CS 4300 Information Retrieval 11:30 AM Dec 4, 2014

Project 2

Alok Kumar Prusty (akp77)
Nikhil Anand Navali (nn259)
Prashanth Basappa (pb476)

Part 1: Pseudo-relevance feedback

1.c

Parameters	CACM	MEDLAR
A = 4, B=8, C=0, K= 5	0.29200796076418545	0.5777024631484999
A = 4, B=16, C=0, K=10	0.2702088792437362	0.599727096804838

1.d

Parameters	CACM	MEDLAR
A = 4, B=8, C=0, K= 5	Improved 23 Got Worse 27 Stayed Same 2	Improved 26 Got Worse 4
A = 4, B=16, C=0, K=10	Improved 23 Got Worse 29	Improved 28 Got Worse 2

1.e

In general there are two specific observations. If we try to run rocchio feedback with less number of documents we will get higher accuracy. During calculating rocchio scores if number of documents are less then we will get better accuracy after we append the terms to main query vector. Secondly if the query strings are relatively small we will get better accuracy with rocchio feedback. This is the reason why MEDLAR is outperforming CACM collection in terms of accuracy with rocchio feedback.

CACM with A=4,B=8,C=0,K=5

This accuracy is less than normal ATC.ATC accuracy as we are trying to overfit the query vector because the documents are more. Hence more terms will make it to the input space of rocchio query vector so number of relevant documents will reduce and hence the accuracy.

CACM with A=4,B=16,C=0,K=10

In this setup we are trying to put more weight in the relevant documents but since K is higher more terms will be

appended from the large document set which will reduce the overall accuracy.

MEDLAR with A=4,B=8,C=0,K=5

In this set up there are two advantages coming from the setup. First the documents are less and the average query length is also low hence more and more relevant documents will make their way to the retrieved list improving the overall accuracy.

MEDLAR with A=4,B=16,C=0,K=10

In this setup we are putting more weight on the relevant terms and adding more terms in the rocchio vectors so more relevant documents will be retrieved hence improving the overall precision.

1.f

While running with MEDLAR collection with A=4, B=16, C=0 and K=10 , we noticed **Query 21's** average precision is getting nearly doubled with rocchio pseudo relevance feedback. With ATC weights we are getting a precision of 0.172086 while with rocchio pseudo relevance feedback we are getting precision of 0.383929. Below are the values for original indexed query with ATC weights and modified with query vector.

Query Vector's ATC values

{languag=0.4504792170148357, infanc=0.49735352664657556, develop=0.2047047266580413, school=0.5252333730170959, pre=0.40178429854165926, ag=0.265524388273925}

Rocchio Weights of new Query Vector with top 10 appended terms

{jaundic=0.813430912275272, causat=1.0080989505792923, children=1.8215130378472988, languag=4.986227574892922, education=0.8066853909853687, empha=1.059519054016172, formal=0.8414470875127434, speech=0.9617519482012798, school=2.1009334920683838, pre=1.607137194166637, infanc=3.363222663180138, commun=1.30562025125691, develop=1.757510735273966, autist=0.8497064889019993, ag=1.261589965132676, disord=0.9422636822383242}

Query: language development in infancy and pre-school age.

Actual Relevant Documents as per the relevance information

MED-0604 MED-0605 MED-0608 MED-0610 MED-0612 MED-0613 MED-0615 MED-0616 MED-0618 MED-0619 MED-0620 MED-0622 MED-0626 MED-0630 MED-0631 MED-0884 MED-0885 MED-0886 MED-0888 MED-0890 MED-0891 MED-0892 MED-0893 MED-0894 MED-0895 MED-0896 MED-0898

Top 100 Retrieval with ATC.ATC weight

{MED-0766=0.19052002766880644, MED-0613=0.16732179660908944, MED-0815=0.14573020379445598, MED-0892=0.1345675559078396, MED-0896=0.13290386572480356, MED-0253=0.1315843734945783, MED-0821=0.13112302402707635, MED-0917=0.11349977792341504, MED-0100=0.10744464051244819, MED-0887=0.10606553792433183, MED-0202=0.1025687726620426, MED-0889=0.1024954956865467, MED-0810=0.10214728340939021, MED-0626=0.10203277456871704, MED-0812=0.10100671400988923, MED-0736=0.09614341919771552, MED-0609=0.09084312206384389, MED-0616=0.08962428061098218, MED-0890=0.08143913307594237, MED-0602=0.08004668156698337, MED-0035=0.0797631187388503, MED-0885=0.07658925197988799, MED-0073=0.0756853503656704, MED-0709=0.07491771272416159, MED-0920=0.07330065336594932, MED-0488=0.07299551314415113, MED-0604=0.07056635906341668, MED-

