Nicholas Radtke
11/26/05
CSE 598/494 Information Integration

Project: Part C

Task 5: K-means

K-means Analysis

In order to analyze the performance of the K-means clustering algorithm, we consider two aspect:

1. The correspondence between the clusters and natural (human interpreted) categories.
2. The effects of varying the number of clusters k.

Each of these is discussed in its own section below.

Note that I am confident that my implementation of K-means forms correct clusters because:

1. I used incremental testing for all the tasks involved. In addition, I created simple test data and calculated the results using a pencil and paper. I then ran and analyzed these test cases on the code to verify it was working properly before developing the next part and moving on to the real data.
2. As shown below, my implementation seems to find meaningful clusters, reinforcing that the clusters are reasonable.

Correspondence Between Clusters and Natural Categories

To analyze the correspondence between the clusters provided by the K-means algorithm and natural categories (where natural refers to categories as interpreted by a human), we run three sample queries and then try to give the resulting categories "names", based on their contents. For all the queries, we use the top 64 (or as many as are returned, if fewer than 64 matches exists) results returned from a vector space search. We hold the number of groups k at 3 for simplicity, in hopes that we are able to name gross categories, rather than categories with minute differences. We also display all documents in each group, rather than just the top three, since it gives a better notion of what category the cluster really represents. Each document in a cluster includes its similarity to its respective centroid, sorted from highest to lowest similarity. Thus, highly similar documents will be given more weight when considering a name for the category. If no obvious category exists, we will say this rather than inventing a category.

The sample queries chosen, based on the domain of the crawl, are:

- parking and transit

- dps
- Lattie Coor

After running each query, we will analyze the cluster and natural category correspondence.

Query: parking and transit

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 4 iterations.
Aggregate cluster dissimilarity: 6.5278707

Group 0: Showing top 21 of 21 in cluster
Intra-cluster dissimilarity: 3.2509604
1. 0.757288 www.asu.edu%%dps%%pts%%event%%football.html
2. 0.7558395 www.asu.edu%%dps%%pts%%visitor%%regulations.html
3. 0.75381887 www.asu.edu%%dps%%pts%%visitor%%designated.html
4. 0.72590554 www.asu.edu%%dps%%pts%%event%%basketball.html
5. 0.7252273 www.asu.edu%%dps%%pts%%event%%cardinals.html
6. 0.72515756 www.asu.edu%%dps%%pts%%visitor%%media.html
7. 0.6714612 www.asu.edu%%dps%%pts%%event%%gammage.html
8. 0.6708616 www.asu.edu%%dps%%pts%%event%%wbasketball.html
9. 0.6466085 www.asu.edu%%asunews%%media_info%%parking.html
10. 0.62616605 www.asu.edu%%dps%%pts%%visitor%%meter.html
11. 0.6238922 www.asu.edu%%dps%%pts%%admin%%initiatives.html
12. 0.601227 www.east.asu.edu%%admin%%pts%%events%%index.htm
13. 0.5921144 www.asu.edu%%dps%%pts%%visitor%%disabled.html
14. 0.58761793 www.asu.edu%%dps%%pts%%event%%sports.html
15. 0.56579125 www.east.asu.edu%%admin%%pts%%visitors%%index.htm
16. 0.54954195 www.asu.edu%%dps%%pts%%service%%index.html
17. 0.5462415 www.east.asu.edu%%admin%%pts%%maps%%index.htm
18. 0.52412003 www.asu.edu%%dps%%pts%%visitor%%temp.html
19. 0.50903535 www.asu.edu%%asunews%%university%%
parking_construction_082403.htm
20. 0.4886849 www.asu.edu%%dps%%pts%%admin%%achievements.html
21. 0.36631703 construction.asu.edu%%..%%introduction%%Parking.shtml

Group 1: Showing top 11 of 11 in cluster
Intra-cluster dissimilarity: 2.3488612
1. 0.7215023 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.69346654 herbergercollege.asu.edu%%calendar%%directions.html
3. 0.65300256 www.asu.edu%%dps%%pts%%decals%%howto.html
4. 0.62773865 www.asu.edu%%dps%%pts%%decals%%options.html
5. 0.62243605 www.asu.edu%%dps%%pts%%faq.html
6. 0.5625434 www.asu.edu%%hr%%new_employee%%parking_decal.html
7. 0.55980223 www.east.asu.edu%%admin%%pts%%residences%%index.htm
8. 0.55826116 www.east.asu.edu%%admin%%pts%%appeals%%index.htm
```

```
9. 0.4029011 www.asu.edu%%aad%%manuals%%dps%%index.html
10. 0.35971546 www.asu.edu%%tour%%main%%towers.html
11. 0.34499595 www.asu.edu%%admissions%%steps%%parking.html

Group 2: Showing top 32 of 32 in cluster
Intra-cluster dissimilarity: 0.9280491
1. 0.95531714 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
2. 0.9537557 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.9428633 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
4. 0.9402671 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
5. 0.93702155 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
6. 0.9293429 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
7. 0.92903864 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
8. 0.9282424 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
9. 0.92373425 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
10. 0.9213455 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
11. 0.9208369 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
12. 0.91904664 www.asu.edu%%dps%%pts%%maps%%wfa.html
13. 0.9168156 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
14. 0.9161384 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
15. 0.9146358 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
16. 0.91377074 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
17. 0.91347176 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
18. 0.9112341 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
19. 0.9091442 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
20. 0.9036501 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
21. 0.9023945 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
22. 0.9023646 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
23. 0.90122676 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
24. 0.89889973 www.asu.edu%%dps%%pts%%maps%%visitormap.html
25. 0.8926463 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.8616015 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.8499535 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.7239227 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.65320444 www.asu.edu%%dps%%pts%%citation%%index.html
30. 0.600893 www.asu.edu%%dps%%pts%%admin%%index.html
31. 0.57822907 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
32. 0.5691376 www.asu.edu%%dps%%pts%%flash.html
```

Group 0 seems to correspond to the category of “instructions for visitor parking during events.” Group 1 is a little less clear, but seems to somewhat fit the category “parking decals,” as it includes FAQs about parking and buying decals, instructions how to buy decals, information about which lots require decals, etc. Group 2 is the easiest to name and could be called “maps.” Note that more documents in this category have higher similarities to the centroid than group 0 and group 1. No doubt, this influences the ease of giving this group a natural category.

Note that in this implementation of K-means, the initial centroids are chosen randomly by selecting k random documents from the vector space search and using them for the initial centroids. Because the K-means algorithm is sensitive to the initial centroids when forming clusters, we rerun this first query to see if we get different categories. Note that for brevity's sake, we will not do this on subsequent queries.

Query: parking and transit

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 6 iterations.
Aggregate cluster dissimilarity: 6.78096

Group 0: Showing top 22 of 22 in cluster
Intra-cluster dissimilarity: 4.6812973
1. 0.70814985 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.69583404 www.asu.edu%%dps%%pts%%visitor%%regulations.html
3. 0.6700875 www.asu.edu%%dps%%pts%%visitor%%media.html
4. 0.65571684 www.asu.edu%%dps%%pts%%visitor%%designated.html
5. 0.6231766 www.asu.edu%%asunews%%media_info%%parking.html
6. 0.6214984 www.asu.edu%%dps%%pts%%decals%%options.html
7. 0.61900353 herbergercollege.asu.edu%%calendar%%directions.html
8. 0.6087854 www.asu.edu%%dps%%pts%%faq.html
9. 0.6026495 www.east.asu.edu%%admin%%pts%%visitors%%index.htm
10. 0.59139717 www.asu.edu%%dps%%pts%%visitor%%meter.html
11. 0.5841731 www.east.asu.edu%%admin%%pts%%events%%index.htm
12. 0.58211213 www.asu.edu%%dps%%pts%%decals%%howto.html
13. 0.57614607 www.asu.edu%%dps%%pts%%visitor%%disabled.html
14. 0.55693305 www.east.asu.edu%%admin%%pts%%maps%%index.htm
15. 0.5406165 www.east.asu.edu%%admin%%pts%%appeals%%index.htm
16. 0.5193051 www.east.asu.edu%%admin%%pts%%residences%%index.htm
17. 0.48794377 www.asu.edu%%asunews%%university%%
parking_construction_082403.htm
18. 0.4878825 www.asu.edu%%hr%%new_employee%%parking_decal.html
19. 0.38033807 www.asu.edu%%aad%%manuals%%dps%%index.html
20. 0.3648494 construction.asu.edu%%..%%introduction%%Parking.shtml
21. 0.34387994 www.asu.edu%%admissions%%steps%%parking.html
22. 0.32179612 www.asu.edu%%tour%%main%%towers.html

Group 1: Showing top 6 of 6 in cluster
Intra-cluster dissimilarity: 0.2784325
1. 0.8913516 www.asu.edu%%dps%%pts%%event%%football.html
2. 0.86540914 www.asu.edu%%dps%%pts%%event%%cardinals.html
3. 0.8480794 www.asu.edu%%dps%%pts%%event%%basketball.html
4. 0.7950914 www.asu.edu%%dps%%pts%%event%%wbasketball.html
5. 0.69750994 www.asu.edu%%dps%%pts%%event%%gammage.html
6. 0.69677454 www.asu.edu%%dps%%pts%%event%%sports.html
```

```
Group 2: Showing top 36 of 36 in cluster
Intra-cluster dissimilarity: 1.8212303
1. 0.940609 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
2. 0.93860275 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
3. 0.9353517 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
4. 0.924186 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
5. 0.9203567 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
6. 0.91222835 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
7. 0.91179955 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
8. 0.90904456 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
9. 0.9071534 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
10. 0.9054971 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
11. 0.9049824 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
12. 0.9015175 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
13. 0.9013506 www.asu.edu%%dps%%pts%%maps%%wfa.html
14. 0.89936334 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
15. 0.89861923 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
16. 0.898526 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
17. 0.89783144 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
18. 0.89255506 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
19. 0.89100796 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
20. 0.8908886 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
21. 0.89020556 www.asu.edu%%dps%%pts%%maps%%visitormap.html
22. 0.88837284 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
23. 0.8855815 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
24. 0.88402677 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
25. 0.8720759 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.8469881 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.83025247 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.73554814 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.6890625 www.asu.edu%%dps%%pts%%citation%%index.html
30. 0.6444359 www.asu.edu%%dps%%pts%%admin%%index.html
31. 0.6005662 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
32. 0.5930439 www.asu.edu%%dps%%pts%%flash.html
33. 0.58478093 www.asu.edu%%dps%%pts%%service%%index.html
34. 0.53751075 www.asu.edu%%dps%%pts%%admin%%initiatives.html
35. 0.5081316 www.asu.edu%%dps%%pts%%admin%%achievements.html
36. 0.46583 www.asu.edu%%dps%%pts%%visitor%%temp.html
```

In this case, two of the groups are very clear categories while one is not. Group 0 is the hardest to define. It's not a particularly tight cluster, based on individual document's similarities to the centroid. It seems to consist of instructions regarding who can park where (i.e. media parking, visitor parking, student parking, disabled parking, etc.). Group 1 is a tighter cluster and can be given the title “event parking” or, ignoring the document 5, perhaps even “sports event parking.” Group 2 is once again “maps.” Since maps appeared in both this and the previous version of the “parking and transit” query, it indicates that this is probably a relatively solid cluster. That is, the cluster is relatively tight itself and likely far away from other potential clusters (low intra-cluster distance and high inter-cluster distance).

