This repository contains descriptions of Linked Data datasets using VoID vocabulary and prepared data to perform the empirical experiments for evaluating dataset ranking models. The directory “dataset” contains the dataset descriptions, serialized as an nQuad RDF file, and the directory “prepared\_test\_data” contains test data for experiments, serialized as \*.csv files.

The datasets descriptions include Linksets, classes, properties and topic categories. It mashes up data from DataHub, dataset dumps, VoID files and DBpedia. The DBpedia Spotlight allowed the recognition of named entities in literal values and the linking of each RDF Entity with a list of topic categories. Named entities are directly linked to categories through the predicate dcterms:subject and each topic category is subsumed by others through the predicate skos:broader. A category c is a topic category of an RDF entity iff there exists a property path {e dcterms:subject/skos:broader\* c.} from a named entity e of the dataset to c in DBpedia.

Files in the "prepared\_test\_data" directory are organized in two subdirectories. The first directory called "bayesian-social\_network" contains files for evaluating ranking models using algorithms based on Social Networks Analysis and Bayesian Classifiers. The second directory named "cos-j48-jrip" contains files for evaluating ranking models using algorithms based on JRip and J48 classifiers and Cosine Similarity. Each these two directories, in turn, is also organized in subdirectories that splits test data according to the types of features used in the experiments. The directory "5L" contains test data using dataset representations with five linksets, the directory "12C" contains test data using dataset representations with twelve topic categories, "5L12C" contains test data using dataset representations with five linksets and twelve topic categories, and so on.

The \*.csv files in "cos-j48-jrip" directory are formatted as follows. The first column indicates the datasets and the remaining columns indicate the TF-IDF value of each feature. The set of test data is split into three subsets which contain three files: Test[i].csv, Traning[i].csv and Relevants[i].csv. The test[i].csv contains the target datasets to which the datasets in Training Finally, relevance files indicate the relevance of a test dataset in relation to a training dataset, where it has a value of 1 if relevant and otherwise has a value of 0.

For the data in the "nominal\_methods" directory, the csv files are organized as follows: the first column of the test and training files corresponds to the dataset used in its due partition (test or training) and the other columns are the features used to perform the algorithms. Finally, relevance files indicate the relevance of a test dataset in relation to a training dataset, where it has a value of 1 if relevant and otherwise has a value of 0.