0888=0.07035121339495597, MED-0321=0.06997058507626147, MED-0914=0.06974877077755302, MED-0618=0.0691164276105454, MED-0633=0.06801661462195999, MED-0252=0.0652376112791867, MED-0963=0.06402943355485698, MED-0203=0.06279562549034641, MED-0625=0.062305443554678326, MED-0891=0.05966696013686602, MED-0619=0.05654150849329174, MED-0351=0.05653387383068733, MED-0806=0.05554515271159105, MED-1008=0.05520309107693322, MED-0798=0.05442477135736517, MED-0151=0.052954731737264206, MED-0620=0.050148774461695474, MED-0915=0.04872450172814092, MED-1009=0.04745795800184433, MED-0742=0.04706903627354639, MED-0288=0.04672053231493398, MED-0180=0.04638849889018253, MED-0756=0.044927503414465705, MED-0201=0.04467447519694433, MED-0762=0.044366009614715284, MED-0267=0.04430581103117917, MED-0864=0.044274418090017406, MED-0149=0.04272620738566948, MED-0190=0.04263450302067205, MED-0713=0.04081973977351311, MED-0733=0.04070394037456494, MED-0771=0.04049307632797385, MED-0545=0.04018983268716837, MED-0593=0.039015937910636816, MED-1024=0.03858322475132209, MED-0027=0.037042945501507045, MED-0722=0.03691645051913497, MED-0317=0.036376090005209326, MED-0302=0.03607291479294864, MED-0780=0.03538203912163843, MED-0373=0.03512295031089402, MED-0907=0.0349400743487578, MED-0208=0.0345983381676614, MED-0697=0.03385346337433373, MED-0708=0.03319972614772792, MED-0781=0.03318045641025072 6, MED-0933=0.03307614248399537, MED-0009=0.03279697149491481, MED-0928=0.03272408680288615, MED-0937=0.032368489457183666, MED-0487=0.03206001727512572, MED-0472=0.03191202294476745, MED-0732=0.03176399739776501, MED-0223=0.031552041146653, MED-0853=0.031372001326146065, MED-0693=0.03127859162739916, MED-0349=0.031055451141583375, MED-0421=0.03069290899191664, MED-1033=0.030689495014037386, MED-0662=0.030494340735555577, MED-0072=0.029501082236349535, MED-0188=0.02897558114560171, MED-0589=0.028919321604989445, MED-0499=0.027521123680355926, MED-0678=0.027067534829790633, MED-0454=0.026976107072799158, MED-0158=0.026532109360922873, MED-0386=0.026475774253080137, MED-0452=0.026378454595431648, MED-0462=0.02616699981029147, MED-0504=0.02610502292758598, MED-0807=0.025995800424129224, MED-0010=0.02599064974977375}

Top 100 Retrieval with Rocchio pseudo relevance feedback

{MED-0892=2.8439777223042264, MED-0613=2.4059478814642583, MED-0815=2.3889518516817208, MED-0821=2.3190179405501667, MED-0766=2.0689512562638526, MED-0896=1.9000742679142022, MED-0616=1.5957518055379387, MED-0887=1.496030996680846, MED-0810=1.4384808183277864, MED-0914=1.1515098854615708, MED-0889=1.1344937737221887, MED-0885=1.103130916307935, MED-0253=0.9989304786024459, MED-0890=0.9680655792728086, MED-0604=0.9357506662594508, MED-0888=0.9018802828241106, MED-0915=0.8891716191103023, MED-0620=0.8351335898486186, MED-0618=0.8334164603348658, MED-0100=0.8109293136372993, MED-0626=0.7909398598739484, MED-0619=0.7796489497109395, MED-0488=0.7576294588469538, MED-0920=0.7510237121069926, MED-0812=0.7432499430754675, MED-0351=0.6800207294423923, MED-0891=0.675959627645224, MED-0963=0.6743993494568403, MED-0202=0.6478457392341547, MED-0610=0.6380659202709065, MED-0897=0.623269715595782, MED-0625=0.6146692479528345, MED-0203=0.6050470112221995, MED-0924=0.5995060623130981, MED-0602=0.5412946709629929, MED-0605=0.5360211931613019, MED-0073=0.5356111596210342, MED-0035=0.523118575279897, MED-0609=0.521418312464425, MED-0894=0.5097279319212398, MED-0803=0.49319306282448827, MED-0922=0.49209587563177803, MED-0925=0.48026151719013, MED-0709=0.4731958833263663, MED-0917=0.45399911169366014, MED-0742=0.45185876984763984, MED-1008=0.4511820675729805, MED-0805=0.446030587728959, MED-0321=0.4419487943249237, MED-0853=0.4310210800345072, MED-0898=0.4300563846854175, MED-0884=0.4277491910891362, MED-0822=0.4143695485015199, MED-0798=0.4111494788672613, MED-0607=0.3893128496712101, MED-0736=0.3845736767908621, MED-0245=0.3828631434633409, MED-0849=0.3822549364058780 4, MED-0633=0.3785653375002517, MED-0923=0.37763478070658973, MED-0928=0.37538203766941247, MED-0818=0.3707098301798347, MED-0809=0.36580762354802354, MED-

0918=0.3598162327176534, MED-0807=0.3564299684487179, MED-0036=0.3303128641152101, MED-0927=0.3255180685701372, MED-1009=0.32092198275187, MED-1007=0.31947007801311084, MED-0799=0.31884605330479254, MED-0802=0.3168393001741934, MED-0589=0.312856086473544, MED-0926=0.3018302638879251, MED-0848=0.2995942656150817, MED-0811=0.29778948443463604, MED-0886=0.2970628790209896, MED-0493=0.2953840511447533, MED-0222=0.293321180140795, MED-0762=0.2860052337119021, MED-0252=0.2808985335206492, MED-0705=0.28026889257668314, MED-0814=0.2801220688479287, MED-0484=0.27953930611112127, MED-0349=0.2782139255581531, MED-0622=0.2724253063362865, MED-0200=0.27077569660802836, MED-0593=0.27056806962868285, MED-0984=0.26855569993515405, MED-0611=0.26390225675702017, MED-0115=0.26320530634455214, MED-0893=0.26157697949688186, MED-0745=0.25662534543729754, MED-0722=0.25600852598706897, MED-0959=0.2544216804846459, MED-0322=0.25361978765781734, MED-0907=0.25019061495070183, MED-0631=0.24361185933059343, MED-0623=0.24078228697565757, MED-0732=0.23624341792966916, MED-0288=0.23589814053481817}

As per the above data it can be clearly seen that more number of retrieved documents are matching with the relevant documents with rocchio feedback. Secondly all the matching documents as per rocchio feedback are ranked relatively higher than normal ATC ranks. Some documents (highlighted) which are not part of ATC rank appeared in Rocchio weighted rank. So rocchio retrieval did significantly better by adding new terms such as speech(0.9617519482012798), commun(1.30562025125691) etc which are part of relevant documents. So scanning all the documents and then giving the query positive feedback about the related documents works well for rocchio pseudo relevance feedback model.

1.g

Below are the combinations of parameters we tried to come up with better MAP.

