

Introduction to Text as Data

BA3 Text as Data · Session 1 · 18 September 2026

Steven Denney

Korean Studies, Leiden University

Class objective

By the end of class, you can distinguish a broad topic from an answerable text-as-data question. You can also describe what belongs in a corpus and organize the first version of your course repository.

Attendance and starter questions

Scan the QR code with your phone. Answer the questions, then put your phone away. No (smart) phone? Ask me for a piece of paper.



Today moves from a visible pattern to a reproducible project

When	What	Minutes
09:15	Opening survey	10
09:25	Sources, course overview, and tools	25
09:50	Setup clinic	25
10:15	Repository workshop	30
10:45	Checklist and next-class contract	15

Listen first, test the tools together, then leave with work visible on GitHub.

Screens away

No computers or phones

What matters in South Korean presidential speeches?

시대 관계 감사
노력 발전 정부 국가 사람
사회 세계 국민 경제 협력
평화 한국 나라 생각 대통령
문제 민족

What question does this pattern make you want to ask?

A first split changes what becomes prominent

Authoritarian

문제 자유 건설
한국 발전 경제 노력 대통령
생각 정부 국민 나라 평화
사회 민족 사람 국가 세계
통일 본인

Rhee · Park Chung-hee
Choi Kyu-hah · Chun Doo-hwan
2,928 speeches

Democratic

문제 기업 시대
감사 생각 한국 국가 노력
발전 세계 국민 경제 평화
대통령 나라 정부 협력 사회
산업 대한민국

Kim Young-sam through Moon Jae-in
six presidents · 5,242 speeches

Size shows frequency per 10,000 retained nouns, on a shared scale.

Yun Posun and Roh Tae-woo are excluded transition cases.

The clouds show (weighted) frequencies only

What you can see

The most frequent retained nouns.

A word cloud alone cannot establish meaning, statistical significance, or causation.

What you still need

Context, comparable speech types, and other text-as-data insights.

A corpus

Corpus · 말뭉치

A collection of documents that contains information you can use for analysis

Document A presidential speech

Metadata President, date, title, and type

We will principally use a corpus of presidential speeches in this class from the Presidential Archives, via the `nlp_corpora` collection on my GitHub.

To the primary source!

Moon Jae-in, presidential inauguration

“국민은 이게 나라냐고 물었습니다. 대통령 문재인은 바로 그 질문에서 시작하겠습니다.”

May 2017 · Presidential Archives

How is **국민** used in inaugural addresses across political eras and by different presidents?

Your domain expertise helps determine what to do: which presidents to consider, what types of addresses, the meaning of the words, and so on.

From presidential speeches to your own research

The crash course introduces you to methods for analyzing primary sources (e.g., text)

Six sessions + a final in-class paper

Ask a question, organize texts, apply a method, and check the result against the sources.

Session 1	Sep. 18	Introduction to Text as Data
Session 2	Sep. 25	Text Preprocessing
Session 3	Oct. 2	Descriptive Patterns
Session 4	Oct. 9	Clustering and Similarity
Session 5	Oct. 16	Classification and Sentiment Analysis
Session 6	Oct. 30	Topic Modeling
Hackathon	Nov. 13	In-class text-as-data hackathon

Each week connects preparation, practice, and submission

Before class Read, watch, and complete assigned R foundations.

In class Discuss a method, then build and inspect a workflow.

After class Finish the individual demo and push the required files.

Demo checklists and R exercises are separate requirements.

Weekly practice prepares you for the hackathon

This strand is 25% of the full course grade.

Within this strand

30% weekly deliverables & attendance, 70% in-class hackathon.

Weekly Individual demos, R foundations, and attendance.

Hackathon Four hours, groups of two or three, and a single report.

Nov. 13 · 09:15–13:15

Two tracks, same research goals

Standard track

Use Orange Data Mining (ODM) for demos and analysis. No prior programming experience is expected.

Everyone

Complete R assignments through Swirl and DataCamp. Submit course work through GitHub.

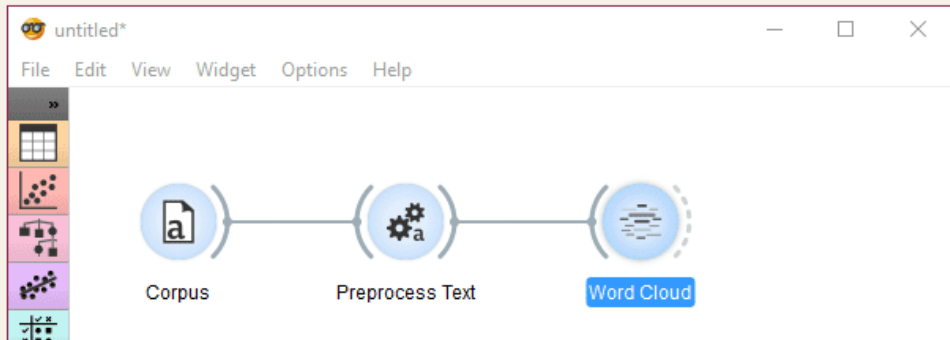
BA2 review track

If you took BA2 Digital Korea, use R for demos and analysis.

Your toolkit

Three tools, different jobs

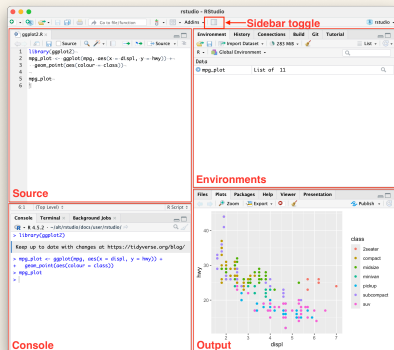
ODM does analysis through connecting widgets



A **widget** performs a task. A **connection** connects data and outputs.

