Document Clustering with Daily Kos

Contents

Introduction

Document clustering, or text clustering, is a very popular application of clustering algorithms. A web search engine, like Google, often returns thousands of results for a simple query. For example, if you type the search term "jaguar" into Google, around 200 million results are returned. This makes it very difficult to browse or find relevant information, especially if the search term has multiple meanings. If we search for "jaguar", we might be looking for information about the animal, the car, or the Jacksonville Jaguars football team.

Clustering methods can be used to automatically group search results into categories, making it easier to find relevant results. This method is used in the search engines PolyMeta and Helioid, as well as on FirstGov.gov, the official Web portal for the U.S. government. The two most common algorithms used for document clustering are Hierarchical and k-means.

In this problem, we'll be clustering articles published on Daily Kos, an American political blog that publishes news and opinion articles written from a progressive point of view. Daily Kos was founded by Markos Moulitsas in 2002, and as of September 2014, the site had an average weekday traffic of hundreds of thousands of visits.

The file dailykos.csv contains data on 3,430 news articles or blogs that have been posted on Daily Kos. These articles were posted in 2004, leading up to the United States Presidential Election. The leading candidates were incumbent President George W. Bush (republican) and John Kerry (democratic). Foreign policy was a dominant topic of the election, specifically, the 2003 invasion of Iraq.

Each of the variables in the dataset is a word that has appeared in at least 50 different articles (1,545 words in total). The set of words has been trimmed according to some of the techniques covered in the previous week on text analytics (punctuation has been removed, and stop words have been removed). For each document, the variable values are the number of times that word appeared in the document.

Exercices

1. Hierarchical Clustering

Problem 1.1 Let's start by building a hierarchical clustering model. First, read the data set into R. Then, compute the distances (using method="euclidean"), and use hclust to build the model (using method="ward.D"). You should cluster on all of the variables.

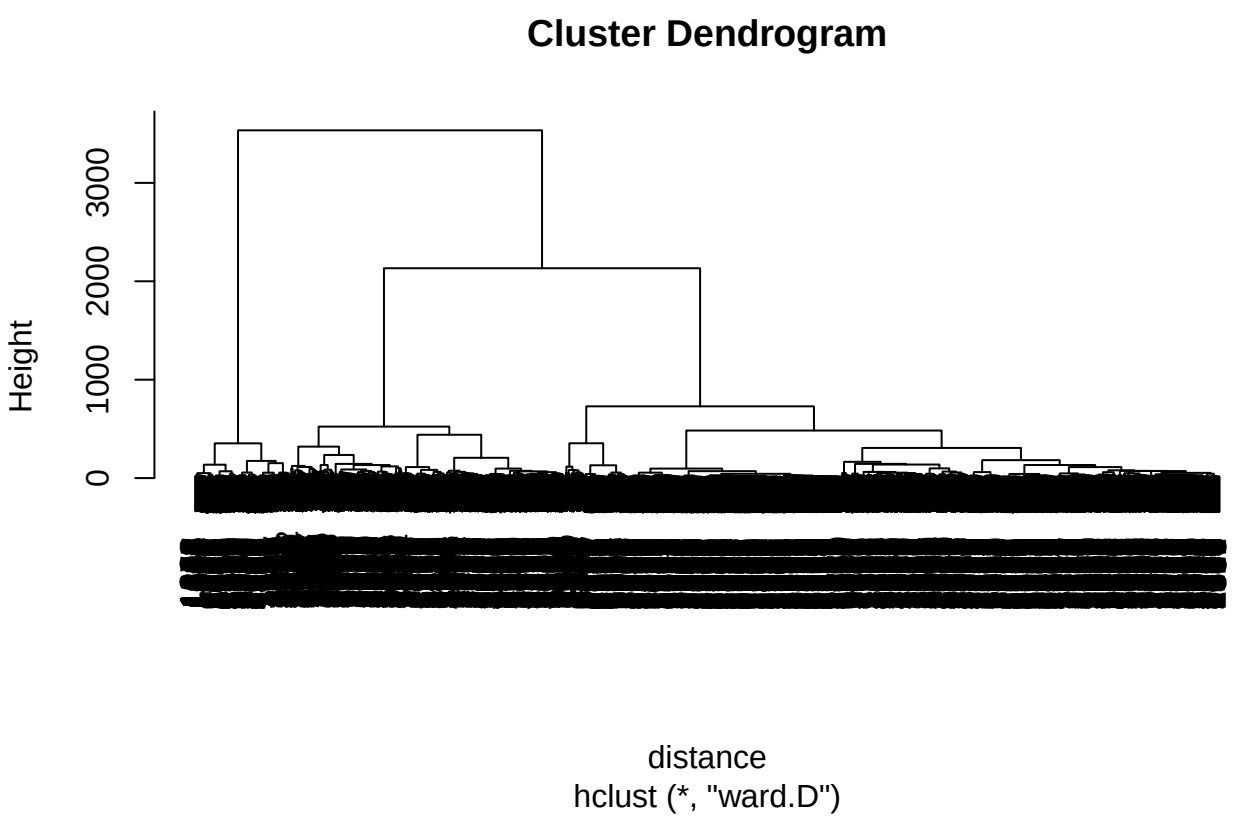
Running the dist function will probably take you a while. Why? Select all that apply.