A = 4, B = 24, C = 0, K = 15;
MEAN AVERAGE PRECISION:0.6075368296712955

Queries Improved	27
Queries got worse	3

A = 4, B = 36,C = 0, K = 20;
MEAN AVERAGE PRECISION:0.6097371198161853

Queries Improved	27
Queries got worse	3

The best run we could get by varying the parameter space is listed above with a MAP of 0.6097371198161853
With Giving more values to B we are actually trying to put more weight on relevant documents hence query efficiency will improve which means more relevant documents will be chosen in the retrieved documents. Seeing the high number of improved queries(27) we can pick any of these and do an analysis. With the above setup lets analyse **Query 26** whose precision is getting improved from 0.291339 to 0.558453.

query 26: methods for experimental production of and known causes of hydrocephalus in animals and humans.

relevant documents:
MED-0708 MED-0712 MED-0713 MED-0714 MED-0715 MED-0716 MED-0717 MED-0719 MED-0721 MED-0722 MED-0723 MED-0724 MED-0725 MED-0726 MED-0959 MED-0960 MED-0961 MED-0962 MED-0963 MED-0964 MED-0965 MED-0966 MED-0967 MED-0968 MED-0969 MED-0970 MED-0972 MED-0973

ATC-ATC retrieval top 100 retrieval
{MED-0967=0.1494511589802416, MED-0966=0.1190915845300388, MED-0703=0.117681009622541, MED-0971=0.11004295698062037, MED-0710=0.10946378385845225, MED-1018=0.10039700690432765, MED-0970=0.09933875902903828, MED-0717=0.09903852937841683, MED-0859=0.09697106583537468, MED-0739=0.09376648556129825, MED-0723=0.0936005626030952, MED-0657=0.09349626109155765, MED-0215=0.09225877671976093, MED-0724=0.0894690500054338, MED-0089=0.08895145575873742, MED-0716=0.08764936848958678, MED-0648=0.08632625883044558, MED-0078=0.08348901883797415, MED-0465=0.08212481252425341, MED-0737=0.08129259293236844, MED-0334=0.08025383446231395, MED-0968=0.07997601079830674, MED-0634=0.07984559841497395, MED-0704=0.07946214194368192, MED-0909=0.0790823393767799, MED-0960=0.07819083239527967, MED-0161=0.07768387159674042, MED-0973=0.0771495291274805, MED-0173=0.07695252010105202, MED-0959=0.0754138748978993, MED-0400=0.07411204343617457, MED-0961=0.07330268307047458, MED-0382=0.07232925815949888, MED-0037=0.07196669561601841, MED-0900=0.07194667723643974, MED-0093=0.07192157171771989, MED-0379=0.07189558230442397, MED-0231=0.07115762600232699, MED-0449=0.07033393402990253, MED-0962=0.06979145781786499, MED-0007=0.069122488012009, MED-0337=0.06909007460412288, MED-0127=0.06805649893704255, MED-0719=0.06801940465286699, MED-0910=0.06675172195210373, MED-0630=0.06632566024772668, MED-0721=0.06628152703579435, MED-0600=0.06541837398775244, MED-0163=0.06541417958511603, MED-0069=0.06531175183832982, MED-0659=0.06348165299891012, MED-0746=0.06315009983774075, MED-0598=0.06283047852307208, MED-0982=0.06269739233335606, MED-0421=0.06174925419880996, MED-0151=0.060881541339936576, MED-0816=0.060138881902250314, MED-0608=0.05973708995543138, MED-0014=0.05956805127467041, MED-0824=0.05917255430559431, MED-0706=0.058941824689119064, MED-0034=0.05885106349582176, MED-0726=0.05881204596554092, MED-0274=0.05878198737712924, MED-0734=0.05834631219192750 6, MED-0702=0.058282550979578046, MED-0658=0.057306128873352 75, MED-0095=0.057238139471343374, MED-0768=0.0568032604523784, MED-0486=0.05621671335006399, MED-0958=0.05612294193629622, MED-0965=0.05602582234637916, MED-0445=0.0559903513573968, MED-0940=0.0559370353763438, MED-0827=0.05579652273504461, MED-0712=0.0554865403764207 1, MED-0370=0.05530599755299367, MED-0647=0.054567763028154304, MED-0154=0.05439899823193177, MED-0051=0.05416360710163152 5, MED-0030=0.054084133674764626, MED-1020=0.05397179854404249, MED-1011=0.0537776161152585 1, MED-0587=0.0535565712754868, MED-0074=0.05340070851105285, MED-0852=0.052752140660824245, MED-0580=0.05263000158938042, MED-0404=0.05261173285316423, MED-0606=0.052558058448250206, MED-0255=0.052436204468541296, MED-0126=0.05234400552912224 5, MED-0865=0.05233737950245226 4, MED-0915=0.052118259025569116, MED-0725=0.05201860814175819, MED-0208=0.05197926255307883, MED-0963=0.05194335004466827, MED-0542=0.05191262741995928, MED-0583=0.0516535184693597, MED-0652=0.051650319838527636, MED-0372=0.0511876465302891}

Rocchio pseudo-relevance top 100 retrieval

{MED-0967=5.8549722218897635, MED-0703=5.05470715733718, MED-0710=4.912044081382181, MED-0966=4.771149738970902, MED-0971=3.0107078428915686, MED-0970=2.919978984921536, MED-

