Information Retrieval

Assignment 2, summer semester 2011

by Benjamin Bachhuber (1028430), Patrick Marschik (0625039)

Table of Contents

An important part of this assignment is that you perform an analysis of the results you obtain with your retrieval system. Interpret the differences you obtain in the retrieved documents when using different text indices, and when using different metrics.
Provide the output of your retrieval systems in the report to back up your findings.

Used queries (set 1)

- 20-newsgroups
    - comp.graphics/38863
    - talk.politics.guns/55082
    - soc.religion.christian/21409
    - talk.politics.mideast/76075
    - sci.med/59297
    - talk.politics.guns/54831
    - rec.sport.baseball/104988
    - sci.crypt/15879
    - misc.forsale/76937
    - sci.crypt/16074
- banksearch
    - A/A0020.txt
    - B/B0414.txt
    - C/C0259.txt
    - D/D0615.txt
    - E/E0853.txt
    - F/F0274.txt

We divided the document into several observations and our interpretation of that observation.

Observation 1

N-Gram indices yield lower distances than Bag-of-Words indices.

Interpretation

Since the number of attributes is much lower when bi-gram or tri-gram indices are used the distance between the documents is much lower.
One exception are 5-grams that is because for the 5-gram index not words but letters were used.

Example

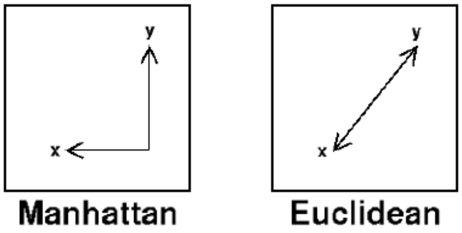
```
query: sci.crypt/15879

news_BoW_Boolean_Thresh_1_40.arff         news_BoW_TFIDF_Thresh_1_40.arff
+-----------------------------------------+------------------------------
comp.windows.x/68218           127.000000 comp.windows.x/68218
misc.forsale/76495             128.000000 rec.sport.baseball/104905
misc.forsale/76596             128.000000 sci.med/59367
rec.autos/101589               128.000000 talk.politics.misc/178690
rec.sport.baseball/104905      128.000000 misc.forsale/76596
sci.med/59367                  128.000000 rec.sport.baseball/105148
talk.politics.misc/178690      128.000000 rec.autos/101589
comp.sys.mac.hardware/52035    129.000000 misc.forsale/76495
rec.sport.hockey/52587         129.000000 comp.sys.mac.hardware/52045
sci.electronics/54080          129.000000 comp.sys.ibm.pc.hardware/60581

news_word-2grams_tf-idf_0.01-0.4.arff     news_word-3grams_tf-idf_0.01-0.4.arff
+-----------------------------------------+----------------------------------------+
comp.os.ms-windows.misc/9941    14.236047 sci.crypt/15769                  2.130621
comp.os.ms-windows.misc/9983    14.236047 sci.crypt/15201                  2.381423
comp.os.ms-windows.misc/9987    14.236047 sci.crypt/15262                  2.381423
comp.sys.ibm.pc.hardware/60581  14.236047 sci.crypt/15265                  2.381423
comp.windows.x/68218            14.236047 sci.crypt/15389                  2.381423
misc.forsale/75889              14.236047 sci.crypt/15494                  2.381423
misc.forsale/75935              14.236047 sci.crypt/15500                  2.381423
misc.forsale/76342              14.236047 sci.crypt/15570                  2.381423
misc.forsale/76474              14.236047 sci.crypt/15594                  2.381423
misc.forsale/76495              14.236047 sci.crypt/15675                  2.381423
```

Observation 2

Distances with L2 are lower than distances with L1.

Interpretation

L2 (Euclidean) distances are lower then L1 (Manhattan) distances because L1 calculates the distances using a grid-like path and L2 calculates the distances using a straight line connection between the instances.