Query: dps

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: dps

Vector space search returned 159 hits. Using top 64.
K-means algorithm required 2 iterations.
Aggregate cluster dissimilarity: 11.530652

Group 0: Showing top 16 of 16 in cluster
Intra-cluster dissimilarity: 6.1464334
1. 0.6920688 www.asu.edu%%aad%%manuals%%acd%%acd123.html
2. 0.6282302 www.asu.edu%%aad%%manuals%%spp%%spp812.html
3. 0.5286359 www.asu.edu%%workingatasu%%index%%index.html
4. 0.48019648 www.west.asu.edu%%asuw2%%staff.shtml
5. 0.4588555 www.west.asu.edu%%asuw2%%faculty.shtml
6. 0.39676055 www.asu.edu%%emergency%%resources.html
7. 0.39409256 www.asu.edu%%aad%%manuals%%sta%%sta104-02.html
8. 0.3940505 www.asu.edu%%asuremembers%%resources.htm
9. 0.38811955 property.asu.edu%%\equipment\maintenance.html
10. 0.37193608 www.west.asu.edu%%asuw2%%students.shtml
11. 0.34289587 www.east.asu.edu%%admin%%facilities.htm
12. 0.32062563 www.asu.edu%%provost%%asenate%%problem.html
13. 0.28945607 www.asu.edu%%president%%cet%%ref.htm
14. 0.22625567 www.asu.edu%%provost%%committees%%ASUWCSC.html
15. 0.20813024 www.asu.edu%%aad%%manuals%%manual-coordinators.html
16. 0.20176598 www.asu.edu%%provost%%committees%%CRBG.html

Group 1: Showing top 35 of 35 in cluster
Intra-cluster dissimilarity: 1.9863174
1. 0.9473797 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
2. 0.94720715 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
3. 0.9383118 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
4. 0.93347037 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
5. 0.9298689 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
6. 0.921805 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
7. 0.9217556 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
8. 0.91983753 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
9. 0.9157982 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
10. 0.91420794 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
11. 0.9126218 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
12. 0.9113253 www.asu.edu%%dps%%pts%%maps%%wfa.html
13. 0.9110815 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
14. 0.9080186 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
15. 0.9067915 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
16. 0.90512204 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
17. 0.9047928 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
18. 0.9034383 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
19. 0.90310305 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
```

```
20. 0.89732814 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
21. 0.89588404 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
22. 0.8947091 www.asu.edu%%dps%%pts%%maps%%visitormap.html
23. 0.8942362 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
24. 0.8928967 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
25. 0.8835234 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.8552143 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.84121025 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.73169255 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.6707524 www.asu.edu%%dps%%pts%%citation%%index.html
30. 0.617856 www.asu.edu%%dps%%pts%%admin%%index.html
31. 0.5914302 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
32. 0.5817246 www.asu.edu%%dps%%pts%%flash.html
33. 0.5620612 www.asu.edu%%dps%%pts%%service%%index.html
34. 0.49068108 www.asu.edu%%dps%%pts%%decals%%removal.html
35. 0.21082926 www.asu.edu%%xed%%linkto%%index.html

Group 2: Showing top 13 of 13 in cluster
Intra-cluster dissimilarity: 3.3979018
1. 0.7011615 www.asu.edu%%dps%%index.html
2. 0.60237515 westcgi.west.asu.edu%%dps%%index.htm
3. 0.5992695 www.asu.edu%%dps%%pts%%psac%%Code%20Perspective.html
4. 0.56609607 www.asu.edu%%tour%%main%%dps.html
5. 0.55494857 www.asu.edu%%emergency%%emergency.html
6. 0.5448182 www.asu.edu%%aad%%manuals%%dps%%index.html
7. 0.50987166 www.asu.edu%%provost%%committees%%PSAC.html
8. 0.5088827 www.asu.edu%%dps%%police%%police.htm
9. 0.49791124 www.asu.edu%%dps%%pts%%psac%%cms.html
10. 0.4781748 www.asu.edu%%dps%%police%%directory.htm
11. 0.4690917 www.asu.edu%%dps%%police%%keys.htm
12. 0.3623961 www.asu.edu%%dps%%orgchart.htm
13. 0.16205357 www.asu.edu%%dps%%mission.htm
```

Group 0 seems to be the category “manuals and resources available at ASU.” Group 1 is, once again “maps,” indicating there is quite a bit of overlap between parking and transit and dps. Finally, group 2 seems to mostly be pages “about dps,” although arguably some of these pages fit better into the “manuals and resources” category.

Query: Lattie Coor

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: lattie coor

Vector space search returned 230 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 20.748936

Group 0: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.5067873
```

```
1. 0.91201144 www.asu.edu%%clas%%shs%%pages%%clinic.htm
2. 0.6611179 www.asu.edu%%clas%%shs%%pages%%deptinfo.htm
3. 0.38015786 www.eoaa.asu.edu%%policy.htm

Group 1: Showing top 27 of 27 in cluster
Intra-cluster dissimilarity: 12.34062
1. 0.64245313 www.asu.edu%%it%%fyi%%sites%%annualschedule.html
2. 0.6337979 www.asu.edu%%it%%fyi%%sites%%coorcomputingcommons.html
3. 0.55013746 www.asu.edu%%it%%fyi%%sites%%maps.html
4. 0.5249831 www.asu.edu%%it%%fyi%%sites%%custwebform.html
5. 0.52450013 www.asu.edu%%it%%fyi%%sites%%gwc185.html
6. 0.5237952 www.asu.edu%%it%%fyi%%sites%%cpcomatrium.html
7. 0.5114331 www.asu.edu%%it%%fyi%%sites%%bac16.html
8. 0.50274813 www.asu.edu%%it%%fyi%%sites%%index.html
9. 0.48984128 www.asu.edu%%feature%%coor.html
10. 0.47546884 www.asu.edu%%it%%fyi%%sites%%ecg150.html
11. 0.42655322 www.asu.edu%%it%%fyi%%sites%%teamwork.html
12. 0.42214954 www.asu.edu%%asunews%%university%%
coorhalldedicate_010904.htm
13. 0.3660191 www.asu.edu%%tour%%main%%coor.html
14. 0.3201793 www.asu.edu%%tour%%main%%a-z.html
15. 0.31895584 clasdean.la.asu.edu%%student%%resources%%advising%%
deptAdvisors.htm
16. 0.23588799 aspin.asu.edu%%asura%%calendar.htm
17. 0.23005514 www.asu.edu%%clas%%sociology%%colloquium%%colloquium.html
18. 0.21166243 aspin.asu.edu%%asura%%current.htm
19. 0.18602714 www.asu.edu%%clas%%shs%%pages%%advising.htm
20. 0.17961478 www.kaet.asu.edu%%cet%%index.htm
21. 0.1769862 www.asu.edu%%clas%%philosophy%%staff.htm
22. 0.17301954 www.asu.edu%%alumni%%oldmain%%information.html
23. 0.16480447 www.asu.edu%%clas%%chicana%%index.html
24. 0.16079684 math.asu.edu%%~chavez%%seminar1.html
25. 0.13625959 www.asu.edu%%clas%%philosophy%%colloquia.htm
26. 0.12613964 www.asu.edu%%ia%%cleanandbeautiful%%trashy.html
27. 0.11722561 www.asu.edu%%lib%%archives%%preslist.htm

Group 2: Showing top 34 of 34 in cluster
Intra-cluster dissimilarity: 7.9015293
1. 0.9898117 www.asu.edu%%news_to_know%%academics%%wilkinson_100703.htm
2. 0.9898117 www.asu.edu%%news_to_know%%academics%%
todd_gitlin_011604.htm
3. 0.9898117 www.asu.edu%%news_to_know%%arts%%artventures_021704.htm
4. 0.9898117 www.asu.edu%%news_to_know%%arts%%dartII_022004.htm
5. 0.9898117 www.asu.edu%%news_to_know%%academics%%wuthrich_lecture.htm
6. 0.9898117 www.asu.edu%%news_to_know%%academics%%
levy_lecture_013004.htm
7. 0.9898117 www.asu.edu%%news_to_know%%academics%%
construction_conference_012004.htm
8. 0.9898117 www.asu.edu%%news_to_know%%academics%%online_eng_020404.htm
9. 0.9898117 www.asu.edu%%news_to_know%%arts%%ceramictour_012604.htm
10. 0.9898117 www.asu.edu%%news_to_know%%academics%%
mba_online_012704.htm
11. 0.9898117 www.asu.edu%%news_to_know%%arts%%ceramics_gala_012304.htm
```

```
12. 0.9898117 www.asu.edu%%news_to_know%%academics%%
urbaneco_symposium_022004.htm
13. 0.9898117 www.asu.edu%%news_to_know%%academics%%
stupidcupid_021004.htm
14. 0.9898117 www.asu.edu%%news_to_know%%arts%%buyarug_012204.htm
15. 0.9898117 www.asu.edu%%news_to_know%%academics%%
usaid_grants_021304.htm
16. 0.9898117 www.asu.edu%%news_to_know%%academics%%statedept_021704.htm
17. 0.9898117 www.asu.edu%%news_to_know%%academics%%
kelloggrant_01260.htm
18. 0.9898117 www.asu.edu%%news_to_know%%academics%%
lewis_franklecture_01204.htm
19. 0.989717 www.asu.edu%%news_to_know%%newsknow_index.htm
20. 0.558534 www.asu.edu%%asunews%%university%%university_index.htm
21. 0.4665401 www.asu.edu%%asunews%%archives%%jan2004.htm
22. 0.36716142 www.asu.edu%%asunews%%media_info%%history.htm
23. 0.33157432 www.asu.edu%%asunews%%archives%%apr2002.htm
24. 0.31702572 www.asu.edu%%asunews%%archives%%may2002.htm
25. 0.29698312 www.asu.edu%%asunews%%archives%%jan2002.htm
26. 0.28489006 www.asu.edu%%asunews%%university%%
east_convocation_051403.htm
27. 0.2672035 www.asu.edu%%asunews%%arts%%lattiehall_020304.htm
28. 0.2568491 www.asu.edu%%asunews%%university%%coorhall_041803.htm
29. 0.25110784 www.asu.edu%%asunews%%university%%
art_arch_coorhall_011304.htm
30. 0.21567634 herbergercollege.asu.edu%%public_art%%public_art.html
31. 0.19086759 www.asu.edu%%ia%%inauguration%%address%%introduction.htm
32. 0.18901657 www.east.asu.edu%%ia%%html%%journeys.html
33. 0.1470463 herbergercollege.asu.edu%%keys%%index.html
34. 0.1080644 www.asu.edu%%clas%%dll%%kor%%events%%eindex.htm
```

Group 0 is very small with only three document in it, all of which seem to have very little to do with each other. Therefore, we do not believe group 0 corresponds to a natural category. Group 1 seems to correspond to “the Lattie Coor building,” although many of the top entries correspond to “the Lattie Coor computing site,” indicating this cluster may have benefited from being split into two subgroups. Group 2 corresponds to “news.” A little further digging shows that most of the documents include the announcement for the following event: “Art and architecture intersect at Lattie F. Coor Hall,” which perhaps would make this a better title for the category. Note that the top 18 documents in this cluster have identical similarity to the centroid; that is because these 18 documents, although they have different URLs, are the same document.

Note that in many of the above cases, the grouping seems to be based on the URL. This can easily be explained based on how people build web pages. For a given directory on a given server, it is likely that similar documents are contained within this directory. The point of directories, after all, is to provide a level of structure to a file system. Thus, humans tend to collect similar files within a single directory. Furthermore, often all the web pages within a certain department/company/group/etc. are all built using an identical template (i.e. menu on the side, contact info at the bottom, etc.), which leads k-means to cluster documents of the domain into a single cluster. A good example of this is seen in the “maps” category in the “parking and transit” and “dps” query results, where the pages are essentially identical except for a graphic representing the specific map.

Thus, after attempting to name natural categories of clusters for three queries, we conclude that the K-means clustering is producing decent natural categories. Some of the categories were a bit of a stretch of the imagination, and one of the categories could not be named, but in general there was a semblance of order in the clusters produced. In cases where the categories were poor, part of the problem could be the value of k – that is K-means was asked to produce too many or too few clusters for the underlying natural categories, resulting in bad entries within a cluster (k too low) or random clusters (k too high).

Varying the Number of Clusters

To analyze the effects of changing the number of clusters, we will take the same three queries and run them on k=3, 6, and 10. As before, we will hold the number of results to use from vector space search constant at 64. For each query, we will measure the following aspects:

- The correspondence of clusters to natural categories
- The aggregate dissimilarity
- The execution times

We define the aggregate dissimilarity as follows: Let the dissimilarity of a document with respect to its centroid be 1-similarity of the document to its respective centroid. The aggregate dissimilarity is the sum of the squares of the dissimilarities for all documents to their respective centroids. A lower dissimilarity corresponds to tighter clusters. A higher dissimilarity corresponds to larger intra-cluster distances. Thus, we are interested in minimizing the aggregate dissimilarity.

We measure the correspondence of the clusters to natural categories for the following reason: Presumably, for a given search, there is a good number of clusters such that each cluster represents a natural category. Too few clusters will mean that multiple natural categories fall into a single cluster. Too many clusters will cause natural categories to be split into multiple clusters. To measure this, we will attempt to name the natural category of each cluster. To keep the defining of categories simple, we will only consider the top 3 documents (or fewer, if there are not 3) in each cluster when naming the category. We also reserve the category “misc” to correspond to clusters that don't seem to have a natural category. If we end up with multiple clusters in the same category, chances are k is set too high. On the other hand, if we end up with many “misc” categories, it might indicate that the number of clusters is set too low (i.e. splitting the cluster again may produce natural categories). It also might indicate a bad choice of initial centroids for the k-means algorithm. Finally, we count the number of clusters that only have a single document. While these documents might be outliers that really belong to their own category, if there are many such groups, it probably indicates too many clusters are being used. That is, we are losing the utility of clustering because our clusters are tending to only have one document in them.

