

Sentiment Analysis and Clustering

BA3 Text as Data · Session 5 · 16 October 2026

Aron van de Pol

Korean Studies, Leiden University

Class objective

By the end of class, you can apply a documented sentiment-scoring workflow, check its scores against the source texts, and compare the tone of clusters.

Attendance

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

No (smart) phone? Ask me for a piece of paper.



Cosine Two police speeches, fourteen years apart, point almost the same way.

Dendrogram The tree of joins. You choose where to cut it.

Cluster Only a proposal until you read the speeches inside it.

Today a new method meets last week's. Sentiment analysis measures tone. Clustering shows what the sentences are about.

Roadmap for today

When	What	Minutes
09:15	Opening survey and review	10
09:25	Sentiment analysis	25
09:50	Clustering meets sentiment	15
10:05	Orange demonstration	40
10:45	Checklist and next week	15

Screens away

No computers or phones

Positive or negative? Which year?

National Assembly minutes

“북한을 화해 협력의 방향으로 끌고 가기 위해 적극적으로 설득하고 때로는 달래기도 하고 ...”

Minister of Unification · Foreign Affairs and Unification Committee

... to steer North Korea toward reconciliation and cooperation, we actively persuade it and at times placate it ...

And this one?

National Assembly minutes

“북한에서 일어나는 고문, 공개처형, 정치범 수용소 감금, 매춘, 강제북송된 영아 살해, 외국인 납치 등을 하지 못하도록 ...”

Saenuri legislator · House Steering Committee

... so that the torture, public executions, political prison camps, prostitution, killing of forcibly returned infants, and abductions of foreigners in North Korea stop ...

You just judged the tone of two texts

First Minister 박재규, October 2000, four months after the first inter-Korean summit.

Second 권석창, October 2016, on a UN human-rights resolution.

Reading 2,990 sentences this way takes days. Today the computer does it, with two methods.

Today's corpus, one row per sentence

Column	Example	What it tells you
sentence	내용에는 북한에서 일어나는 고문, ...	What was said
keyword	북한	North Korea or Japan
date, year	2016-10-20, 2016	When
speaker, role	권석창, legislator	Who, legislator or minister
party, bloc	새누리당, conservative	The speaker's party that day
ruling_opposition	ruling party	In power or in opposition
president	박근혜	Who was president
committee_group	국회운영 house steering	Where it was said
nouns	내용 북한 고문 공개 처형 ...	The nouns, ready for clustering

Today's corpus in numbers

assembly-nk-japan.csv	Sentences
about 북한	1,495
about 일본	1,495
by legislators	2,362
by ministers	628
ruling party, opposition	933, 1,359

Drawn from every committee and plenary sitting, 2000 to 2026, record.assembly.go.kr.

65 sentences per keyword per year, so no year dominates.

2008 to mid-2012 is missing. Those minutes exist only as scans.

Two methods today

Sentiment analysis · new

How positive or negative is a text?

A dictionary of positive and negative words gives each sentence a score.

Clustering · Session 4

What is a text about?

A tree joins sentences that share nouns. Nobody labels anything.

Together they show which topics sound warm and which sound cold.

Sentiment analysis

Measuring tone with two word lists

What is sentiment analysis?

Sentiment analysis · 감성 분석

Measuring whether a text expresses a positive, negative, or neutral attitude.

Our two questions today.

1. Does the tone on North Korea and Japan rise and fall with events?
2. Do ruling-party members speak more positively than the opposition?

Sentiment arithmetic

Words	Score
평화 peace + 희망 hope + 행복 happiness	?
걱정 worry + 위험 danger + 고통 pain	?
감사 thanks + 걱정 worry + 희망 hope	?
정부 government + 회의 meeting + 북한	?

A positive word adds +1. A negative word subtracts 1. Other words count 0.

Sentiment arithmetic

Words	Score
평화 peace + 희망 hope + 행복 happiness	+3
걱정 worry + 위험 danger + 고통 pain	-3
감사 thanks + 걱정 worry + 희망 hope	+1
정부 government + 회의 meeting + 북한	0

A positive word adds +1. A negative word subtracts 1. Other words count 0.

The third sentence mixes feelings. The last one carries no feeling words, so it scores 0.

A sentiment dictionary is two word lists

File	Words	Examples
positive_stems.txt	695	평화 희망 감사
negative_stems.txt	1,399	걱정 위험 고통

The KNU Korean sentiment lexicon (Park et al. 2018). Annotators at Kunsan National University decided word by word.

The files hold stems. 걱정 matches 걱정하다 and 걱정스럽다, because Kiwi cuts our sentences into the same stems.

