Part 1: Reading and Parsing

- Read the dataset IMDb-sample.csv
(Tip: Drag and drop the dataset from the explorer to the Workflow Editor)
- Use the Strings to Document node to create documents
- Use the Column Filter node to delete all columns except the document column

Optional: Use the Document Viewer node to take a look at the documents

Part 2: Enrichment

- Use the POS Tagger node to assign part of speech tags

Optional: Use the Document Viewer node to visualize the tags

Part 3: Preprocessing

- Use the Punctuation Erasure node to remove punctuation
- Use the Number Filter node to remove numbers
- Use the N Chars Filter node to remove words with less than 3 characters
- Use the Stop Word Filter node to delete words with very little meaning, such as "and", "the", "a"...
- Use the Case Converter node to lower case all words
- Use the Snowballer Stemmer node to reduce words to the stem
- Use the Tag Filter node to delete all words besides adjective, adverbs and nouns

Optional: Use the Document Viewer node to take a look at the preprocessed document

Part 4: Transformation and Frequencies

- Use the Bag Of Words Creator to create a bag of words
- Use the TF node to calculate the relative term frequencies
- Use the Document Vector node to get a vector representation of each document

Optional: Calcualute the TF-IDF frequency

Part 5: Classification

- Use the Category To Class node to extract the class labels from the documents
- Use the Partitioning node to create a training and test set
- Use the Decision Tree Learner node to train a model on the training set
- Use the Decision Tree Predictor node to apply the trained decision tree model on the test set
- Use the Scorer node to evaluate the model

Optional: Use other algorithms to train a model. Use the ROC Curve to evaluate the model.