1754
    at 0.60       0.1621
    at 0.70       0.0802
    at 0.80       0.0574
    at 0.90       0.0227
    at 1.00       0.0046
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1770
Precision:
  At    5 docs:   0.2400
  At   10 docs:   0.1867
  At   15 docs:   0.1844
  At   20 docs:   0.1617
  At   30 docs:   0.1278
  At  100 docs:   0.0583
  At  200 docs:   0.0292
  At  500 docs:   0.0117
  At 1000 docs:   0.0058
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1958
```

Okapi

```
Queryid (Num):      30
Total number of documents over all queries
    Retrieved:    3000
    Relevant:      442
    Rel_ret:       175
Interpolated Recall - Precision Averages:
    at 0.00       0.3734
    at 0.10       0.3223
    at 0.20       0.3015
    at 0.30       0.2424
    at 0.40       0.2054
    at 0.50       0.1949
    at 0.60       0.1488
    at 0.70       0.0864
    at 0.80       0.0605
    at 0.90       0.0127
    at 1.00       0.0076
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1694
Precision:
  At    5 docs:   0.2333
  At   10 docs:   0.2067
  At   15 docs:   0.1800
  At   20 docs:   0.1633
  At   30 docs:   0.1367
  At  100 docs:   0.0583
```

```
  At  200 docs:   0.0292
  At  500 docs:   0.0117
  At 1000 docs:   0.0058
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1701
```

Custom

```
Queryid (Num):      30
Total number of documents over all queries
    Retrieved:    3000
    Relevant:      442
    Rel_ret:       177
Interpolated Recall - Precision Averages:
    at 0.00       0.3617
    at 0.10       0.3281
    at 0.20       0.3105
    at 0.30       0.2587
    at 0.40       0.1975
    at 0.50       0.1720
    at 0.60       0.1546
    at 0.70       0.1033
    at 0.80       0.0632
    at 0.90       0.0268
    at 1.00       0.0079
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.1698
Precision:
  At    5 docs:   0.2200
  At   10 docs:   0.1933
  At   15 docs:   0.1756
  At   20 docs:   0.1550
  At   30 docs:   0.1333
  At  100 docs:   0.0590
  At  200 docs:   0.0295
  At  500 docs:   0.0118
  At 1000 docs:   0.0059
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1713
```