Finally, we measure the amount of time each query takes to see if changing the number of clusters significantly impacts execution time.

The queries and mapping of clusters to natural categories are listed below:

Query: parking and transit, k=3

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 5 iterations.
Aggregate cluster dissimilarity: 6.839459

Group 0: Showing top 3 of 25 in cluster
Intra-cluster dissimilarity: 5.4506326
1. 0.70760065 www.asu.edu%%dps%%pts%%visitor%%regulations.html
2. 0.69559467 www.east.asu.edu%%admin%%pts%%faq%%index.htm
3. 0.67353135 www.asu.edu%%dps%%pts%%visitor%%media.html

Group 1: Showing top 3 of 33 in cluster
Intra-cluster dissimilarity: 1.1103941
1. 0.953461 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
2. 0.95268315 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.94294447 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html

Group 2: Showing top 3 of 6 in cluster
```

```
Intra-cluster dissimilarity: 0.27843234
1. 0.8913517 www.asu.edu%%dps%%pts%%event%%football.html
2. 0.86540926 www.asu.edu%%dps%%pts%%event%%cardinals.html
3. 0.8480794 www.asu.edu%%dps%%pts%%event%%basketball.html
```

Group Number	Natural Category
0	parking info and regulations
1	maps
2	sports event parking

Query: parking and transit, k=6

```
Using up to top 64 documents
Number of clusters (k) is 6
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 5.3996134

Group 0: Showing top 3 of 12 in cluster
Intra-cluster dissimilarity: 1.7711482
1. 0.8178613 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.74004984 www.asu.edu%%dps%%pts%%decals%%options.html
3. 0.7310724 www.asu.edu%%dps%%pts%%decals%%howto.html

Group 1: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.13811338
1. 0.8998055 www.asu.edu%%dps%%pts%%admin%%initiatives.html
2. 0.7803267 www.asu.edu%%dps%%pts%%admin%%achievements.html
3. 0.71747905 www.asu.edu%%dps%%pts%%admin%%index.html

Group 2: Showing top 3 of 16 in cluster
Intra-cluster dissimilarity: 2.6474712
1. 0.718759 www.asu.edu%%dps%%pts%%event%%football.html
2. 0.6992005 www.asu.edu%%dps%%pts%%visitor%%regulations.html
3. 0.69573873 www.asu.edu%%dps%%pts%%visitor%%designated.html

Group 3: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.0012829283
1. 0.97734165 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
2. 0.97225964 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html

Group 4: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.07287244
1. 0.8759156 www.asu.edu%%dps%%pts%%citation%%index.html
2. 0.8330706 www.asu.edu%%dps%%pts%%flash.html
3. 0.82792425 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
```

```
Group 5: Showing top 3 of 28 in cluster
Intra-cluster dissimilarity: 0.7687253
1. 0.9574642 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
2. 0.9548224 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.94457376 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
```

Group Number	Natural Category
0	parking decals
1	parking and transit goals
2	visitor parking
3	golf course maps
4	misc
5	parking maps

Query: parking and transit, k=10

```
Using up to top 64 documents
Number of clusters (k) is 10
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 4 iterations.
Aggregate cluster dissimilarity: 3.698783

Group 0: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 0.74273163
1. 0.7825871 www.asu.edu%%dps%%pts%%admin%%initiatives.html
2. 0.7528485 www.asu.edu%%dps%%pts%%citation%%index.html
3. 0.7475325 www.asu.edu%%dps%%pts%%faq.html

Group 1: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 1.4210855E-14
1. 1.0000001 construction.asu.edu%%..%%introduction%%Parking.shtml

Group 2: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 1.4210855E-14
1. 1.0000001 www.asu.edu%%tour%%main%%towers.html

Group 3: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 1.1574967
1. 0.83529973 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.76269346 www.asu.edu%%dps%%pts%%decals%%howto.html
3. 0.74349254 www.asu.edu%%dps%%pts%%decals%%options.html

Group 4: Showing top 3 of 6 in cluster
Intra-cluster dissimilarity: 0.27843246
1. 0.8913517 www.asu.edu%%dps%%pts%%event%%football.html
```

```
2. 0.86540926 www.asu.edu%%dps%%pts%%event%%cardinals.html
3. 0.8480794 www.asu.edu%%dps%%pts%%event%%basketball.html

Group 5: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 0.78703564
1. 0.8076815 www.asu.edu%%dps%%pts%%visitor%%designated.html
2. 0.80394393 www.asu.edu%%dps%%pts%%visitor%%regulations.html
3. 0.75440294 www.asu.edu%%dps%%pts%%visitor%%media.html

Group 6: Showing top 3 of 15 in cluster
Intra-cluster dissimilarity: 0.0646056
1. 0.9662652 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
2. 0.9580682 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
3. 0.958015 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html

Group 7: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.0012829283
1. 0.97734165 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
2. 0.97225964 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html

Group 8: Showing top 3 of 4 in cluster
Intra-cluster dissimilarity: 0.55735
1. 0.91946226 herbergercollege.asu.edu%%calendar%%directions.html
2. 0.60816514 www.asu.edu%%asunews%%media_info%%parking.html
3. 0.573718 www.asu.edu%%dps%%pts%%visitor%%meter.html

Group 9: Showing top 3 of 11 in cluster
Intra-cluster dissimilarity: 0.10984792
1. 0.9658493 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2. 0.95496017 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
3. 0.9442444 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
```

Group Number	Natural Category
0	misc
1	parking at DEWSC
2	parking at the Towers Complex
3	parking decals
4	sports event parking
5	visitor parking
6	maps
7	golf course maps
8	visitor parking
9	maps

Query: dps, k=3

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: dps

Vector space search returned 159 hits. Using top 64.
K-means algorithm required 7 iterations.
Aggregate cluster dissimilarity: 11.618886

Group 0: Showing top 3 of 34 in cluster
Intra-cluster dissimilarity: 1.3505524
1. 0.9508654 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
2. 0.9506304 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.94160867 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html

Group 1: Showing top 3 of 10 in cluster
Intra-cluster dissimilarity: 2.6072195
1. 0.6389543 www.west.asu.edu%%asuw2%%faculty.shtml
2. 0.61735666 www.asu.edu%%emergency%%resources.html
3. 0.6167311 www.west.asu.edu%%asuw2%%staff.shtml

Group 2: Showing top 3 of 20 in cluster
Intra-cluster dissimilarity: 7.661114
1. 0.64618975 www.asu.edu%%aad%%manuals%%acd%%acd123.html
2. 0.57400775 www.asu.edu%%aad%%manuals%%spp%%spp812.html
3. 0.54587007 www.asu.edu%%dps%%index.html
```

Group Number	Natural Category
0	maps
1	contact info
2	manuals

Query: dps, k=6

```
Using up to top 64 documents
Number of clusters (k) is 6
Searching for: dps

Vector space search returned 159 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 10.969845

Group 0: Showing top 3 of 15 in cluster
Intra-cluster dissimilarity: 4.9630566
1. 0.6056353 www.asu.edu%%dps%%index.html
2. 0.55804425 www.asu.edu%%tour%%main%%dps.html
3. 0.5242126 www.asu.edu%%dps%%pts%%psac%%Code%20Perspective.html

Group 1: Showing top 3 of 4 in cluster
```

```
Intra-cluster dissimilarity: 0.012518359
1. 0.95325655 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
2. 0.9528377 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
3. 0.94562536 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html

Group 2: Showing top 3 of 13 in cluster
Intra-cluster dissimilarity: 0.7316461
1. 0.9610993 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
2. 0.9560767 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
3. 0.9543053 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html

Group 3: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 1.1317661
1. 0.7279288 www.asu.edu%%dps%%pts%%citation%%index.html
2. 0.71344525 www.asu.edu%%dps%%pts%%admin%%index.html
3. 0.66718155 www.asu.edu%%dps%%pts%%service%%index.html

Group 4: Showing top 3 of 12 in cluster
Intra-cluster dissimilarity: 0.11784662
1. 0.96586573 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2. 0.9543757 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
3. 0.9440208 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html

Group 5: Showing top 3 of 12 in cluster
Intra-cluster dissimilarity: 4.0130105
1. 0.7593818 www.asu.edu%%aad%%manuals%%acd%%acd123.html
2. 0.67527807 www.asu.edu%%aad%%manuals%%spp%%spp812.html
3. 0.53690827 www.asu.edu%%workingatasu%%index%%index.html
```

Group Number	Natural Category
0	misc
1	maps
2	maps
3	parking and transit contact info
4	maps
5	misc

Query: dps, k=10

```
Using up to top 64 documents
Number of clusters (k) is 10
Searching for: dps

Vector space search returned 159 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 6.1977224
```

```
Group 0: Showing top 3 of 15 in cluster
Intra-cluster dissimilarity: 0.06951092
1. 0.96624655 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
2. 0.9580831 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
3. 0.95802796 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html

Group 1: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.20460932
1. 0.9179465 www.asu.edu%%dps%%orgchart.htm
2. 0.55516684 www.asu.edu%%dps%%mission.htm

Group 2: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 2.1175501
1. 0.6493491 www.asu.edu%%workingatasu%%index%%index.html
2. 0.6001948 www.west.asu.edu%%asuw2%%staff.shtml
3. 0.5875635 www.west.asu.edu%%asuw2%%faculty.shtml

Group 3: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.0012665294
1. 0.9751705 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
2. 0.9745044 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html

Group 4: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.16445579
1. 0.84863037 www.asu.edu%%emergency%%resources.html
2. 0.79677194 www.asu.edu%%asuremembers%%resources.htm
3. 0.6833908 www.asu.edu%%president%%cet%%ref.htm

Group 5: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 1.3501587
1. 0.7418827 www.asu.edu%%dps%%index.html
2. 0.63245976 westcgi.west.asu.edu%%dps%%index.htm
3. 0.6099873 www.asu.edu%%emergency%%emergency.html

Group 6: Showing top 3 of 17 in cluster
Intra-cluster dissimilarity: 0.89944154
1. 0.93845975 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2. 0.91560894 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
3. 0.9110874 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html

Group 7: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.33274406
1. 0.9538696 www.asu.edu%%aad%%manuals%%acd%%acd123.html
2. 0.88714886 www.asu.edu%%aad%%manuals%%spp%%spp812.html
3. 0.43619093 www.asu.edu%%aad%%manuals%%sta%%sta104-02.html

Group 8: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 5.684342E-14
1. 1.0000002 www.asu.edu%%xed%%linkto%%index.html

Group 9: Showing top 3 of 5 in cluster
Intra-cluster dissimilarity: 1.0579855
1. 0.78593427 www.asu.edu%%dps%%pts%%psac%%Code%20Perspective.html
```

```
2. 0.6851231 www.asu.edu%%provost%%committees%%PSAC.html
3. 0.6266238 www.asu.edu%%dps%%pts%%psac%%cms.html
```

Group Number	Natural Category
0	maps
1	misc
2	resources and contact info
3	Noble Science Library maps
4	resources and contact info
5	emergency
6	maps
7	manuals
8	resources and contact info
9	misc

Query: Lattie Coor, k=3

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: lattie coor

Vector space search returned 230 hits. Using top 64.
K-means algorithm required 9 iterations.
Aggregate cluster dissimilarity: 18.818779

Group 0: Showing top 3 of 20 in cluster
Intra-cluster dissimilarity: 8.178141
1. 0.6357145 www.asu.edu%%feature%%coor.html
2. 0.62509644 www.asu.edu%%asunews%%university%%
coorhalldedicate_010904.htm
3. 0.5878861 www.asu.edu%%tour%%main%%coor.html

Group 1: Showing top 3 of 10 in cluster
Intra-cluster dissimilarity: 1.4989594
1. 0.81129295 www.asu.edu%%it%%fyi%%sites%%annualschedule.html
2. 0.64802027 www.asu.edu%%it%%fyi%%sites%%maps.html
3. 0.6247444 www.asu.edu%%it%%fyi%%sites%%cpcomatrium.html

Group 2: Showing top 3 of 34 in cluster
Intra-cluster dissimilarity: 9.141678
1. 0.99231124 www.asu.edu%%news_to_know%%arts%%buyarug_012204.htm
2. 0.99231124 www.asu.edu%%news_to_know%%arts%%ceramics_gala_012304.htm
3. 0.99231124 www.asu.edu%%news_to_know%%academics%%
kelloggrant_01260.htm
```

Group Number	Natural Category
0	Lattie Coor Hall dedication
1	computing sites
2	art and architecture at Lattie Coor Hall

Query: Lattie Coor, k=6

```
Using up to top 64 documents
Number of clusters (k) is 6
Searching for: lattie coor