0968=2.3700433858484646, MED-0717=1.9833547804831042, MED-0961=1.8451071033740833, MED-0716=1.770872475064417, MED-0704=1.5913162289521452, MED-0724=1.5707347863848606, MED-0723=1.5672400105492776, MED-0973=1.5450036300756604, MED-0960=1.4706939493325546, MED-0382=1.4484724363314436, MED-0718=1.4276925992193958, MED-0721=1.3744036829793536, MED-0722=1.3110548963728614, MED-0964=1.3036354232648657, MED-0701=1.2882336652657052, MED-0706=1.2430142319972666, MED-0959=1.2290490824091895, MED-0712=1.211713999130839, MED-0962=1.1829799452799288, MED-0702=1.167171774758679, MED-0965=1.1616185228167173, MED-0720=1.0909153269710667, MED-0366=1.0476640739062582, MED-0725=1.0417294741708598, MED-0963=1.0402223485343005, MED-0719=1.0164037143102649, MED-0715=0.998963977624484, MED-0726=0.9487049426289595, MED-0214=0.9315995327200123, MED-0114=0.8176187444561069, MED-0768=0.8005627818278386, MED-0623=0.7809185252944595, MED-0311=0.7405503685726983, MED-0622=0.7370459030441309, MED-0767=0.7224636764254875, MED-1018=0.7212951820695263, MED-0761=0.648626197362691, MED-0317=0.6345442049202006, MED-0253=0.6142303561550625, MED-0707=0.612532610523248, MED-0853=0.6012404333383757, MED-0816=0.5950608093620533, MED-0749=0.5809749787216322, MED-0465=0.563940163771987, MED-0239=0.5590916742313516, MED-0859=0.5576412687528254, MED-0081=0.54671736864141, MED-0743=0.5456611460985803, MED-0657=0.544994783682241, MED-0685=0.5253476533548191, MED-0089=0.5233131456831077, MED-0267=0.522621001399797, MED-0379=0.5118637376041949, MED-0827=0.5112306124505022, MED-0007=0.5059595426493917, MED-0754=0.5034712100729045, MED-0082=0.5005726325657601, MED-0614=0.49458316786019896, MED-0400=0.49067492647028266, MED-0337=0.48918826770352397, MED-0811=0.4867439887645712 6, MED-0273=0.4862406175836116 5, MED-0634=0.4792206736178001, MED-0225=0.4743317154237425 3, MED-0255=0.4737223706937585, MED-0051=0.4679383566807195 4, MED-0607=0.4658533398185002, MED-0680=0.4656987365795177, MED-0218=0.463481423964279, MED-0112=0.4623023252294657, MED-0770=0.4620242620157396 5, MED-0934=0.4609007210413111, MED-0732=0.4593471741688737, MED-0600=0.4579281407193762 6, MED-0852=0.4572088580666377, MED-0608=0.4545922651087499 5, MED-0215=0.4542798463091862, MED-0093=0.4509096382608766, MED-0956=0.4462294169296479 6, MED-0742=0.4377654058710880 3, MED-0173=0.4376831911329175 4, MED-0713=0.4351880681892472 7, MED-0799=0.4345108494791316, MED-0739=0.4324095388082777, MED-0737=0.4288851347670113, MED-0334=0.4231010123709317, MED-0658=0.4176577275327043 6, MED-0648=0.4172961092780264 6, MED-0772=0.4098523561828882 3, MED-0161=0.4093503130649397 4, MED-0625=0.4072405221294246 3, MED-0413=0.4048524443351422, MED-0122=0.4025543241152158, MED-0445=0.39849554172718415}

It can be seen that by putting more weights (higher B) more retrieved documents are matching with the relevant documents and they are also ranked relatively higher than normal ATC retrieval.

Part 2: Complete Link Clustering

2.c

MAP Values	CACM	MEDLAR
2.a – Highest Similarity – K = 20	0.3030215250892239	0.5140779108714725
2.b – Average Similarity – K = 10	0.2828706820855185	0.5067190506708493

2.d

<table>
<tr><th></th><th>CACM</th><th>MEDLAR</th></tr>
<tr><td>2.a – Highest Similarity – K = 20</td><td><table><tr><td>Improved</td><td>19</td></tr><tr><td>Worse</td><td>26</td></tr><tr><td>Same</td><td>7</td></tr></table></td><td><table><tr><td>Improved</td><td>24</td></tr><tr><td>Worse</td><td>6</td></tr><tr><td>Same</td><td>0</td></tr></table></td></tr>
<tr><td>2.b – Average Similarity – K = 10</td><td><table><tr><td>Improved</td><td>18</td></tr><tr><td>Worse</td><td>30</td></tr><tr><td>Same</td><td>4</td></tr></table></td><td><table><tr><td>Improved</td><td>16</td></tr><tr><td>Worse</td><td>14</td></tr><tr><td>Same</td><td>0</td></tr></table></td></tr>
</table>

Improved -> Number of queries whose average precision improved after clustering.
Worse -> Number of queries whose average precision improved after clustering.
Same -> Number of queries whose average precision did not change after clustering.

2.e

<u>K = 20 and highest similarity measure:</u>

The final ranking of the clusters also depends on the original ATC.ATC values. If the ATC.ATC has ranked the relevant documents with better rank, the clustering would just add similar documents which might be relevant to that higher ranked document, thus bettering the ranks for relevant documents. But if ATC.ATC has ranked some non-relevant documents with a better rank, the clustering would similar documents which might not be relevant to query and give them a better rank. This would bring the ranking of relevant documents down hence having negative impact. Now with highest similarity, the best document will make a cluster with other similar documents and all these documents will have a better rank with clustering. Now since not always, the best document is relevant document, this cluster of (non-relevant)documents will push the ranking of other documents thus effecting negatively on the MAP values. Similarly, if the original ATC.ATC rankings were good, relevant clustered documents would be clustered and got a better ranking. We can assume that clustering just amplifies the effect of ATC.ATC ranking. Also, since this is complete clustering, not always the most similar documents are picked. Since the CACM collection more generic compared to Medlars, the clusters are formed with relatively less similar documents. ATC.ATC without clustering for CACM collection had MAP value 0.3095 and with clustering is 0.3030. But ATC.ATC performed very well on Medlars collection, that is again reflected in clustered MAP values for Medlars. The MAP value increased by a marginal value.