Orange looks up every stem of a sentence in both lists.

For sentiment, keep the verbs

In Session 2 the course script kept nouns only. Feelings often sit in verbs and adjectives, so today we add **VV** and **VA**.

Setting	What Kiwi keeps of the minister's sentence	Hits
Nouns only	자세 주도 입장 북한 화해 협력 방향 적극 설득 입 장 배려 생각	2 positive
With verbs	... 화해 협력 방향 끝 위하 적극 설득 달래 ...	4 positive

Without verbs, 37% of our sentences match no word at all. With verbs, 30%.

Turning matches into a score

Step	Minister, 2000	Legislator, 2016
1. Count the content words	16	19
2. Positive hits	화해, 위하, 적극, 달래 (4)	none (0)
3. Negative hits	none (0)	감금, 강제, 살해, 납치 (4)
4. $100 \times (\text{pos} - \text{neg}) / \text{words}$	$100 \times 4 / 16$	$100 \times (-4) / 19$
5. Score	+25.0	-21.1
6. Label (sign of the score)	positive	negative

Dividing by the number of words makes long and short sentences comparable.
Our sentences are 36% positive, 41% neutral (score 0), 23% negative.

The score says negative, and it is

강득구, Democratic Party, 2021

“일본의 우익 역사관을 세우기 위한 몸부림은 ... 받아들이기 힘들 정도로 노골적이고 그리고 이성적 틀을 벗어났습니다.”

Special Committee on Budget and Accounts

Japan's struggle to build a right-wing view of history ... has become so blatant it is hard to accept, and has left the bounds of reason.

Score –12.5. The sentence mentions Japan and puts it in a negative light. Which part of Japan's conduct, and why, needs other methods.

Let op! Every dictionary has blind spots

The **Assembly lists** are KNU after reading this corpus. 183 words out, 39 in.

16 words each	KNU hits	Score	Assembly hits	Score
Minister, 2000	+ 화해 위하 적극 달래	+25.0	+ 화해 달래	+12.5
강득구, 2021	+ 위하 이성 — 받아들이 벗어나 의미 힘들	−12.5	none	0.0

Better. 위하 is grammar (위해서, in order to). 적극 only strengthens.

Worse. 받아들이기 힘들다 and 틀을 벗어났다 carry the anger, but the Assembly lists dropped 받아들이 and 벗어나 as usually neutral.

No list is right everywhere. In Orange you run both.

Let op! Kiwi cuts off the 불

김관진, Minister of National Defense, 2013

“... 그것이 북한 내부의 불안정으로 이어질지 ...”

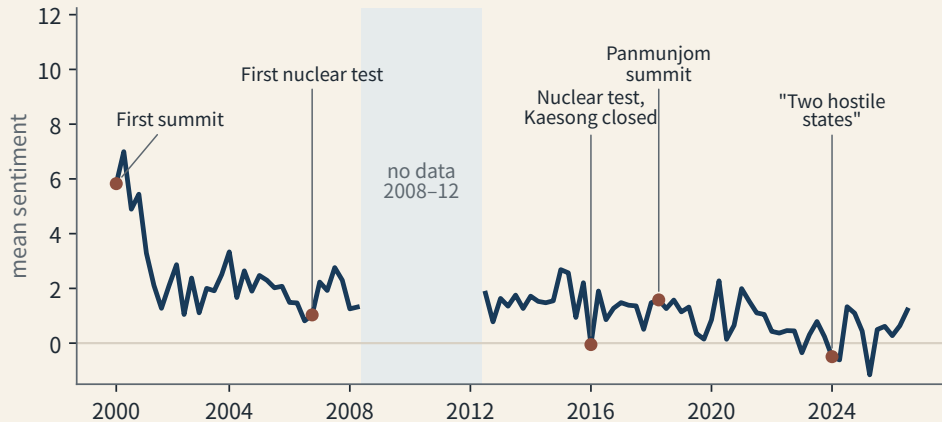
National Defense Committee

... whether it leads to instability inside North Korea ...

