

ABSTRACT

An automated computer implemented method of deduplicating point of interest of computer implemented database including: standardizing data of a pair of records, wherein equivalent words with different spelling may be normalized to a single spelling; classifying the pair of records into a category of pre-defined address categories; carrying out a category specific determination of non-duplicate and failing to determine that the pair of records may be a category specific non-duplicate; tokenizing the data of the pair of records into a pair of tokenized records; determining that core words of the pair of records are linked by computing scores of cross information and distance; and determining whether the scores pass a pre-determined threshold. An aspect of the disclosure relates to an electronic data processing system, a data processing apparatus, a non-transitory computer-readable medium, and a computer executable code for deduplicating point of interest of computer implemented database.

FIG. 1