<u>K = 10 and average similarity measure:</u>

With less number of clusters, each cluster will have relatively more number of documents which might impact on the final MAP values. So if the ATC.ATC is doing bad on a particular collection, clustering would again amplify this effect and might cluster non-relevant documents thus pushing down the rankings of relevant documents. An example of this is shown in the analysis part 2.f. We can see that for CACM collection, the MAP decreased from 0.3095 to 0.2828 and results of 30 queries performed badly with clustering.

2.f

For CACM Collection with Average similarity - K =10 the MAP value for Query 32 changes substantially.

Average precision without clustering: 0.447712 **Difference in average precision = 0.253268**

with clustering: 0.194444 **So, query 32 did worse with clustering**

For Query 32, the relevant documents are CACM-3139, CACM-1145 and CACM-0366. These documents appear in the rank 2, 3 and 17 respectively after atc.atc run. But the top most document in atc.atc run is not a relevant document. Clusters are formed after the run of complete linkage with average similarity. CACM-1791, CACM-2408 and CACM-1839(top most document in ATC.ATC run) together form a cluster since the vocabulary of CACM-1791 and CAC-2408 closely match with the document CACM-1839. But neither, CACM-1791 or CACM-2408 is the relevant document for query 32. What clustering assumes is that similar documents will behave similarly with respect to relevance to information need[1] Here, though we use average similarity, the final score for CACM-1791, CACM-2408 and CACM-1839 documents ((0.24 + 0.13 + 0.11)/ 3 = 0.16) is greater than the clustered documents (CACM-3139, CACM-3018, CACM-2619, CACM-3040, CACM-2059) where the relevant document is present. For the cluster which has the relevant document the average similarity score is 0.14. Hence documents CACM-1791, CACM-2408 and CACM-1839 are promoted to top 3 ranks pushing the CACM-3139 down by 2 positions. So in the original ATC.ATC run, CACM-3139 was in position 2, but after clustering it's in position 4.

Similar explanation can be given for the documents CACM-1145 and CACM-0366 which had better ranking in the ATC.ATC run but got pushed down in clustering due to the way clustering works. CACM-1145 was in position 3 in ATC.ATC and in position 9 in clustering. CACM-0366 was in position 17 in ATC.ATC run and in position 27 in clustering. So in this case, what clustering did was bought similar documents based on ATC.ATC rankings together thus pushing the relevant documents down. Thus average precision will be lower comparatively.

The highlighted rows are the relevant documents for the query 32 of CACM collection.

	Original ranking with similarity score			**Final Ranking with similarity score**
1	CACM-1839=0.24682754033902246		1	CACM-1839 = 0.166961446010045
2	CACM-3139=0.2204013915677121		2	CACM-1791=0.166961446010045
3	CACM-1145=0.19155683968355924		3	CACM-2408=0.166961446010045
4	CACM-2177=0.14126726641857296		4	CACM-3139=0.145997819925752
5	CACM-2275=0.13664841686752358		5	CACM-3018=0.145997819925752
6	CACM-2408=0.13432223999804338		6	CACM-2619=0.145997819925752
7	CACM-3018=0.13160183121031419		7	CACM-3040=0.145997819925752
8	CACM-1721=0.12964621643309254		8	CACM-2059=0.145997819925752
9	CACM-2619=0.12913491773572067		9	CACM-1145=0.142124431392169

10	CACM-3040=0.1289407530008598		10	CACM-1045=0.142124431392169
11	CACM-0222=0.12218827863951867		11	CACM-0222=0.142124431392169
12	CACM-2059=0.11991020611415562		12	CACM-2275=0.124755224882991
13	CACM-1791=0.11973455769306908		13	CACM-1721=0.124755224882991
14	CACM-0070=0.11520966458521698		14	CACM-2606=0.124755224882991
15	CACM-1905=0.11314296775122965		15	CACM-1504=0.114270414455132
16	CACM-1045=0.1126281758534294		16	CACM-2177=0.114270414455132
17	CACM-0366=0.11188099849062912		17	CACM-2155=0.114270414455132
18	CACM-1804=0.11022704998868672		18	CACM-2524=0.114270414455132
19	CACM-1529=0.11007662219756004		19	CACM-1804=0.110227049988686
20	CACM-1504=0.10871638391928354		20	CACM-1905=0.109012432659117
21	CACM-2606=0.107971041348359		21	CACM-1899=0.109012432659117
22	CACM-1563=0.10753927187621254		22	CACM-1529=0.109012432659117
23	CACM-2630=0.107164773408931		23	CACM-0070=0.108456677292176
24	CACM-2155=0.10711387486650638		24	CACM-2695=0.108456677292176
25	CACM-1954=0.10478986229941561		25	CACM-1563=0.107352022642571
26	CACM-2043=0.10463316607115275		26	CACM-2630=0.107352022642571
27	CACM-1899=0.1038177080285636		27	CACM-0366=0.105417457528091
28	CACM-2695=0.10170368999913669		28	CACM-1954=0.105417457528091
29	CACM-1576=0.10036580325116756		29	CACM-2043=0.105417457528091
30	CACM-2524=0.0999841326161683		30	CACM-1576=0.105417457528091

10 Clusters for the top 30 documents

Cluster1 = {CACM-1839, CACM-1791, CACM-2408}
Cluster2 = {CACM-3139, CACM-3018, CACM-2619, CACM-3040, CACM-2059}
Cluster3 = {CACM-1145, CACM-1044, CACM-0222}
Cluster4 = {CACM-2275, CACM-1721, CACM-2606}
Cluster5 = {CACM-0070, CACM-2695}

Cluster6 = {CACM-1905, CACM-1899, CACM-1529}
Cluster7 = {CACM-0366, CACM-1954, CACM-2043, CACM-1576}
Cluster8 = {CACM-1804}
Cluster9 = {CACM-1504, CACM-2177, CACM-2155, CACM-2524,}
Cluster10 = {CACM-1563, CACM-2630}

2.g **Cluster Hypothesis** states that documents that are clustered together behave similarly with respect to relevance to information needs.[1]

From the statistics in 2.d, we can see that for the CACM collection, the results for more than half of the number of queries are worse with clustering for both higher and average similarity. For CACM collection, the vocabulary is not too much tight, in the sense the there are more common and generic terms present. Complete linkage forms clusters with relatively less similar documents. Also, the original ATC.ATC MAP values itself are low. So in this case non-relevant documents are clustered and hence those rendered non relevant to information needs.

