

Clustering and Similarity

BA3 Text as Data · Session 4 · 9 October 2026

Steven Denney

Korean Studies, Leiden University

Class objective

By the end of class, you can compare documents by vocabulary, group them with hierarchical clustering. You'll also be introduced to k-means.

Attendance and review questions

Scan the QR code with your phone and take the survey for attendance and review.



TF-IDF 국민 is in 83 of 100 speeches, so its IDF is low, and so is its TF-IDF.

DTM One row per speech, one column per word.

R `c(3, 5, 7)` creates a vector of three numbers.

Roadmap for today

When	What	Minutes
09:15	Opening survey and review	10
09:25	Similarity, clustering, and k-means	40
10:05	Orange demonstration, clustering	40
10:45	Checklist and next week	15

Screens away

No computers or phones

Three questions for today

Session 3 asked which words set a speech apart. Today we compare whole speeches.

1. How do we measure whether two speeches are alike?

Cosine similarity and cosine distance.

2. How do we use those distances to group speeches?

Hierarchical clustering.

3. Is there another way to group them?

k-means, as a bonus.

Today's corpus has 300 speeches

President	Speeches	Years
Lee Myung-bak	100	2008–2012
Park Geun-hye	100	2013–2016
Moon Jae-in	100	2017–2022

The 100 speeches you know, plus 200 more drawn at random from the same archive.

Same columns and the same metadata.

Question 1. When are two speeches alike?

We want to group speeches by the words they use. First we need to measure how alike two speeches are.

The basic idea

Speeches that use the same words in similar proportions are alike. Speeches that use different words are not.

To measure this, each speech has to become numbers.

Turning speeches into numbers

Speech	경찰	치안	국민	백신	...
Lee, Police Day 2008	89.3	7.2	7.7	0	...
Park, Police Day 2015	95.9	54.4	5.9	0	...
Moon, vaccine hub 2021	0	0	0	96.7	...

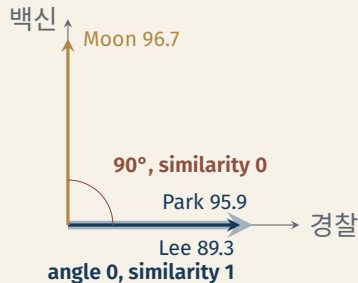
Document-term matrix with TF-IDF values, 300 speeches and 7,265 words.

Each speech is a **vector**, a row of numbers with one value for each word.

The two police speeches give high weight to the same word (**blue**). The vaccine speech weights a different word (**gold**).

How can we measure this difference with a single number?

Cosine similarity measures how alike two speeches are



Two of the 7,265 word columns, as axes.

Cosine similarity

Each speech is an arrow. Its direction depends on the relative weights of its words. Cosine similarity turns the angle between two arrows into a score from 0 to 1.

Small angle, high similarity = Similar vocabulary.

Large angle, low similarity = Different vocabulary.

With all 7,265 words, Lee and Park score **0.77**.

From similarity to distance, four tiny speeches

Speech	경찰	안전	백신
A	2	1	0
B	4	2	0
C	1	2	0
D	0	0	3

Comparison	Similarity	Distance
A and B, same proportions	1.0	0.0
A and C, other proportions	0.8	0.2
A and D, no shared words	0.0	1.0

Cosine distance = $1 - \text{cosine similarity}$. Higher similarity means lower distance. B is A at twice the length, and the distance is still 0.

Similarity and distance in the 300 speeches

Speech 1	Speech 2	Similarity	Distance
Lee, TBS e-FM radio launch	Lee, Busan e-FM radio launch	0.88	0.12
Lee, Police Day 2008	Park, Police Day 2015	0.77	0.23
Park, Police Day 2015	Moon, new police officers 2022	0.76	0.24
Lee, Police Day 2008	Moon, vaccine hub 2021	0.01	0.99

Speeches for the same occasion are close, across presidents and 14 years. A police speech and a vaccine speech share almost no words.

Most speech pairs are not very similar

95%

of all pairs have a similarity
below 0.1, so a distance above
0.9

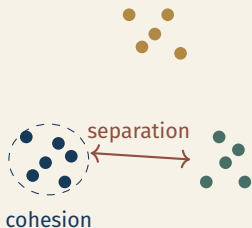
Each of the 300 speeches is compared with the other 299. That gives **44,850 pairs**. Orange's Distances widget computes a distance for each.