1. **We have a lot of observations, so it takes a long time to compute the distance between each pair of observations.**
2. **We have a lot of variables, so the distance computation is long.**
3. Our variables have a wide range of values, so the distances are more complicated.
4. The euclidean distance is known to take a long time to compute, regardless of the size of the data.

Explanation :
You can read in the data set, compute the distances, and build the hierarchical clustering model by using the following commands:

The distance computation can take a long time if you have a lot of observations and/or if there are a lot of variables. As we saw in recitation, it might not even work if you have too many of either!

Problem 1.2 Plot the dendrogram of your hierarchical clustering model. Just looking at the dendrogram.



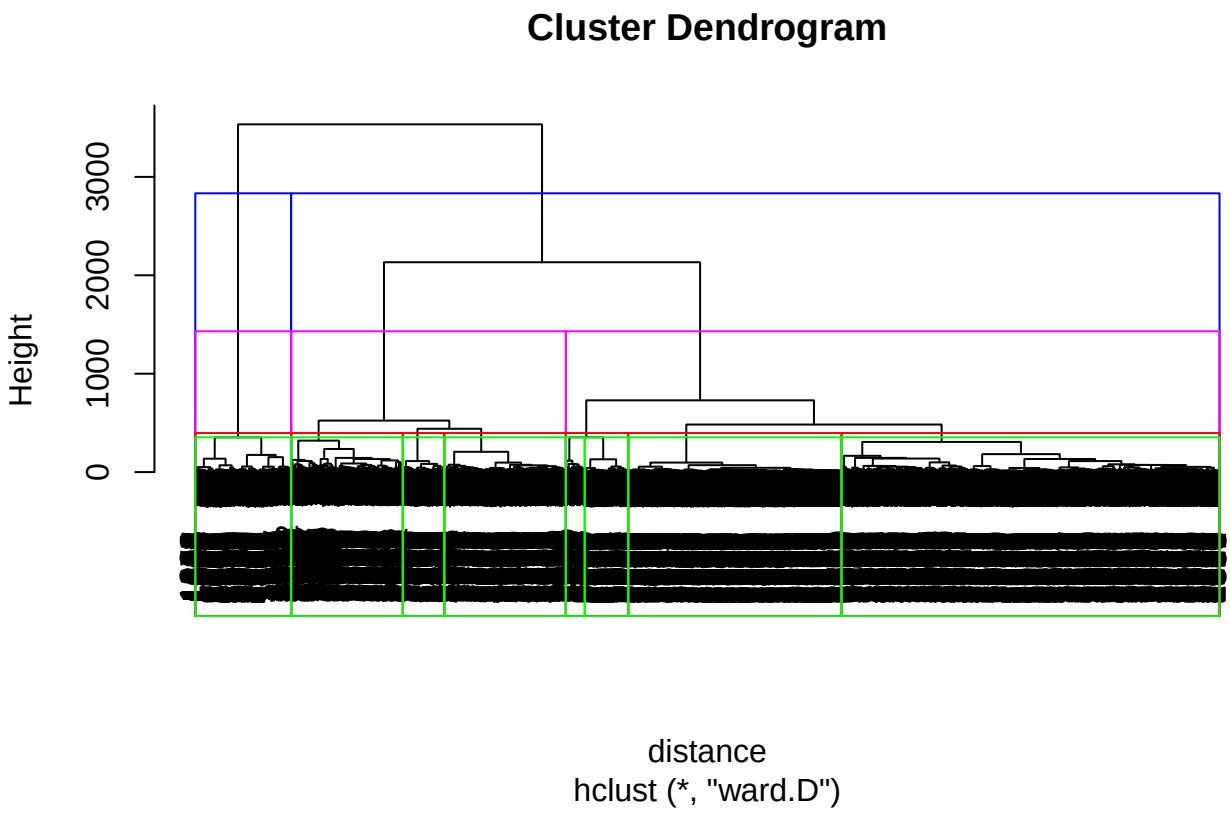
Which of the following seem like good choices for the number of clusters?
Select all that apply.

