# **Documentation Metanomnomnom Algorithm**

The Metanomnomnom-Algorithm tries to rely as much as possible on the supplied Metanome helper utilities and follows the naïve approach to look for column combinations by utilizing Metanomes own PLIBuilder to build the complete position list index and check all possible column combinations for uniqueness.

In order to cope with the apparent complexity of larger datasets the resulting data is pruned by removing all supersets of a unique column combination from the remaining possible column combinations. This however is quite costly as the superset creation is continued until the size of the supersets exceeds the number of column, while continuously comparing the created supersets with all of the remaining column combinations in order to find and remove duplicates. By pruning the column combination candidates this way, the data complexity was supposed to be reduced; however the decreasing amount of data comes at the cost of an increased computational complexity and unnecessary temporary data creation.

Since this algorithm does not convince with its processing capabilities and does not showcase parallelism or the efficient usage of appropriate data structures, publishing the algorithm as Metanome algorithm might not be advisable, except as an example for the possibilities of the enclosed tools and data structures.

Using this approach on the WDC\_planets dataset resulted in 7 unique column combinations, in 204.24ms. The ncvoter-1k dataset already caused a serious delay until the results appeared, preventing us from running it with the whole ncvoter dataset. All tests were obtained using an Intel i5/4300U running at 1.9GHz with 8 GB of main memory and a Hynix solid state drive.

Since the Metanomnomnom algorithm already takes a considerable amount of time to analyze the small fraction of the ncvoter dataset the runtime limitation of our algorithm is apparent. To support larger datasets the pruning would have to be implemented in a more efficient way. Especially the creation of the supersets needs to be optimized to achieve shorter runtimes. Due to the already long runtimes for the smaller datasets, limitations through high memory consumptions have not been observed so far.

NULL values are treated as different values, since the implementation relied on the Metanome tools implementation the NULL value behavior has neither been changed nor benchmarked.