



Data Science: Introduction to Text Mining with R (S41)

12 – 14 July 2021

Course location:

Course Director: Dr. Ayoub Bagheri

E-mail: ms.summerschool@uu.nl

Saturday 13 & Sunday 14 July 2021		
Time	Activity	Description
12.00 – 18.00	Key pick up	<i>You will find the exact key pick up location in the pre-departure information, which becomes available after you have paid the course fee.</i>

Day	Time	Type	Description	Location
Monday	09:00 – 10:30	Lecture	Introduction to R & RStudio: <ul style="list-style-type: none"> - Basic commands Introduction to text mining: <ul style="list-style-type: none"> - Definitions - Regular expressions 	
	10:45 – 11:45	Computer Lab	R intro and processing textual data in R	
	11:45 – 12:30	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	
	14:00 – 15:30	Lecture	Preprocessing text: <ul style="list-style-type: none"> - Cleaning - Tokenization (n-grams) - Stop words removal - Stemming Language detection and spell checking POS tagging, Named Entity Recognition	
	15:45 – 16:30	Computer Lab	Text preprocessing Text visualisation N-gram tokenization	
	16:30 – 17:00	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	



Day	Time	Type	Description	Location
Tuesday	09:00 – 10:30	Lecture	Text as features: <ul style="list-style-type: none"> - Vector space model - Representation - Feature selection methods Text classification	
	10:45 – 11:45	Computer Lab	Training and test data K-fold cross validation Naïve Bayes and SVM	
	11:45 – 12:30	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	
	14:00 – 15:30	Lecture	Sentiment analysis <ul style="list-style-type: none"> - Multi-class classification 	
	15:45 – 16:30	Computer Lab	Sentiment classification	
	16:30 – 17:00	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	

Day	Time	Type	Description	Location
Wednesday	09:00 – 10:30	Lecture	Text clustering: <ul style="list-style-type: none"> - K-Means - Topic modelling (latent Dirichlet allocation) 	
	10:45 – 11:45	Computer Lab	Text clustering	
	11:45 – 12:30	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	
	14:00 – 15:30	Lecture	Word embedding <ul style="list-style-type: none"> - Word2Vec, GloVe, P2Vec 	
	15:45 – 16:30	Computer Lab	Training word vectors Vector similarities Using pre-trained word vectors	
	16:30 – 17:00	Plenary Discussion	Students and teachers discuss and present their solutions to the computer lab	

For information about the Social Programme, please visit the [Utrecht Summer School website!](#)