But for Medlars collection, the results for more than half of the number of queries have improved for both higher and average similarity. For Medlar collections, the vocabulary are more arcane and the cluster will more tightly bound that is more of the similar documents will be together and also the original ATC.ATC MAP values are high for MEDLAR collection.

Thus, we can accept the cluster hypothesis.

Part 3: Pseudo-relevance feedback

3.c

Parameters	CACM	MEDLAR
A=4, B=8, C=4, K=5	0.37643539154659145	0.5552409626971925
A=4, B=16, C=0, K=5	0.35684756136545914	0.5534035220735672

3.d

Parameters	CACM		MEDLAR	
A=4, B=8, C=4, K=5	Improved	34	Improved	21
	Got Worse	16	Got Worse	9
	Stayed Same	2	Stayed Same	0

A=4, B=16, C=0, K=5	Improved 32 Got Worse 15 Stayed Same 5	Improved 21 Got worse 8 Stayed Same 1

3.e

CACM with A=4, B=8, C=4, K=5
When we consider non relevant documents into account while deciding the terms to append in the query vector we will have both parameters for adding relevant document as well as removing non relevant documents from the query retrieval. Hence the overall precision will improve significantly. That's why we are getting a MAP score which is better than normal ATC.ATC score.

CACM with A=4, B=16, C=0, K=5
By setting C=0 we are not considering non relevant document into account for deciding the precision. Hence our MAP is getting reduced. But it is still better than normal ATC score as we have set B higher so more number of relevant documents will be considered while appending the terms in the query vector for rocchio relevance feedback calculation.

MEDLAR with A=4, B=8, C=4, K=5
With MEDLAR the overall MAP is higher because of less number of documents and smaller average query length. It is even better than Rocchio pseudo relevance as we are taking non relevant documents into account also while calculating the new terms to be appended in the query vector.

MEDLAR with A=4, B=16, C=0, K=5
There is a declination in MAP as compared to previous parameter setup because of C parameter. As we are not considering relevant documents into account for calculating the augmented rocchio query vector the overall precision will reduce, but with setting higher B we are trying to put more weight on relevant documents so the overall dip won't be that much.

3.f
while running with CACM collection with A=4,B=8,C=4,K=5 we noticed that **query 36**'s score is significantly improving from 0.149956 to 0.353152. Below are the query vector's ATC value query's rocchio weights, top 100 documents retrieved w.r.t ATC ,top 100 documents retrieved w.r.t rocchio relevance feedback and the relevant documents as per the original query.

Query
What is the type of a module? (I don't want the entire literature on Abstract Data Types here, but I'm not sure how to phrase this to avoid it. I'm interested in questions about how one can check that a module "matches" contexts in which it is used.)

Query ATC vector
{abstract=0.21856078090187955, check=0.23022500505203625, data=0.10356747579408569, don't=0.41144372453943084, type=0.21226903180940002, avoid=0.22603326837876675, literatur=0.2494559709813734, interest=0.2010806333665937, phrase=0.26136349267887365, context=0.19562818533876955, match=0.22470927863074497, question=0.2110356648819674, i'm=0.0, modul=0.34848465690516484, entir=0.23641085766670056, sure=0.37611350023932394}

Query Vector with rocchio weights

{satterthwait=1.8817185084535357, abstract=1.873820567899564, check=0.920900020208145, ca770802=1.8817185084535357, data=0.9826630742046234, geschk=1.8817185084535357, don't=1.6457748981577234, mesa'=1.8817185084535357, type=1.5730278746120463, avoid=0.904133073515067, mesa=2.63440591183495, literatur=0.9978238839254936, interest=0.8043225334663748, phrase=1.0454539707154946, context=0.7825127413550782, match=0.8988371145229799, question=0.8441426595278696, i'm=0.0, entir=0.9456434306668022, modul=2.8283389859931605, sure=0.3323311691152817}

Actual relevant documents as per the query

{CACM-2265 CACM-2558 CACM-2625 CACM-2632 CACM-2651 CACM-2868 CACM-2939 CACM-2940 CACM-2941 CACM-2956 CACM-2957 CACM-2958 CACM-2960 CACM-3031 CACM-3103 CACM-3150}

