Einführung in NLP-Anwendungen mit R und Python. Unter besonderer Berücksichtigung von Twitterdaten

	Tag 1	Tag 2
09:00-10:30	Vorstellung Kurze theoretische Einführung NLP Quanteda universe Analyse-Pipeline Beispieldatensatz Technisches Set-up	Theorie Word Embeddings (kurz) Feature-Generierung für Klassifikation ML-Background (Learner, Metrics, Train-Test, Split, Tuning, ...) mlr3 universe
10:30-10:45	Pause	Pause
10:45-12:00	Kurze Einführung Scraping Basic text cleaning I (Regex, Symbole)	Klassifikation Ergebnisanalyse Visualisierung Kleiner Ausblick mlr3-Pipeline (Erklärung, warum eigentlich nested resampling erforderlich)
12:00-13:00	Pause	Pause
13:00-14:30	Basic text cleaning II (Stemming, Lemmatization) Static feature extraction (Dictionary-based, POS Tagging)	Einführung in Deep Learning Transfer Learning BERT - fundamental tasks, basic & advanced application
14:30-14:45	Pause	Pause
14:45-16:00	Theorie Topic Modeling (Simon & Patrick?) Implementierung STM & keywordbasierte Variante (eher Präsentation)	Praxis: Implementierung BERT SA (Fine Tuning) Feedback & Wrapup

30.04.2021 (Fr) + 07.05.2021 (Fr)

- Allgemeines
- R
- BERT
- überwiegend Frontalunterricht
- überwiegend Ausprobieren

