Cleaning and Visualizing a dirty set of restaurant data

Florian Loher
Technical University of Applied Science Regensburg
florian.loher@st.oth-regensburg.de

***Abstract*—This document shows a possible approach to cleaning and visualizing the dirty dataset provided at https://hpi.de/naumann/projects/repeatability/datasets/restaurants-dataset.html. It describes how the data is first audited, then cleaned in MongoDB, removing duplicates, using a common search engine to find correct restaurant names and standardizing road and city names. Lastly the data is visualized by generating a website that contains an OSM map and markers indicating the location of each restaurant.**

***Index Terms*—MongoDB, Data cleaning**

I. Introduction

Big data is a rapidly growing field of research that already gained overwhelming interest in the general public. The amount of data is increasing at an exponential rate an is likely to grow further at this rate. To be able to leverage the power of data, the need for ways to clean is as high as ever.

Data cleaning, also referred to as data scrubbing or data cleansing, is a research field concerned with improving the quality of faulty data. Typical aspects that are sought to be improved are the amount of duplicates, type errors or inconsistencies[1].

In this article I am going to outline a possible approach to detecting duplicates in a dataset of 864 restaurants[1]. I first audit the data. Then I generate a training set, standardize fields and choose *SoftTF-IDF* string matching measure with *Jaro-Distance* as sub measure as the algorithm for duplicate detection. After creating a gold standard for the training set I train the thresholds for restaurant name, phone number and address similarity that determine if two records will be declared duplicates. Lastly the detection algorithm is run on the test data and compared to the gold standard[2]. Using and comparing different sizes of training data I show that even with a training set 10% the size of the test data (86 records, 22 of which are part of duplicates) a recall of $\approx 86\%$ is with a precision of $\approx 95\%$ typically achieved.

[1]Restaurant data provided at https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/projekte/repeatability/Restaurants/restaurants.tsv

[2]Gold standard provided at https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/projekte/repeatability/Restaurants/restaurants_DPL.tsv

In section II I am going to introduce basic terminology concerning data cleaning and describe the string-matching algorithms I use. Section

II. Basics

A. Duplicate Detection Fundamentals

B. String matching

Duplicate detection relies on the matching of strings i.e. comparing fields of distinct records in a given dataset and calculating how similar they are, based on a defined measure.

Token-based measures are typically optimized to handle rotation errors. Frequent differences between fields that represent the same entity are for example swapping first and last name, or having a title prepended or put at the end. By splitting the strings into tokens and comparing the resulting sets of tokens such rotations cease to impact those measures. The token-based string matching algorithm *SoftTF-IDF* (soft term frequency, indirect document frequency)

C. Matching algorithms for names and addresses

III. Auditing and data preparation

Before trying to find duplicate data in any given dataset most data cleaning approaches start with a phase of auditing and preparing the data. This includes but is not limited to discovering which types of fields are present in the data and removing unnecessary characters from fields.

IV. Generation of training data

V. Choice of string-matching algorithm SoftTF-IDF

VI. Training with training data

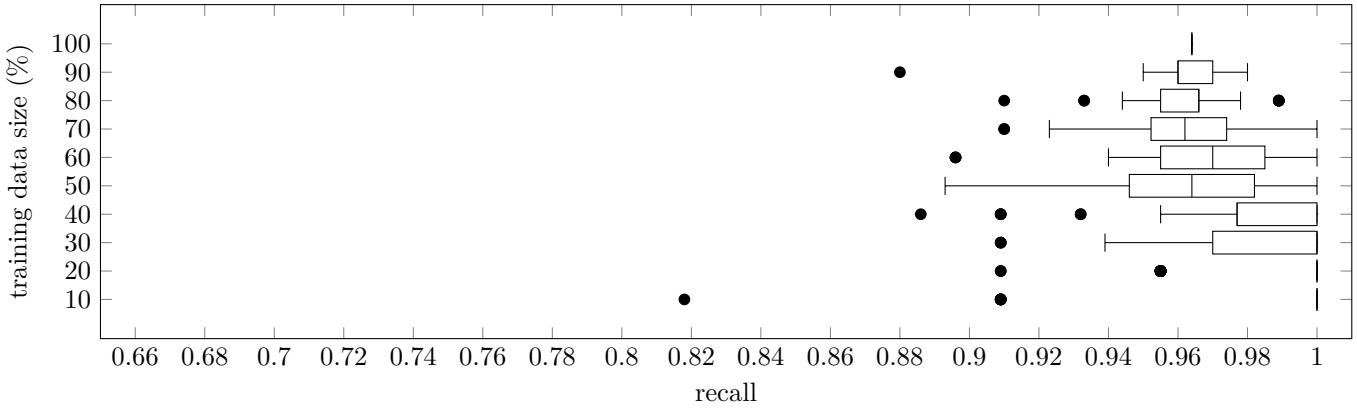
VII. Detection of duplicates in restaurant data