Top 100 documents as per query ATC weights

{CACM-1625=0.13361605843642824, CACM-2084=0.12817833607703746, CACM-2941=0.1280609423466922, CACM-2082=0.12413719104491998, CACM-1737=0.1215858347220591, CACM-3105=0.11890283573660743, CACM-2867=0.11864151161565167, CACM-2265=0.11643214062394436, CACM-1569=0.11619660789975433, CACM-1768=0.11325287384435417, CACM-1364=0.1120317592800612, CACM-3087=0.11157063536062242, CACM-2939=0.10436835814077097, CACM-0483=0.10269820704986835, CACM-2958=0.10100457937304042, CACM-1281=0.09994171793467105, CACM-3103=0.09993284194763705, CACM-1717=0.09959279876573364, CACM-2705=0.09916589315036713, CACM-2582=0.09629952408735491, CACM-3148=0.09607811045633073, CACM-0463=0.09435172749709127, CACM-1359=0.0916170630489419, CACM-3030=0.09150119478567012, CACM-2746=0.08805140395914395, CACM-1787=0.08722425172129568, CACM-2184=0.08642711319749373, CACM-1683=0.08466202931410491, CACM-2247=0.08304466595682068, CACM-2002=0.08293254031350349, CACM-2572=0.08290068775890128, CACM-2815=0.0825958876339278, CACM-0897=0.08163254143712634, CACM-3100=0.08123255801251414, CACM-2684=0.077050319430632, CACM-3031=0.07602471499050475, CACM-2356=0.07495928415587874, CACM-0583=0.07484127688777922, CACM-2859=0.07426629019465007, CACM-2940=0.07345318161921663, CACM-2651=0.07228975595052767, CACM-0820=0.07154679147887007, CACM-2960=0.07137808138964373, CACM-1907=0.07022138479413084, CACM-2470=0.0698158787901519, CACM-2109=0.06819589558032303, CACM-2537=0.06765404191822527, CACM-0144=0.0675500576131773, CACM-1806=0.06730246388857133, CACM-2127=0.0656178389718512, CACM-0024=0.0653174264737998, CACM-0167=0.06522955828880762, CACM-3201=0.0647592402102601, CACM-2575=0.06394991088182692, CACM-0847=0.0636893504787104, CACM-2898=0.06347109897298127, CACM-2530=0.06330098096024706, CACM-1698=0.06263985012561026, CACM-2798=0.06258568353557767, CACM-0242=0.06230798752126979, CACM-1595=0.06210897705355672, CACM-2957=0.061356303387573884, CACM-2497=0.061184032065454474, CACM-1930=0.060983140489793905, CACM-2113=0.06053016935888432, CACM-1012=0.06004355857347381, CACM-0740=0.059643354377383065, CACM-2469=0.058735136519101006, CACM-1033=0.057583776704592846, CACM-0670=0.0573138696213768, CACM-1601=0.057111184653365885, CACM-1614=0.056784993797004794, CACM-1939=0.056667635006845685, CACM-3162=0.05666700747691845, CACM-2030=0.05664341420860377, CACM-2931=0.056473181741682074, CACM-1087=0.05604312931349783, CACM-

0356=0.05603878800770461, CACM-1141=0.05556662069589758, CACM-0136=0.05492092516193659, CACM-1706=0.054822195477186475, CACM-2527=0.05455123860931811, CACM-1112=0.054511440259017885, CACM-2910=0.053966544665652726, CACM-1695=0.053295615147638545, CACM-1131=0.0528635842208316, CACM-1825=0.05265717169796234, CACM-1096=0.05149319206230583, CACM-2305=0.05119367619624776, CACM-2299=0.05117358192357542, CACM-1456=0.0511112678801824, CACM-1824=0.05108666328388787, CACM-2289=0.051063331737717405, CACM-2956=0.05106015685756135, CACM-0233=0.05082197508098115, CACM-1290=0.050622397426142024, CACM-2254=0.05033235601618377, CACM-0253=0.05030592905348219, CACM-2966=0.050228902712211726, CACM-0964=0.04999191949153896}

Top 100 documents as per rocchio augmented query

{CACM-2941=3.678076253861388, CACM-2867=0.9569294350456155, CACM-2958=0.8347328481493731, CACM-3103=0.8262741189686675, CACM-3105=0.8030462786753019, CACM-3030=0.7650008301926678, CACM-2939=0.7454599743160644, CACM-1364=0.6955028366336227, CACM-2265=0.676439229857673, CACM-2247=0.6739994477529431, CACM-3148=0.669142078633754, CACM-1359=0.663088822177659, CACM-2815=0.6622949477143562, CACM-2582=0.6398472779515604, CACM-3031=0.6282915811112902, CACM-2960=0.6239286248990459, CACM-2084=0.6140086554092058, CACM-2940=0.6106812035594226, CACM-2356=0.6083776187538347, CACM-2859=0.5885727225062038, CACM-2002=0.584729686548036, CACM-2184=0.5757074125331916, CACM-2082=0.5683862047178254, CACM-2705=0.5672649850898414, CACM-1768=0.5418172761405764, CACM-2898=0.524313332429273, CACM-3087=0.5198841631359646, CACM-2798=0.5169992142545268, CACM-0483=0.5101064876653725, CACM-2957=0.5078207523401257, CACM-2470=0.494477270349151, CACM-3100=0.48870315186160407, CACM-1737=0.4863433388882364, CACM-2497=0.4848936473369928, CACM-2651=0.4811290870338548, CACM-1087=0.48048313134565723, CACM-2469=0.4767006901822136, CACM-1281=0.4699052692342751, CACM-1569=0.46478643159901734, CACM-0233=0.4357198112811219, CACM-0253=0.43129551495348145, CACM-1787=0.429277403764399, CACM-2684=0.42910289636015697, CACM-1290=0.42128410469475, CACM-2289=0.40399299744961054, CACM-2956=0.4005732088412158, CACM-1717=0.39837119506293456, CACM-1907=0.3954266700411529, CACM-1323=0.3868750534933581, CACM-2931=0.38543634684500927, CACM-0463=0.37740690998836507, CACM-2812=0.3774019825444337, CACM-2113=0.36989973482817085, CACM-2738=0.35683253329222087, CACM-2937=0.35290423200032245, CACM-2746=0.3522056158365758, CACM-1614=0.34656816273281665, CACM-1683=0.33864811725641963, CACM-1033=0.3301914795274836, CACM-2243=0.32925228986002486, CACM-0897=0.32653016574850535, CACM-0676=0.3246889933742514, CACM-2299=0.32285931853316674, CACM-1678=0.32109755457422373, CACM-2244=0.3182785919664299, CACM-0100=0.3125539359631327, CACM-2972=0.31183273415717877, CACM-0457=0.3086362155693784, CACM-2379=0.30769694570414874, CACM-2254=0.30615610940895954, CACM-2959=0.3024056853054301, CACM-0583=0.2993651075511169, CACM-2527=0.2959139580523674, CACM-3171=0.2956790530409334, CACM-2030=0.2934783960931988, CACM-1112=0.2925841702399806, CACM-1700=0.29140281505003224, CACM-0794=0.2913008182788593, CACM-3133=0.2906963159508391, CACM-1749=0.2900439826033804, CACM-0329=0.28917424726972035, CACM-0820=0.28618716591548027, CACM-2242=0.2861293387072555, CACM-2241=0.2847191701948289, CACM-0618=0.28339695736511433, CACM-2722=0.2786539883081658, CACM-2109=0.27278358232129213, CACM-1625=0.2708139511706128, CACM-2537=0.2706161676729011, CACM-1012=0.2703051926565949, CACM-0492=0.27029050373426855, CACM-0144=0.2702002304527092, CACM-1806=0.26920985555428534, CACM-2032=0.26818707495864996, CACM-2950=0.2679675271266807, CACM-2060=0.2639840104651539, CACM-2701=0.2634664830864036, CACM-0549=0.26345420896998384, CACM-2127=0.2624713558874048, CACM-

0024=0.2612697058951992}

As we can see from the above data , ranks of the highlighted documents are getting improved with the rocchio feedback. So the retrieval is actually performing better by adding additional terms in the query vector.