We now have a distance for every pair of speeches. *We do not yet have groups.*

Question 2. Clustering

Clustering

Sorting documents into groups by how similar their words are, so that documents in the same group are more alike than documents in different groups.



No labels. Nobody gives the groups in advance.

Cohesion. Speeches in a group are close together.

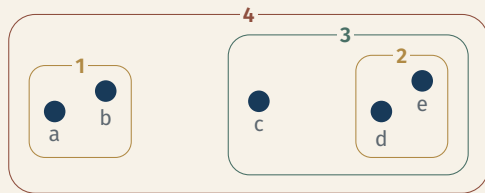
Separation. The groups are far apart.

Close and far are cosine distances.

Hierarchical clustering in three steps

1. Start with every speech as its own group.
2. Join the two closest groups.
3. Repeat until all speeches are in one group.

Step 2 needs a rule for how far apart two groups are. That rule is the **linkage**.



The numbers give the order of the joins.

Linkage, the distance between two groups

Once a group holds several speeches, we need a distance between groups, not only between speeches.

Single The closest pair of speeches.

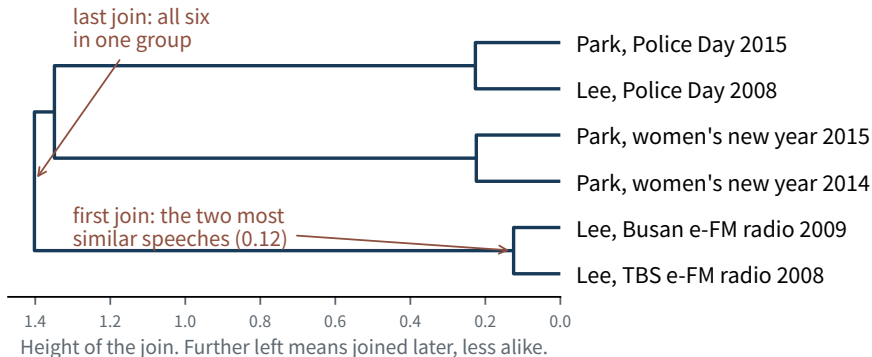
Complete The farthest pair of speeches.

Average The average over all pairs.

Ward Favours joins that keep groups compact.

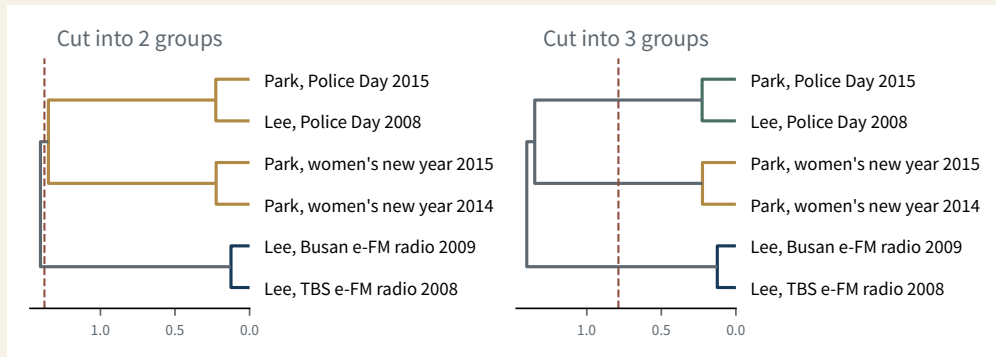
We use **Ward**, Orange's default. It gives compact groups of a size you can read. The appendix compares all four on our speeches.