1. **2**
2. **3**
3. 5
4. 6

Explanation :
You can plot the dendrogram with the command:

Where "kosHierClust" is the name of your clustering model. The choices 2 and 3 are good cluster choices according to the dendrogram, because there is a lot of space between the horizontal lines in the dendrogram in those cut off spots (draw a horizontal line across the dendrogram where it crosses 2 or 3 vertical lines). The choices of 5 and 6 do not seem good according to the dendrogram because there is very little space.

Problem 1.3 In this problem, we are trying to cluster news articles or blog posts into groups. This can be used to show readers categories to choose from when trying to decide what to read. Just thinking about this application.



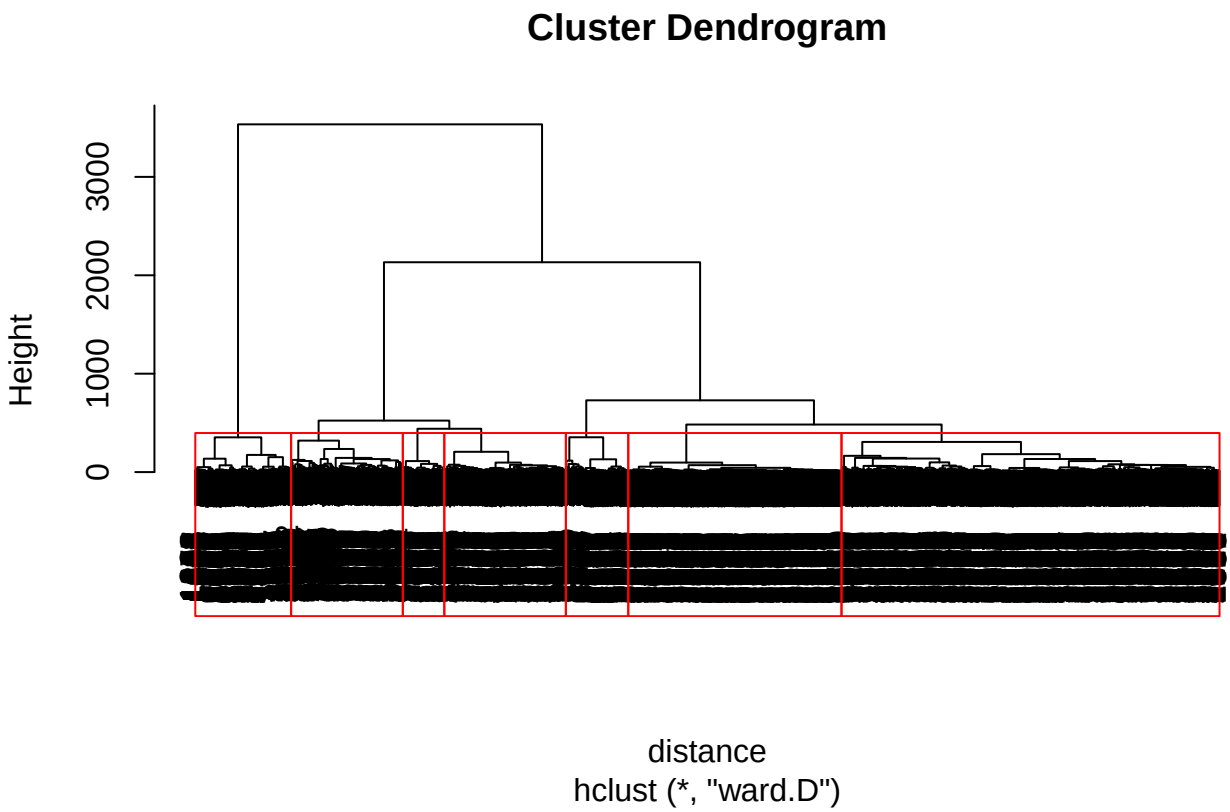
What are good choices for the number of clusters?
Select all that apply.

