Text Mining 2022-2023

Teacher: Suzan Verberne
Teaching assistants: Amin Abolghasemi, Juan Bascur Cifuentes, Pavlos Zakkas, Kamand Hajiaghapour
Contact address: tmcourse@liacs.leidenuniv.nl

Course schedule

The course weeks consist of: a lecture, literature to read, and either a practical exercise (tutorial style) or a hand-in assignment.

The lectures are on Wednesday, 9.00-10.45.
Location:

- September: CORPUS / 2.02 (Corpus Nederland)
- October-December: GORL / 01 (Gorlaeus building)

The literature will be distributed on Brightspace. The majority of the chapters comes from this book, abbreviated as J&M in the course schedule below.

- J&M: Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed), 2021

Week	Lecture	Literature	
1 (7 Sept)	Introduction		
2 (14 Sept)	Text processing	J&M chapter 2. Regular Expressions, Text Normalization, Edit Distance	Exercise: Chapter 1 of "Advanced NLP
3 (21 Sept)	Vector Semantics	J&M chapter 6. Vector Semantics	Exercise: Word Embedding Tutorial: W
4 (28 Sept)	Text categorization	J&M chapter 4.1-4.3. Naive Bayes Classification	Exercise: Text classification tutorial (sk
5 (5 Oct)	Data collection and annotation	Finin (2010). Annotating Named Entities in Twitter Data with Crowdsourcing McHugh (2012). Interrater reliability: the kappa statistic	**Assignment 1. Text classification** (dea
6 (12 Oct)	Information Extraction	J&M chapter 8. Sequence Labeling for Parts of Speech and Named Entities J&M chapter 17. Information Extraction	Exercise: Sequence labelling tutorial (cr
7 (19 Oct)	Neural NLP and transfer learning	J&M chapter 7. Neural Nets and Neural Language Models J&M chapter 9. Deep Learning Architectures for Sequence Processing	Exercise: Fine-Tuning BERT (Hugging
(26 Oct)	No lecture		
8 (2 Nov)	Text summarization	To be decided	**Assignment 2. Information Extractio**
9 (9 Nov)	Sentiment analysis	To be decided	Exercise: to be added
10 (16 Nov)	Biomedical text mining	Lee et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining	
11 (23 Nov)	Industrial Text Mining	Guest lecture	Paper reading for the final assignment
12 (30 Nov)	Conclusions		**Final assignment** (deadline 8 Jan)
13 (7 Dec)	Online lab session		**Final assignment** (deadline 8 Jan)
(3 Jan)	Exam		
(3 Feb)	Re-sit		

The assessment of the course consists of a written exam (50% of course grade) and practical assignments (50% of course grade). The practical assignments comprise two smaller assignments (10% each) and one more substantial, final assignment (30%). The grade for the written exam should be 5.5 or higher in order to complete the course. The average grade for the practical assignments should be 5.5 or higher in order to complete the course. If one of the tasks is not submitted the grade for that task is 0. Each assignment has a re-sit opportunity (a later submission). The maximum grade for a re-sit assignment is 6.

Earlier editions of this course

Link to the course page for this course in 2021-2022
Link to the course page for this course in 2020-2021
Link to the course page for this course in 2019-2020
Link to the course page for this course in 2018-2019

Ontwerp: Free CSS Templates.