Build visual workflows for Korean text. We start using them next week.

R is the language. RStudio is where you work.



Source

Write and save a script.

Console

Run commands and see results.

Swirl runs inside R. DataCamp provides online practice.

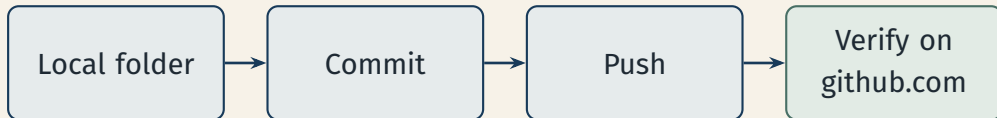
GitHub stores your work. Desktop sends it there.

GitHub

The website that stores your private repository, its files, and their version history.

GitHub Desktop

The application that records local changes and sends them to GitHub.



The course website tells you what to do next

Start here Installations, setup help, and the template.

Sessions Preparation, class tasks, and demo checklists.

Assignments Deadlines and R foundations.

Course guide Tracks, assessment, and policies.

<https://scdenney.github.io/ba3tad-26>

Laptops open

Check the tools, then build your course repository

Five checks

- ☐ Orange opens with Text Mining widgets.
- ☐ RStudio returns 2 for $1 + 1$.
- ☐ GitHub Desktop is signed in.
- ☐ GitHub opens in your browser.
- ☐ Your course folder and template are ready.

DataCamp invitations forthcoming. BA2 review people, see me after class.

If something fails, follow these routes

Check	Route
Orange	Check the Text Mining add-on. Ask about a lab workstation.
RStudio	Record the error, then continue with repository setup.
GitHub Desktop	Sign in through the browser with the same account.
Course folder	Use a permanent folder inside Documents.
Template	Download from Start here and extract once.

Create the private repository, invite, and clone

1. On github.com create ba3tad-26-surname. Set it to Private and add a README.
2. Under Settings and Collaborators, add scdenney.
3. In GitHub Desktop, clone the repository into your course folder inside Documents.

Let op!

Confirm the repository is private and the local path is permanent before adding files.

Build this structure at the repository root

```
ba3tad-26-surname/  
├── README.md  
├── .gitignore  
├── data/  
│   └── README.md  
├── demos/  
│   ├── 01-setup/  
│   ├── 02-preprocessing/  
│   ├── 03-descriptive-patterns/  
│   ├── 04-clustering/  
│   ├── 05-sentiment/  
│   └── 06-topic-modeling/  
└── r-learning/
```

Each demo folder holds README.md, workflow, and output.

Fill it, commit, push, and verify

4. Move the template contents into the cloned folder, with no nested wrapper folder.
5. Fill the Student line in `README.md`. In GitHub Desktop, commit as `Complete initial repository setup`.
6. Push origin, open the repository on github.com, and check that the README and folders open.

STOP Show the pushed README to a neighbor.

Record your sources and commit again

In data/README.md

- Where did the texts come from?
- What do they cover?
- What may you share?

Use not yet chosen for now.

Commit as Add provenance record, then push.

FAIR

- Findable
- Accessible
- Interoperable
- Reusable

The BA2 review track adds an RStudio Project and a script

Orange is not permitted for assessed work on this track.

1. Create an RStudio Project from the cloned directory.
2. Save `session-01.R` under `demos/01-setup/workflow/`.
3. Run the starter script. It prints a three-row data frame and `sessionInfo()`.
4. Commit the project and script, push, and verify.

Before you leave

What must be finished, submitted, and prepared next

Demo 1 checklist

- ☐ The repository is private and scdenney is a collaborator. You must do this so I can track your progress.
- ☐ Every required folder exists and opens on github.com.
- ☐ The README tells a new reader what the repository contains.
- ☐ No passwords, personal data, restricted data, or large corpora are committed.
- ☐ The final commit is pushed and visible on github.com.

Demo 1 is due Sep. 25, 09:15

Complete the required files and two pushed commits. Wording can be refined after class.

Next week we start in Orange

Demo 1

Finish the repository checklist.

Programming exercises

Swirl: six assigned lessons, six screenshots.

DataCamp: chapter 1, checked online.

These are separate requirements. Due Sep. 25 at 09:15.

Also prepare: watch the Orange tutorial, download the speeches, and read *Text as Data in Six Steps*. Bring a question. Details and screenshot filenames are on the Session 2 page.

Appendix

Key terms and definitions

Appendix · Sources and questions

Primary source	Material from the people or period being studied.
Corpus	A selected collection of documents assembled for analysis.
Document	One text treated as a unit, such as a speech.
Metadata	Information describing a document, such as its author, date, or type.
Research question	A focused question that the selected sources can help answer.

Appendix · Reading patterns

Frequency	How often a word occurs. Relative frequency adjusts for the amount of text.
Word cloud	A display in which word size represents frequency or another specified weight.
Preprocessing	Preparing text for analysis through explicit choices about what to retain or change.
Close reading	Interpreting a passage with attention to language, audience, and context.

Appendix · Tools and workflows

Workflow	A connected sequence of steps from input to output.
Widget	An Orange component that performs a task.
R / RStudio	R is the programming language. RStudio is an application for working with it.
Script / Console	A script saves instructions. The Console runs commands and displays results.
GitHub / Desktop	GitHub hosts repositories online. Desktop manages local changes and synchronization.

Appendix · Keeping a research record

Repository	A project's files and their recorded version history.
Clone	Create a local copy of a repository.
Commit / Push	Record a version locally / send commits to the remote repository.
README	A file explaining a project, its contents, and how to use it.
Provenance	A record of where data came from and how they were handled.
FAIR	Findable, Accessible, Interoperable, Reusable: principles for managing research data.