Vector space search returned 230 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 14.685906

Group 0: Showing top 3 of 10 in cluster
Intra-cluster dissimilarity: 1.9920292
1. 0.8149367 www.asu.edu%%asunews%%arts%%lattiehall_020304.htm
2. 0.8001178 www.asu.edu%%asunews%%university%%coorhall_041803.htm
3. 0.79283893 www.asu.edu%%asunews%%university%%
art_arch_coorhall_011304.htm

Group 1: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.48296034
1. 0.7944422 www.asu.edu%%asunews%%archives%%apr2002.htm
2. 0.75630033 www.asu.edu%%asunews%%archives%%may2002.htm
3. 0.38249147 www.asu.edu%%ia%%cleanandbeautiful%%trashy.html

Group 2: Showing top 3 of 8 in cluster
Intra-cluster dissimilarity: 3.2552001
1. 0.76973814 www.asu.edu%%it%%fyi%%sites%%annualschedule.html
2. 0.43468606 www.asu.edu%%feature%%coor.html
3. 0.39237702 www.asu.edu%%clas%%shs%%pages%%clinic.htm

Group 3: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.3672876
1. 0.7292059 www.asu.edu%%clas%%sociology%%colloquium%%colloquium.html
2. 0.6677579 www.asu.edu%%clas%%philosophy%%colloquia.htm
3. 0.57154536 www.asu.edu%%clas%%philosophy%%staff.htm

Group 4: Showing top 3 of 31 in cluster
Intra-cluster dissimilarity: 8.244655
1. 0.9935666 www.asu.edu%%news_to_know%%arts%%ceramics_gala_012304.htm
2. 0.9935666 www.asu.edu%%news_to_know%%arts%%ceramictour_012604.htm
3. 0.9935666 www.asu.edu%%news_to_know%%academics%%kelloggrant_01260.htm

Group 5: Showing top 3 of 9 in cluster
Intra-cluster dissimilarity: 0.34377477
```

```
1. 0.8674054 www.asu.edu%%it%%fyi%%sites%%maps.html
2. 0.84589446 www.asu.edu%%it%%fyi%%sites%%cpcomatrium.html
3. 0.84106576 www.asu.edu%%it%%fyi%%sites%%index.html
```

Group Number	Natural Category
0	art and architecture at Lattie Coor Hall
1	misc
2	misc
3	offices/classes in Lattie Coor Hall
4	art and architecture at Lattie Coor Hall
5	computing sites

Query: Lattie Coor, k=10

```
Using up to top 64 documents
Number of clusters (k) is 10
Searching for: lattie coor

Vector space search returned 230 hits. Using top 64.
K-means algorithm required 4 iterations.
Aggregate cluster dissimilarity: 11.717463

Group 0: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 5.684342E-14
1. 1.0000002 math.asu.edu%%~chavez%%seminar1.html

Group 1: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.0650512
1. 0.83768755 www.asu.edu%%asunews%%archives%%apr2002.htm
2. 0.80326194 www.asu.edu%%asunews%%archives%%may2002.htm

Group 2: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 1.4210855E-14
1. 1.0000001 www.asu.edu%%alumni%%oldmain%%information.html

Group 3: Showing top 3 of 4 in cluster
Intra-cluster dissimilarity: 0.9669083
1. 0.86833185 www.asu.edu%%clas%%shs%%pages%%clinic.htm
2. 0.67073566 www.asu.edu%%clas%%shs%%pages%%deptinfo.htm
3. 0.5665764 www.asu.edu%%clas%%shs%%pages%%advising.htm

Group 4: Showing top 3 of 9 in cluster
Intra-cluster dissimilarity: 0.34377483
1. 0.8674054 www.asu.edu%%it%%fyi%%sites%%maps.html
2. 0.84589446 www.asu.edu%%it%%fyi%%sites%%cpcomatrium.html
3. 0.84106576 www.asu.edu%%it%%fyi%%sites%%index.html
```

```
Group 5: Showing top 3 of 16 in cluster
Intra-cluster dissimilarity: 7.9974427
1. 0.79128844 www.asu.edu%%asunews%%university%%university_index.htm
2. 0.50992554 www.asu.edu%%asunews%%media_info%%history.htm
3. 0.38237113 clasdean.la.asu.edu%%student%%resources%%advising%%
deptAdvisors.htm

Group 6: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 5.684342E-14
1. 1.0000002 www.asu.edu%%it%%fyi%%sites%%annualschedule.html

Group 7: Showing top 3 of 19 in cluster
Intra-cluster dissimilarity: 0.31186476
1. 0.999337 www.asu.edu%%news_to_know%%academics%%
urbaneco_symposium_022004.htm
2. 0.999337 www.asu.edu%%news_to_know%%arts%%dartII_022004.htm
3. 0.999337 www.asu.edu%%news_to_know%%academics%%wilkinson_100703.htm

Group 8: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 3.6379788E-12
1. 1.0000019 www.asu.edu%%news_to_know%%academics%%
usaid_grants_021304.htm

Group 9: Showing top 3 of 10 in cluster
Intra-cluster dissimilarity: 2.0324206
1. 0.7432616 www.asu.edu%%asunews%%university%%coorhall_041803.htm
2. 0.7264299 www.asu.edu%%asunews%%arts%%lattiehall_020304.htm
3. 0.7147448 www.asu.edu%%asunews%%university%%
art_arch_coorhall_011304.htm
```

Group Number	Natural Category
0	lecture in Lattie Coor Hall
1	misc
2	Old Main
3	Speech and Hearing department
4	computing sites
5	misc
6	computing sites
7	art and architecture at Lattie Coor Hall
8	art and architecture at Lattie Coor Hall
9	art and architecture at Lattie Coor Hall

A summary of the number of identical categories, misc categories (not considered when counting identical categories), and single document categories is shown in Table 1.

k	Aspect	"parking and transit"	"dps"	"Lattie Coor"	Averages
k=3	identical	0	0	0	0
	misc	0	0	0	0
	single	0	0	0	0
k=6	identical	0	3	2	1.67
	misc	1	2	2	1.67
	single	0	0	0	0
k=10	identical	4	5	5	4.67
	misc	1	2	2	1.67
	single	2	1	4	2.33

Table 1: Summary of the number of clusters with identical categories, number of misc categories, and number of single document clusters with respect to three values of k and three queries.

All three queries performed well at k=3. Note that this may be misleading, however, since we only looked at the top 3 documents in each cluster. A high percentage of the remaining documents in each cluster may be "junk;" that is, they may not fit well into the categories we named. Thus, we are interested in finding the highest value of k that still produced decent results. By doing so, we reduce the amount of "junk" in each cluster that doesn't fit the actual category.

For "parking and transit," k=6 performed well. k=10 was too many clusters, since 4 of the clusters represented the same category and a suspicious 2 categories had only a single document.

Both "dps" and "Lattie Coor" started to have multiple clusters with the same category by k=6, indicating 6 was too many clusters. On the other hand, they also both had several misc categories, indicating either that more clusters were needed to break the misc clusters into meaningful categories, or the original centroids in the k-means algorithm were poor choices, leading to lousy clusters. We therefore believe that both of these queries should be run around k=6, or possibly k=5 or 4. Further testing would need to be conducted to verify this.

Clearly, the correct value for k is query dependent. However, if looking for a general rule of thumb, we can consider the averages for guidance. Based on these values, k should be set between 3 and 6 to produce decent results for most queries.

Table 2 shows the aggregate dissimilarity for each query with respect to number of clusters. While part of the aggregate dissimilarity is influenced by the initial centroids given to k-means, it is also affected by the value of k. As k increases, there are more clusters, meaning it is more likely that we can put each document closer to a centroid. This is particularly apparent when we have outliers that can now be put in a cluster all by themselves, thus contributing 0 to the dissimilarity metric, when earlier they may have contributed significant values to the aggregate dissimilarity. The general trend in Table 2 supports this argument; as the number of clusters increases, the aggregate dissimilarity decreases. Indeed, if we were to increase the number of clusters to 64, matching the number of results we are taking from vector space search, the dissimilarity would be zero, as each document would be in its own cluster.

	Aggregate Dissimilarity			
k	**“parking and transit”**	**“dps”**	**“Lattie Coor”**	**Averages**
k=3	6.84	11.62	18.82	12.43
k=6	5.40	10.97	14.69	10.35
k=10	3.70	6.20	11.72	7.21

Table 2: Aggregate dissimilarity with respect to query and k.

Finally, Table 3 contains the execution time for each query. Based on this, there is no evidence to support that adjusting k significantly affects execution time.

	Execution Time (seconds)		
k	**“parking and transit”**	**“dps”**	**“Lattie Coor”**
k=3	2	2	6
k=6	2	2	4
k=10	2	2	6

Table 3: Execution time with respect to query and k.

Conclusion

This paper analyzed the K-means algorithm. First, we looked at the ability of K-means to form clusters that corresponded to natural categories, as interpreted by a human. In doing this, K-means was victim of the value of k we provided, but all and all seemed to produce clusters that associated with natural categories. Secondly, we considered the effects of varying the number of clusters. We found that each query has its own optimal value for k, which will produce decent clusters corresponding to unique natural categories. We also observed that increasing k results in a lower aggregate dissimilarity. Finally, we noted that varying k had no significant affect on the execution time of the algorithm.

Nicholas Radtke
11/26/05
CSE 598/494 Information Integration

Project: Part C

Task 6: Buckshot Algorithm

Buckshot Algorithm Analysis

In order to analyze the performance of the Buckshot algorithm, we consider two aspect:

1. The correspondence between the clusters and natural (human interpreted) categories.
2. The ability of the Buckshot algorithm to produce better (or worse) clusters than the K-means algorithm in task 5.

Note that we will not be reanalyzing the effects of varying the number of clusters k. We see no reason to redo this analysis since 1) it wasn't asked for and 2) all that has changed is how we choose the initial centroids used by K-means. We argue that the Buckshot algorithm will have the following affects on the metrics measured in task 5:

- **The correspondence of clusters to natural categories:** The actual number of natural categories is unaffected by the Buckshot algorithm since we still use the top 64 vector space search results. The only difference is that the Buckshot algorithm may provide better centroids to seed K-means with, resulting in fewer misc categories.
- **The aggregate dissimilarity:** If the Buckshot algorithm provides centroids that lead to better clusters, the average dissimilarities for a query will simply shift downward. However, the general trend of more clusters yielding lower aggregate dissimilarities will remain true.
- **The execution times:** It is possible that the Buckshot algorithm will produce better initial centroids for K-means, possibly causing K-means to require fewer iterations before stabilizing. However, this will not change that the K-means algorithm is not significantly sensitive to the number of clusters k with respect to execution time. Thus, any change in execution time would need to be attributed to performing the pre-K-means algorithm (i.e. the HAC). Although not presented here, some quick tests confirmed that the value of k has no significant impact on the time it takes for HAC to run.

Since there is no reason to believe there will be significant changes in these factors for the Buckshot algorithm, we do not remeasure them here with respect to varying k.

Note that I am confident that my implementation of the Buckshot algorithm forms correct clusters because:

1. I used incremental testing for all the tasks involved. In addition, I created simple test data and calculated the results using a pencil and paper. I then ran and analyzed these test cases on the code to verify it was working properly before developing the next part and moving on to the real data.
2. As shown below, my implementation seems to find meaningful clusters, reinforcing that the clusters are reasonable.

Correspondence Between Clusters and Natural Categories

To analyze the correspondence between the clusters provided by the Buckshot algorithm and natural categories (where natural refers to categories as interpreted by a human), we run three sample queries and then try to give the resulting categories "names", based on their contents. For all the queries, we use the top 64 (or as many as are returned, if fewer than 64 matches exists) results returned from a vector space search. We hold the number of groups k at 3 for simplicity, in hopes that we are able to name gross categories, rather than categories with minute differences. We also display all documents in each group, rather than just the top three, since it gives a better notion of what category the cluster really represents. Each document in a cluster includes its similarity to its respective centroid, sorted from highest to lowest similarity. Thus, highly similar documents will be given more weight when considering a name for the category. If no obvious category exists, we will say this rather than inventing a category.

The sample queries chosen, based on the domain of the crawl, are:

- parking and transit
- dps
- Lattie Coor

After running each query, we analyze the cluster and natural category correspondence.

Query: parking and transit

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 7 iterations.
Aggregate cluster dissimilarity: 7.302911
```

```
K-means Group 0: Showing top 47 of 47 in cluster
Intra-cluster dissimilarity: 4.7867374
1. 0.86733365 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2. 0.85645664 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.84680915 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
4. 0.82703793 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
5. 0.82584524 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
6. 0.82413495 www.asu.edu%%dps%%pts%%maps%%visitormap.html
7. 0.8233125 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
8. 0.8228899 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
9. 0.8212932 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
10. 0.8186657 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
11. 0.81610286 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
12. 0.8157883 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
13. 0.8133115 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
14. 0.8126773 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
15. 0.8102337 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
16. 0.8095762 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
17. 0.80758333 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
18. 0.8068171 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
19. 0.8031797 www.asu.edu%%dps%%pts%%maps%%wfa.html
20. 0.80190456 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
21. 0.8006351 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
22. 0.7994697 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
23. 0.7950814 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
24. 0.7922902 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
25. 0.77074033 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.7644299 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.73389786 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.6692489 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.6602731 www.asu.edu%%dps%%pts%%event%%gammage.html
30. 0.64657456 www.asu.edu%%dps%%pts%%visitor%%regulations.html
31. 0.6314108 www.asu.edu%%dps%%pts%%event%%football.html
32. 0.6213011 www.asu.edu%%dps%%pts%%event%%wbasketball.html
33. 0.61768574 www.asu.edu%%dps%%pts%%event%%basketball.html
34. 0.61236703 www.asu.edu%%dps%%pts%%visitor%%designated.html
35. 0.61157185 www.asu.edu%%dps%%pts%%citation%%index.html
36. 0.59297365 www.asu.edu%%dps%%pts%%service%%index.html
37. 0.59119594 www.asu.edu%%dps%%pts%%event%%cardinals.html
38. 0.5864194 www.asu.edu%%dps%%pts%%admin%%index.html
39. 0.5802553 www.asu.edu%%dps%%pts%%visitor%%meter.html
40. 0.5787867 www.asu.edu%%dps%%pts%%admin%%initiatives.html
41. 0.55155605 www.asu.edu%%dps%%pts%%event%%sports.html
42. 0.53887194 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
43. 0.53463227 www.asu.edu%%dps%%pts%%flash.html
44. 0.5026568 www.asu.edu%%dps%%pts%%visitor%%temp.html
45. 0.49823275 www.asu.edu%%dps%%pts%%admin%%achievements.html
46. 0.4745951 herbergercollege.asu.edu%%calendar%%directions.html
47. 0.32101455 construction.asu.edu%%..%%introduction%%Parking.shtml

K-means Group 1: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.10821293
```

```
1. 0.92881876 www.asu.edu%%asunews%%media_info%%parking.html
2. 0.85691214 www.asu.edu%%dps%%pts%%visitor%%media.html
3. 0.71247256 www.asu.edu%%asunews%%university%%
parking_construction_082403.htm

K-means Group 2: Showing top 14 of 14 in cluster
Intra-cluster dissimilarity: 2.4079607
1. 0.8209717 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.7328134 www.asu.edu%%dps%%pts%%decals%%options.html
3. 0.72107416 www.asu.edu%%dps%%pts%%decals%%howto.html
4. 0.6861637 www.east.asu.edu%%admin%%pts%%residences%%index.htm
5. 0.6711426 www.asu.edu%%dps%%pts%%faq.html
6. 0.6598166 www.east.asu.edu%%admin%%pts%%appeals%%index.htm
7. 0.6228333 www.east.asu.edu%%admin%%pts%%maps%%index.htm
8. 0.6184191 www.east.asu.edu%%admin%%pts%%visitors%%index.htm
9. 0.6015198 www.asu.edu%%hr%%new_employee%%parking_decal.html
10. 0.5955999 www.east.asu.edu%%admin%%pts%%events%%index.htm
11. 0.5948738 www.asu.edu%%dps%%pts%%visitor%%disabled.html
12. 0.44329947 www.asu.edu%%aad%%manuals%%dps%%index.html
13. 0.3820629 www.asu.edu%%tour%%main%%towers.html
14. 0.33797097 www.asu.edu%%admissions%%steps%%parking.html
```

