Unit 5 - additional notes

KMeans Example 1

Sample data set of 5 documents, and 5 terms:

	applications	binary	computer	graph	||x||
Doc1	1	4	1	0	4.24
Doc2	2	2	2	0	3.46
Doc3	2	1	0	4	4.58
Doc4	3	1	1	3	4.47
Doc5	4	0	0	6	7.21

1. Let K = 2. Select two rows at random to represent the initial cluster mean: Doc1 and Doc5
2. Calculate the distance between each document and the cluster centre (using cosine measure)

Sim(Doc1, Doc2) =	0.82	Sim(Doc5, Doc2) =	0.32
Sim(Doc1, Doc3) =	0.31	Sim(Doc5, Doc3) =	0.97
Sim(Doc1, Doc4) =	0.42	Sim(Doc5, Doc4) =	0.93

Doc 2 is allocated to cluster 1, along with Doc 1 which is the current cluster mean
Docs 3 & 4 are allocated to cluster 2, along with Doc 5 which is the current cluster mean

3. A new cluster mean is calculated for each cluster:

Cluster 1:

	applications	binary	computer	graph	||x||
Doc1	1	4	1	0	4.24
Doc2	2	2	2	0	3.46
Mean:	1.5	3	1.5	0	3.67

Cluster 2:

	applications	binary	computer	graph	||x||
Doc3	2	1	0	4	4.58
Doc4	3	1	1	3	4.47
Doc5	4	0	0	6	7.21
Mean:	3	0.67	0.33	4.33	5.32

4. The distances between each document and the new cluster means are calculated:

sim(Doc1, mean1) =	0.96	sim(Doc1, mean2) =	0.27
sim(Doc2, mean1) =	0.94	sim(Doc2, mean2) =	0.43
sim(Doc3, mean1) =	0.36	sim(Doc3, mean2) =	0.98
sim(Doc4, mean1) =	0.55	sim(Doc4, mean2) =	0.97
sim(Doc5, mean1) =	0.23	sim(Doc5, mean2) =	0.99

5. All documents remain in the same cluster so the algorithm stops.

Unit 5 - additional notes

KMeans Example 2

Same sample data set of 5 documents, and 5 terms:

	applications	binary	computer	graph	||x||
Doc1	1	4	1	0	4.24
Doc2	2	2	2	0	3.46
Doc3	2	1	0	4	4.58
Doc4	3	1	1	3	4.47
Doc5	4	0	0	6	7.21

1. Let K = 2. Select two rows at random to represent the initial cluster means: Doc1 and Doc2
2. Calculate the distance between each document and the cluster centre (using cosine measure)

Sim(Doc1, Doc3) =	0.31	Sim(Doc2, Doc3) =	0.38
Sim(Doc1, Doc4) =	0.42	Sim(Doc2, Doc4) =	0.65
Sim(Doc1, Doc5) =	0.13	Sim(Doc2, Doc5) =	0.32

All documents are more similar to Doc2, so cluster 1 just has doc1, while cluster 2 has the remaining clusters.

3. A new cluster mean is calculated for each cluster:

Cluster 1:

	applications	binary	computer	graph	||x||
Doc1	1	4	1	0	4.24
Mean:	1	4	1	0	4.24

Cluster 2:

	applications	binary	computer	graph	||x||
Doc2	2	2	2	0	3.46
Doc3	2	1	0	4	4.58
Doc4	3	1	1	3	4.47
Doc5	4	0	0	6	7.21
Mean:	2.75	1	0.75	3.25	4.44

4. The distances between each document and the new cluster means are calculated:

sim(Doc1, mean1) =	1	sim(Doc1, mean2) =	0.4
sim(Doc2, mean1) =	0.82	sim(Doc2, mean2) =	0.59
sim(Doc3, mean1) =	0.31	sim(Doc3, mean2) =	0.96
sim(Doc4, mean1) =	0.42	sim(Doc4, mean2) =	1
sim(Doc5, mean1) =	0.13	sim(Doc5, mean2) =	0.95

5. Document 2 moves from cluster 2 to cluster 1
6. A new cluster mean is calculated for each cluster
Note(from this point on, the figures are the same as Example 1)

KMeans Example 2

Cluster 1:

	applications	binary	computer	graph	||x||
Doc1	1	4	1	0	4.24
Doc2	2	2	2	0	3.46
Mean:	1.5	3	1.5	0	3.67

Cluster 2:

	applications	binary	computer	graph	||x||
Doc3	2	1	0	4	4.58
Doc4	3	1	1	3	4.47
Doc5	4	0	0	6	7.21
Mean:	3	0.67	0.33	4.33	5.32

