Shirin Mohebbi

Text mining Assignment

-Platform:

KNIME: KNIME is an open-source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining "Building Blocks of Analytics" concept . A graphical user interface allows assembly of nodes blending different data sources, including preprocessing, modeling, data analysis and visualization without, or with only minimal, programming. in this project i used KNIME to train a decision tree for a text mining classification problem. here is the list of nodes i used for the classification and a description of what each of them does.

CSV Reader: Reads our input CSV file and save it in a table. i used this table for further operations.

Column filter: This node allows columns to be filtered from the input table while only the remaining columns are passed to the output table.

Strings To Document: Converts the specified strings to documents. For each row a document will be created and attached to that row. The strings of the specified columns will be used as title, authors, and full text. Furthermore the defined category, source, type, and date will be set.

-Preprocessing Nodes:

Category To Number: This node takes columns with nominal data and maps every category to an integer.

Punctuation Erasure: Removes all punctuation characters of terms contained in the input documents.

Number Filter: Filters all terms contained in the input documents that consist of digits, including decimal separators "," or "." and possible leading "+" or "-".

Case Converter: This node converts alphanumeric characters to lowercase or UPPERCASE.

Stop Word Filter: Filters all terms of the input documents, which are contained in the specified stop word list and/or in the second input table. The node provides built-in stop word lists for various languages. Additionally, stop words can be passed to the second input port.

Snowball Stemmer: Stems terms contained in the input documents with the Snowball stemming library.

-Feature Extraction Nodes:

Bag Of Words Creator: This node creates a bag of words (BoW) of a set of documents. A BoW consists of at least one column containing the terms occurring in the corresponding document. All term related columns like the document column can be selected in the node dialog and will be copied to the output table.

TF: Computes the relative term frequency (tf) of each term according to each document and adds a column containing the tf value. The value is computed by dividing the absolute frequency of a term according to a document by the number of all terms of that document.

IDF: The normalized idf is defined by: idf(t) = log(f(D) / f(d,t)). The probabilistic idf is defined by: idf(t) = log((f(D) - f(d,t)) / f(d,t)), where f(D) is the number of all documents and f(d,t) is the number of documents containing term t.

Math Formula: This node evaluates a mathematical expression based on the values in a row. The computed results can be either appended as new column or be used to replace an input column. Available variables are the values in the corresponding row of the table (left list in the dialog). Commonly used functions are shown in the list "Mathematical Functions".

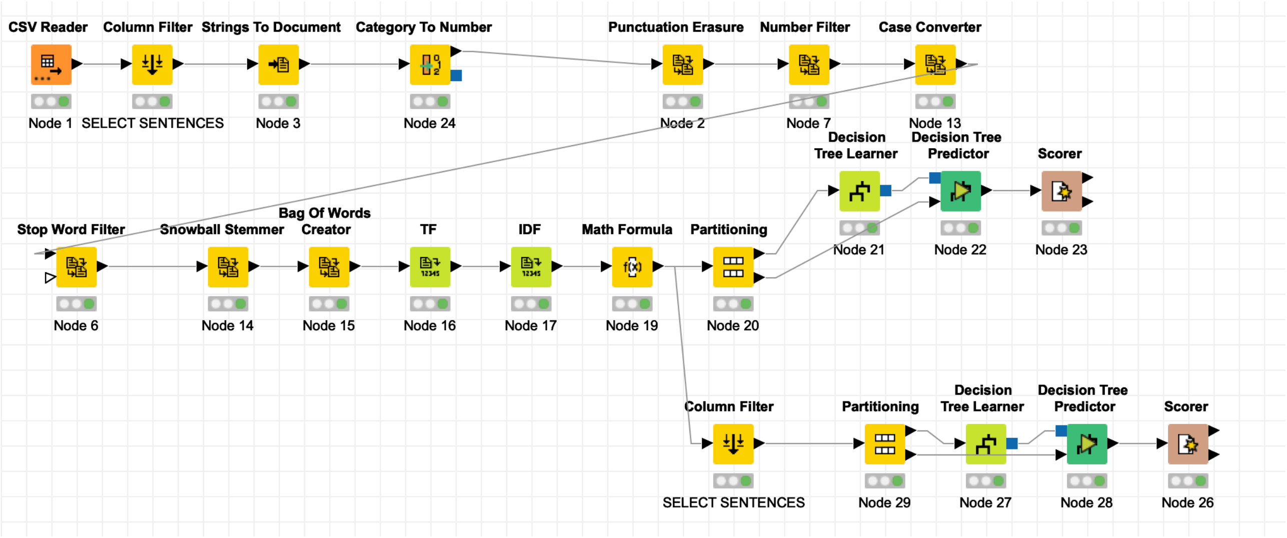
-Training Nodes:

