# BA2: Digital Korea

## Week 5: Practice & Deepen

### From Preprocessing to Description

Steven Denney

Korean Studies
Leiden University

March 2, 2026

## Today's Agenda

1. Research in action: CTA and Korean Studies
2. Review: Core concepts
3. R programming check-in
4. Descriptive exploration tools in Orange
5. Break
6. Hands-on: Exploring presidential speeches in Orange
7. Looking ahead: Week 6 assessment

# Research in Action

## CTA in Korean Studies

### Your corpus, your questions

**Research questions you can ask**

- Which presidents talked most about 경제?
- When did 통일 disappear?
- How did rhetoric shift across 30 years?

**The tools you already have**

- A preprocessed Korean corpus
- Bag of Words / TF-IDF
- Word clouds, bar charts, metadata

**Today's goal**

Learn to **explore descriptively** — the step between preprocessing and deeper analysis.

# Review: Core Concepts

# Key Terms Review

## Make sure you can explain each of these

| Term | Definition | Description | Think of it as... |
|---|---|---|---|
| **BoW** | Bag of Words | Represent text by word counts, ignoring order and grammar | Cutting words out, dropping them in a bag |
| **TF** | Term Frequency | Count of a word in one document (can be normalized by length) | One cell of the matrix |
| **DF** | Document Frequency | Number of documents in which a word appears | One column's spread across rows |
| **DTM** | Document-Term Matrix | The table (docs $\times$ words) that BoW creates | The spreadsheet the computer "sees" |
| **IDF** | Inverse Doc. Freq. | $\log(N/\text{DF})$; how rare a word is across the corpus | High IDF = rare = informative |
| **TF-IDF** | TF $\times$ IDF | Weighting that highlights words frequent here but rare across the corpus | "Frequent here, rare elsewhere" |

## Matrix vs. DTM vs. Dataframe

**Matrix**
A grid of numbers. No labels, no column names — just rows and columns of values.

General math concept.

**DTM**
A *specific* matrix: rows = documents, columns = terms, cells = counts (or TF-IDF weights).

What BoW creates.

**Dataframe**
A table with named columns of mixed types (text, numbers, dates). What you see in the Corpus widget.

Your CSV is this.

**Key distinction**

Your CSV is a **dataframe**. After Bag of Words, it becomes a **DTM**. Both are tables — but the DTM is all numbers, built for computation.

## What the BoW Widget Actually Does

The Bag of Words widget builds a DTM and gives you three options that change the cell values:

1. **Count** — each cell is the raw integer count of a word in a document
2. **IDF toggle** — multiplies each cell by the word's IDF weight. Common words get down-weighted (e.g., 국민: IDF $\approx 0.19$)
3. **Regularization** (= L2 normalization) — scales each document row to unit length, removing the effect of document length differences

### Why the Word Cloud shows decimals

The Word Cloud displays the **mean value across all documents** for each word. With Count, a word with 3,977 total occurrences across 749 speeches shows as $\approx$5.31. Toggle IDF on and that same word might drop to $\approx$1 because its high frequency is down-weighted.

## Self-Check

Check your understanding with these review questions:

1. What information does the Bag of Words model **discard**?
2. If a word has high TF but low TF-IDF, what does that tell you?
3. What is the difference between TF and DF?
4. Why is the DTM mostly zeros (sparse)?
5. A word appears in 748 out of 749 speeches. Is its IDF high or low?

# R Programming Check-In

## DataCamp Check-In

### Introduction to R: Chapters 4 (Factors) & 5 (Matrices)

**Current assignment:** Factors & Matrices — extended deadline **March 9**.

- Factors: categorical data, levels, ordering
- Matrices: 2D structures, row/column operations
- These connect directly to the DTM concept

**Extra credit opportunity**

Complete **all supplementary DataCamp exercises** (beyond required chapters) → **+0.25** on your final assignment grade.

Any questions or difficulties with R so far?

# Descriptive Exploration in Orange

## Why Start with Descriptive Analysis?

**Before you dive deeper, explore**

- Validates your preprocessing
- Reveals corpus structure
- Generates hypotheses
- Catches data quality issues early

**What it gives you**

- Most frequent terms overall
- Distinctive terms per group
- Distribution patterns
- Keywords in context

**Descriptive analysis is not optional**

Skipping exploration means you won't notice problems until later — wrong column, leftover stopwords, unexpected patterns. Always look at your data first.

# Our Toolkit

| Widget | What it shows | When to use it |
|---|---|---|
| Word Cloud | Visual overview of frequent terms | First look at your data |
| Bar Plot | Precise term frequencies or TF-IDF scores | Comparing values across terms |
| Concordance | Keywords in their original context | Understanding word usage |
| Statistics | Per-document metrics (word count, etc.) | Spotting outliers, data quality |
| Select Rows | Filters corpus by metadata | Creating subsets (by president, year) |

## Statistics Widget

The **Statistics** widget reports per-document metrics:

- **Word count** per document
- **Unique words** per document
- Useful for spotting **outliers**: unusually short or long documents
- Helps identify **data quality issues**: empty documents, duplicates

**What to look for**

If one speech has 50 words and another has 5,000, that difference will affect your analysis. The Statistics widget makes this visible *before* you start interpreting results.