Script	What reaches the word lists	Assembly score
Course script	안정 “stability” (the prefix 불 is dropped)	+6.67
JOIN_PREFIXES = True	불안정 “instability”	0.0

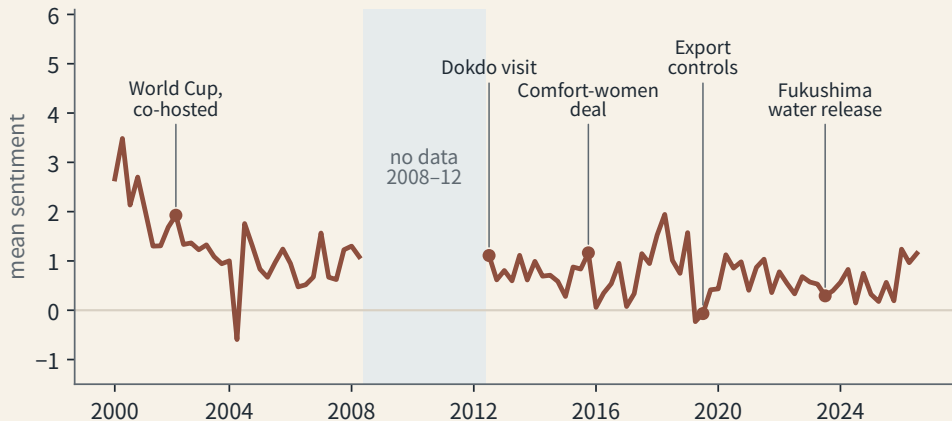
Kiwi tags 불, 비, 미, 무, 반, 물, 부 as prefixes, and the script keeps only nouns and verbs. Your session 5 script glues them back on. It changes 72 of our 2,990 sentences.

Question 1 • Tone on North Korea follows events



Mean sentiment of all 140,679 북한 sentences in the minutes, by quarter.

Tone on Japan dips with each dispute



Mean sentiment of all 93,425 일본 sentences in the minutes, by quarter.

Question 2 · Who sounds more positive?

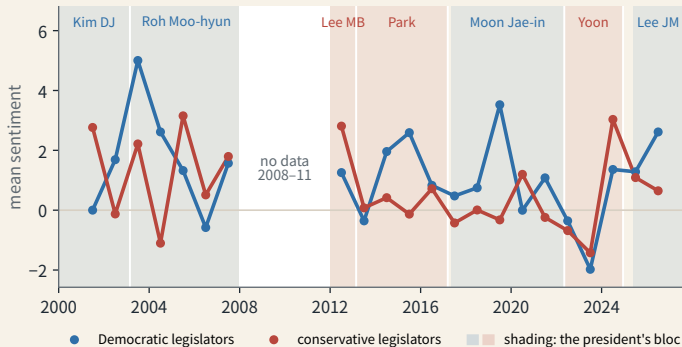
Speaker	북한	일본
Ministers (government)	3.28	2.35
Ruling party	1.29	1.20
Opposition	0.69	0.85

Mean scores in our 2,990 sentences. You will make this table in Orange today.

Ministers sound warmest. The ruling party is slightly warmer than the opposition, on average.

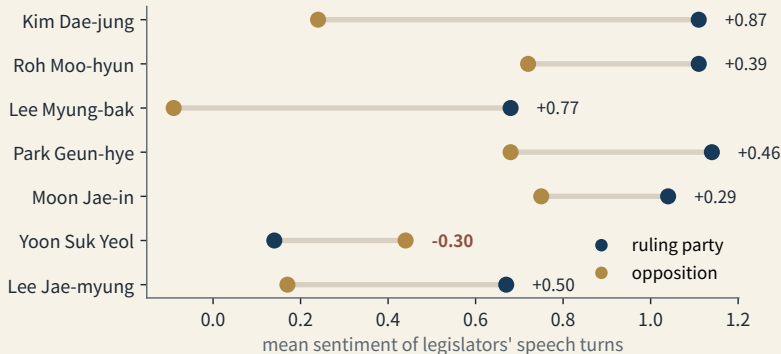
Half a point is small. Single sentences run from -50 to $+40$. Does the gap hold up?

In our small file, the gap wobbles



Legislators in our file, 북한 and 일본 only, about 100 sentences a year. Shading shows the president's bloc. The governing side sounds warmer in 14 of 22 years. Under Park it flips.

In the full minutes, the gap holds



Legislators' speech turns on all topics, full minutes. The ruling party is warmer under six of seven presidents, Park included. A small gap needs a big sample. Why is Yoon's party the exception?

Clustering meets sentiment

What is the cold talk about?

Sentiment says how, not about what

Every sentence now has a score. The score does not say what the sentence is about.

Last week's tree finds that, without labels.

1. Turn each sentence into a vector of its nouns.
2. Cosine distance and Ward linkage build the tree.
3. Cut it into groups, then average the scores in each group.

Two methods, two kinds of preprocessing

	Sentiment	Clustering
Keep	Nouns, verbs, adjectives	Nouns only
Why	Tone sits in 우려하다 worry, 좋다 good	Topic sits in 미사일 missile, 회담 talks
Also drop	Nothing more	Assembly filler such as 생각 thought, 말씀 remark, 위원 member, and 북한 itself
Text column	processed_text, you make it	nouns, ready in the file

북한 is in every sentence, so it tells the tree nothing.

