**14.11.18**

- arff-file guter Ansatz

- so viele Kommentare wie möglich in Datensatz (außer nicht-englischsprachig)

- 50% einer Klasse sollten kein Problem sein, erst ab 90% wird es kritisch, dass nur noch eine Klasse vorhergesagt wird

- interessante WekaClassifier: meta>MultiClass, trees>RandomTree und Ordinal Classifier (->Mark A. Hall, Eibe Frank PDF)

- Klassifizierungsprobleme: Jeweils ansehen: Klassifizierung des Zuges (+3 = sehr gut, -3 = sehr schlecht), Klassifizierung der Spielsituation (+3 = sehr gut für Weiß, -3 = sehr gut für Schwarz)

- für den Anfang sind zwei Klassen ausreichend (Ergebnisse unbedingt in Arbeit aufnehmen), später auf jeden Fall auf Ordinalsystem mit mehr Klassen wechseln

- Fokus auf Kontrolle der Lerner und Fehleranalyse (most\_informative\_features)

- Stopwords ja oder nein?

- one-against-all vs. pairwise classification

- für nächstes Mal: Tabellen vorbereiten mit Ergebnissen

**28.11.18**

- interessante WekaClassifier: meta>MultiClass mit Optionen -M=0 (one-against-all) und =3 (pairwise) sowie -P-Optionen, trees>RandomTree und Ordinal Classifier (->Mark A. Hall, Eibe Frank PDF)

- Word2Vec ansehen

- TF-IDF-Vektormodell nutzen, um Wortgewicht mit einzubeziehen (Abwertung von z.B. white und black), COUNT vs WEIGHT

- neue Daten von CD ziehen

- Cross-Validation mit 10 Folds ausreichend

- Tabellen für allgemeine Erkenntnisse mit ~10 Klassifizierern, für weitere Untersuchungen sind 2-3 ausreichend

- Bigrams und Trigrams-Mehrwert austesten

- Parameter: Stopwords ja/nein, Stemming ja/nein, Hapax ja/nein, Lowercase ja/nein, Tokenizer-Varianten, Word-Threshold (unterschiedliche Rankings)

- Neben Accuracy kann auch ROC-Area betrachtet werden