Limes, Silk, and CSV dedupe require a configuration file, describing the metrics used to compare a record pair as well as the acceptance thresholds which are used for deciding whether two records are the same or not. In SERF, the user is required to implement a decision maker function by himself. In all cases, the frameworks take out a pair of records and decides whether they are the same or not. the outcome of the entity resolution task highly depends on specified metric function and thresholds. In our experiments we tried to choose the best fitted metric function based on the tools each framework offers, and tune the thresholds to get the best result possible.

In Limes, user first specifies all the attributes he wants to use for record comparison. Limes requires all the specified attributes to be available in both records, so it only considers the records which have all the attributes. For product domain we had to ignore the manufacturer name of the product, since more than 90% of the records in Google product dataset do not have the manufacturer name. The configuration in Silk is almost the same as Limes, however Silk lets user to specify which attributes are not required to be present and can be ignored while comparing two records which may not have them. For SERF, we implemented a separate record comparator class for each domain and used the similarity measures available in their source code to compare attributes. Table XXX illustrates a summary of the acceptance metrics for different domain and frameworks:

Table product domain

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Name | Price | Description | Manufacturer | Acceptance Metric |
| Limes | trigrams | difference | cosine | trigrams | N>0.6 AND P<20 AND D>0.5 |
| Silk | trigrams | 20N+10M+5P+D |
| Serf | Normalized difference |  |  |
| Dedupe |  |  |  |  |  |

Table publication domain

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Papers | | Author | | Acceptance Metric | |
|  | title | authors | name | publications | Papers | Authors |
| Limes | trigrams | Soft jaccard | cosine | Soft jaccard | T>0.7 AND A>0.6 | N>0.8 AND P>0.6 |
| Silk | trigrams | 6T+A | 6N+P |
| Serf |  |  |  |
| Dedupe |  |  |  |  |  |  |

Table human trafficking domain

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | images | Phone number | Eye color | ethnicity | email | Provider name |
| Limes | soft jaccard on image SHA1 | Exact match | Levenshtein | Jaro | Exact match | Jaro |
| Silk |
| Serf | Soft jaccard on image similarity |
| Dedupe |  |  |  |  |  |  |