Last week Speeches of hundreds of words. IDF lifted the words that set one speech apart.

Today A sentence keeps about ten nouns. With IDF, one rare word decides where it goes, and the tree splits on oddities.

So Bag of Words with Count, IDF (**None**), L2. Every word in a sentence counts the same.

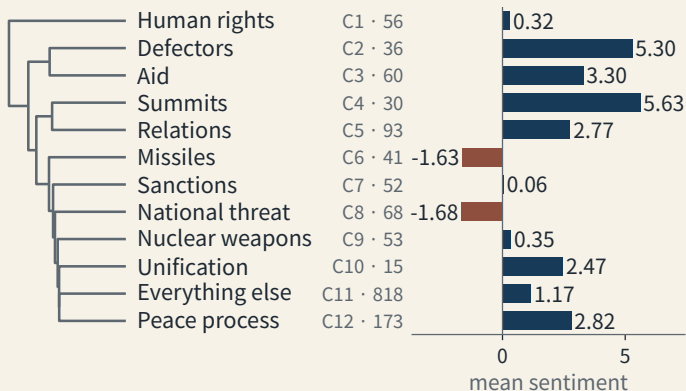
Twelve groups in the 북한 sentences

	n	Top words	Name
C1	56	인권 보고서	Human rights
C2	36	주민 이탈	Defectors
C3	60	지원 인도	Aid
C4	30	회담 정상	Summits
C5	93	남북 관계	Relations
C6	41	미사일 발사	Missiles

	n	Top words	Name
C7	52	제재 조치	Sanctions
C8	68	위협 국민	National threat
C9	53	핵무기 보유	Nuclear weapons
C10	15	통일 준비	Unification
C11	818	도발 가능	Everything else
C12	173	평화 비핵	Peace process

Ward on 1,495 sentences, cut at twelve. At ten, the peace process stays inside the leftover group.

The tree splits engagement from security



Below human rights, the tree divides in two. The branch with defectors, aid, summits, and relations is warm. Missiles and threats sit on the other.

Let op! A cluster is a proposal

C11, 818 sentences Provocation, China, and the president sit together. More than half the sentences mention North Korea in passing, like last week's C6.

C8, threat Threats to the nation, and also attacks on “pro-North forces” at home.

Small groups C10 has 15 sentences. One sentence can move its mean by a point.

Read the sentences in a group before you name it.

Seven terms you must know

Sentiment analysis	Measuring whether a text is positive, negative, or neutral.
Sentiment dictionary	Two word lists, positive and negative.
Sentiment score	$100 \times (\text{positive hits} - \text{negative hits}) / \text{content words}.$
Stopword	A word dropped because it says nothing about the topic, such as 말씀.
Cluster	A group of texts the tree joins from shared words.
Cluster mean	The average sentiment score of the texts in one cluster.
Neutral	Score 0. No feeling, or no word from the lists.

Laptops open

Open session05-starter.ows. It holds the preprocessing, Sentiment Analysis and Formula. We build the rest together.

What we build together

Widget	Settings
In the starter	File, Corpus, Python Script (Kiwi, VV, VA, JOIN_PREFIXES), Corpus on processed_text , Preprocess Text, Sentiment Analysis, Formula (label)
Sentiment Analysis	Orange does not save the word lists. Browse to <code>positive_stems.txt</code> and <code>negative_stems.txt</code>
Data Table, Viewer, Save	Sort by sentiment , read the extremes, save <code>output/text-checks.csv</code>
Pivot Table, Box Plot	Mean sentiment by ruling_opposition and keyword
Select Rows, Columns	북한 sentences. Everything into Metas
Corpus, Preprocess	Text feature nouns . Stopwords <code>stopwords-ko-assembly.txt</code>
Bag of Words, Distances	Count, IDF (None), L2. Cosine
Hierarchical Clustering	Ward, Top N = 12 . Pivot Table and Data Table by Cluster