The first 28 entries of group 0 seem to be maps, with the remaining documents all lying in the same directory but having a less obvious connection. Thus “maps” is probably a good name for this category. Group 1 seems to have two documents describing where the media is able to park, along with a document describing a parking structure being erected in Lot 59. The category for this group is therefore not clear. Finally, group 2 seems to be summed up as the “parking policy FAQ” category.

Because the Buckshot algorithm is randomly selecting the documents it uses to do HAC, we run the same query a second time to see if we get different results. K-means proved very sensitive to the random selection. It is our hope that the Buckshot algorithm reduces some of this sensitivity and we end up with similar categories most of the time.

Query: parking and transit

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: parking transit

Vector space search returned 1295 hits. Using top 64.
K-means algorithm required 6 iterations.
Aggregate cluster dissimilarity: 7.1344385

K-means Group 0: Showing top 48 of 48 in cluster
Intra-cluster dissimilarity: 4.3206134
```

```
1. 0.8834723 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2. 0.8687122 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.86349833 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
4. 0.84575015 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
5. 0.8413222 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
6. 0.8408307 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
7. 0.8390831 www.asu.edu%%dps%%pts%%maps%%visitormap.html
8. 0.8389105 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
9. 0.837216 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
10. 0.833288 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
11. 0.8305809 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
12. 0.8299735 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
13. 0.82951695 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
14. 0.82730514 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
15. 0.82649475 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
16. 0.8212385 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
17. 0.8201209 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
18. 0.81770635 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
19. 0.81710327 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
20. 0.81690395 www.asu.edu%%dps%%pts%%maps%%wfa.html
21. 0.8145275 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
22. 0.8138397 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
23. 0.8128743 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
24. 0.8050961 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
25. 0.7834659 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.7733165 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.7459902 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.69581455 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.65829563 www.asu.edu%%dps%%pts%%visitor%%regulations.html
30. 0.654346 www.asu.edu%%dps%%pts%%event%%gammage.html
31. 0.6517345 www.asu.edu%%dps%%pts%%citation%%index.html
32. 0.6464899 www.asu.edu%%dps%%pts%%event%%football.html
33. 0.6310959 www.asu.edu%%dps%%pts%%event%%wbasketball.html
34. 0.6281309 www.asu.edu%%dps%%pts%%event%%basketball.html
35. 0.6205834 www.asu.edu%%dps%%pts%%visitor%%designated.html
36. 0.62037396 www.asu.edu%%dps%%pts%%admin%%index.html
37. 0.61086375 www.asu.edu%%dps%%pts%%service%%index.html
38. 0.60718274 www.asu.edu%%dps%%pts%%event%%cardinals.html
39. 0.590028 www.asu.edu%%dps%%pts%%admin%%initiatives.html
40. 0.57854694 www.asu.edu%%dps%%pts%%visitor%%media.html
41. 0.56636703 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
42. 0.56264526 www.asu.edu%%dps%%pts%%flash.html
43. 0.5568996 www.asu.edu%%dps%%pts%%faq.html
44. 0.54758275 www.asu.edu%%dps%%pts%%visitor%%meter.html
45. 0.54090285 www.asu.edu%%dps%%pts%%event%%sports.html
46. 0.52696097 www.asu.edu%%dps%%pts%%admin%%achievements.html
47. 0.52181983 www.asu.edu%%dps%%pts%%visitor%%temp.html
48. 0.5077476 www.asu.edu%%dps%%pts%%visitor%%disabled.html

K-means Group 1: Showing top 12 of 12 in cluster
Intra-cluster dissimilarity: 2.0024009
1. 0.8278789 www.east.asu.edu%%admin%%pts%%faq%%index.htm
2. 0.73149985 www.asu.edu%%dps%%pts%%decals%%howto.html
```

```
3. 0.7138399 www.east.asu.edu%%admin%%pts%%residences%%index.htm
4. 0.7128367 www.asu.edu%%dps%%pts%%decals%%options.html
5. 0.69088334 www.east.asu.edu%%admin%%pts%%appeals%%index.htm
6. 0.62994874 www.east.asu.edu%%admin%%pts%%maps%%index.htm
7. 0.62890404 www.east.asu.edu%%admin%%pts%%visitors%%index.htm
8. 0.6220123 www.asu.edu%%hr%%new_employee%%parking_decal.html
9. 0.6085076 www.east.asu.edu%%admin%%pts%%events%%index.htm
10. 0.464573 www.asu.edu%%aad%%manuals%%dps%%index.html
11. 0.4057958 www.asu.edu%%tour%%main%%towers.html
12. 0.34408283 www.asu.edu%%admissions%%steps%%parking.html

K-means Group 2: Showing top 4 of 4 in cluster
Intra-cluster dissimilarity: 0.81142426
1. 0.9283642 herbergercollege.asu.edu%%calendar%%directions.html
2. 0.61258346 www.asu.edu%%asunews%%media_info%%parking.html
3. 0.54312944 www.asu.edu%%asunews%%university%%parking_constr
uction_082403.htm
4. 0.3310678 construction.asu.edu%%..%%introduction%%Parking.shtml
```

Amazingly, the groups produced by the second sample run are very similar to those produced by the first run. Indeed, the second group 0 is very similar to the first group 0 and can also be called “maps.” Group 1 of the second run is similar to group 2 of the first run, again belonging loosely to the category of “parking policy FAQ.” Finally, group 2 seems to be a hodge-podge of documents with no clear category, just like group 1 in the first run. This evidence seems to suggest that the Buckshot algorithm is less sensitive to the randomization of the initial document selection than the K-means algorithm. As has been pointed out in class, sometimes stability is more important than “correct” results.

Query: dps

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: dps

Vector space search returned 159 hits. Using top 64.

K-means algorithm required 2 iterations.
Aggregate cluster dissimilarity: 12.686776

K-means Group 0: Showing top 34 of 34 in cluster
Intra-cluster dissimilarity: 1.3505528
1. 0.95086527 www.asu.edu%%dps%%pts%%maps%%visitorcenter.html
2. 0.9506304 www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3. 0.94160867 www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
4. 0.93657106 www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
5. 0.93299514 www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
6. 0.925038 www.asu.edu%%dps%%pts%%maps%%memorialunion.html
```

```
7. 0.9249315 www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
8. 0.9232709 www.asu.edu%%dps%%pts%%maps%%haydenlibrary.html
9. 0.9191346 www.asu.edu%%dps%%pts%%maps%%computingcommons.html
10. 0.9177033 www.asu.edu%%dps%%pts%%maps%%vhaydenlibrary.html
11. 0.9161012 www.asu.edu%%dps%%pts%%maps%%noblesciencelibrary.html
12. 0.9146533 www.asu.edu%%dps%%pts%%maps%%wfa.html
13. 0.9144067 www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
14. 0.91137415 www.asu.edu%%dps%%pts%%maps%%gammageauditorium.html
15. 0.91033185 www.asu.edu%%dps%%pts%%maps%%vnoblesciencelibrary.html
16. 0.90850085 www.asu.edu%%dps%%pts%%maps%%vcomputingcommons.html
17. 0.9084687 www.asu.edu%%dps%%pts%%maps%%baseballstadium.html
18. 0.9071477 www.asu.edu%%dps%%pts%%maps%%footballstadium.html
19. 0.90667385 www.asu.edu%%dps%%pts%%maps%%vvisitorcenter.html
20. 0.9007254 www.asu.edu%%dps%%pts%%maps%%vgammageauditorium.html
21. 0.89953023 www.asu.edu%%dps%%pts%%maps%%vkarstengolfcourse.html
22. 0.89795345 www.asu.edu%%dps%%pts%%maps%%vbaseballstadium.html
23. 0.8972754 www.asu.edu%%dps%%pts%%maps%%visitormap.html
24. 0.89664716 www.asu.edu%%dps%%pts%%maps%%vfootballstadium.html
25. 0.8870202 www.asu.edu%%dps%%pts%%maps%%karstengolfcourse.html
26. 0.8573152 www.asu.edu%%dps%%pts%%maps%%asumap.html
27. 0.8445506 www.asu.edu%%dps%%pts%%maps%%ross-blakelylawlibrary.html
28. 0.7318823 www.asu.edu%%dps%%pts%%maps%%index.html
29. 0.6701244 www.asu.edu%%dps%%pts%%citation%%index.html
30. 0.61726016 www.asu.edu%%dps%%pts%%admin%%index.html
31. 0.5897662 www.asu.edu%%dps%%pts%%shuttle%%usbshuttle.html
32. 0.58151644 www.asu.edu%%dps%%pts%%flash.html
33. 0.5611668 www.asu.edu%%dps%%pts%%service%%index.html
34. 0.49046633 www.asu.edu%%dps%%pts%%decals%%removal.html

K-means Group 1: Showing top 27 of 27 in cluster
Intra-cluster dissimilarity: 11.1256895
1. 0.6041399 www.asu.edu%%aad%%manuals%%acd%%acd123.html
2. 0.535062 www.asu.edu%%aad%%manuals%%spp%%spp812.html
3. 0.5254255 www.asu.edu%%dps%%index.html
4. 0.51843524 www.asu.edu%%workingatasu%%index%%index.html
5. 0.47067782 www.asu.edu%%emergency%%resources.html
6. 0.46388015 www.asu.edu%%emergency%%emergency.html
7. 0.45490217 www.west.asu.edu%%asuw2%%staff.shtml
8. 0.44672215 www.asu.edu%%asuremembers%%resources.htm
9. 0.42793888 www.west.asu.edu%%asuw2%%faculty.shtml
10. 0.4243292 westcgi.west.asu.edu%%dps%%index.htm
11. 0.42208642 www.asu.edu%%tour%%main%%dps.html
12. 0.4022689 www.asu.edu%%aad%%manuals%%dps%%index.html
13. 0.37841117 www.west.asu.edu%%asuw2%%students.shtml
14. 0.36278874 www.asu.edu%%aad%%manuals%%sta%%sta104-02.html
15. 0.36144665 property.asu.edu%%\equipment\maintenance.html
16. 0.34774965 www.asu.edu%%dps%%police%%keys.htm
17. 0.33707324 www.asu.edu%%provost%%asenate%%problem.html
18. 0.32466257 www.east.asu.edu%%admin%%facilities.htm
19. 0.3126478 www.asu.edu%%dps%%police%%police.htm
20. 0.3020934 www.asu.edu%%dps%%police%%directory.htm
21. 0.29499918 www.asu.edu%%president%%cet%%ref.htm
22. 0.2717327 www.asu.edu%%xed%%linkto%%index.html
```

```
23. 0.2482216 www.asu.edu%%dps%%orgchart.htm
24. 0.20567565 www.asu.edu%%provost%%committees%%ASUWCSC.html
25. 0.1958046 www.asu.edu%%aad%%manuals%%manual-coordinators.html
26. 0.18495019 www.asu.edu%%provost%%committees%%CRBG.html
27. 0.12839805 www.asu.edu%%dps%%mission.htm

K-means Group 2: Showing top 3 of 3 in cluster
Intra-cluster dissimilarity: 0.2105342
1. 0.87268454 www.asu.edu%%dps%%pts%%psac%%Code%20Perspective.html
2. 0.69840145 www.asu.edu%%dps%%pts%%psac%%cms.html
3. 0.6784984 www.asu.edu%%provost%%committees%%PSAC.html
```