7. The distances between each document and the new cluster means are calculated:

sim(Doc1, mean1) =	0.96	sim(Doc1, mean2) =	0.27
sim(Doc2, mean1) =	0.94	sim(Doc2, mean2) =	0.43
sim(Doc3, mean1) =	0.36	sim(Doc3, mean2) =	0.98
sim(Doc4, mean1) =	0.55	sim(Doc4, mean2) =	0.97
sim(Doc5, mean1) =	0.23	sim(Doc5, mean2) =	0.99

8. All documents remain in the same cluster so the algorithm stops

Aglomerative clustering example

Same sample data set of 5 documents, and 5 terms:

	applications	binary	computer	graph	||x||	
Doc1	1	4	1	0	4.24	Cluster 1
Doc2	2	2	2	0	3.46	Cluster 2
Doc3	2	1	0	4	4.58	Cluster 3
Doc4	3	1	1	3	4.47	Cluster 4
Doc5	4	0	0	6	7.21	Cluster 5

1. Create 5 clusters, each containing just one document:

Cluster 1 – doc 1
Cluster 2 – doc 2
Cluster 3 – doc 3
Cluster 4 – doc 4
Clsuter 5 – doc 5

2. Find the closest clusters and merge them.

This means calculating the distance between all points (using cosine measure)

Documents:	Similarity	Order of proximity
Sim(Doc1, Doc2) =	0.82	4
Sim(Doc1, Doc3) =	0.31	10
Sim(Doc1, Doc4) =	0.42	7
Sim(Doc1, Doc5) =	0.13	11
Sim(Doc2, Doc3) =	0.38	8
Sim(Doc2, Doc4) =	0.65	5
Sim(Doc2, Doc5) =	0.32	9
Sim(Doc3, Doc4) =	0.93	2
Sim(Doc3, Doc5) =	**0.97**	1
Sim(Doc4, Doc5) =	0.93	2

Doc 3 & 5 are closest with a similarity measure of 0.98. Therefore these are merged into one cluster giving:

Cluster 1 – doc 1
Cluster 2 – doc 2
Cluster 3 – doc 3 & doc 5
Cluster 4 – doc 4

3. Again find the closest clusters and merge.

From the distances above, the next highest similarity measure at 0.93 is between doc 4 and cluster 3 (either doc 5 or doc 3). Doc 4 is added to cluster 3 giving:

Cluster 1 – doc 1
Cluster 2 – doc 2
Cluster 3 – doc 3, doc 4 & doc 5

4. Again find the closest clusters and merge.

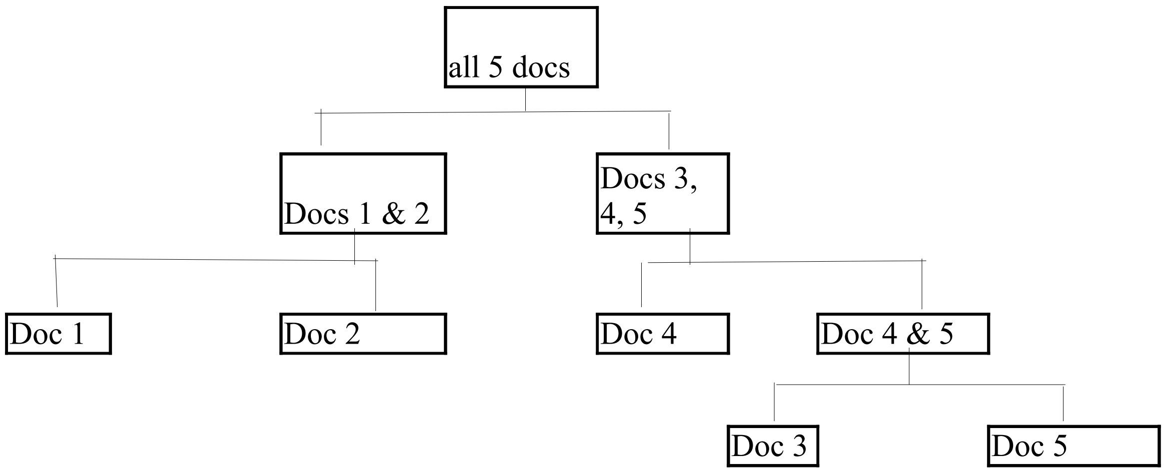
Aglomerative clustering example

The next shortest distance at 0.82 is between doc 1 and doc 2 giving:
Cluster 1 – doc 1 & doc 2
Cluster 3 – doc 3, doc 4 & doc 5

Note: at this points the clusters are the same as for the k-means example

5. Again find the closest clusters and merge.
The next shortest distance is between doc 4 and doc 2, resulting in the final two clusters being merged giving:
Cluster 1 – doc 1, doc 2, doc 3, doc 4 & doc 5

Final result:



Note: Setting a threshold similarity level at around 0.75 would suggest that **two** is the optimal number of clusters for this data.