Extension of Frost/Snowman for tag-based linkage result analysis

Background

Record linkage is an essential part of many data integration processes and aims to detect records in multiple databases that correspond to the same real-world entity, e.g., a person that was treated in multiple hospitals. However due to privacy concerns the identifying data is not allowed to be shared in plaintext for the linkage. Therefore privacy-preserving techniques have been developed. Some methods for Privacy-preserving Record Linkage (PPRL) encode the plaintext before sharing it with a third party, that does the linkage on the encoded data only. Typically there are some limitations regarding the linkage quality due to the predefined and therefore static encoding. In order to analyse the advantages and disavantages of different techniques it is beneficial to analyse the properties of correct and incorrectly matched records.

Recently a framework called Frost with its reference implementation Snowman has been proposed to compare matching solutions on different datasets. The tool already supports some exploration techniques such as showing record pairs filtered by certain critieria, e.g. the similarity range. However it does not allow to annotate links with tags, so that these can be used for the exploration.

There are three types of tags, hoewever the first two are probably identical in the implementation:

- Plaintext-based, e.g., if the firstname is rare/common
- Encoding-based, e.g., the fillrate of a Bloom Filter encoding
- link-based, e.g. the difference is minor (typo) or major (single vs double lastname)

Tags can be simple Strings, e.g. "PERSON_MOVED", or key-value-pairs, e.g. "FIRSTNAME_FREQUENCY_RANK -> 2.

The goal of this thesis is the extension of Snowman for tag-based linkage result analysis. The tags/annotations can be pre-generated, e.g., from the generator of synthetic testdata itself, or provided by an external link analyzer (web)service or computed internally by Snowman.

The GUI should provide views to use the tags for linkage result analysis, e.g.

- overview of linkage quality measures (recall, precision, F1-score) for subsets filtered based on these tags
- explanation view for wrong links, e.g. what are the most common tags of false positives/negatives
- comparison of tags of corresponding links in results of different matching solutions

Tasks

- Make yourself familiar with Snow/Frostman
- Extend the data model to support link annotations/tags
- Build importer for pre-computed tags (from a CSV file)
- Design and implement GUI and backend for analyzing linkage results based on these tags
- Implement a link annotator for the Snowman Backend

Literature

- Short (PP)RL Background
- Snowman paper
- Git repository