Reading a dendrogram



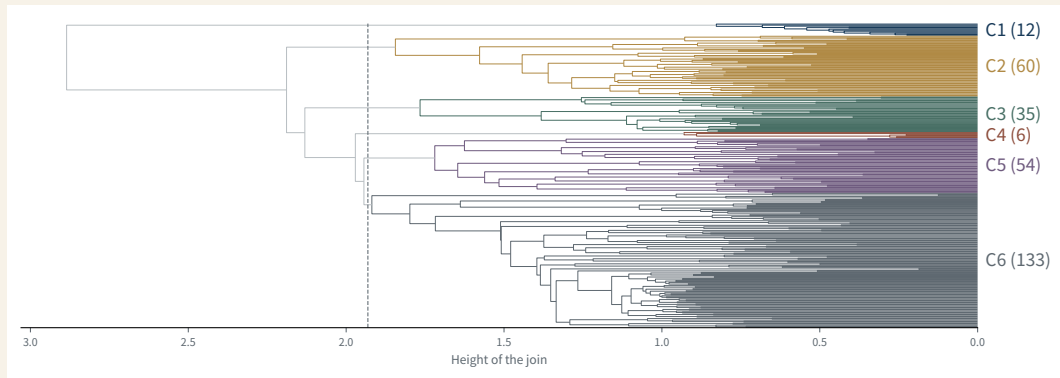
Each name on the right is one speech. Read from right to left. Where two branches meet, two groups join. The further left the join, the less alike the two groups.

Cutting the tree



A cut is a vertical line. Each branch it crosses becomes one group. The tree does not choose the cut. You do, and you report it. These slides cut the 300 speeches into six groups. In the demo you choose three to five.

All 300 speeches, cut into six groups



Ward linkage on cosine distance, as Orange draws it. The dashed line is the cut. C1 to C6 are Orange's names, from top to bottom, not ranks.

What the six groups contain

	Size	Top TF-IDF words	Proposed name	Lee/Park/Moon
C1	12	여성, 가정, 지도자, 양립	Women	1 / 10 / 1
C2	60	코로나, 백신, 방역, 위기, 기업	Economy and crises	19 / 7 / 34
C3	35	양국, 동포, 협력, 한국, 중국	Other countries	4 / 16 / 15
C4	6	경찰, 경찰관, 치안, 범죄	Police	2 / 2 / 2
C5	54	통일, 민주주의, 평화, 북한	Unification and democracy	26 / 15 / 13
C6	133	창조, 방송, 문화, 혁신	Everything else	48 / 50 / 35

Do speeches group by topic or by president?

- Topic** C4 has two police speeches from each president. Police ceremonies share their words, whoever speaks.
 - Topic over time** C2 joins Lee's economic speeches from the financial crisis (G20, trade, budget) with Moon's pandemic speeches. Both use 위기, 기업, and 금융.
 - One president** Ten of the twelve speeches in C1, on women, are Park's.
- Mostly topic.** The algorithm saw only the words, never the president or the date.

The algorithm finds groups. We interpret them.

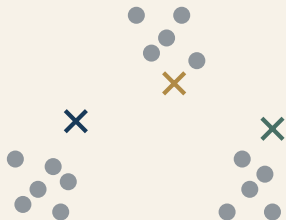
The algorithm groups speeches that share words. It cannot say what a group means. To name a group, read its speeches and use what you know about the data and the topic. This is the qualitative step of a mixed-methods analysis.

C1, women Ten of its twelve speeches are Park's. A Park group? The titles show all twelve are about women.

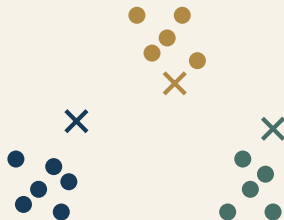
C4, police Moon's cabinet and aides' meetings fall into four groups. The one in C4 discussed police and prosecution reform.

Question 3. Bonus: k-means in three steps

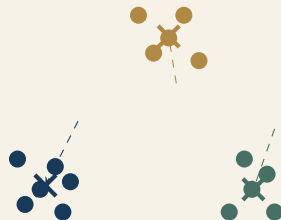
1. Place $k = 3$ centres



2. Each speech joins its nearest centre



3. Each centre moves to the middle of its group



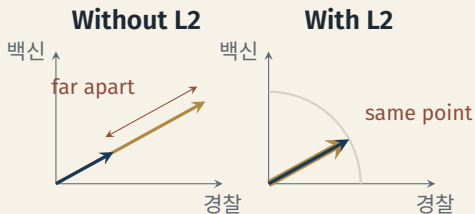
Repeat steps 2 and 3 until no speech changes group. You choose k , the number of groups, before you start. A centre is the average of the speeches in its group. For each word, it holds that word's average value in those speeches.

Hierarchical clustering and k-means

	Hierarchical clustering	k-means
Number of groups	Chosen after, by cutting the tree	Chosen before, as k
Result	A tree of joins	k groups, no tree
Distance	Any, here cosine distance	Straight-line distance to a centre
Demo 4	Required	Optional

k-means is informative to compare, and you may try it. It is not required.

