Determining Myers-Briggs Personality Types from Internet Postings

Jose Reyes - Data Science Career Track - Winter 2018

What is the problem you are trying to solve?
The Myers–Briggs Type Indicator (MBTI) is an introspective self-report questionnaire with the purpose of indicating differing psychological preferences in how people perceive the world around them and make decisions.[1] My research is focused on classifying users to 1 of the 16 Myers-Briggs personality types using textual data found in internet comment sections and user forums.

Who is your client and why do they care about this problem?
My clients are primarily educational companies and/or academic institutions that need to deliver complex information to targeted audiences. Additional clients also include organizations that need to assess the personality types of their clients/employees quickly and without requiring them to the take actual Myers-Briggs (MB) personality assessment. My research would allow these organizations to asses their users' MB personality types from written text from either online or offline, which would allow for better customization of the delivery of information. For example, if a user is identified as one of the extrovert MB types, the material can be presented in a livelier manner so as to maintain their attention and improve retention, whereas if a user is deemed more introvert, a more tamed delivery would be applied.

Because the Myers-Briggs Type Indicator (MBTI) has to be taken and re-taken regularly to assess the official indicator (as personality types tend to waver over time), my research can lead to the formation of a machine learning algorithm which can reassess the users' MBTIs via written content on an ongoing basis which can then provide an up-to-date and appropriate method of delivery of educational information without the constant need to retake the official MBTI assessment. If a user's personality shifts from one type to another, the machine learning algorithm would be able to detect and account for the change. Ultimately, my research would be able to provide these organizations with a more robust method to connect with their students, allowing for the customization of the learning experience so that it aligns with the student's method of processing information. This in turn would provide better material retention, improved satisfaction from both students and instructors which in turn can lead to better educational company bottom line.

What data are you going to use for this? How will you acquire this data?
I will be using the MBTI-Type dataset from the Kaggle website.

In brief, outline your approach to solving this problem.
The dataset consists of 8676 rows and 2 variables. I will clean and separate the data set by into URL/comments for easier analysis of each and I proportionally allocate a to-be-determined number of records to a training and test sets.

Utilizing the training set, I will conduct a Natural Language Processing analysis of the user posts and identify the significant variables that best predict user's Myers-Briggs

[1] https://en.wikipedia.org/wiki/Myers–Briggs_Type_Indicator

personality type. My dependent variable for this study will be the "type" variable which denotes the user's Myers-Briggs Personality Indicator. As part of the data wrangling process, I will extract the nouns, adjectives and verbs from the "posts" column along with the websites visited; these will serve as my independent variables. However, with further exploration and analysis of this dataset, this list of independent variables may change depending on their level of significance in predicting a successful outcome.

Utilizing the findings of the training set, I will apply the machine learning algorithm to the testing set and compare the predicted outcome versus the actual outcome to measure the success and strength of the model.

What are your deliverables?

I will provide the code that illustrates my analysis, I will also provide a slide presentation along with a paper that shows my empirical findings along with recommendations for implementation of my proposed solutions along with a discussion and additional ideas for further research.