Example

```
query: sci.crypt/15879

L1
rank news_5grams_001_04_stm_stw.arff           news_BoW_Boolean_Thresh_1_40.arff
-----+-----------------------------------------+----------------------------------------
#  0 comp.os.ms-windows.misc/9987   160.430601 comp.windows.x/68218           127.000000
#  1 comp.os.ms-windows.misc/9983   160.446238 misc.forsale/76495             128.000000
#  2 sci.med/59367                  172.368580 misc.forsale/76596             128.000000
#  3 comp.windows.x/68218           175.368580 rec.autos/101589               128.000000
#  4 comp.sys.ibm.pc.hardware/60581 177.959972 rec.sport.baseball/104905      128.000000
#  5 misc.forsale/76573             182.282892 sci.med/59367                  128.000000
#  6 rec.autos/101589               182.400952 talk.politics.misc/178690      128.000000
#  7 misc.forsale/76495             185.157462 comp.sys.mac.hardware/52035    129.000000
#  8 talk.politics.misc/178690      185.184196 rec.sport.hockey/52587         129.000000
#  9 rec.sport.hockey/53822         185.657334 sci.electronics/54080          129.000000


L2
rank news_5grams_001_04_stm_stw.arff           news_BoW_Boolean_Thresh_1_40.arff
-----+-----------------------------------------+----------------------------------------
#  0 comp.os.ms-windows.misc/9987     6.350798 comp.windows.x/68218            11.269428
#  1 comp.os.ms-windows.misc/9983     6.350806 misc.forsale/76495              11.313708
#  2 rec.sport.hockey/52630           6.430515 misc.forsale/76596              11.313708
#  3 comp.graphics/38853              6.467570 rec.autos/101589                11.313708
#  4 comp.sys.mac.hardware/51892      6.536839 rec.sport.baseball/104905       11.313708
#  5 sci.med/59183                    6.545552 sci.med/59367                   11.313708
#  6 comp.graphics/38778              6.556626 talk.politics.misc/178690       11.313708
#  7 comp.graphics/39078              6.575926 comp.sys.mac.hardware/52035     11.357817
#  8 sci.electronics/53569            6.662111 rec.sport.hockey/52587          11.357817
#  9 rec.sport.hockey/53822           6.706946 sci.electronics/54080           11.357817
```

Observation 3

For the news corpus using the 3-grams and the BoW Boolean index the top 10 items are exactly the same for L1 and L2.

Interpretation

The distances using the L2 metric are generally lower but the top results in the news corpus are still from the same class even though the order might be slightly different.
The phrases in the news corpus are very distinct and form a tight cluster.

Example

```
query: talk.politics.guns/55082

L1
news_BoW_Boolean_Thresh_1_40.arff
+-----------------------------------------
rec.autos/103227               160.000000
talk.politics.misc/178690      160.000000
comp.sys.ibm.pc.hardware/60288 161.000000
comp.windows.x/68218           161.000000
talk.politics.guns/54315       161.000000
talk.politics.mideast/76312    161.000000
talk.politics.mideast/76505    161.000000
talk.politics.misc/178683      161.000000
comp.windows.x/68322           162.000000
rec.autos/101589               162.000000

news_word-3grams_tf-idf_0.01-0.4.arff
+-----------------------------------------
talk.politics.guns/54278         0.348820
talk.politics.guns/54315         0.348820
talk.politics.guns/54375         0.348820
talk.politics.guns/54402         0.348820
talk.politics.guns/54405         0.348820
talk.politics.guns/54478         0.348820
talk.politics.guns/54602         0.348820
talk.politics.guns/54643         0.348820
talk.politics.guns/54659         0.348820
talk.politics.guns/54723         0.348820

L2
news_BoW_Boolean_Thresh_1_40.arff
+-----------------------------------------
rec.autos/103227                12.649111
talk.politics.misc/178690       12.649111
comp.sys.ibm.pc.hardware/60288  12.688578
comp.windows.x/68218            12.688578
talk.politics.guns/54315        12.688578
talk.politics.mideast/76312     12.688578
talk.politics.mideast/76505     12.688578
talk.politics.misc/178683       12.688578
comp.windows.x/68322            12.727922
rec.autos/101589                12.727922

news_word-3grams_tf-idf_0.01-0.4.arff
+-----------------------------------------
talk.politics.guns/54278         0.204511
talk.politics.guns/54315         0.204511
talk.politics.guns/54375         0.204511
talk.politics.guns/54402         0.204511
talk.politics.guns/54405         0.204511
talk.politics.guns/54478         0.204511
talk.politics.guns/54602         0.204511
talk.politics.guns/54643         0.204511
talk.politics.guns/54659         0.204511
talk.politics.guns/54723         0.204511
```

Observation 4

The average rank and the average distance correlate except for the news corpus using the L1 metric.

