Sungkyunkwan University Winter International Student Experience 2026

[Statistical Foundations for Text Mining with R]

Prof. [Farman Ali], [Department of Applied AI]

COURSE INFORMATION

Credits: 3 credits
Class Hours: 42 contact hours in total]
Class Period: 22 December 2026 – 7 January 2027

SHORT COURSE DESCRIPTION

This course provides a comprehensive, practical foundation for extracting insights from text data using statistical methods implemented in the R programming language. The first half of the course covers key statistical principles needed for data inference, such as conditional probability, probability distributions, confidence intervals, hypothesis testing, and regression analysis.

The second half directly applies these concepts to essential tasks in text mining. Students will learn to manage and process raw text data, convert it into quantitative features, and develop statistical models to recognize patterns, topics, and sentiments. Key topics include the Bag-of-Words model, TF-IDF weighting, topic modelling with Latent Dirichlet Allocation (LDA), sentiment analysis, and text classification using algorithms like Naive Bayes and Regression.

LEARNING OBJECTIVES

Upon completing this course, students will be able to apply essential statistical concepts, preprocess raw text data, utilize both unsupervised and supervised learning methods, and perform text classification.

READING MATERIALS

The lecture materials will be prepared as PPT slides or a PDF file. Both PPT slides and a whiteboard will be used during the lecture to discuss topics. There is no need to purchase a textbook for the first half of the course. The following textbook is required for the second half of the course: Tidy Text Mining with R by Julia Silge and David Robinson.

COURSE REQUIREMENTS AND GRADING

Familiarity with basic statistical concepts, such as mean, median, variance, and standard deviation, is expected. A prior introductory statistics course is strongly recommended. No previous knowledge of R is assumed, but students must have experience programming in at least one language (e.g., Python, Java, or R itself). Basic mathematics is highly advisable.

Students are advised to review the lecture materials in advance. This course is graded on a Pass/Fail basis. Sungkyunkwan University (SKKU) regulations require a minimum score of 60 out of 100 points to pass. The final grade is based on the following components: Attendance 10%, Midterm 30%, Assignments or Quiz 30%, and Final term or Project presentation 30%.

COURSE SCHEDULE

– WEEK I –

1. **Tuesday (22 Dec):** Introduction to this Course (Structured and Unstructured Data), Conditional Probability, Probability Foundations, and Discrete Distributions

2. **Wednesday (23 Dec):** Continuous Distributions and the Normal Distribution

3. **Thursday (24 Dec):** Inference I - The Central Limit Theorem and Estimation, and Quiz

– WEEK II –

4. **Monday (28 Dec):** Statistical Inference II - Introduction to Hypothesis Testing, and Hypothesis Testing in Practice

5. **Tuesday (29 Dec):** Modeling Relationships: Correlation and Regression

6. **Wednesday (30 Dec):** Midterm

7. **Thursday (31 Dec)** – Introduction to Text Mining, Levels of Natural Language, and Installation of R libraries

8. **Friday (1 Jan):** Unsupervised Text Mining: Sentiment analysis and Opinion Mining

– WEEK III –

9. **Monday (4 Jan):** Unsupervised Text Mining: Topic Modeling & Clustering

10. **Tuesday (5 Jan):** Supervised Text Mining - Text Classification

11. **Wednesday (6 Jan):** Text Representation (Lecture), and Student Final Project Presentation

12. **Thursday (7 Jan):** Final Exam or Student Final Project Presentation