Malware analysis is a growing research area but with still many open problems [1]. For example T signatures for anti-virus toolkits are created manually using some malware-analysis techniques and tools, that can analyze programs either by executing them (dynamic analysis), or by inspecting them (static analysis). Static analysis can extract information from the binary representation of the program. Data mining techniques for detecting malware were first introduced by [2] on three different static features: Portable Executable (PE), strings and byte sequences. Interpretable text is a high-level specification of malicious behavior, for example: window.open('readme.eml') always occur in worms of type Nimda [3]. Text Mining classification can be useful, and be however prohibitive because of the tokenization process than may either produce a very high dimensionality of features or lose relevant information by the use of a standard text IR tokenization. Nevertheless, Big Data technologies and massive clustering techniques are now possible so that the release of a TREC style collection, that is still missing, will help the IR and the cyber security community to deeply explore at what extent Information Retrieval and Text Mining classification can be effective and useful to malware detection. Our text collection contains about 650K documents with the text extracted from malware and will be extended with a similar size of malware-free collection.