Bonus: why k-means needs L2



Two speeches with the same mix of words, one twice as long.

L2 normalisation

Divides each row by its length, so every speech has length 1. In Orange, Bag of Words with Regularization **L2**.

k-means uses straight-line (Euclidean) distance, which grows with length. With L2, only the mix of words counts.

Without L2, 291 of the 300 speeches fall into one group. Cosine ignores length, so the tree needs no L2.

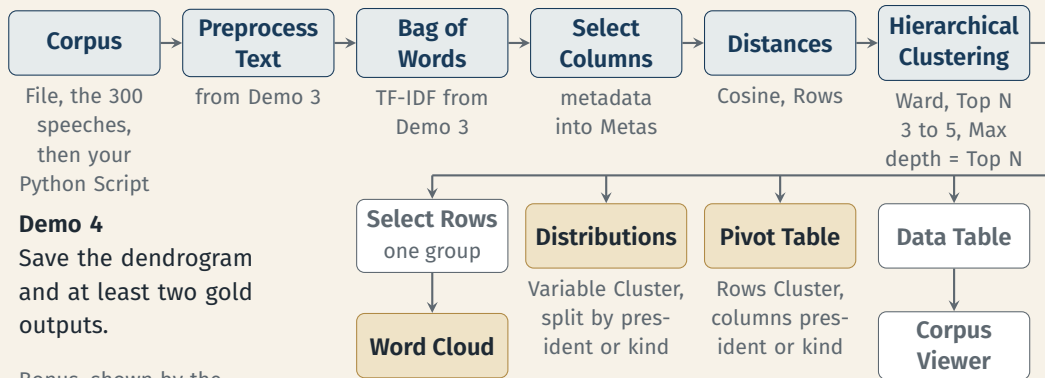
Eight terms you must know

Vector	A speech as a row of numbers, one per word.
Cosine similarity	The cosine of the angle between two vectors. 1 means the same direction, 0 means no shared words.
Cosine distance	$1 - \text{cosine similarity}$. Higher similarity means lower distance.
Clustering	Grouping similar documents, without groups given in advance.
Hierarchical clustering	Joins the two closest groups, step by step, until one group is left.
Linkage	The rule for how far apart two groups are. We use Ward.
Dendrogram	The tree of those joins. Cutting it gives the groups.
k-means	Puts each document with the nearest of k centres, with k chosen in advance.

Laptops open

Open your Demo 3 workflow to build towards Demo 4.

Today's workflow in Orange



Demo 4

Save the dendrogram and at least two gold outputs.

Bonus, shown by the instructor. Bag of Words with L2, Select Columns, k-Means.

Let op! Distances needs a document-term matrix

No matrix Preprocess Text gives cleaned words, not numbers. Connected straight to Distances, every distance is 0.

Counts Raw counts work, but 국민 and 정부 make every speech look alike. The average similarity is 0.14, against 0.03 with TF-IDF.

TF-IDF Words found in almost every speech count for little, so distinctive words decide which speeches are alike.

In Orange, Distances comes after your TF-IDF Bag of Words and Select Columns.

Settings in Hierarchical Clustering

Linkage

Ward ▼

changes the groups Keep **Ward**. Other linkages give other groups.

Annotations

president ▼

drawing only The label on each branch. Pick one metadata column, such as president or kind, and say why in your README.

Pruning

⦿ Max depth

5

drawing only, at Top N or higher Draws only the top of the tree. Set Max depth to your Top N for a readable image.

Selection

⦿ Top N

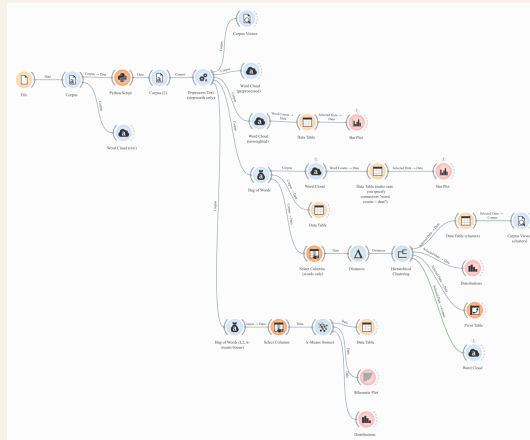
5

changes the groups **Top N** sets the number of groups, 3, 4, or 5. Clicking a branch switches to Manual.

