Digital Humanities: Text-as-Data

Week 1 – Introduction to DH, GitHub, and Orange Data Mining

Steven Denney
Leiden University

BA3 Korean Studies
October 10, 2025

Welcome to the DH Strand

- This is the **Digital Humanities (DH)** component of the BA3 course *Contemporary Korea and Digital Humanities*.
- Focus: **text-as-data methods** using **Orange Data Mining (ODM)**.
- Goal: Learn to prepare, analyze, and interpret text data relevant to **Korean Studies research**.
- The DH strand complements the topical reading seminar by providing practical, research-ready skills.

Course Structure (6 Weeks + Final Project)

Week	Topic
1	Introduction to DH, GitHub & Orange Data Mining
2	Text Preprocessing (tokenization, stopwords, normalization)
3	Descriptive Patterns (frequency, keywords, clustering)
4	Classification & Prediction (supervised methods)
5	Clustering & Similarity (unsupervised methods)
6	Topic Modeling & Project Design
Final	*Hackathon: Group Project Presentation (Nov. 28)*

Learning Goals

- Understand how DH methods apply to humanities and social science questions.
- Gain hands-on experience with Orange Data Mining (no coding required).
- Learn to manage data responsibly using GitHub and FAIR principles.
- Connect digital methods to your BA thesis and research interests.
- *Optional:* Learn the basics of R Programming

Assignments and Deliverables

- Each week: **Deliverables to reinforce learning**
- Submit to your GitHub repository (see our class repor for more).
- **Grading:**
  - 2 Complete & accurate
  - 1 Attempted but incomplete
  - 0 Missing or late
- Instructor reviews repos directly — no Brightspace upload needed.

Assessment Overview

- DH strand = 25% of full course grade.
- Breakdown:
  - Weekly Deliverables & Attendance – 30%
  - Final Project (Hackathon) – 70%
- Hackathon: In-person, 4-hour text analysis project in small groups (Nov. 28).

Optional R Programming Track

- Optional way to extend DH work into coding.
- Platforms:
  - **Swirl** (in-R interactive tutorials)
  - **DataCamp** course: *Introduction to Text Analysis in R*
- Assessment policy:
  - Extra credit: up to +0.25 points for completion
  - Penalty for opting in but not completing
- Opt in only if you plan to finish — totally optional. But I very strongly encourage you to do it!

Week 1 Objectives

Main goals for today:

1. Set up your personal GitHub repository.
2. Install and open Orange Data Mining.
3. Explore the course repository structure.
4. Submit your repo link via Google Sheet.
5. Commit to learning some R?

Deliverables: (see assignments/week01.md)

- README file with reflection:
  - What do you expect to learn in this course?
  - How might DH support your thesis research?
- Screenshot of:
  - Orange installation.
  - Local folder structure (showing repo + cloned course repo).

Setting Up GitHub

- Create a GitHub account (if you don't have one).
- Make a **private repository** named: `DH-TopicalReading-<Surname>`.
- Add instructor (`scdenney`) as collaborator.
- Clone the course repo.
- Copy the `data/nikh` folder into your own repo.
- Submit your repo URL to the shared Google Sheet.

Introducing Orange Data Mining

- Open-source, drag-and-drop data mining suite.
- Ideal for visual and exploratory analysis — no coding required.
- We'll use it to analyze Korean text corpora.
- Key features:
  - Modular workflows ("widgets")
  - Immediate visual feedback
  - Easy export of results (for screenshots & reflection)
- Download: `https://orangedatamining.com`

Tutorials to Watch *Before* Class

- Watch the short tutorials from the official Orange playlist on YouTube.
- Watching these first helps us use class time for skills reinforcement and demos.

The Corpus: National Institute of Korean History (NIKH)

- Our main practice dataset: **NIKH history textbook corpus**.
- State-authorized history textbooks (1895–2015).
- Used to study how the state constructs narratives of nationhood.
- Enables exploration of:
  - How concepts like *minjok*, *kungmin*, and *simin* evolve.
  - Continuity and change across curricula and political eras.
- Data located in the course repo: `data/nikh/nikh.csv`

Why NIKH Textbooks?

- Central to the formation of collective memory and national identity.
- Offer a long-term lens on civic education in South Korea.
- Help bridge theory (national identity, state discourse) and DH practice.
- Also useful for comparative questions — other corpora (e.g. Kaebyŏk, Kyŏngje Yŏngu) available later.

Our Weekly Class Rhythm

1. Watch Orange tutorials before class.
2. Build a small workflow together in class.
3. Discuss what the output tells us.
4. Reflection and discussion.
5. Pull changes from class repo to your local clone (e.g., updates, presentations)

Hands-On Demo: First Workflow

- Today's goal: load and explore `nikh.csv`.
- In Orange:
  1. Use the **File** widget to import data.
  2. Connect to **Corpus Viewer** or **Word Cloud**.
  3. Experiment with basic visualization and inspection.
- This will be our pattern each week — learn by doing, then reflecting.

DEMO

Next Steps

- Complete your Week 1 deliverable by Monday, 17:00.
- Upload all screenshots and reflections to your GitHub repo.
- Optional: Start Swirl R tutorials if you want to explore coding.
- Get ready for Week 2: Text preprocessing!

More on Text Preprocessing – For Next Week

- **Video:** *Getting Started with Orange 16 – Text Preprocessing* youtube.com/watch?v=V70UwJZWkZ8 Demonstrates the Preprocess Text widget, stopwords, and tokenization.
- **Blog:** *Text Preprocessing – Tips & Tricks* orangedatamining.com/blog/text-preprocessing-tricks Explains why preprocessing is "the key and the door" to text mining. Highlights step ordering, stopword effects, and lemmatization.
- **Widget Documentation:** *Preprocess Text (Orange3-Text)* orange3-text.readthedocs.io Describes how Orange applies transformations → tokenization → normalization → filtering → n-grams.
- **Blog:** *Observing Word Distribution* orangedatamining.com/blog/observing-word-distribution Shows how preprocessing choices alter word clouds and key terms.
- **Conceptual overview:** *Text Preprocessing (Old Orange Blog)* oldorange.biolab.si/blog/preprocessing Reflects on why no one-size-fits-all preprocessing exists—especially for non-English languages.