As clearly as ever, group 0's category is “maps.” Group 1 is less defined, almost seeming to be two categories: “emergency resourse directories” and “pages falling in the dps homepage.” Finally, group 2 seems to fit the category “committees changing public safety policies.”

Query: Lattie Coor

```
Using up to top 64 documents
Number of clusters (k) is 3
Searching for: lattie coor

Vector space search returned 230 hits. Using top 64.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 23.136074

K-means Group 0: Showing top 51 of 51 in cluster
Intra-cluster dissimilarity: 20.207314
1. 0.9794857 www.asu.edu%%news_to_know%%academics%%
construction_conference_012004.htm
2. 0.9794857 www.asu.edu%%news_to_know%%academics%%kelloggrant_01260.htm
3. 0.9794857 www.asu.edu%%news_to_know%%academics%%
levy_lecture_013004.htm
4. 0.9794857 www.asu.edu%%news_to_know%%academics%%
lewis_franklecture_01204.htm
5. 0.9794857 www.asu.edu%%news_to_know%%academics%%mba_online_012704.htm
6. 0.9794857 www.asu.edu%%news_to_know%%academics%%online_eng_020404.htm
7. 0.9794857 www.asu.edu%%news_to_know%%academics%%statedept_021704.htm
8. 0.9794857 www.asu.edu%%news_to_know%%academics%%
stupidcupid_021004.htm
9. 0.9794857 www.asu.edu%%news_to_know%%academics%%
todd_gitlin_011604.htm
10. 0.9794857 www.asu.edu%%news_to_know%%academics%%
urbaneco_symposium_022004.htm
11. 0.9794857 www.asu.edu%%news_to_know%%academics%%
usaid_grants_021304.htm
12. 0.9794857 www.asu.edu%%news_to_know%%academics%%wilkinson_100703.htm
13. 0.9794857 www.asu.edu%%news_to_know%%academics%%wuthrich_lecture.htm
14. 0.9794857 www.asu.edu%%news_to_know%%arts%%artventures_021704.htm
15. 0.9794857 www.asu.edu%%news_to_know%%arts%%buyarug_012204.htm
```

```
16. 0.9794857 www.asu.edu%%news_to_know%%arts%%ceramics_gala_012304.htm
17. 0.9794857 www.asu.edu%%news_to_know%%arts%%ceramictour_012604.htm
18. 0.9794857 www.asu.edu%%news_to_know%%arts%%dartII_022004.htm
19. 0.9794002 www.asu.edu%%news_to_know%%newsknow_index.htm
20. 0.5681827 www.asu.edu%%asunews%%university%%university_index.htm
21. 0.4658978 www.asu.edu%%asunews%%archives%%jan2004.htm
22. 0.3811199 www.asu.edu%%asunews%%media_info%%history.htm
23. 0.34088156 www.asu.edu%%asunews%%archives%%apr2002.htm
24. 0.3316812 www.asu.edu%%asunews%%university%%
coorhalldedicate_010904.htm
25. 0.32223782 www.asu.edu%%asunews%%archives%%may2002.htm
26. 0.30060092 www.asu.edu%%asunews%%archives%%jan2002.htm
27. 0.29884446 www.asu.edu%%asunews%%university%%
east_convocation_051403.htm
28. 0.2791541 www.asu.edu%%asunews%%arts%%lattiehall_020304.htm
29. 0.27371484 www.asu.edu%%asunews%%university%%coorhall_041803.htm
30. 0.2656932 www.asu.edu%%feature%%coor.html
31. 0.2643759 www.asu.edu%%asunews%%university%%
art_arch_coorhall_011304.htm
32. 0.24360783 www.asu.edu%%tour%%main%%coor.html
33. 0.23395954 herbergercollege.asu.edu%%public_art%%public_art.html
34. 0.19844036 www.east.asu.edu%%ia%%html%%journeys.html
35. 0.19718413 www.asu.edu%%ia%%inauguration%%address%%introduction.htm
36. 0.18662266 www.asu.edu%%tour%%main%%a-z.html
37. 0.15028656 herbergercollege.asu.edu%%keys%%index.html
38. 0.1403725 www.asu.edu%%alumni%%oldmain%%information.html
39. 0.13806504 clasdean.la.asu.edu%%student%%resources%%advising%%
deptAdvisors.htm
40. 0.13794878 aspin.asu.edu%%asura%%calendar.htm
41. 0.13721792 www.asu.edu%%clas%%shs%%pages%%deptinfo.htm
42. 0.130339 www.kaet.asu.edu%%cet%%index.htm
43. 0.12819105 www.asu.edu%%clas%%shs%%pages%%clinic.htm
44. 0.12747493 aspin.asu.edu%%asura%%current.htm
45. 0.11338684 www.asu.edu%%clas%%dll%%kor%%events%%eindex.htm
46. 0.101647176 www.asu.edu%%lib%%archives%%preslist.htm
47. 0.09555133 www.asu.edu%%clas%%philosophy%%staff.htm
48. 0.08259188 www.asu.edu%%ia%%cleanandbeautiful%%trashy.html
49. 0.08074667 www.eoaa.asu.edu%%policy.htm
50. 0.07341925 www.asu.edu%%clas%%shs%%pages%%advising.htm
51. 0.071087316 www.asu.edu%%clas%%philosophy%%colloquia.htm

K-means Group 1: Showing top 12 of 12 in cluster
Intra-cluster dissimilarity: 2.9287596
1. 0.80872434 www.asu.edu%%it%%fyi%%sites%%annualschedule.html
2. 0.6431497 www.asu.edu%%it%%fyi%%sites%%maps.html
3. 0.62854695 www.asu.edu%%it%%fyi%%sites%%coorcomputingcommons.html
4. 0.61935467 www.asu.edu%%it%%fyi%%sites%%cpcomatrium.html
5. 0.61888087 www.asu.edu%%it%%fyi%%sites%%gwc185.html
6. 0.6072227 www.asu.edu%%it%%fyi%%sites%%bac16.html
7. 0.6042188 www.asu.edu%%it%%fyi%%sites%%custwebform.html
8. 0.59784144 www.asu.edu%%it%%fyi%%sites%%index.html
9. 0.5637042 www.asu.edu%%it%%fyi%%sites%%ecg150.html
10. 0.4800628 www.asu.edu%%it%%fyi%%sites%%teamwork.html
```

```
11. 0.1902954 www.asu.edu%%clas%%sociology%%colloquium%%colloquium.html
12. 0.13527317 www.asu.edu%%clas%%chicana%%index.html

K-means Group 2: Showing top 1 of 1 in cluster
Intra-cluster dissimilarity: 5.684342E-14
1. 1.0000002 math.asu.edu%%~chavez%%seminar1.html
```

All of the top-ranked documents in group 0 refer to the same category, making this the "art and architecture at Lattie Coor Hall" category. Group 1 is the "computing sites" category. Finally, group 2 has a single document in it, making it its own category: "lecture in Lattie Coor Hall."

Based on the sample queries above, it appears that the Buckshot algorithm produces clusters that correspond to natural categories. An advantage that Buckshot has over K-means in this experiment is that the Buckshot algorithm is less sensitive to the documents that are randomly selected as seeds to the algorithm.

K-means versus Buckshot Algorithm

With clustering algorithms, we are interested in minimizing the intra-cluster distance and maximizing the inter-cluster distance. Therefore, in order to compare the K-means and Buckshot algorithm, we will use the aggregate dissimilarity metric, which provides a notion of the tightness of the clusters.

We define the aggregate dissimilarity as follows: Let the dissimilarity of a document with respect to its centroid be 1-similarity of the document to its respective centroid. The aggregate dissimilarity is the sum of the squares of the dissimilarities for all documents to their respective centroids. A lower dissimilarity corresponds to tighter clusters. A higher dissimilarity corresponds to larger intra-cluster distances. Thus, we are interested in minimizing the aggregate dissimilarity.

For both the K-means and Buckshot algorithm, we run the queries "parking and transit," "dps," and "Lattie Coor" for k=3, 6, and 10. We will use the average dissimilarities over k for comparison to minimize the impact of high aggregate dissimilarities due to the incorrect number of clusters being used. Due to space constraints, we do not include the list of documents for each of our sample runs; we only provide the aggregate dissimilarities. In all cases, the top 150 results returned from vector space search are used in the clustering. (We increase the number of results to use to 150 so that the Buckshot algorithm can sample sqrt(150)=12 to accommodate k=10. That is, the sqrt should be greater than the value of k or the HAC becomes trivial, leading the Buckshot algorithm to perform identically to K-means.)

Table 1 contains the aggregate dissimilarities from our experiment. The smaller average values are printed in boldface.

	K-means			Buckshot		
k	“parking and transit”	“dps”	“Lattie Coor”	“parking and transit”	“dps”	“Lattie Coor”
k=3	52.075040	54.319397	22.034801	56.034832	56.792020	19.251820
k=6	44.226517	42.013996	16.683573	47.779930	42.387580	13.549209 15.115963
k=10	37.637928	36.430830	15.292011	35.730858	34.890285	15.139688
average	**44.646495**	**44.254741**	18.003462	46.515207	44.689962	**16.502490**

Table 1: Aggregate dissimilarities for three queries and three values of k.

For the query “parking and transit,” K-means produced tighter clusters. In fact, of all three queries, this one shows the largest difference between the two algorithms. While one would expect Buckshot to nearly always outperform K-means, there is always the chance the K-means got lucky in this particular case with its initial seeds (or, as the case may be, Buckshot got unlucky with its choice of seeds for HAC).

For the queries “dps”, K-means produced just slightly better clusters. For all intents and purposes, this is basically a tie. If the test were run again, it is likely that the results in these categories would change.

For the query “Lattie Coor,” Buckshot produced better clusters, although the margin of difference for this query is not as large as for the “parking and transit” query.

With one win, one loss, and one tie, there is no real evidence that the Buckshot algorithm produces better clusters than K-means. Even so, the Buckshot algorithm has the advantage that it is more consistent with its results when submitting a query multiple times, as was shown in the previous section. This is worth something as consistency is sometimes more important the the producing the best answer.