Interpretation

Trivially the best matches for each index also have the lowest distance respectively.
The reason for the discrepancy using the L1 metric with the news corpus is that here the difference between the distances of the BoW & n-gram metrics are bigger than in the bank corpus or using the L2 metric. Therefore the BoW indices have more weight on the ranking.

Example

```
query: talk.politics.guns/55082

L2

document                                 #occur  avg rank        avg dist
comp.windows.x/68218                           5          53.600           4.968
sci.med/59367                                  5          88.200           4.969
comp.sys.ibm.pc.hardware/60581                 5          92.600           5.155
rec.autos/101589                               5         102.800           5.184
misc.forsale/76495                             5         107.000           5.209
rec.sport.baseball/104905                      5         129.600           5.203
comp.graphics/38775                            5         149.200           5.367
misc.forsale/75889                             5         175.800           5.365
rec.sport.baseball/105148                      5         179.800           5.324
misc.forsale/76573                             5         229.800           5.330
misc.forsale/76440                             5         239.200           5.412
rec.sport.hockey/53707                         5         261.800           5.404
rec.sport.hockey/53822                         5         270.000           5.173
rec.sport.baseball/102629                      5         280.600           5.470
comp.graphics/38681                            5         288.400           5.454
rec.sport.hockey/54229                         5         294.400           5.444
rec.sport.hockey/54093                         5         314.000           5.271
talk.politics.guns/54590                       5         321.200           5.379
misc.forsale/76937                             5         332.200           5.530

L1

document                                 #occur  avg rank        avg dist
talk.politics.guns/54375                       5          22.800          84.049
comp.sys.ibm.pc.hardware/60581                 5          30.800          77.462
comp.windows.x/68218                           5          31.600          75.414
comp.graphics/38775                            5          38.200          80.663
misc.forsale/76495                             5          45.800          78.728
rec.autos/101589                               5          47.400          77.681
comp.graphics/38716                            5          48.800          87.728
talk.politics.guns/54315                       5          52.600          89.879
rec.sport.baseball/104905                      5          56.600          78.587
rec.sport.baseball/105148                      5          60.200          81.433
```

```
misc.forsale/76573                             5          60.600          79.246
misc.forsale/76440                             5          61.400          81.207
talk.politics.guns/54659                       5          63.200          91.679
comp.graphics/39050                            5          68.400          85.990
sci.med/59367                                  5          70.800          75.214
misc.forsale/76358                             5          72.000          87.077
comp.os.ms-windows.misc/9941                   5          73.200          89.331
comp.graphics/38664                            5          79.600          88.759
rec.sport.baseball/104333                      5          83.200          88.213
rec.autos/103049                               5          92.800          88.820
```

Observation 5

The tri-gram index always returns documents from the same class.

Interpretation

It seems tri-grams are very distinct for different domains.

Example

```
query: misc.forsale/76937
news_word-3grams_tf-idf_0.01-0.4.arff
+-----------------------------------------
misc.forsale/74780               0.000000
misc.forsale/75856               0.000000
misc.forsale/75889               0.000000
misc.forsale/75896               0.000000
misc.forsale/75935               0.000000
misc.forsale/75964               0.000000
misc.forsale/75987               0.000000
misc.forsale/76042               0.000000
misc.forsale/76181               0.000000
misc.forsale/76182               0.000000

query: F/F0274.txt
bank_word-3grams_tf-idf_0.01-0.4.arff
+----------------------------------------
F/F0118.txt                      4.208317
F/F0720.txt                      4.422636
F/500                            4.758317
F/F0060.txt                      4.758317
F/F0147.txt                      4.758317
F/F0171.txt                      4.758317
F/F0223.txt                      4.758317
F/F0251.txt                      4.758317
F/F0258.txt                      4.758317
F/F0297.txt                      4.758317
```

Observation 6

For the news corpus with the L2 metric it seems that the BoW boolean & TF*IDF indices always contain the document with the lowest average distance/rank. For the bank corpus all indices contain the document with the lowest average distance/rank.