Partitioning: The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports.

Decision Tree Learner: This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point.

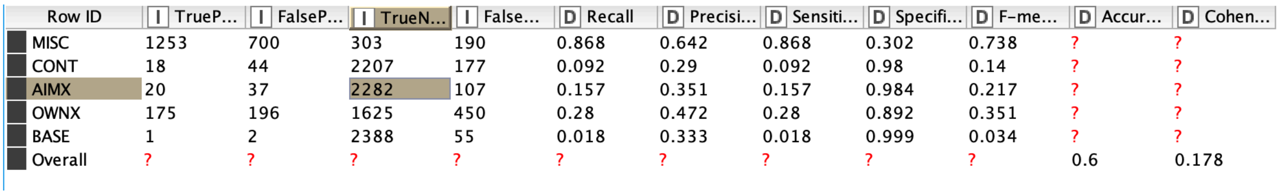
Decision Tree Predictor: This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns.

Scorer: Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match.



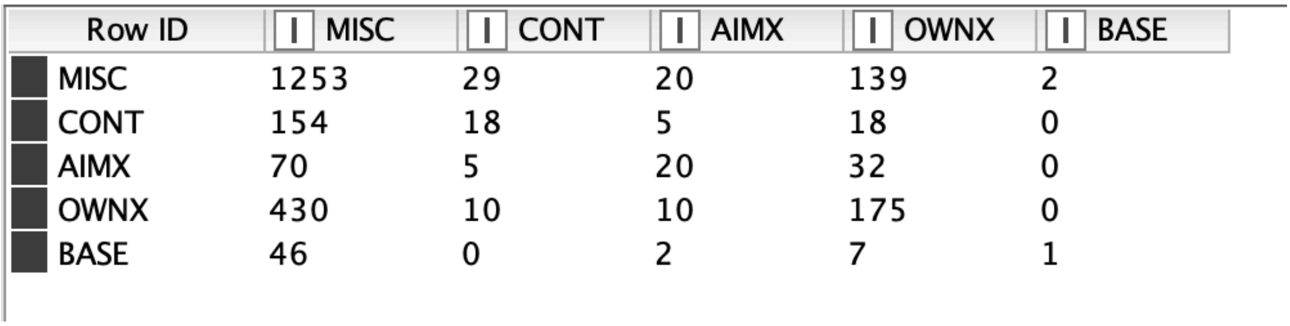
-Results:

Accuracy table:



Row ID	TrueP…	FalseP…	TrueN…	False…	Recall	Precisi…	Sensiti…	Specifi…	F-me…	Accur…	Cohen…
MISC	1253	700	303	190	0.868	0.642	0.868	0.302	0.738	?	?
CONT	18	44	2207	177	0.092	0.29	0.092	0.98	0.14	?	?
AIMX	20	37	2282	107	0.157	0.351	0.157	0.984	0.217	?	?
OWNX	175	196	1625	450	0.28	0.472	0.28	0.892	0.351	?	?
BASE	1	2	2388	55	0.018	0.333	0.018	0.999	0.034	?	?
Overall	?	?	?	?	?	?	?	?	?	0.6	0.178

Confusion matrix:



Row ID	MISC	CONT	AIMX	OWNX	BASE
MISC	1253	29	20	139	2
CONT	154	18	5	18	0
AIMX	70	5	20	32	0
OWNX	430	10	10	175	0
BASE	46	0	2	7	1