Capstone Project Data Wrangling

Emily McCormick

1/13/2019

I was actually fortunate enough that the data for my Capstone project was relatively cleaned already. To begin, the demogrpahic information and was already provieed in a clean csv file. There is a unique identification number for each happy moment (hmid), as well as a unique identifer for each individual replying to the survey. From the source, the dataset schemeas are as follows:

Cleaned\_hm.csv cleaned\_hm.csv contains cleaned-up happy moments and some additional information in addition to original happy moments.

hmid (int): Happy moment ID wid (int): Worker ID reflection\_period (str): Reflection period used in the instructions provided to the worker (3m or 24h) original\_hm (str): Original happy moment cleaned\_hm (str): Cleaned happy moment modified (bool): If True, original\_hm is “cleaned up” to generate cleaned\_hm (True or False) predicted\_category (str): Happiness category label predicted by our classifier (7 categories. Please see the reference for details) ground\_truth\_category (str): Ground truth category label. The value is NaN if the ground truth label is missing for the happy moment num\_sentence (int): Number of sentences in the happy moment

demographic.csv demographic.csv contains demographic information of the workers who contributed to the happy moment collection.

wid (int): Worker ID age (float): Age country (str): Country of residence (follows the ISO 3166 Country Code) gender (str): {Male (m), Female (f), Other (o)} marital (str): Marital status {single, married, divorced, separated, or widowed} parenthood (str): Parenthood status {yes (y) or no (n)}

The cleaned\_hm.csv file was cleaned from an original file by spell-checking the data as well as removing empty/one-word statements.