1. 2
2. 3
3. **7**
4. **8**

Explanation :
Thinking about the application, it is probably better to show the reader more categories than 2 or 3. These categories would probably be too broad to be useful. Seven or eight categories seems more reasonable.

Problem 1.4 Let's pick 7 clusters. This number is reasonable according to the dendrogram, and also seems reasonable for the application. Use the cutree function to split your data into 7 clusters.

Now, we don't really want to run tapply on every single variable when we have over 1,000 different variables. Let's instead use the subset function to subset our data by cluster. Create 7 new datasets, each containing the observations from one of the clusters.




```
## [1] 374
```

How many observations are in cluster 3?

Answer : 374

For all clusters :

```
## [1] 1266  321  374  139  407  714  209
```

Which cluster has the most observations?

1. **Cluster 1**
2. Cluster 2
3. Cluster 3
4. Cluster 4
5. Cluster 5
6. Cluster 6
7. Cluster 7

Which cluster has the fewest observations?

1. Cluster 1
2. Cluster 2
3. Cluster 3

4. **Cluster 4**
5. Cluster 5
6. Cluster 6
7. Cluster 7

Explanation :
You can split your data into clusters by first using the cutree function to compute the cluster numbers:

Then, you can use the subset function 7 times to split the data into the 7 clusters:

If you use the nrow function on each of these new datasets, you can see that cluster 3 has 374 observations, cluster 1 has the most observations, and cluster 4 has the fewest number of observations.

Alternatively, you could answer these questions by looking at the output of table(hierGroups).

More Advanced Approach:
There is a very useful function in R called the "split" function. Given a vector assigning groups like hierGroups, you could split dailykos into the clusters by typing:

Then cluster 1 can be accessed by typing HierCluster[[1]], cluster 2 can be accessed by typing HierCluster[[2]], etc. If you have a variable in your current R session called "split", you will need to remove it with rm(split) before using the split function.

Problem 1.5 Instead of looking at the average value in each variable individually, we'll just look at the top 6 words in each cluster. To do this for cluster 1, type the following in your R console (where "HierCluster1" should be replaced with the name of your first cluster subset):

This computes the mean frequency values of each of the words in cluster 1, and then outputs the 6 words that occur the most frequently. The colMeans function computes the column (word) means, the sort function orders the words in increasing order of the mean values, and the tail function outputs the last 6 words listed, which are the ones with the largest column means.

```
##    kerry     bush
## 1.062401 1.705371
```

What is the most frequent word in this cluster, in terms of average value? Enter the word exactly how you see it in the output:

Answer : bush

Explanation :
After running the R command given above, we can see that the most frequent word on average is "bush". This corresponds to President George W. Bush.