Tag	Block	Typ	Part	Inhalt	Von	Bis	Dauer	Gesamt	Moderator						
1	1	Organisatorisches	Allgemein	**Vorstellungsrunde & Seminarablauf**	09:00	09:10	10	90	Beide		Zeitanteil			**Offene Fragen**	
1	1	Theorie	Allgemein	**Kurzeinführung NLP** Was gibt es generell für Tasks? Was verbirgt sich grob hinter Sentimentanalyse & Topic Modeling?	09:10	09:25	15	90	Asmik		Asmik	47%		Wie läuft die Anmeldung etc., müssen wir da irgendwie tätig werden?	
1	1	Theorie	Allgemein	**Beispieldatensatz** Wie sehen unsere Daten aus? Was ist speziell bei Twitter-Daten zu beachten?	09:25	09:40	15	90	Asmik		Lisa	45%		Kenntnisstand der Leute bzgl. NLP & R?	
1	1	Theorie	Allgemein	**Analyse-Pipeline** Welche Aufgaben müssen hier konkret erledigt werden?	09:40	09:50	10	90	Lisa		Beide	8%		Sprache?	
1	1	Theorie	Allgemein	**Quanteda universe** Was sind Dokumente, Corpus, Tokens, DFM?	09:50	10:10	20	90	Lisa					Code-Ausführung: Google Colab oder so was? (idealerweise unabhängig von lokalen Geräten)	
1	1	Praxis	R	**Technisches Set-up** Sind alle in der Lage, die Übungen auszuführen?	10:10	10:30	20	90	Lisa					Topic Modeling theoretischer Part: Dürfen wir uns da Simon & Patrick ins Boot holen?	
1	2	Theorie	R	**Scraping** Wie funktioniert Scraping prinzipiell? Was sind die most basic steps?	10:45	11:00	15	75	Asmik					Evaluation als Berichtsbestandteil: klassischer LMU-Bogen oder selber Feedback abfragen?	
1	2	Praxis	R	**Scraping** Greife auf Website zu und ziehe Information XY	11:00	11:30	30	75	Asmik						
1	2	Theorie	R	**Basic Text Cleaning: Regular Expressions** Was sind Regex, wie können wir damit umgehen?	11:30	11:50	20	75	Lisa						
1	2	Praxis	R	**Regular Expressions** Einstiegsaufgabe: gegebene Patterns verschiedenen Fragestellungen zuordnen	11:50	12:00	10	75	Lisa						
1	3	Praxis	R	**Regular Expressions** Selbst Patterns schreiben Inklusive stringr Basics (Finden, Entfernen, Ersetzen)	13:00	13:30	30	90	Lisa						
1	3	Theorie	R	**Basic Text Cleaning: Stemming, Lemmatization** Was ist das alles, wofür brauchen wir das?	13:30	13:40	10	90	Asmik						
1	3	Theorie	R	**Static Feature Extraction: Dictionary-based features, lexikalische Features, POS Tags** Welche Features sind generell sinnvoll? Was verbirgt sich hinter ausgewählten?	13:40	13:50	10	90	Asmik						
1	3	Praxis	R	**Dictionary-based features** Selber Dictionary erstellen und Look-up durchführen	13:50	14:10	20	90	Lisa						
1	3	Praxis	R	**Lexikalische features** Eigene Kategorie definieren (z. B. Verneinungen zeigen und Intensivierungen machen lassen) und Match durchführen	14:10	14:30	20	90	Lisa						
1	4	Theorie	R	**Topic Modeling** Kurzer Abriss: was ist TM, welche Möglichkeiten existieren?	14:45	14:55	10	75	Asmik						
1	4	Theorie	R	**Structural Topic Model** Was ist die Idee und wie sieht die Implementierung aus? Simon & Patrick? (wenn ja, Implementierung für die beiden aufbereiten)	14:55	15:25	30	75	Beide						
1	4	Theorie	R	**Topic Modeling in unserer Pipeline** Wie wurde STM integriert (inklusive Pooling-Problematik)? Wie kann man, basierend auf Keywords, gezielt nach Topics suchen?	15:25	15:40	15	75	Lisa						
1	4	Praxis	R	**Ergebnisanalyse Topic Modeling** Häufigste Wörter und Topic-Anteile aus Modelloutput ermitteln (stark unterstützt mit vorgegebenem Code-Gerüst)	15:40	16:00	20	75	Lisa						
2	5	Theorie	R	**Word Embeddings** Was ist das? Wie setzen wir das in Kombination mit Topic Modeling um?	09:00	09:20	20	90	Asmik	X					
2	5	Theorie	R	**Sentiment Classification** Wie kommen wir von unseren Features zu einem Label? Welche Aspekte müssen beachtet werden? (Train-Test-Split, passende Metrik, ...) Welche Learner bieten sich an?	09:20	09:45	25	90	Lisa	X					
2	5	Theorie	R	**mlr3 universe** Wie können wir das in R mit mlr3 umsetzen?	09:45	10:10	25	90	Lisa						
2	5	Praxis	R	**mlr3 universe** Basic steps selber durchführen: Task definieren, Learner definieren und trainieren, prädiktieren	10:10	10:30	20	90	Lisa						
2	6	Theorie	R	**Evaluation** Welche Evaluationsmetriken bieten sich an?	10:45	10:55	10	75	Lisa						
2	6	Praxis	R	**Evaluation** Learner evaluieren (confusion matrix etc.) Rumprobieren, ob bessere Ergebnisse erzielbar sind mit anderen Hyperparametern	10:55	11:15	20	75	Lisa						
2	6	Theorie	R	**Overfitting & Bias** Worin besteht die Problematik von wiederholten Analysen auf denselben Daten? Wie könnten wir das hier lösen (nur ganz kurzer Teaser mlr3-Pipeline)?	11:15	11:30	15	75	Lisa						

2	6	Praxis	R	**Visualisierung** Wie können wir unsere Ergebnisse darstellen?	11:30	11:40	10	75	Asmik	X					
2	6	Praxis	R	**Visualisierung** Paar Plots erstellen (Balkendiagramme, Word Clouds)	11:40	12:00	20	75	Asmik	X					
2	7	Theorie	BERT	- **Einführung in Deep Learning**: Basics, Optimierung, RNN - **Transfer Learning** - **BERT - fundamental tasks**: Masked Language Modeling, Next Sentence Prediction - **BERT - basic application**: Input Preprocessing , Pretraining, Fine-Tuning - **BERT - advanced application**: Postprocessing, Aspect Extraction, Aspect Based Sentiment Classification	13:00	14:30	90	90	Asmik	X					
2	8	Praxis	BERT	- Einrichtung Google Collab, GPU - Durchführung von BERT SA (Fine Tuning) auf Basis von germeval und unseren Tweeter-Daten - Hauptaufgaben: Fokus auf eigenständiger Implementierung von Input Preprocessing & Evaluierung von Ergebnissen (der Code für den eigentlichen Trainingsprozess wird zur Verfügung gestellt) - Besprechung der Lösung	14:45	15:50	65	75	Asmik	X					
2	8	Organisatorisches	Allgemein	**Wrapup & Feedback**	15:50	16:00	10	75	Beide						