Interpretation

The news corpus is seemingly best indexed with BoW algorithms. This might be because phrases used in newsgroups appear often in different contexts and do not help to distinguish the documents (hence the low occurence of the top average ranked in the n-gram algorithms).
For the bank corpus this could mean that the documents are equally well classified by phrases and words.

Example

```
query: talk.politics.guns/55082

document                                 #occur  avg rank        avg dist
comp.windows.x/68218                           5          53.600           4.968

news_BoW_Boolean_Thresh_1_40.arff         news_BoW_TFIDF_Thresh_1_40.arff
+-----------------------------------------+----------------------------------------
rec.autos/103227                12.649111 comp.windows.x/68218             2.527604
talk.politics.misc/178690       12.649111 talk.politics.misc/178690        2.693690
comp.sys.ibm.pc.hardware/60288  12.688578 rec.sport.baseball/104905        2.718231
comp.windows.x/68218            12.688578 sci.med/59367                    2.718231
talk.politics.guns/54315        12.688578 talk.politics.guns/55087         2.798417
talk.politics.mideast/76312     12.688578 talk.politics.guns/54590         2.805905
talk.politics.mideast/76505     12.688578 alt.atheism/54163                2.827434
talk.politics.misc/178683       12.688578 rec.sport.hockey/52630           2.840195
comp.windows.x/68322            12.727922 talk.politics.guns/54561         2.841365
rec.autos/101589                12.727922 talk.politics.guns/55238         2.849393
```

Observation 7

It also seems the document `A/A0019.txt` from the bank corpus was always ranked high for each query.

Interpretation

The document must contain many occurrance of very common words and phrases that appear in many bank documents.

Example

```
news_BoW_Boolean_Thresh_1_40.arff         news_BoW_TFIDF_Thresh_1_40.arff

Bank Corpus:

query: E/E0853.txt
document                                 #occur  avg rank        avg dist
E/E0752.txt                                    6           5.333           0.183
E/E0871.txt                                    6           6.333           0.086
E/E0773.txt                                    6           6.667           0.269
E/E0907.txt                                    6           6.833           0.000
E/E0943.txt                                    6           7.833           0.000
A/A0019.txt                                    6          18.667          17.804

query: B/B0414.txt
document                                 #occur  avg rank        avg dist
B/B0368.txt                                    6           1.000           6.017
B/B0818.txt                                    6           2.000          13.692
B/B0459.txt                                    6           3.333          18.210
B/B0730.txt                                    6           4.333          20.666
B/B0595.txt                                    6           4.667          26.007
B/B0505.txt                                    6           5.667          28.086
B/B0274.txt                                    6           7.000          30.578
B/B0023.txt                                    6          16.833          57.719
B/B0042.txt                                    6          17.833          57.719
B/B0136.txt                                    6          19.167          57.719
B/B0183.txt                                    6          20.167          57.719
B/B0211.txt                                    6          21.167          57.719
B/B0230.txt                                    6          22.167          57.719
B/B0277.txt                                    6          23.167          57.719
B/B0305.txt                                    6          24.333          57.719
B/B0324.txt                                    6          25.333          57.719
B/B0470.txt                                    6          26.667          57.719
B/B0553.txt                                    6          28.000          57.719
B/B0614.txt                                    6          29.167          57.719
B/B0625.txt                                    6          30.167          57.719
A/A0019.txt                                    6          31.000          58.385

query: F/F0274.txt
document                                 #occur  avg rank        avg dist
A/A0019.txt                                    6          12.167          15.729
```

Observation 8

A query for similar documents to “`E/E0853.txt`” of the bank corpus often results in distances of zero for other documents of the class “`E`”.