It is a bit of a surprise that the differences are not more pronounced. Buckshot is expected to perform better because it has a more intelligent way of choosing the seed for the K-means part of the algorithm. It is not, however, able to avoid still randomly choosing documents – it just chooses them for HAC instead of K-means. From the data above, it appears that, while HAC isn't sensitive to the data it is given, the clusters it produces, and the resulting centroids, do not seem to be particularly good for priming K-means. Rather, they seem about as good as using K-means without HAC, although HAC does seem to add some stability in regards to repeatability. A possible way to improve the performance of the Buckshot algorithm would be to use a sample size larger than the sqrt(n), where n is the number of documents being clustered. However, since the Buckshot algorithm is $O(n^2)$, the cost of using a larger n is significant.

Conclusion

In this paper, we have analyzed the Buckshot algorithm. First, we examined the clusters produced by the Buckshot algorithm to see if they corresponded to natural (human interpretable) categories. Although they were not perfect, and potentially influenced by our incorrect choice of k, the clusters did indeed seem to correspond to natural categories. It was noted that the Buckshot algorithm also tended to produce more consistent categories when run multiple times for the same query. Finally, we compared the goodness of the clusters produced by K-means and the Buckshot algorithm by comparing the aggregate dissimilarities. The performance of both algorithms was very similar, making it difficult to declare one algorithm as superior.

Nicholas Radtke
11/26/05
CSE 598/494 Information Integration

Project: Part C

Task Extra Credit: Similar Pages

I have implemented a similar pages feature which works as follows: Given a filename that exists in the web crawl, I construct a new query. This query simply contains the terms that were indexed by the Lucene package for this particular page. Note that if a term appears multiple times in the document, it is listed multiple times in the query. The query is then subjected to a vector space similarity search. This way, the results are the documents that are nearest the original document, or the most similar pages. Note that the page used to build the query is explicitly removed from the results. After all, there is little sense in listing it since the user already has seen this page.

One peculiarity should be noted about the similarity scores. It is expected that the query, as built above, would have a cosine similarity of 1.0 with the corresponding page in the web crawl when performing the vector space search. This is not the case. The reason is because the documents in the index are using TF-IDF. The query, however, is only using TF. Thus the cosine similarity is not what may be expected. This has the strange side effect that the identical page in the crawl may not have the highest similarity to the query. However, in all the test cases I ran, this page ranked very near, if not at the top.

To demonstrate that the similar pages feature works, we provide a sample query below:

```
Searching for pages similar to:
www.asu.edu%%fastt%%scholarships%%index.html

7014 total matching documents

1. 0.87865084 www.asu.edu%%fa%%types.html
2. 0.8666596 www.asu.edu%%fa%%scholarships%%departmental.html
3. 0.8659849 www.asu.edu%%fastt%%scholarships%%departmental.html
4. 0.8591669 www.asu.edu%%fa%%scholarships%%general.html
5. 0.8590927 www.asu.edu%%fastt%%scholarships%%general.html
6. 0.85584354 www.asu.edu%%fa%%apply%%index.html
7. 0.85529923 www.asu.edu%%fastt%%apply%%index.html
8. 0.83560586 www.asu.edu%%fa%%fan.html
9. 0.83483565 www.asu.edu%%fa%%studemp%%index.html
10. 0.79066104 www.asu.edu%%fa%%scholarships%%national.html
```

By the URLs alone, it can be seen that the resulting pages are likely very similar to the original page. A perusal of these pages confirms that indeed the pages are similar.

To show the results were not a fluke, below is another sample query and the results.

```
Searching for pages similar to:
www.asu.edu%%clas%%womens_studies%%pages%%events2.html

7143 total matching documents

1. 0.4323179 www.asu.edu%%clas%%womens_studies%%pages%%degrees.html
2. 0.38584986 www.asu.edu%%provost%%smis%%clas%%ba%%wsba.html
3. 0.37580672 www.west.asu.edu%%wsteam%%geninfo.htm
4. 0.33618492 www.asu.edu%%clas%%womens_studies%%pages%%wstclasses.htm
5. 0.3169282 www.west.asu.edu%%academic%%jobs%%WmnStudiesLect.html
6. 0.31298 www.west.asu.edu%%wsteam%%resource.htm
7. 0.30990443 www.asu.edu%%aad%%catalogs%%general%%womens-studies.html
8. 0.3051193 www.west.asu.edu%%wsteam%%courses.htm
9. 0.30480862 www.asu.edu%%clas%%womens_studies%%pages%%aboutus.html
10. 0.30158433 www.west.asu.edu%%wsteam%%corerequirmnts.htm
```

In this case, it is a little harder to tell based on the URLs, but visiting the pages confirms that both the original and resulting pages are all tied into the Women's Studies department.

Nicholas Radtke
11/29/05
CSE 598/494 Information Integration

Project: Part C

Extra Credit: Alternative Merging Functions for HAC

For the extra credit, I have implemented both single link and complete link merging functions in addition to the centroid distance function from task 6. I will compare the results of using these three methods when merging in HAC.

With clustering algorithms, we are interested in minimizing the intra-cluster distance and maximizing the inter-cluster distance. Therefore, in order to compare the centroid distance, single link, and complete link merging functions, we will use the aggregate dissimilarity metric, which provides a notion of the tightness of the clusters.

We define the aggregate dissimilarity as follows: Let the dissimilarity of a document with respect to its centroid be 1-similarity of the document to its respective centroid. The aggregate dissimilarity is the sum of the squares of the dissimilarities for all documents to their respective centroids. A lower dissimilarity corresponds to tighter clusters. A higher dissimilarity corresponds to larger intra-cluster distances. Thus, we are interested in minimizing the aggregate dissimilarity.

The merging functions are used in the first stage of the Buckshot algorithm, during the HAC. We could consider the aggregate dissimilarity metric, therefore, for the clusters generated by HAC. However, what we are most interested in is the tightness of the clusters produced by K-means, which are directly influenced by the clusters formed by HAC. Therefore, we measure the aggregate dissimilarity only on the clusters generated by K-means. In other words, we do not care how tight the initial HAC clusters are – we are more interested in whether these clusters provide good seeds for the K-means algorithm, causing the resulting K-means clusters to be tight.

For all three merging functions, we run the queries q_1="Michael Crow," q_2="sun devil," and q_3="gammage" for k=3, 6, and 10. We will use the average dissimilarities over k for comparison to minimize the impact of high aggregate dissimilarities due to the incorrect number of clusters being used. Since the initial seeds are being chosen randomly for HAC, we run each query 3 times and then average the results, thus hopefully eliminating the effects of particularly good or bad seeds due to luck. Due to space constraints, we do not include the list of documents for each of our sample runs; we only provide the aggregate dissimilarities. In all cases, the top 150 results returned from vector space search are used in the clustering.

Table 1 contains the aggregate dissimilarities from our experiment. The smallest average values for each query are printed in boldface.

		Centroid Distance			**Single Link**			**Complete Link**		
k		**q_1**	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**
k=3	Run 1	22.84	65.03	51.58	22.88	65.91	50.68	23.07	65.11	49.99
	Run 2	19.65	60.73	67.64	20.28	64.34	57.61	22.73	65.03	50.02
	Run 3	22.84	68.59	57.22	22.84	68.46	56.11	18.85	67.60	50.08
k=6	Run 1	16.09	51.75	34.35	16.07	57.02	37.15	17.78	48.48	37.05
	Run 2	17.30	49.71	38.25	16.74	56.18	43.84	19.90	52.98	33.04
	Run 3	16.46	56.16	35.88	16.86	55.15	38.30	22.18	51.23	33.38
k=10	Run 1	14.42	40.58	28.07	16.14	44.5	31.48	15.61	42.52	26.06
	Run 2	15.07	41.80	24.62	18.76	40.75	29.56	14.71	45.28	23.19
	Run 3	15.32	44.03	29.61	15.03	42.08	24.41	15.05	41.68	28.21
Avg		**17.78**	**53.15**	40.80	18.40	54.93	41.02	18.88	53.32	**36.78**

Table 1: Comparison of aggregate dissimilarities for resulting K-means clusters based on centroid distance, single link, and complete link HAC merging functions.

Observing the averages, the centroid distance merging function had the lowest aggregate dissimilarity for queries q_1 and q_2, although it should be noted that the difference between centroid distance and complete link for q_2 is minimal. For query q_3, not only did complete link produce the smallest aggregate dissimilarity, it also produced the largest difference between its average and the next best average over all three queries.

The fact the centroid distance performed the best overall is perhaps one of the reasons that it is the most common merging function. Note that single link is prone to finding chain-like clusters while complete link is biased to find spherical clusters. Thus, the performance of all three of these merging functions depends on the data to which they are being applied. As is shown by the experiment above, even within a single domain, there may not be a clear winner. Apparently query q_3 and spherical clusters were a good match, while q_1 and q_2 did not group as well as spheres. To get the best clusters, it may be necessary to run several merging functions on the data and see which produces the best clusters.

Nicholas Radtke
11/30/05
CSE 598/494 Information Integration

Project: Part C

Extra Credit: Recomputing the Centroid After Several Changes and Using a Heuristic to Pick the Centroids in K-means

For extra credit, I have augmented my K-means algorithm to recompute the centroids after every c changes, c>=0, where a change is defined as a document moving from its current cluster to a different cluster. I have also augmented my K-means clustering algorithm to use a heuristic to pick the centroids. Details of both modifications, as well as experiments showing the effects of these changes, are described below.

Recomputing the Centroids After c Changes

I have augmented my K-means algorithm so that it may recompute the centroids after c changes. Regardless of the value of c, the algorithm always recomputes the centroids at the end of a K-means iteration. If c is set to 0, then the centroids are ONLY recalculated at the end of each K-means iteration.

We are interested in the effect c has on the following two items:

1. The tightness of the clusters
2. The number of iterations before K-means stabilizes

In order to measure the tightness of the clusters, we use the aggregate dissimilarity. We define the aggregate dissimilarity as follows: Let the dissimilarity of a document with respect to its centroid be 1-similarity of the document to its respective centroid. The aggregate dissimilarity is the sum of the squares of the dissimilarities for all documents to their respective centroids. A lower dissimilarity corresponds to tighter clusters. A higher dissimilarity corresponds to larger intra-cluster distances. Thus, we are interested in minimizing the aggregate dissimilarity.

We are also interested in minimizing the number of iterations necessary for K-means to stabilize. Fewer iterations means less computation time.

To see the effects of varying c, we will run the following queries: q_1=”Michael Crow,” q_2=”sun devil,” and q_3=”gammage.” We will hold k at a constant 5. This value is chosen because it seemed to work relatively well in prior clustering experiments within this domain and thus we do not expect it to cause poor results due to the data not clustering well into 5 groups. Since the initial seeds for K-means are chosen randomly, we will run each query 3 times and average the results. In all cases, we will use the top 150 results returned from vector space search for clustering. Finally, we will vary c over the values 0 (one extreme), 1 (another extreme), and 3 (a middle value).

For each run, we will record the aggregate dissimilarity and number of iterations required by K-means.

The aggregate dissimilarities for the experiment are shown in Table 1. The smallest average value for each query is shown in boldface.

	c=0			**c=1**			**c=3**		
	q_1	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**
Run 1	18.02	54.20	36.88	22.29	56.82	43.15	16.73	55.05	39.42
Run 2	20.52	55.52	41.84	17.88	53.69	38.67	18.10	53.49	42.15
Run 3	22.47	54.71	43.26	22.01	50.88	43.75	18.62	56.44	37.94
Avg.	20.34	54.81	40.66	20.73	**53.80**	41.86	**17.82**	54.99	**39.84**

Table 1: The aggregate dissimilarities for three values of c.

In two of the three queries, the middle value (c=3) performed better than either extreme (c=0 and c=1). However, both c=1 and c=3 performed better than c=0. This seems to imply that occasionally recalculating the centroids, after some number of changes, is beneficial in producing tighter clusters. Intuitively, this seems to make some sense. When c>0, the clusters do a better job of tracking the documents within them, versus when c=0, there may be many changes to a cluster before the centroids are recomputed, leading to an increased chance of rather dissimilar documents being added to the cluster. Unfortunately, enough dissimilar documents may have the effect of altering the centroid enough that the dissimilar documents stay in the group for subsequent iterations. While this may still happen when c>0, the chances are reduced.

It is interesting that the extreme c=1 performed worse than c=3. This indicates that while recomputing the centroids after some number of changes leads to tighter clusters, a value of c that is too low leads to a hypersensitivity. This hypersensitivity may mean that noise (i.e. documents that fall right between centroids) effects the clustering. A slightly larger value of c is more immune to this noise and thus may end up producing better clusters. There is likely an optimal c (or range of c) that could be found by experimentation for this particular domain.