VIII. Comparison of different sizes of training data

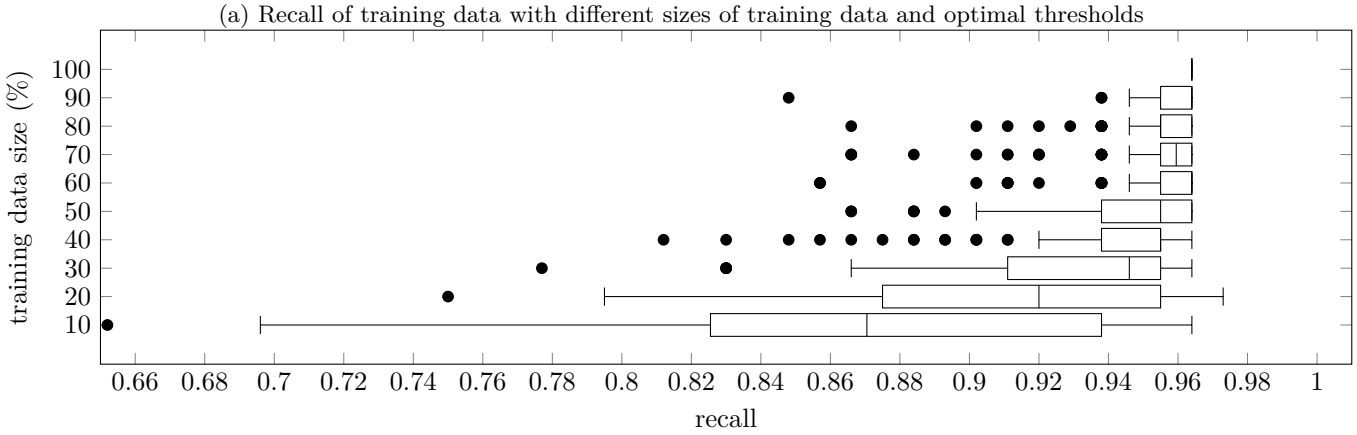
IX. Conclusions

References

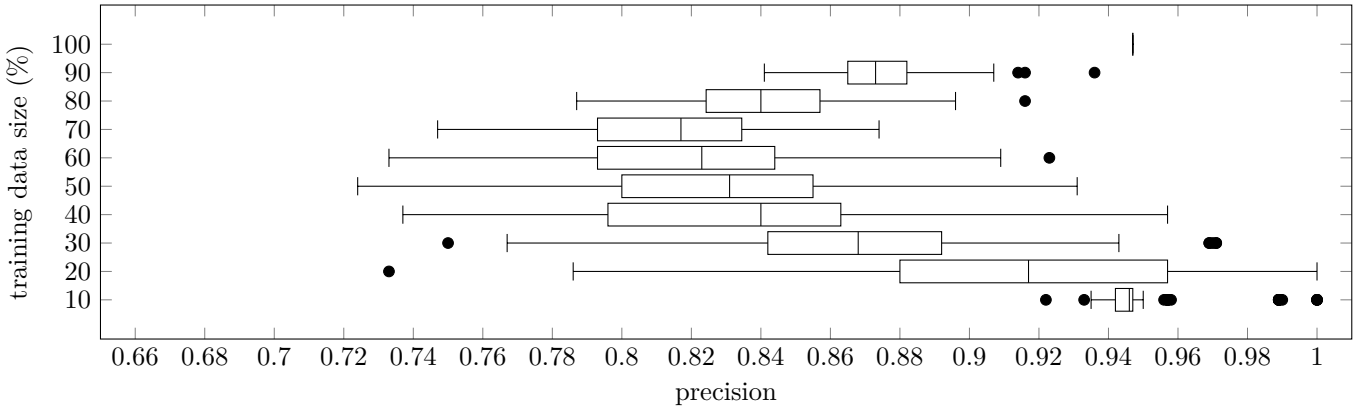
[1] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive name matching in information integration", *IEEE Intelligent Systems*, vol. 18, no. 5, pp. 16–23, 2003, ISSN: 1541-1672. DOI: 10.1109/MIS.2003.1234765.



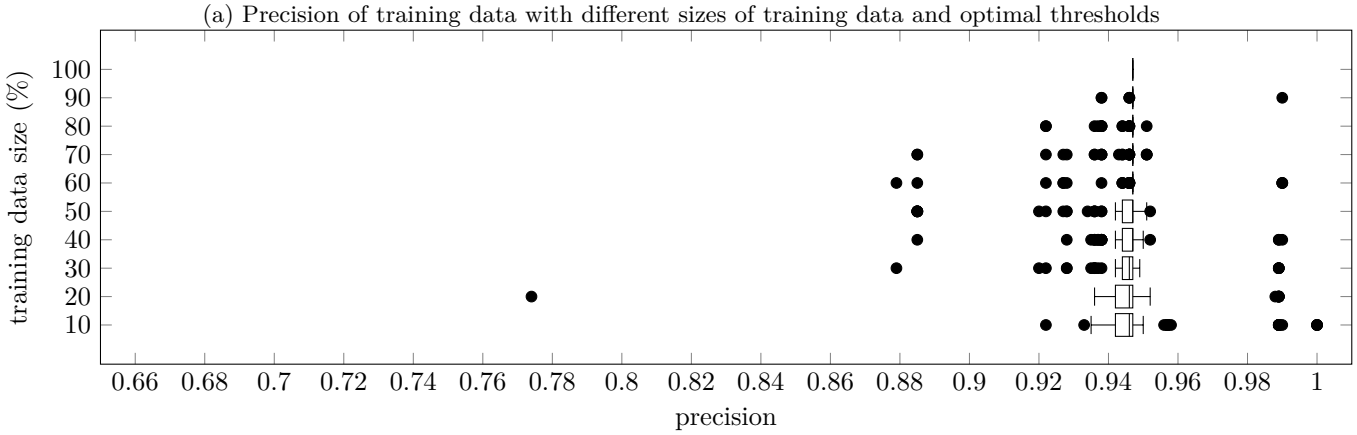
(a) Recall of training data with different sizes of training data and optimal thresholds



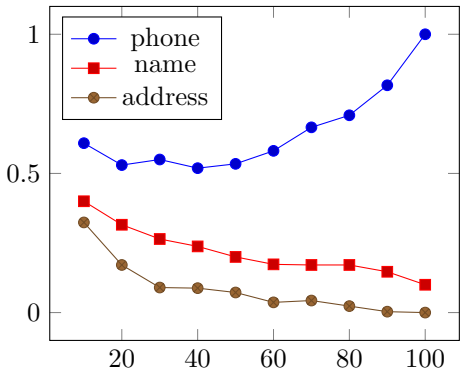
(b) Recall of test data with different sizes of training data and optimal thresholds



(a) Precision of training data with different sizes of training data and optimal thresholds



(b) Precision of test data with different sizes of training data and optimal thresholds


phone
name
address
1
0.5
0
20
40
60
80
100