Code Instructions

The code in this project is all done in python and only requires one file (the mapper/combiner) to be run with the included dataset “survey.csv”. It should be noted that this program was written with Python 3.5 so I can’t say for sure whether or not it will work with other versions of the Python interpreter.

Ensure that “mapper.py” and “survey.csv” are in the same folder when you run the program. “combineOutput.txt” and “mapOutput.txt” contain output from a previous run of the program and will simply be overwritten if they are in the same folder as “mapper.py” when it gets run (which doesn’t matter because the dataset is the same), if these files are not in the same location as “mapper.py” the program will just generate these files when it runs.

“datasetInfo.txt” just contains info about the dataset being used “Mental Health in Tech Survey” which is from Kaggle. The info in this file is directly copied from the “Description” section of <https://www.kaggle.com/osmi/mental-health-in-tech-survey>. “datasetInfo.txt” does not need to be in the same folder as the running mapper

“survey.csv” must be in the same folder as “mapper.py” when the mapper runs because it is the data source file being read. “survey.xlsx” is a modified version of the dataset that still contains column header information as well as some extra columns that contain summary information about the ranges of values entered by survey participants. An example of this is the extra column on gender classification information which contains a row for every unique value entered in the gender section of the survey. The derived gender classification information was used in my code to break survey users into three groups Male, Female, and Gender Queer (It should be noted that this term doesn’t perfectly describe the group because this group includes both those who identify as trans men and trans women who aren’t really genderqueer by most definitions. A better term would’ve been “Those who do not strictly identify as cisgender men or women” however that is very long. It should also be noted that I assumed those who identified as males, females, men, or women were cisgender if they did not mention otherwise, which is probably not entirely accurate). “survey.xlsx” was important to the design of the program and reading through the extra columns may be interesting, but it is not needed to run the program.

Once you have the right files in a folder to run the program you have to make sure Python is installed. I would initially attempt to run it with whatever version of Python 3 you have installed, but if there are issues I would install Python 3.5 specifically to run it.

Run “mapper.py” with python 3 and look in the resulting content in “combineOutput.txt” and “mapOutput.txt”. “combineOutput.txt” contains a summarized count of all of the key value pairs being summarized while “mapOutput.txt” contains the same data in an un-summarized state.

Project Report

The Problem

For this project I decided to do an analysis on “Mental Health in Tech Survey” from kaggle.com. This dataset contains some demographic information (Gender, Country, Age) about those taking the survey as well as the answers to some questions about their experience with mental health in the IT industry. I was particularly interested in the relationship between gender, mental health treatment and work interference caused by mental illness (I also looked at age but didn’t find much interesting). I though it would be particularly interesting if there was a significant difference in the number of people from one gender group compared to another who reported no treatment but some work interruption.

I thought a correlation here would be interesting because it might say something about how we think about mental illness differently for different demographics of people. For example, if one group suffers with significant work interruptions due to mental health issues but has very low treatment rates it might suggest that that it is less socially acceptable for members of that group to have mental health issues. Alternatively, it could suggest that treatment is simply less available to those people, possibly for reasons of cost for example (unfortunately the survey did not have salary information).

Method

To solve this problem, I implemented a system of deriving specific key value pairs from the dataset with a MapReduce program written in Python 3. The dataset was originally stored in a comma separated value format so I used the Python csv library to go through the file record by record. For each record the program writes a treatment/gender (treatment: ‘Yes’ or ‘No’, gender: male, female or genderqueer) key value pair to an external file (“mapOutput.txt”). Treatment/age key value pairs are also written to this file for each record in the data file. A final key value pair is recorded on the file for treatment/gender/workInterfere key value pair (workInterference being a survey question on how much mental heath issues interrupt work for the survey taker).

Gender information has been summarized by breaking all the different entries into three groups which are essentially cisgender males, cisgender females and everyone else. This was done by using Microsoft excel to determine all the unique entries in the gender column (can be seen in ‘survey.xlxs’) and then classifying misspellings in the code for mapper.py. It should be noted that all entries that didn’t specifically mention whether the person was cisgender or transgender were assumed to be cisgender (which probably isn’t correct, but transgender people are a fairly small demographic so this misclassification shouldn’t skew the results too much).

