CS513 Data Cleaning Project - Phase 2

Ehab Aleid ealeid2@illinois.edu

Diana Arita dianama2@illinois.edu

Kevin Zhong Kzhong7@illinois.edu

7/31/2021

1. **Data Cleaning Performed**
2. Transformation on FMID
   1. Remove any duplicate values by blanking duplicate values and removing them from the dataset
3. Transformation on Marketname
   1. Use text face to cluster and merge similar values using key-collision method
4. Transformation on City, County, and State
   1. Remote whitespace and special Characters
   2. Use text fa​​cet to cluster and merge similar values using key-collision method
5. Transformation on website, youtube, twitter, and facebook
   1. Blank any invalid values/links
   2. Add add relevant http protocol and domain name to any value with missing incorrect hyperlink so that all links follow similar format
      1. Used refex to filter for invalid hyperlink values and added missing string as header using python
6. Transformation on SeasonXDate (X = 1-4)
   1. Split each season column using “ to “ as separator so we get two columns, one for from date and the second for to date
   2. Convert time format to ISO standard Y-M-DT format
7. Transformation on Update time
   1. Convert time format to ISO standard Y-M-DT format
8. **Document Data Q​​uality Changes**

a.

b.

1. **Create a Workflow Model**

a. The outer workflow will be created using YesWorkflow

b. The inner workflow for the Open Refine history will be created using Yes Workflow and the or2yw tool.

1. **Summary**

Modeling the project using YesWorkflow shows provenance.