BA2: Digital Korea

Week 1: Introduction & Getting Started

Steven Denney & Aron van de Pol

Korean Studies
Leiden University

February 2, 2026

Today's Agenda

1. Welcome & introductions
2. Course overview
3. What is computational text analysis?
4. Tools and technical setup
5. In-class assignments
6. Looking ahead

Welcome

About This Course

The basics

- 12 sessions
- Mondays, 15:15–17:00
- Huizinga 0.09 (DH Lab) & Arsenaal B0.05

Assessment

- Participation (15%)
- Research Methods Project (35%)
- Final Paper (50%)

What you'll learn

- Treat text as data
- Preprocess, analyze, visualize
- Clustering & classification
- Topic modeling
- Foundational R programming

About Us

Dr. Steven Denney

- Assistant Professor, Korean Studies & IR
- Research: Korean studies, computational social science, DH, comparative politics
- `s.c.denney@hum.leidenuniv.nl`

Aron van de Pol

- PhD Candidate, Centre for Digital Humanities (LUCDH)
- Research: Korean studies, computer vision, modern Korean print culture
- `a.m.van.de.pol@hum.leidenuniv.nl`

Digital Humanities Lab

LUCDH offers support and resources for digital methods in the humanities—a valuable resource for your studies and research.

About You

Next week - when I’m there in person!

Course Overview

Why Computational Text Analysis?

The challenge

- Vast amounts of text data
- Historical archives
- News, social media, government documents
- Too much to read manually

The opportunity

- Discover patterns at scale
- Systematic, reproducible analysis
- New research questions
- Complement close reading

Key insight

Computational methods don’t replace careful reading—they augment it.

Course Trajectory

To the course website!

Learning Objectives

By the end of this course, you will be able to:

1. **Apply** text preprocessing, descriptive analysis, clustering, classification, and topic modeling
2. **Practice** data management and transparency best practices
3. **Establish** a foundation in the R programming language
4. **Reflect** on the strengths and limitations of computational methods and how they apply to the study of Korea and area studies generally

What is Computational Text Analysis?

Text as Data

Core idea

Written language can be transformed into structured data that computers can process and analyze.

This means:

- Words become numbers
- Meaning becomes measurable
- Thousands of texts (“documents”) become manageable
- Hidden patterns become discoverable

What is a Corpus?

Corpus (pl. *corpora*): A structured collection of texts assembled for analysis.

Examples

- Presidential speeches
- Newspaper articles
- Social media posts
- Historical documents
- Interview transcripts

Key considerations

- Selection criteria
- Time period
- Source(s)
- Language(s)
- Metadata

Our Corpora

Course materials

We will work with curated Korean-language corpora spanning historical texts, periodicals, political speeches, social media, and interview data.

- Repository: `https://github.com/scdenney/nlp_corpora`
- Focus on Korea-relevant content
- Truncated versions for classroom use
- Accommodations for non-Korean readers
- We will grow this repository

Tools & Technical Setup

Our Toolkit

Primary analysis

- **Orange Data Mining**
- Visual, drag-and-drop interface
- No programming required
- Powerful text analysis widgets

Programming foundation

- **R + RStudio**
- Industry-standard for data science
- **Swirl** for interactive tutorials
- **DataCamp** for guided courses

Why Orange Data Mining?

- Widget-based: Build workflows visually
- Accessible: Focus on concepts, not coding
- Powerful: Real analysis capabilities

Why R?

Practical reasons

- Free and open source
- Huge ecosystem of packages
- Strong text analysis tools
- Reproducible research

Career reasons

- Widely used in academia
- Growing in industry
- Transferable skill
- Gateway to Python, etc.

Note

No prior programming experience required. We learn together.

GitHub for Version Control

Why GitHub?

- Track changes to your work
- Collaborate and share
- Industry-standard workflow
- Portfolio for future work

Course website: `https://scdenney.github.io/ba2_digital-korea`

In-Class Assignments

Today’s Tasks

Do today, from ”Getting Started”

These steps ensure you have the technical foundation for the semester.

1. GitHub setup
2. Create class repo and share with ’scdenney’ (that’s me)
3. Confirm DataCamp enrollment (check email)
4. Verify installations: RStudio, Swirl, Orange Data Mining (next slide)

Software Verification

Check that you have installed:

- □ **R** — `r-project.org`
- □ **RStudio** — `posit.co/download/rstudio-desktop`
- □ **Swirl** — Run in R: `install.packages("swirl")`
- □ **Orange Data Mining** — `orangedatamining.com/download`

Trouble?

Not to worry. We'll troubleshoot together.

Looking Ahead

For Next Week

R Programming (due by start of Week 2):

- Complete Swirl R Programming lessons:
  - 1: Basic Building Blocks
  - 2: Workspace and Files
  - 4: Vectors
  - 6: Subsetting Vectors
  - 7: Matrices and Data Frames
  - 12: Looking at Data

Week 2 topic: Foundations of Computational Text Analysis

For Next Week

For Week 2:

- Grimmer, Roberts, and Stewart — Chapter 2: “Social Science Research and Text Analysis” (will be provided via email)
- Markdown explainer (on course website, under “Getting Started”)

Orange Data Mining Tutorials:

- Getting Started 01–04
- `https://www.youtube.com/playlist?list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy`

Bookmark this:

- Orange Widget Catalog: `https://orangedatamining.com/widget-catalog/`
- Your go-to reference when learning new widgets or troubleshooting