Excersice Sheet 11

Block 1: Develop a Naive Bayesian classifier that can detect spam SMS. The learning record contains the text and the label for each SMS: Spam SMS are marked as `spam` and normal SMS as `ham`. The record is to be converted into a Document-Term Matrix[1], which serves as input for the Naive Bayes classifier.

1. Determine the number of `spam` and `ham` messages in the record. Perform a word tokenization[2]. For example, you can use `tidytext::unnest_tokens()`. Convert all uppercase letters to lowercase letters and remove punctuation marks like ".", "," and ";". Remove stop words like "and", "of" and "or" from the SMS text. You can use stop dictionaries like `tidytext::stop_words` or `tm::stopwords()`.
2. Identify the 10 most common words for Spam and Ham SMS. Remove words that occur less than 2 times in total in all SMS. Create a Document-Term Matrix. The rows of the matrix correspond to the SMS and the columns correspond to all words that occur in all SMS. Each value in the matrix indicates whether a particular word occurs in a particular SMS (`TRUE`/`FALSE`).
3. Divide the data set into a training and a test quantity in the ratio 70%:30%. Make sure that the distribution of `spam` and `ham` is approximately the same in both quantities. Use `set.seed()` for reproducibility. Learn a Naive Bayes classifier on the training set, e.g. with `e1071:naiveBayes()`. Use the learned model to predict spam in the test set. Create a Confusion Matrix and calculate Accuracy, Sensitivity and Specificity. Calculate the improvement or deterioration in accuracy, sensitivity and specificity of the model compared to a simpler classifier that would always predict the majority class (`ham`) for each SMS.

Block 2: Since 1946, all member states of the United Nations have come together at the United Nations General Assembly to discuss and vote on resolutions, among other things. Currently 193 states belong to the United Nations. Each of these member states has exactly one vote in the General Assembly's resolution votes on issues such as disarmament, international security, humanitarian aid and human rights.
The record for this task contains the complete voting process at the General Assembly of each country. Is it possible to predict whether Germany will vote "yes" or "no" in a resolution vote?

4. Display the number of resolutions voted on each year in a line chart. In which year were there the most votes and how many were there? Calculate between Germany and the USA for each year the proportion of equal votes (variable `vote`) for resolutions, hereinafter referred to as `agreement`. For the year 2006, the agreement between the two states was only about 25% of a total of 87 votes. (*Note: until 1989 "Federal Republic of Germany"; from 1989 "Germany"*)
5. Create a linear regression model that predicts the agreement between the two states based on the year (`agreement ~ year`). Interpret the trend and the p-value of the regression coefficient for `year`. Check the statement of the model graphically. Create a distance matrix between all pairs of states based on their voting history. Only consider states that have cast a vote in at least 70% of all votes. Determine the 5 states that are most similar or most dissimilar to Germany with regard to the voting history at UN General Assemblies.
6. Divide the data set into a training and test set at a ratio of 75%:25%. Create a kNN classifier with $k = 3$ (`caret::knn3Train()`) to predict the vote of Germany in a vote based on the vo-

tes of the countries `'Italy', 'Netherlands', 'United States of America', 'Israel', 'Cuba', 'India'`. Remove votes in which Germany abstained (`vote=2` ("Abstain")) to get a binary target variable for `vote=1` ("Yes") and `vote=0` ("No"). Create the Confusion Matrix and calculate the Accuracy for the model. On the same data, create a logistic regression model (`glm(..., family = "binomial")`) and compare the accuracy with that of the kNN classifier.

Dataset for Block 1: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/spam.csv
(adaptiert von http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/)

Dataset for Block 2: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/UNVotes.rds
(adapted by https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/12379)
- Data Dictionary / Codebook: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/UNVotes_Codebook.pdf

[1] https://en.wikipedia.org/wiki/Document-term_matrix
[2] https://de.wikipedia.org/wiki/Tokenisierung, http://tidytextmining.com/tidytext.html