After all these key value pairs (all values are 1) are recorded in “mapOutput.txt” a combiner reads through “mapOutput.txt” line by line and adds all the values together for each key and the stores the totals in another file “combineOutput.txt”. The user looks at the results in “combineOutput.txt” to analyse the data. There is no reducer in this program because on a single node (like what I have) the reducer would essentially do the same thing as the combiner (the code would be nearly identical in fact). It should also be noted that this program outputs key value pairs in a file and not in standard output (I thought it would be more convenient in a file) and is therefore not compatible with Hadoop.

Analysis of Data in Output File

If you would like to see all of the output data you can check out “combineOutput.txt” I’m only including data here if I can derive information from it that is both statistically significant (significance level = 0.10, two tailed p values) and interesting. It should be noted that I have never taken a statistics course and am learning this as I go.

treatedMales 447, treatedFemales 168 , treatedGQ 22, untreatedMales 538, untreatedFemales 76, untreatedGQ 8

447 of the 985 (45.38%) Males in the survey received treatment while 168 of the 244 cis Females in the survey (68.85) received treatment suggesting that cis females in the tech sector are more likely to have received treatment for mental health issues than cis males.

22 of the 30 (73.33) genderqueer people in the survey received treatment which is statistically significant when compared to the treatment rates for cis males, suggesting that genderqueer people in the tech sector are more likely to have received treatment for mental health issues than cis males.

untreatedMalesOften 19, untreatedMalesSometimes 88, untreatedMalesRarely 45, untreatedMalesNever 162, untreatedFemalesOften 2, untreatedFemalesSometimes 17, untreatedFemalesRarely 5, untreatedFemalesNever 17

treatedMalesOften 85, treatedMalesSometimes 257, treatedMalesRarely 79, treatedMalesNever 23, treatedFemalesOften 32, treatedFemalesSometimes 91, treatedFemalesRarely 37,   
treatedFemalesNever 7

88 of the 314 (28.03%) of the untreated males that indicated how much of a work disruption was caused by their mental illness sated that they were interrupted “sometimes”, while 17 of the 41 (41.46%) untreated females that indicated how often their work was interrupted by their mental illness said “sometimes”.

185 of the 768 (24.41%) of the cis males that indicated how much of a work disruption was caused by their mental illness sated that they were interrupted “never”, while 24 of the 208 (11.54%) cis females that indicated how often their work was interrupted by their mental illness said “never”.

The results of this analysis (in regards to gender) may have been much more interesting if the dataset was larger as there were a whole bunch of results that popped up that seemed interesting but unfortunately had p values above my significance level (which I was already stretching at 0.10).

I think there is potential to derive interesting information about how age demographics play into this however the question about whether one has received treatment or not would have to change on the survey for age data to be useful (perhaps: have you received treatment within the last 5 years?). Otherwise the likelihood that someone has received treatment (at any point in their lives) rises fairly consistently with age. This made age information pretty useless for what I was trying to analyze.

In regards to problems, I was originally planning on analysing an entirely different dataset but it was using a weird codec, and I spend a ton of time trying codec after codec to get my program to read it properly but it would always fail at a different line. As soon as I switched to this dataset it ran with utf-8 without any problems.

Summary

While the information derived from this program is interesting, it would really be better applied to a larger dataset (1259 rows just isn’t enough) and this algorithm wouldn’t have issues with a dataset that was 100 times larger than that. With a dataset of this size (small) when you break down simple groups into more specific groups, the groups quickly become too small to make statistically significant conclusions with. That being said the program did confidently determine that that cis males in the tech sector are less likely to have received treatment for mental illness which might mean that they struggle less with mental illness or it might mean that they are less likely to seek treatment for the mental illness they have (though it should be noted that this could’ve been partially-confirmand by the data about the frequency of work interruptions for these cis males and it was not). It could also be said with decent certainty that cis females are more likely to say that they are moderately(“sometimes”) interrupted by their work than cis males. I’m not really sure how this should be interpreted though as I initially thought that it might mean that cis females prefer to identify with less extreme descriptions of their mental health conditions, however there wasn’t a significant difference in the proportion of cis males and cis females that described frequent (“often”) interruptions. The fact that cis males generally report “never” having work interruptions due to mental health concerns more often than cis females also isn’t surprizing considering that cis males are less likely to have ever received treatment.

Overall I was hoping to find more interesting information and I think I would’ve with a larger dataset, but it was still an interesting project to work on.