Problem 1.6 Now repeat the command given in the previous problem for each of the other clusters, and answer the following questions.

```
## [[1]]
##      poll  democrat     kerry      bush
## 0.9036335 0.9194313 1.0624013 1.7053712
##
## [[2]]
## challenge      vote      poll  november
##  4.096573  4.398754  4.847352 10.339564
##
## [[3]]
```

```
##      state republican   democrat       bush
##   2.320856   2.524064   3.823529   4.406417
##
## [[4]]
## presided     poll     bush    kerry
## 1.625899 3.589928 7.834532 8.438849
##
## [[5]]
## administration            war           iraq           bush
##       1.230958       1.776413       2.427518       3.941032
##
## [[6]]
##     kerry     elect  democrat      poll
## 0.5168067 0.5350140 0.5644258 0.5812325
##
## [[7]]
##   edward     poll    kerry     dean
## 2.607656 2.765550 3.952153 5.803828
```

Which words best describe cluster 2?

1. november, vote, edward, bush
2. kerry, bush, elect, poll
3. **november, poll, vote, challenge**
4. bush, democrat, republican, state

Which cluster could best be described as the cluster related to the Iraq war?

1. Cluster 1
2. Cluster 2
3. Cluster 3
4. Cluster 4
5. **Cluster 5**
6. Cluster 6
7. Cluster 7

In 2004, one of the candidates for the Democratic nomination for the President of the United States was Howard Dean, John Kerry was the candidate who won the democratic nomination, and John Edwards with the running mate of John Kerry (the Vice President nominee).

Given this information, **which cluster best corresponds to the democratic party?**

1. Cluster 1
2. Cluster 2
3. Cluster 3
4. Cluster 4
5. Cluster 5
6. Cluster 6
7. **Cluster 7**

Explanation :
You can repeat the command on each of the clusters by typing the following:

You can see that the words that best describe Cluster 2 are november, poll, vote, and challenge. The most common words in Cluster 5 are bush, iraq, war, and administration, so it is the cluster that can best be described as corresponding to the Iraq war. And the most common words in Cluster 7 are dean, kerry, poll, and edward, so it looks like the democratic cluster.

2. K-Means Clustering

Problem 2.1 Now, run k-means clustering, **setting the seed to 1000** right before you run the kmeans function. Again, pick the number of clusters equal to 7. You don't need to add the iters.max argument.

Subset your data into the 7 clusters (7 new datasets) by using the "cluster" variable of your kmeans output.

How many observations are in Cluster 3?

```
## [1] 300
```

Answer : 277

For all clusters :

Which cluster has the most observations?

```
## [1] 4
```

1. Cluster 1
2. Cluster 2
3. Cluster 3
4. **Cluster 4**
5. Cluster 5
6. Cluster 6
7. Cluster 7

Which cluster has the fewest number of observations?

```
## [1] 5
```

1. Cluster 1
2. **Cluster 2**
3. Cluster 3
4. Cluster 4
5. Cluster 5
6. Cluster 6
7. Cluster 7

Explanation :
You can run k-means clustering by using the following commands:

Then, you can subset your data into the 7 clusters by using the following commands:

Alternatively, you could answer these questions by looking at the output of table(KmeansCluster$cluster).

More Advanced Approach: There is a very useful function in R called the "split" function. Given a vector assigning groups like KmeansCluster$cluster, you could split dailykos into the clusters by typing:

Then cluster 1 can be accessed by typing

cluster 2 can be accessed by typing

etc. If you have a variable in your current R session called "split", you will need to remove it with %rm(split)% before using the split function.

Problem 2.2 Now, output the six most frequent words in each cluster, like we did in the previous problem, for each of the k-means clusters.

```
## [[1]]
##           time           iraq          kerry administration       presided
##       1.586667       1.640000       1.653333       2.620000       2.726667
##           bush
##      11.333333
##
## [[2]]
##  democrat      bush challenge      vote      poll  november
##  2.899696  2.960486  4.121581  4.446809  4.872340 10.370821
##
## [[3]]
##    voter presided campaign     poll     bush    kerry
## 1.326667 1.336667 1.403333 2.816667 5.963333 6.613333
##
## [[4]]
## republican      elect      kerry       poll   democrat       bush
##  0.5772077  0.5786877  0.6581154  0.7380365  0.7409965  1.1588555
##
## [[5]]
## primaries  democrat    edward     clark     kerry      dean
##  2.333333  2.708333  2.826389  3.083333  5.041667  8.236111
##
## [[6]]
## administration          iraqi       american           bush            war
##       1.396364       1.621818       1.694545       2.607273       3.036364
##           iraq
##       4.094545
##
## [[7]]
##       race     senate      state    parties republican   democrat
##   2.341463   2.409756   2.995122   3.243902   4.200000   6.185366
```

Which k-means cluster best corresponds to the Iraq War?