## Word Clouds: What They Show (and Don't)

**What you can read**

- Most frequent terms at a glance
- Relative size = relative frequency
- Quick preprocessing sanity check
- Color by metadata variable

**Limitations**

- Imprecise — size differences are hard to judge
- No exact counts visible
- Layout is random each time
- Can be misleading for comparison

**Rule of thumb**

Word clouds are useful for a **first look**. Always follow up with bar charts for anything you want to claim or compare precisely.

## Bar Charts: Precise Comparison

The **Bar Plot** widget gives you what word clouds cannot: exact values and clear ranking.

- Both Word Cloud and Bar Plot display the **mean value per word** from the DTM
- What those values *are* depends on your BoW settings:
  - **Count** → mean raw frequency across documents
  - **Count + IDF** → mean IDF-weighted frequency
- **Color by metadata** (e.g., president) for comparison

**Try this**

Switch your BoW widget between Count and Count + IDF — watch how the rankings change. Terms that rank high under Count but drop with IDF are common across the corpus (high DF), not distinctive.

## Concordance: Keywords in Context

The **Concordance** widget shows every occurrence of a word with its surrounding text — a KWIC (keyword in context) display.

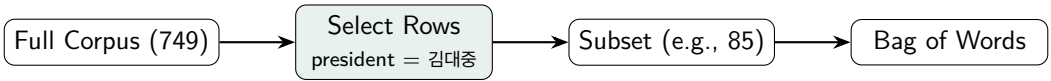
| Left context | Keyword | Right context |
|---|---|---|
| …평화와 | **통일** | 을 위해 노력… |
| …남북 경제 | **통일** | 시대를 준비… |
| …기본질서에 의한 | **통일** | 정책을 추진… |

**Bridge between quantitative and qualitative**

A word count tells you *how often*. Concordance tells you *how* — same word, different meanings across presidents and decades.

## Subsetting: Select Rows

The **Select Rows** widget filters your corpus by metadata values.




- Filter by president, year, speech type, or any metadata column
- Place **before** Bag of Words to focus your analysis on one group
- Connect visualization widgets to the BoW output for focused exploration

## New Corpus: Korean History Textbooks

### NIKH (국사편찬위원회) History Textbook Corpus — demo subset

| Era | Books | Period covered |
|---|---|---|
| Colonial | 3 | Japanese colonial rule (1910–1945) |
| Authoritarian | 3 | Rhee–Park–Chun (1954–1987) |
| Democratic | 3 | 6th–7th Curriculum (1995–2002) |

- 9 textbooks, ~1.8 MB — loads quickly in Orange
- Key metadata: `era`, `period`, `level`, `year`
- Text column: `full_text` (raw) → `processed_text` (after script)

**Research angle**

How did the teaching of Korean history change across political eras?

# Demo

Exploring Korean history textbooks in Orange

# Break

We will resume in 10 minutes.

# Hands-On Activity

## Hands-On: Exploring History Textbooks

**Corpus:** 9 Korean history textbooks from 3 eras (Colonial, Authoritarian, Democratic).

1. Load the CSV → **File** → **Corpus** (select `full_text`)
2. Run the **Python Script** (preprocessing) → **Corpus** (select `processed_text`)
3. Add **Select Rows** → filter by `era`
4. Connect to **Bag of Words** → start with Count
5. Add **Word Cloud** and **Bar Plot** → explore top terms
6. Switch the era in Select Rows → compare

**Work in pairs**

Follow the same workflow we used for presidential speeches — just a new corpus.

## Guiding Questions

1. What are the **top terms** for each era? Do they match what you'd expect?
2. How do rankings change when you switch from **Count** to **TF-IDF**?
3. Compare Colonial vs. Democratic — which terms are **shared** vs. **unique**?

**If you finish early:**

- Use **Concordance** — pick a term and see how its context differs across eras
- Try enabling **Regularization** (L2 normalization) in BoW — how does it affect results?
- Can you find a term that ranks high under Count but drops under TF-IDF? Why?

## Debrief

Let's discuss as a class:

1. What was the most **surprising** finding from your exploration?
2. Did switching from Count to TF-IDF change your interpretation?
3. What **limitations** did you notice in this approach?
4. How might you **follow up** on what you found? What additional analysis would help?

**Remember**

Descriptive exploration generates **hypotheses**. The methods we learn in Weeks 7–10 will help you **test** them.

# Looking Ahead

## Week 6: Assessment Preview

**Part 1: Online Quiz** (∼20 min)

- **10 multiple-choice questions**, each worth **1 point** (10 pts total)
- Covers concepts from Weeks 1–5

**Part 2: Hands-On Task** (∼20 min)

- Download a corpus, preprocess it in Orange, produce a Word Cloud
- Upload to your GitHub repo: `.ows` file, Word Cloud image, short `.md` reflection
- Graded on a 3-point scale:
  - **0** — Incomplete or wrong
  - **1** — Attempted but not fully correct
  - **2** — Completed and fully correct

### Scoring

The average of both parts is rescaled to **10**. A study guide is posted on the course website (Assignments, Week 5).

## Closing

**This week:**

- Review all key concepts (study guide on course website)
- Practice building workflows in Orange
- Complete DataCamp Factors & Matrices (due Mar. 9)

Thank you!

s.c.denney@hum.leidenuniv.nl