3.g
As per the requirements mentioned in the question we took top 1 relevant and top 1 non relevant document into account while considering the rocchio weights. We feel always considering the top 1 document from each set might not work out. As an experiment we did the entire run considering all relevant and all non relevant documents. As a result we got MAP as below.

A=4,B=8,C=4,K=5
CACM: MEAN AVERAGE PRECISION:0.43147470047503933
MEDLAR: MEAN AVERAGE PRECISION:0.628758383206966

A=4,B=16,C=0,K=5
CACM: MEAN AVERAGE PRECISION:0.42619283919483575
MEDLAR: MEAN AVERAGE PRECISION:0.6442846167587912

Comparing these values with the single document setup we can clearly see a significant improvement. So it is not fair to look at top one document from both the set of relevant and non relevant documents.

Part 4: Retrieving with Document Collection on Disk

The pseudo code that operates on a large direct index collection can be written to disk by using RandomAccessFiles.

Direct indexing is used on IR collections as a supplementary index for relevance feedback,clustering etc.

To write the pseudocode for direct indexing, we need three data structures - term dictionary(token,token ID) , direct file and variable file. We also need term frequency component to calculate the relevance scores. The idf calculation is computed and stored with the other documents. For direct indexing, the docID is mapped to indexed vector of document. The output files are DirectFile <DocID,<offset,vector-length>> and VariableFile <DocID,<TokenID,freq>>. A unique docId is created using a integer hashing function on the pointer location, retrieved using seek(). Every word is parsed through the tokenizer where stopping and stemming is applied and inserts the token and the corresponding tokenID

The following assumptions can be made:

- All files are static
- Here, each direct info entry and variable info entry is of length 2 * sizeof(integer)
- The dictionary hashtable should be small enough to be stored in memory
- Structure for direct file is <offset,length> and for variable file is <tokenID,token freq>

**********Pseudocode for direct indexing using RandomAccessFiles**********

1) Retrieve query from user.
2) For each individual file(static)
3) Apply stopping and stemming to the query to obtain the root token words.
4) While not the end of the document
5) Create three RandomAccessFiles - FileWrite(variableFile), Filewriter(DirectFile), FileWriter(idfFile)
6) For all the documents in the corpus, read() N pages into a intermediary file of P pages initially:
    a) For each document in the corpus, apply stemming and stopping
    b) Find unique words that form all the vocabulary in all of the documents in the corpus and add it to the dictionary
7) For each word in the document:
    a) <tokenID, freq> is added to variableFile
        i) Then reset Vd as D*sizeof(integer) * Length of the index vextor.
        ii) Add the value to the Variable file entry <offset,vector-length>
    b) Update Filewrite(VariableFile) and write <DocID,<tokenID,freq>>
    c) Also, update Filewrite(DirectFile) and write <DocID, offset,vector-length>
    d) Now, calculate the idf value and store it for the term in the whole document frequency by adding an unique tokenID with token as <tokenID, token>.
8) Add all these idf values for each intermediate iteration for that particular word and store it as <tokenID,count>
9) Repeat idf and tf calculation steps till all of the intermediate calculations are written onto the file data structures - Direct File and Variable File
10) If these root words matches the <token> set from the tf-idf calculation i.e., <tokenID,token> data structure.
        i) Use seek() to get idf values from the data structure <tokenID> from idf calculation
        ii) FileRead(DirectFile) and FileRead(Variable File), and using seek()
            (1) Get <DocID, offset,vector-length> from Directfile
            (2) Depending on the offset value, we point to the particular location and using vector-length we get the length of the <tokenID,freq> pairs.
    b) Repeat till step 11) for every document where <tokenID> value matches <tokenID,freq> in the Variable file
        i) calculate the tf-idf for the corresponding token using a particular method(say atc.atc)
        ii) Store the value as <DocID, tf-idf relevance score>
11) Sort the <DocID, tf-idf relevance scores> to retrieve the required top N values needed.
12) Extract the Documents using the DocID's to get the most relevant documents
13) Close the Files

****

Data structures used:

- Record: a collection of data items or fields that all bear some relationship to one another.
- File: a collection of records
- ArrayList
- String
- Hashtables used in the term dictionary and document location
- RandomAccessFile
- BufferedWriter

All the temporary data structures are cleared for every iteration of the document and RandomAccessFile continues to read every document in the connection till all the files are read.

Part 5: Miscellaneous

Below are the contributions of individual team members for this project.

Alok: Wrote the code for calculating Pseudo relevance feedback and Rocchio relevance feedback and did the query analysis for them.

Nikhil: Wrote the code for complete link clustering and did the query analysis for that.

Prashanth: Wrote the pseudo code for extending the setup to retrieving with large document collection on disk

References:

[1] http://en.wikipedia.org/wiki/Cluster_hypothesis

http://en.wikipedia.org/wiki/Search_engine_indexing

http://www.coderookie.com/2006/tutorial/the-pseudocode-programming-process/

http://www.javaworld.com/article/2076333/java-web-development/use-a-randomaccessfile-to-build-a-low-level-database.html#resources to understand direct indexing on database