1. Cluster 1
2. Cluster 2
3. **Cluster 3**
4. Cluster 4
5. Cluster 5
6. Cluster 6
7. Cluster 7

Which k-means cluster best corresponds to the democratic party? (Remember that we are looking for the names of the key democratic party leaders.)

1. Cluster 1
2. **Cluster 2**
3. Cluster 3
4. Cluster 4
5. Cluster 5

6. Cluster 6
7. Cluster 7

Explanation :
You can output the most frequent words in each of the k-means clusters by using the following commands:

By looking at the output, you can see that the cluster best correponding to the Iraq War is cluster 3 (top words are iraq, war, and bush) and the cluster best corresponding to the democratic party is cluster 2 (top words dean, kerry, clark, and edward).

Problem 2.3 For the rest of this problem, we'll ask you to compare how observations were assigned to clusters in the two different methods. Use the table function to compare the cluster assignment of hierarchical clustering to the cluster assignment of k-means clustering.

```
##
## KosGroup    1    2    3    4    5    6    7
##        1    3    0  110 1026   11   62   54
##        2    0  320    1    0    0    0    0
##        3   85    8   24   67    9   42  139
##        4   10    0  123    0    5    0    1
##        5   52    1   34  141    0  171    8
##        6    0    0    0  710    2    0    2
##        7    0    0    8   83  117    0    1
```

Which Hierarchical Cluster best corresponds to K-Means Cluster 2?

1. Hierarchical Cluster 1
2. Hierarchical Cluster 2
3. Hierarchical Cluster 3
4. Hierarchical Cluster 4
5. Hierarchical Cluster 5
6. Hierarchical Cluster 6
7. **Hierarchical Cluster 7**
8. No Hierarchical Cluster contains at least half of the points in K-Means Cluster 2.

Explanation :
From :

We read that 116 (80.6%) of the observations in K-Means Cluster 2 also fall in Hierarchical Cluster 7.

Problem 2.4 Which Hierarchical Cluster best corresponds to K-Means Cluster 3?

1. Hierarchical Cluster 1
2. Hierarchical Cluster 2
3. Hierarchical Cluster 3
4. Hierarchical Cluster 4
5. **Hierarchical Cluster 5**
6. Hierarchical Cluster 6
7. Hierarchical Cluster 7
8. No Hierarchical Cluster contains at least half of the points in K-Means Cluster 3.

Explanation :
From :

We read that 171 (61.7%) of the observations in K-Means Cluster 3 also fall in Hierarchical Cluster 5.

Problem 2.5 Which Hierarchical Cluster best corresponds to K-Means Cluster 7?

1. Hierarchical Cluster 1
2. Hierarchical Cluster 2
3. Hierarchical Cluster 3
4. Hierarchical Cluster 4
5. Hierarchical Cluster 5
6. Hierarchical Cluster 6
7. Hierarchical Cluster 7
8. **No Hierarchical Cluster contains at least half of the points in K-Means Cluster 7.**

Explanation :
From :

We read that no more than 123 (39.9%) of the observations in K-Means Cluster 7 fall in any hierarchical cluster.

Problem 2.6 Which Hierarchical Cluster best corresponds to K-Means Cluster 6?

1. Hierarchical Cluster 1
2. **Hierarchical Cluster 2**
3. Hierarchical Cluster 3
4. Hierarchical Cluster 4
5. Hierarchical Cluster 5
6. Hierarchical Cluster 6
7. Hierarchical Cluster 7
8. No Hierarchical Cluster contains at least half of the points in K-Means Cluster 6.

Explanation :
From :

We read that 320 (97.3%) of observations in K-Means Cluster 6 fall in Hierarchical Cluster 2.