This thesis describes the FAFEFI program that focuses on n-gram and skip-gram extraction from large data sets. The thesis presents two different approaches to passing input data to the program. It also describes the design of data structures for n-gram and skip-gram representation within computer memory, the algorithm of n-gram and skip-gram extraction, memory-friendly options of saving extracted data and their final composition into output feature vectors. It also offers a variety of extra functions such as line filter and line modifier and a great deal of configurable parameters ranging from in-file separators to formatting the names of output files. Moreover, the program provides a differentiation in its activity by enabling saving data just after extraction from the train set and brings tools for cluster parallelization.