Although c=0 performed the worst, it should be noted that recalculating the centroids is an expensive task. During the experiment, all queries for c=0 ran in a few seconds. The same cannot be said for queries with c>0. While some ran just as quickly, others took closer to a minute to complete. It turns out this was not a result of excessive iterations in the K-means algorithm. What actually was causing the slowdown was a poor selection of initial centroids, meaning that many documents had to change groups. This meant that the centroids were frequently recomputed, especially in the case of c=1, where one particularly unlucky query ran for 1:30. This indicates that c should be set as large as possible to minimize centroid recomputation but small enough that the resulting clusters have a low aggregate dissimilarity.

Table 2 shows the number of iterations needed for K-means to stabilize. Although there was not significant variance, it appears that c=0 and c=1 performed better than c=3. However, based on the execution time issues mentioned above, we become less concerned with the number of iterations and more concerned with the number of times the centroids need to be recomputed. This is because centroid recomputation is more expensive than number of iterations.

	c=0			**c=1**			**c=3**		
	q_1	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**	**q_1**	**q_2**	**q_3**
Run 1	3	7	6	3	5	5	3	6	9
Run 2	3	4	3	4	4	7	4	7	3
Run 3	3	9	3	5	6	4	3	5	10
Avg.	**3**	6.67	**4**	4	**5**	5.33	3.33	6	7.33
Avg. over queries	4.56			**4.33**			5.56		

Table 2: The number of iterations before K-means clustering algorithm stabilized.

Using a Heuristic to Select K-means Centroids

In order to improve the clusters formed by K-means, we apply an extremely simple heuristic: We run K-means t times and then pick the run that had the lowest aggregate dissimilarity.

To demonstrate that this is effective, we run the single query “dance team” for t=1 (normal K-means) and t=5 (K-means with heuristic). For each value of t, we run the experiment 3 times to show that the results are consistent. We set k=5 and use the top 150 results from vector space search. We disable recalculating the clusters after c changes by setting c=0.

The results of the experiment are shown in Table 3. While the heuristic we used was extremely simple, it worked. In fact in all three cases, t=5 produced tighter clusters. Note that this was not without cost. Since we are essentially running K-means 5 times, it is expected that the execution time will suffer by a multiple of 5. If tight clusters are essential, this trade-off may be worth it. In other cases, it may be better to use a lower value for t and have cluster tightness suffer.

	t=1	**t=5**
Run 1	48.738773	42.889877
Run 2	53.351097	41.629510
Run 3	47.771816	43.651176

Table 3: Aggregate dissimilarities for normal K-means (t=1) and K-means augmented with heuristic.

Nicholas Radtke
11/30/05
CSE 598/494 Information Integration

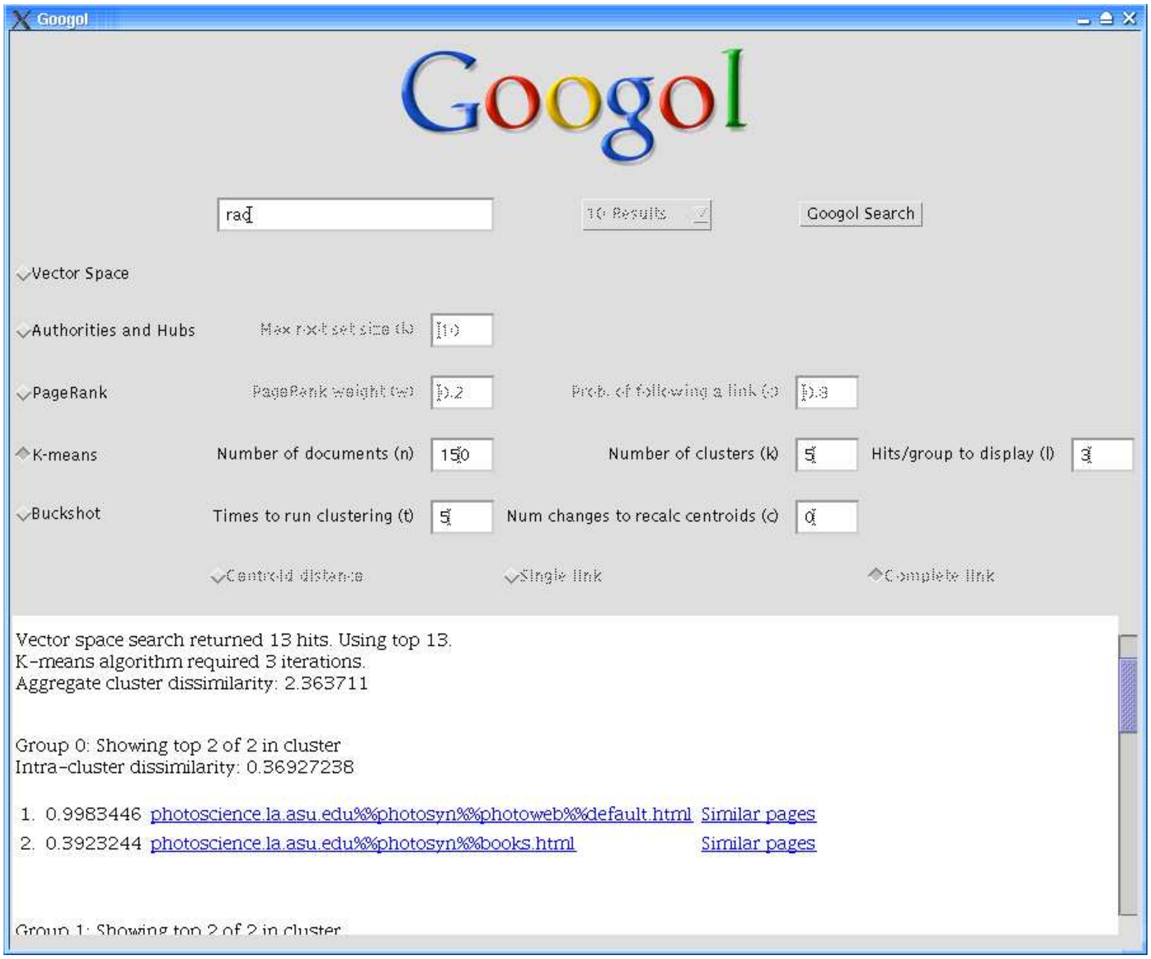
Project: Part C

Extra Credit: GUI

A GUI has been implemented that allows the user to do the following:

1. Search using Vector Similary Model
2. Search using Authorities and Hubs
    1. Includes the ability to change the value of K, the size of the root set.
3. Search using PageRank
    1. Includes the ability to change the value of C, the probability that a random surfer follows a link on the current page.
    2. Includes the ability to change the value of W, the weight assigned to PageRank versus vector similarity
4. Search and cluster via K-means
    1. Includes the ability to adjust the number of documents used from vector space search.
    2. Includes the ability to change number of clusters.
    3. Includes the ability to adjust the number of hits displayed for each cluster.
    4. Includes the ability to change the number of times K-means is run for a single query, where the best run is returned as the results.
    5. Includes the ability to adjust the number of changes before the K-means algorithm recomputes the centroids of the clusters.
5. Search and cluster via Buckshot algorithm
    1. Includes the ability to adjust the number of documents used from vector space search.
    2. Includes the ability to change number of clusters.
    3. Includes the ability to adjust the number of hits displayed for each cluster.
    4. Includes the ability to adjust the number of changes before the K-means algorithm recomputes the centroids of the clusters.
    5. Includes the ability to select from the following merging functions in HAC:
        1. Centroid Distance
        2. Single Link
        3. Complete Link
6. Search for similar pages via hyperlinks
7. Open the HTML in the results via hyperlinks

Sample screen shots are included below:



rad

○ Vector Space

○ Authorities and Hubs　Max root set size (k) 10

○ PageRank　PageRank weight (w) 0.2　Prob. of following a link (q) 0.8

● K-means　Number of documents (n) 150　Number of clusters (k) 5　Hits/group to display (l) 3

○ Buckshot　Times to run clustering (t) 5　Num changes to recalc centroids (c) 0

○ Centroid distance　○ Single link　● Complete link

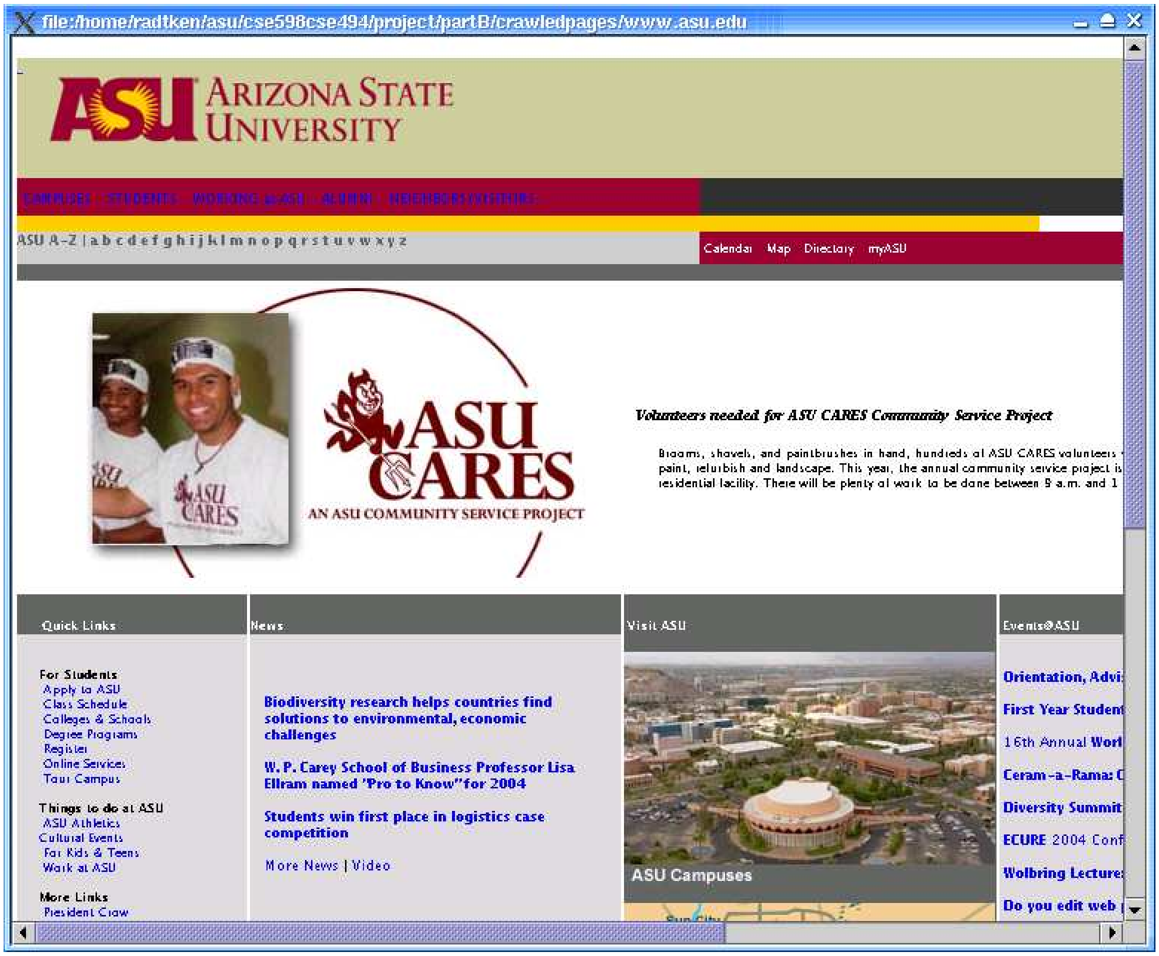
```
Vector space search returned 13 hits. Using top 13.
K-means algorithm required 3 iterations.
Aggregate cluster dissimilarity: 2.363711


Group 0: Showing top 2 of 2 in cluster
Intra-cluster dissimilarity: 0.36927238

 1.  0.9983446  photoscience.la.asu.edu%%photosyn%%photoweb%%default.html  Similar pages
 2.  0.3923244  photoscience.la.asu.edu%%photosyn%%books.html               Similar pages


Group 1: Showing top 2 of 2 in cluster
```

Main Window



file:/home/radtken/asu/cse598cse494/project/partB/crawledpages/www.asu.edu

Volunteers needed for ASU CARES Community Service Project

Brooms, shovels, and paintbrushes in hand, hundreds of ASU CARES volunteers
paint, refurbish and landscape. This year, the annual community service project is
residential facility. There will be plenty of work to be done between 9 a.m. and 1

Biodiversity research helps countries find solutions to environmental, economic challenges

W. P. Carey School of Business Professor Lisa Ellram named 'Pro to Know"for 2004

Students win first place in logistics case competition

Displaying HTML file in separate window