Zoom

drawing only The size of the drawing. Save it with Save Image, bottom left, as output/clusters.png.

Your Session 4 workflow in Orange



The snapshot is also on the Slides page.

Task The same dendrogram, two more figures or tables, and README, made in R. Orange is not permitted.

Sheet The assignment sheet is on the Session 4 page, with every step and the code. You preprocess the 300 speeches with your Session 2 script.

New Cosine similarity with `%*%`, `hclust` with Ward, and `cutree`. `kmeans` is a bonus step.

Before you leave

What must be finished, submitted, and prepared next

Demo 4 checklist

- ☐ The README names the settings (TF-IDF, cosine distance, Ward linkage, the number of groups, and the label).
- ☐ `output/clusters.png` shows the dendrogram with 3 to 5 groups.
- ☐ Two more figures or tables about the groups are in `output/` and shown in the README.
- ☐ The README says what each figure shows and names each group.
- ☐ The workflow or script is in `workflow/`.
- ☐ The final files are pushed to GitHub.

Demo 4 is due Oct. 16, 09:15

Files go in `demos/04-clustering/`. Graded for checklist completion.
k-means is optional.

Next week classifies speeches

Demo 4

Finish the checklist and push.

Due Oct. 16 at 09:15.

Also watch the Orange sentiment tutorial. Details are on the Session 5 page.

Programming exercises

DataCamp Introduction to R, chapters 4, 5, and 6, checked online.

Appendix

Key terms and definitions

Appendix · Similarity and clustering

Vector	A row of the document-term matrix, one number per word.
Cosine similarity	The cosine of the angle between two vectors. Ignores length.
Cosine distance	$1 - \text{cosine similarity}$.
L2 regularization	Rescales each row to length 1. Cosine is unchanged.
Cohesion, separation	Close within groups, far between groups.
Linkage	How the distance between two groups is measured. Ward favours compact groups.
Dendrogram	The tree of joins. The height of a join shows how far apart the two groups were.

Appendix · Four linkages on the 300 speeches

Linkage	Which two groups join next	Six groups on the 300 speeches
Single	The two with the closest pair of speeches	295, and five speeches on their own
Complete	The two whose farthest pair of speeches is closest	159, 76, 30, 25, 6, 4
Average	The two with the smallest average distance over all pairs	288, 8, and four speeches on their own
Ward	The two whose merge adds the least spread inside the new group	133, 60, 54, 35, 12, 6

Ward's rule is defined for Euclidean distance. Orange applies it to cosine distance, so it holds only approximately. Exact Ward on L2 rows gives 175, 57, 46, 12, 6, 4. The slides and demo use Orange's version.

Appendix · k-means on the 300 speeches, $k = 6$

	Size	Top TF-IDF words	Speeches shared with hierarchical clustering
C4	12	여성, 가정, 지도자, 양립	The same 12 as C1, women
C2	35	양국, 아세안, 협력, 한국, 중국	23 of the 35 in C3, other countries
C3	63	코로나, 백신, 방역, 위기	47 of the 60 in C2, economy and crises
C5	51	경찰, 통일, 평화, 민주주의	30 from C5 and 5 of the 6 police speeches
C6	74	창조, 혁신, 센터, 지역	61 from C6. 45 are Park's, her creative economy
C1	65	농업, 마을, 방송, 불교	36 from C6, the rest mixed

With L2. k-means numbers its groups in its own order, so its C4 is the tree's C1.

Appendix · Orange widgets and settings

Select Columns	Chooses which columns are features. Metas are shown but not measured.
Distances	Computes the distance between every pair of rows. We use Cosine.
Hierarchical Clustering	Draws the dendrogram. Top N selects that many groups and adds a Cluster column. Pruning with Max depth draws only the top of the tree.
Pivot Table	Counts rows by two columns, for example Cluster by president.
Distributions	A bar chart of one column, which can be split by another.
Select Rows	Keeps the rows that meet a condition, for example Cluster is C1.
Word Cloud	Shows the words of the speeches it receives. After TF-IDF, a larger word has a higher average TF-IDF value.
k-Means	Groups rows around k centres and adds its own Cluster column.