

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

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 in the data[1]

II. DATA CLEANING FUNDAMENTALS

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.

REFERENCES

- [1] IEEE Data Eng. Bull., S. Sarawagi, ed., special issue on data cleaning, vol. 23, no. 4, Dec. 2000.
- [2] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions”, *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
- [3] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [5] K. Elissa, “Title of paper if known,” unpublished.
- [6] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.*, in press.
- [7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [8] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.