Automatic detection of name variations

Toon Calders (ULB)

This project concerns a data cleaning and integration task with real data. For this project a large data collection consisting of historical birth, death, and marriage certificates of the province of North-Brabant in the Netherlands is available. This collection contains certificates for about 3 million people, from 1580 until 1955. This collection of paper documents has been indexed by volunteers. For many of the certificates (unfortunately the index is not complete yet), the names of the people involved in it, and their role have been recorded in a database. Consider for instance the following example of an index entry for a death certificate:

Death certificate	
Deceased	Johanna Louise Fredrika Frans
Relation of the deceased	Gerard Cornelius Reincke de Sitter
Father of the deceased	Carl Ludwig Frans
Mother of the deceased	Alida Philippina Zehender
Type of deed	death certificate
Number of deed	5
Place	Beers
Date of decease	26-02-1825
Period	1825
Contains	Overlijdensregister 1825
Number of inventory	50
Record number	456

There are, however, several problems with the data recorded by the volunteers:

1. Volunteers made mistakes when recording the names
2. Natural name variations occur; for instance, during the Napoleonic era, Willem preferred to be called Guillaume. After the French left the Netherlands, Willem became Willem again. Other, less spectacular variations: Fredrika versus Frederika.
3. Another source of variation is the granularity at which locations are reported. Sometimes locations have been reported at suburb or even neighborhood level, whereas in other records only the city is reported.
4. Also the original data contained errors. For instance, the order of names may have been swapped.

The goal of this graduation project is to automatically detect name variations for location and person names, using statistical and data mining methods. Because of the large size of the database it is very likely that most name variations occur frequently. In a pilot study, it was shown that name variations could be detected by finding pairs of full names sharing most surnames, but not all. The differences often were name variations. Your task will be to extend this approach to also include locations, and exploit additional background knowledge such as: for most birth certificates there is a matching death certificate, no one has more than one birth and death certificate, etc. This project has a large research component, so your creative input will be required as well.