Formula turns the score into a label

Variable Definitions

New	label	positive if sentiment > 0 else (negative if sentiment < 0 else neutral)	
Remove	☐ Meta attribute	Select Feature	Select Function
Values (optional)		positive, neutral, negative	

	sentence	sentiment	label
include title			
1	따라서 북한의 지방행정제도에 대해서 시간을 가지고 꾸준히 연구해 나가...	4.7619	positive
2	다음으로는 북한이 주한미군 철수문제 제기시 대책이 무엇이나 하는 질의...	0	neutral
3	그러나 말씀하신 추진일정이라든가 소요예산 또 투자예산의 조달방안 등...	13.6364	positive
4	沈載權 위원님께서 남-북한의 교류 협력을 뒷받침하기 위해 법령을 정비...	8.33333	positive
5	앞으로도 정부는 국제사회와 협력해서 북한인권상황 개선을 위해서 계속 ...	15.3846	positive
6	그러나 북한지역의 도로망에 대해서는 현시점에서 재원조달방안 등 구체...	0	neutral
7	그래서 북한 관광문제에 대해서는 논의를 했지, 또 먼 장래에는 상호 왕래...	0	neutral
8	북한은 그동안 체제유지 차원에서 개혁, 개방을 거부하고 우리 식대로 살...	9.09091	positive
9	또한 북한과의 통계분야 상호교류도 적극 추진할 예정인 바 2001년 국제...	7.14286	positive
10	북한군은 남북정상회담 기간 중 특별근무 강화기간을 설정하여 접경지역 ...	9.30233	positive

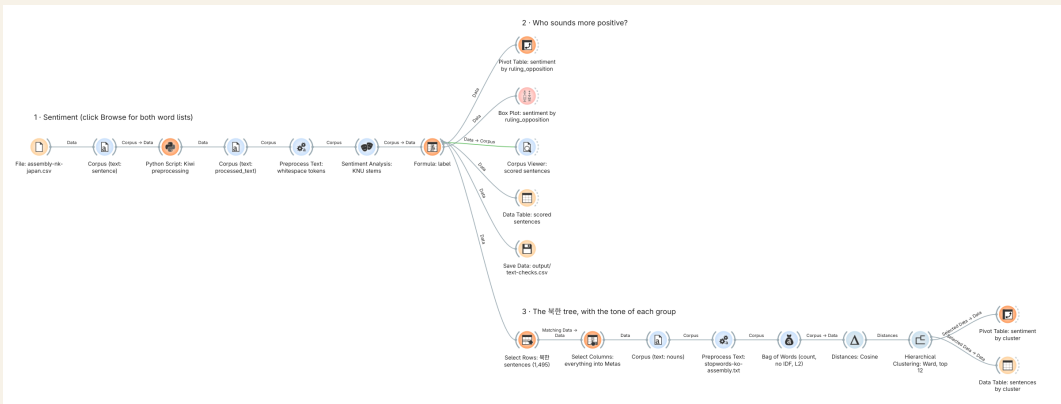
≡ ? | 2990 | 2990

Name the new column, **label**.

Expression. Read it as a sentence.
Above 0 is positive, below 0 negative,
otherwise neutral.

Values lists the categories, so
Distributions and Pivot Table can
count them.

Where we end today



21 widgets. The starter holds the first seven. Row 3 is the tree you run on 일본 for Demo 5.

Demo 5 · What does the Assembly say about Japan?

Run the class workflow on the **일본** sentences. Select Rows keyword 일본, Top N = 12. Then write a paragraph in your Demo 5 README on GitHub.

1. What do members talk about when they talk about Japan? (clustering)
2. With what sentiment do they speak about each topic? (sentiment per cluster)
3. Why are some topics warmer and others colder?

Base it on the clusters and their scores.

Task The same scores, the six checks, and the tree with a mean score per group, made in R. Orange is not permitted.

Sheet The assignment sheet is on the Session 5 page, with every step and the code.

Before you leave

What must be finished, submitted, and prepared next

Demo 5 checklist

- ☐ The scoring method or lexicon is named.
- ☐ The workflow/script and score output are present.
- ☐ The README has a paragraph on what the 일본 tree shows.
- ☐ Six source texts were manually checked.
- ☐ Every checked case has one of the required labels.
- ☐ The final files are pushed to GitHub.

Demo 5 is due Oct. 30, 09:15

Files go in `demos/05-sentiment/`. Check three high and three low scores and mark each reasonable, questionable, or misleading.

Next session finds topics

Demo 5

Finish the checklist and push.

Due Oct. 30 at 09:15.

Also bring one possible hackathon question. Details are on the Session 6 page.

Programming exercises

DataCamp Intermediate R, checked online.

Appendix · Sentiment and clustering

Hit	A word in the text that is on one of the lists.
Neutral	Score 0. No feeling, or no matched words.
Text feature	The column Orange reads as the text. Sentiment uses <code>processed_text</code> , clustering uses nouns.
Stopword file	The course list plus 31 Assembly words, one word per line.
Top words	The words with the highest average Bag of Words weight in a cluster.
Cluster mean	The average score of the texts in one cluster.