Interpretation

A closer examination of the query file makes the problem obvious:

```
Index of /software/lcc/tst

Index of /software/lcc/tst
 Name Last modified Size Description
 Parent Directory 25-Sep-2001 17:56 -
 8q.0 26-Feb-1997 17:42 0k
 8q.c 26-Feb-1997 17:42 1k
 array.0 26-Feb-1997 17:42 0k
 array.c 26-Feb-1997 17:42 1k
 cf.0 26-Feb-1997 17:42 1k
 cf.c 26-Feb-1997 17:42 1k
 cq.0 26-Feb-1997 17:42 0k
 cq.c 26-Feb-1997 17:42 122k
 cvt.0 26-Feb-1997 17:42 0k
 cvt.c 26-Feb-1997 17:42 1k
 fields.0 26-Feb-1997 17:42 0k
 fields.c 26-Feb-1997 17:42 1k
 front.0 26-Feb-1997 17:43 0k
 front.c 26-Feb-1997 17:43 2k
 incr.0 26-Feb-1997 17:43 0k
 incr.c 26-Feb-1997 17:43 1k
 init.0 26-Feb-1997 17:43 0k
 init.c 26-Feb-1997 17:43 1k
 limits.0 26-Feb-1997 17:43 0k
 limits.c 26-Feb-1997 17:43 1k
 paranoia.0 26-Feb-1997 17:43 0k
 paranoia.c 26-Feb-1997 17:43 57k
 sort.0 26-Feb-1997 17:43 0k
 sort.c 26-Feb-1997 17:43 1k
 spill.0 26-Feb-1997 17:43 0k
 spill.c 26-Feb-1997 17:43 1k
 stdarg.0 26-Feb-1997 17:43 0k
 stdarg.c 26-Feb-1997 17:43 1k
 struct.0 26-Feb-1997 17:43 0k
 struct.c 26-Feb-1997 17:43 2k
 switch.0 26-Feb-1997 17:43 0k
 switch.c 26-Feb-1997 17:43 3k
 wf1.0 26-Feb-1997 17:43 2k
 wf1.c 26-Feb-1997 17:43 2k
 yacc.0 26-Feb-1997 17:43 1k
 yacc.c 13-Dec-1999 18:29 13k
Apache/1.3.26 Server at www.cs.princeton.edu Port 80
```

Almost all of the words used in this file are not added to the indices.

Example

```
query: E/E0853.txt
bank_BoW_Boolean_Thresh_1_40.arff
+----------------------------------------
E/E0752.txt                      0.000000
E/E0773.txt                      0.000000
E/E0871.txt                      0.000000
E/E0907.txt                      0.000000
E/E0943.txt                      0.000000
E/500                           41.000000
E/E0621.txt                     41.000000
A/A0019.txt                     42.000000
A/A0128.txt                     42.000000
A/A0169.txt                     42.000000

bank_BoW_FT_0_5-70_0_TFxIDF.arff
+----------------------------------------
E/E0752.txt                      0.000000
E/E0773.txt                      0.000000
E/E0871.txt                      0.000000
E/E0907.txt                      0.000000
E/E0943.txt                      0.000000
A/A0019.txt                     10.039496
A/A0128.txt                     10.039496
A/A0169.txt                     10.039496
A/A0195.txt                     10.039496
A/A0221.txt                     10.039496

bank_BoW_TF_Thresh_1_40.arff
+---------------------------------------
E/E0871.txt                      0.00000
E/E0907.txt                      0.00000
E/E0943.txt                      0.00000
E/E0752.txt                      0.26198
E/E0773.txt                      0.26198
E/500                           17.29108
E/E0621.txt                     17.29108
A/A0019.txt                     18.29108
A/A0128.txt                     18.29108
A/A0169.txt                     18.29108

bank_s_i_0.5_70.0_3_out.arff
+----------------------------------------
E/E0907.txt                      0.000000
E/E0943.txt                      0.000000
E/E0871.txt                      0.513131
E/E0752.txt                      0.814516
E/E0773.txt                      1.327647
B/B0973.txt                     31.889871
A/A0019.txt                     31.926345
A/A0128.txt                     31.926345
A/A0169.txt                     31.926345
A/A0195.txt                     31.926345
```

```
bank_word-2grams_tf-idf_0.01-0.4.arff
+----------------------------------------
E/E0871.txt                      0.000000
E/E0907.txt                      0.000000
E/E0943.txt                      0.000000
E/E0752.txt                      0.023530
E/E0773.txt                      0.023530
E/E0594.txt                      0.972861
E/E0642.txt                      0.972861
E/E0698.txt                      1.963391
E/E0631.txt                      2.008894
E/E0559.txt                      2.023148
```

```
bank_word-3grams_tf-idf_0.01-0.4.arff
+----------------------------------------
E/500                            0.000000
E/E0058.txt                      0.000000
E/E0150.txt                      0.000000
E/E0190.txt                      0.000000
E/E0214.txt                      0.000000
E/E0414.txt                      0.000000
E/E0487.txt                      0.000000
E/E0522.txt                      0.000000
E/E0545.txt                      0.000000
E/E0559.txt                      0.000000
```

