Milestone 1 - R vs. Python Preference Analysis through the RStudio Community Survey

Xiaolong Ge

Objective

The objective of this project is to utilize the information collected from the RStudio community survey to understand the factors influencing the preference for R or Python among people like data scientists. And predict the possible trend of preference of using R or Python in the future.

Data

The data sets for this analysis is accessible via the provided repository of GitHUb, which hosts the survey responses collected by the RStudio community. It includes the data vary from year 2018 to 2020. They conducted the survey through the Internet in English and Spanish. Each year, there are minor updates in the questions of the survey. At the same time, because of the data sets are stored on GitHub, we can use the R package `usethis` to clone the repository and get the data, functions like `use_git_clone()` will helps.

The data sets are divided into 3 parts: `data`, `dictionaries` and `gendercoder`. In the `data` folder, the question and answer forms are split by `tab` and stored in one-to-one correspondence in tsv files, and the answers contain numbers, open text, and fixed text. `How likely are you to recommend R to a colleague, friend, or family member?`, `How often do you write unit tests for your R code?` Questions like these that directly express preference will be key to determining preference. In addition, information about the respondent's background, such as `What industry do you work or participate in?`, can also help to classify and predict their affiliation.In the `dictionaries` and `gendercoder` folders, on the other hand, the coding on gender issues and the corresponding dictionaries for translating Spanish into English are recorded, which may be useful for data cleaning.

Due to the respondents are all coming from RStudio community, it do carry sampling bias and not random, so it may not reflect the true preference. In addition, the lack of information due to language translation and the difficulty of categorizing information due to open text will also be barriers to predicting preferences.

Exploratory Analysis

1. Based on the above objectives, the main goal is to analyse and make predictions about the preferences of a specific group of people for R or Python. For example, based on the survey data obtained in the last three years, it was analysed to find out what group of people prefer R and what related factors play a key role in the group of people who prefer R.
2. Once we get the main result, it is possible to make predictions about future preferences, like the possible situation of 2021.
3. Are preferences affected by gender differences? We can use `gendercoder` to answer this question.
4. At the same time, we can also learn where people like or dislike about R. For example, including packages such as `ggplot`.

Approach

Given the nature of the survey data, I need to process the data files into a format that is easy for R to handle, like like using numbers as weights instead of text responses. Also, I need to process Spanish responses with the help of dictionaries. Also, for open text, features need to be extracted by NLP methods such as keyword matching to facilitate data categorization. Finally, the analysis and prediction of key factors are made by graphing and combining the preference results.

But I don't any ideas of how to use `gendercoder` for processing gender data at present.

Challenges

1. Several challenges are anticipated in this project, including handling missing data and dealing with potentially unbalanced response categories, so that we can ensure that the model can generalize well to unseen data.
2. Additionally, it is also difficult to process and categorize open texts and extract their features because of the existence of open texts, where each person expresses the same degree of attitude differently.