Kurs-Name: Data Analytics & Big Data
WiSe 2024/25 – Prof. Dr. Frank Krickel
Aufgabenblatt zu 05 TEXT MINING

Aufgabe 1: IRIS DATA SET: Lineare Regression

- Ladet die Datei *iris data set.xlsx* aus MS Teams herunter.
- Erstellt einen neuen Knime-Workflow
- Untersucht, welcher der metrischen Werte lässt sich am besten aus den jeweils anderen drei voraussagen? Wie lauten die Koeefizienten?
- Verwendet dazu den LinearRegressionLearner-Node

Aufgabe 2: IRIS DATA SET: Klassifikationsverfahren

- Erstellt drei neue Knime-Workflows und untersucht die IRIS-Daten mit folgenden drei Verfahren:
  - K-Nearest-Neighbours
  - Decision Trees
  - Logistic Regression
- Vergleicht die Ergebnisse mittels des Scorer (JavaScript)-Nodes.

Aufgabe 3: IRIS DATA SET: Clustering

- Erstellt einen KNIME-Workflow, der (nur!) anhand der numerischen Daten drei Cluster bildet.
- Prüft ob, der Algorithmus die bereits vorhandene Klassifizierung abbildet.

Aufgabe 4: Naive Bayes-Beispiel (Wiederholung aus der Vorlesung)

- Lade die Datei *05_DEMO_NAIVEBAYES.xlsx*!
- Analysiere die Trainingsdaten mit dem Naive Bayes-Learner!
- Sage die Klassifizierung (kauft/kauft nicht) für die Fragedaten vorher!

Aufgabe 5: IRIS DATA SET: Naive Bayes

- Erstellt einen neuen Knime-Workflow und untersucht die IRIS-Daten mit dem Naive Bayes Algorithmus
- Vergleicht die Ergebnisse mittels des Scorer (JavaScript)-Nodes mit den Ergebnissen aus Aufgabe 2.

Aufgabe 5: Text Mining mit Naive Bayes: Spam-Daten (Wiederholung aus der Vorlesung)

- Lade die Datei *05_SPAM_DATA.xlsx*!
- Analysiere die Trainingsdaten mit dem Naive Bayes-Learner!
- Sage die Klassifizierung (Spam/kein Spam) für die Fragedaten vorher!

Aufgabe 6: Text Mining mit Naive Bayes: Sentiment-Analyse

- Lade die Datei *05_SENTIMENT_DATA.xlsx*!
- Analysiere die Trainingsdaten mit dem Naive Bayes-Learner!
- Sage die Klassifizierung (Spam/kein Spam) für die Fragedaten vorher!

Aufgabe 7: Bag of Words: Sentiment Analyse (Wiederholung aus der Vorlesung)

- Ladet die Datei *05_SENTIMENT_DATA.xlsx*!
- Erstellt einen „Bag of Words“ (BoW), wie in der Vorlesung dargestellt.
- Vergleicht den BoW mit Eurer händischen Analyse

Aufgabe 8: Bag of Words: AmazonCellPhoneComments

- Ladet die Datei *AmazonCellPhoneComments.csv*!
- Erstellt einen „Bag of Words“ (BoW), wie in der Vorlesung dargestellt
- Beachtet: Die Zielkategorie „Sentiment“ muss ein String sein
- Versucht eine Analyse mit einem Klassifikationsalgorithmus Eurer Wahl
- Das Resultat wird vermutlich immer unbefriedigend sein – warum?

Aufgabe 9: Bag of Words: AmazonFoodReviews

- Ladet die Datei *FoodReviews.csv*!
- Erstellt einen „Bag of Words“ (BoW), wie in der Vorlesung dargestellt
- Vorsicht: Diese Datei ist sehr groß! Schaltet einen Filter vor, um die Durchlaufzeiten zu reduzieren.