

SYSTEMS AND METHODS FOR SUBSET SELECTION AND OPTIMIZATION FOR BALANCED SAMPLED DATASET GENERATION

ABSTRACT

[0001] Methods and systems for data management of documents in one or more data repositories in a computer network or cloud infrastructure are provided. The method includes sampling the documents in the one or more data repositories and formulating representative subsets of the sampled documents. The method further includes generating sampled data sets of the sampled documents and balancing the sampled data sets for further processing of the sampled documents. The formulation of the representative subsets is performed for identification of some of the representative subsets for initial processing.

FIG. 1