Text Mining, Transforming Text into Knowledge: Course Syllabus 2025

Ayoub Bagheri

2025-01-09

1 Introduction

With the rapid growth of digital textual data in many areas of science, there is a growing need for automated tools that can analyse, classify, and interpret this type of data. Text mining techniques can be used to create a structured representation of text, making its content more accessible to users and researchers. Text mining applications are everywhere: social media, web search, advertising, email, customer service, healthcare, marketing, etc. During the course, students will actively learn how to apply text mining methods to data analysis and how to use them together with natural language processing and machine learning techniques on real data problems. The course has a strong practical focus: students will gain hands-on experience in Python by applying the methods to real data during the course and interpreting the results.

Prerequisites

We assume that students who will join the course will have basic knowledge and / or motivation in programming (Python) and data science.

Objectives

The aim of this course is to provide students with an understanding of the principles, problems, techniques, and solutions associated with text mining and to enable them to gain knowledge of how recent advances in text mining relate to innovative approaches to organising, characterizing, finding and exploiting large amounts of textual information in the search for new knowledge. On completion of the course, students should be able to:

- Explain and use text preprocessing and representation techniques.
- Describe a text analysis system and its components, both optional and mandatory.
- Define a text mining pipeline given a practical data science problem.

- Implement all steps of a text mining pipeline: feature extraction, model learning, model evaluation.
- Analyse and reflect on the different techniques used in text mining, the parameters required, and the problem solved.
- Understand and apply some of the state-of-the-art methods in text mining and natural language processing.
- Plan and carry out a text analysis experiment.

2 Course Policy

This course is worth 7.5 ECTS, which means it is designed to give one lecture and on lab per week.

Weekly course flow

A regular week in this course consists of one lecture (Monday at 15:15-17:00) and one lab session (Monday at 17:15-19:00). The material is introduced on a theoretical level in the lectures and then put into practice in the lab sessions. The practical work done in these labs is drawn from real life situations that allow the students to experience how to solve text mining and NLP taks in data science problems.

In addition, students will spend time during the course on two take-home group assignments.

- The **lectures** are in-person. The required readings should be read before the lecture. These are *not* optional.
- The **lab sessions** are in-person interactive sessions in which you apply the methods you learn about in the lectures. The answers to the exercises in the labs are discussed at the end of each session.
- The skills acquired in the lectures and the labs provide the basis for doing the **take-home assignment**. This assignment is made in groups of 3-4 students and handed in via Brightspace.

Synchronous course policy

- 202400006 is an offline-first course, with in-person lectures and lab sessions.
- We find it important for interactive and collaborative learning that the course is offline-first.
- If you miss a session, e.g., due to sickness, you should catch up in the regular way:
  - Read the readings
  - Go through the lecture slides
  - Do the practicals
  - Ask your peers if you have questions

– (after the above) ask the lab teacher and the lecturer for further explanation

Who to ask what

If you have questions, first **ensure the answer isn't in this syllabus** and then follow the table below:

Question type	How to ask
Course proceedings	Email course coordinator (a.bagheri@uu.nl)
Content - general	Email / ask lecturer
Practical content	Email / ask lab teacher (t.h.vanderkuil@uu.nl)
Assignment content	Email / ask lab teacher
Lecture content	Email Lecturer

Grading policy

Your final grade in the course consists of the following grading components:

- Assignments (10%): There are two group assignments. Each assignment is graded and worth 5% of the final grade.
- Final exam (90%): At the end of the course, there is a final exam. The exam consists of TRUE/FALSE, multiple-choice, and open questions.

To pass the course:

- The weighted final grade of all grading components should be greater than or equal to 5.5. There is a minimum required grade of 5.5 for each of the above grading components.

Resit:

- You can only retake one resit and only for the exam.

3 Course materials

Required Software

In this course, we will use Python. Try to install both on your computer by the start of the course.

Installing Python & Jupyter The best choice is to install Python and Jupyter on your computer, and for the easiest complete way could be to install [Anaconda] (https://www.anaconda.com/download). Otherwise you can use Google Colab, which is an interactive online notebook environment; this means no installation is necessary! However,

you do need a Google account, so make sure you have one (or make one specifically for the course).

Required readings

Freely available sections from the following books:

Book	Title (Authors)	URL
`SLP3`	Speech and Language Processing, third edition (Jurafsky & Martin)	link
`NLPE`	Natural Language Processing (Eisenstein)	link
`ISLR`	Introduction to Statistical Learning (James et al.)	link
`IVFS`	Introduction to Variable and Feature Selection (Guyon & Elisseeff)	link
`PTMs`	Probabilistic Topic Models (Blei)	link

And some other freely available articles & chapters

4 Class Schedule

You can find the up-to-date class schedule with locations on mytimetable.uu.nl.

Week	Date	Topic	Type	Reading
1	2025-02-03	Intro to text mining & regular expressions	Lecture	Syllabus
1	2025-02-03	Intro to text mining & regular expressions	Lab	NLPE 1, SLP3 2.1
2	2025-02-10	Text preprocessing	Lecture	SLP3 2.2, 2.3, 2.4, 2.5, 2.6, 2.7
2	2025-02-10	Text preprocessing	Lab	
3	2025-02-17	Text classification	Lecture	SLP3 4.1, 4.2, 4.3, 4.7, 4.8, NLPE 4.4
3	2025-02-17	Text classification	Lab	
4	2025-02-24	10:00 AM assignment 1	Deadline	
4	2025-02-24	Feature selection	Lecture	IVSF

Week	Date	Topic	Type	Reading
4	2025-02-24	Feature selection	Lab	
5	2025-03-03	Clustering & topic modeling	Lecture	ISLR 12.4, PTMs
5	2025-03-03	Clustering & topic modeling	Lab	
6	2025-03-10	Word embedding	Lecture	SLP3 6.3, 6.8, 6.9, 6.10, NLPE 14.5
6	2025-03-10	Word embedding	Lab	
7	2025-03-17	Deep learning & LLMs	Lecture	SLP3 7.1, 7.3, 8.1, 8.2
7	2025-03-17	Deep learning & LLMs	Lab	
8	2025-03-24	10:00 AM assignment 2	Deadline	
8	2025-03-24	Sentiment analysis	Lecture	NLPE 4.1, SLP3 4.4
8	2025-03-24	Sentiment analysis	Lab	
9	2025-03-31	Responsible text mining & applications	Lecture	NLPE 14.6, SLP3 6.11
9	2025-03-31	Responsible text mining & applications	Lab	
10	2025-04-10	Final exam	Exam	
	2025-04-22	Inspecting the final exam	Exam Inspection	
10	2025-05-14	Resit exam	Exam	