Bonus Task

Example Output

Corpus: News
Query: “microsoft”

```
query: 0/0
rank news_5grams_001_04_stm_stw.arff           news_BoW_Boolean_Thresh_1_40.arff
-----+-----------------------------------------+-----------------------------------------
#  0 comp.os.ms-windows.misc/9987        1,000 sci.electronics/53569              34,612
#  1 comp.os.ms-windows.misc/9983        1,000 comp.graphics/38853                34,785
#  2 rec.sport.hockey/52630              1,625 comp.graphics/39078                35,426
#  3 comp.graphics/38853                 2,318 comp.sys.mac.hardware/51892        36,986
#  4 rec.sport.hockey/53822              2,427 comp.graphics/38778                37,510


rank news_BoW_TFIDF_Thresh_1_40.arff           news_word-2grams_tf-idf_0.01-0.4.arff
-----+-----------------------------------------+-----------------------------------------
#  0 comp.windows.x/68218                1,000 alt.atheism/54143                   0,514
#  1 rec.sport.baseball/104905           1,414 alt.atheism/54163                   0,625
#  2 sci.med/59367                       1,414 comp.os.ms-windows.misc/9941        1,000
#  3 talk.politics.misc/178690           1,414 comp.os.ms-windows.misc/9983        1,000
#  4 rec.sport.hockey/52630              1,658 comp.os.ms-windows.misc/9987        1,000
```

```
rank news_word-3grams_tf-idf_0.01-0.4.arff
-----+-----------------------------------------+
#  0 alt.atheism/51219                   0,000
#  1 alt.atheism/53142                   0,000
#  2 alt.atheism/53143                   0,000
#  3 alt.atheism/53219                   0,000
#  4 alt.atheism/53252                   0,000

document                                 #occur  avg rank        avg dist
alt.atheism/54163                              5          12,800           9,680
sci.space/59848                                5         106,600           9,823
comp.graphics/38853                            5         107,400           8,280
rec.sport.hockey/52630                         5         119,400          10,016
sci.med/59347                                  5         133,000           9,948
alt.atheism/53645                              5         143,400          10,250
rec.autos/103450                               5         150,400          10,171
talk.politics.mideast/77390                    5         166,400          10,730
talk.religion.misc/83441                       5         178,200          10,554
comp.windows.x/67995                           5         181,800          10,832
talk.religion.misc/84197                       5         197,600          10,554
talk.politics.mideast/77270                    5         197,600          10,747
talk.politics.mideast/76035                    5         202,600          10,865
talk.politics.mideast/76075                    5         203,200          10,663
talk.politics.mideast/77230                    5         208,000          10,214
sci.crypt/15812                                5         210,400          10,447
alt.atheism/53137                              5         211,400          10,643
rec.motorcycles/105116                         5         211,400          11,000
talk.politics.mideast/77364                    5         212,600           9,868
rec.sport.hockey/52617                         5         214,400          11,033
talk.politics.mideast/77817                    5         215,600          10,843
talk.religion.misc/84147                       5         219,800          10,301
sci.space/61174                                5         221,800          11,047
talk.politics.mideast/77183                    5         227,400          10,877
rec.autos/103352                               5         230,200          10,683
```