Dear Candidate,

This assignment is designed to generally gauge your experience and exposure to concepts and approaches typically used to accomplish tasks that Hearful tackles to extract useful information from mined data.

There are multiple approaches that you may use to accomplish the assignment and any approach you take is fine, as long as you can explain the logic behind your approach. Hints and considerations will be provided along the way. If you have any follow-up questions, do not hesitate to reach out to me at *anya at hearfulhub dot com*

Please complete the assignment and commit your code to a GitHub repository. Subsequently, share the link to your completed assignment with me via email.

Thank you for your time completing this assignment.

Regards,

Anya

# Task 1

This assignment deals with identifying and cleaning product brand strings from raw data obtained from the source website.

When data is obtained (scraped) from a source website, the brand string seen on the web page of a particular product may be written in a multitude of different ways. Some brands may be more complicated to identify than others depending on the variety of ways their products are presented by source websites.

In this task, you will be looking for brand strings that associate with the brand Ralph Lauren in a raw data file called “Distinct\_Brands.txt”.

Download the file and take some time to analyze the various ways the Ralph Lauren brand may be scraped directly from the page. Some examples are below:

ralph lauren

RalphLauren

ralph-lauren

Polo Ralph Lauren

Polo-Ralph-Lauren

Polo Raulph

Polo Ralph Lauren Woven Stripe Pajama Pants

There are many other strings in this file identifying products with the brand Ralph Lauren. The above list is just a small subset of those strings to give you an idea.

Your task is to study the raw file and use your best judgement to identify all the strings that are affiliated with the brand Ralph Lauren. Subsequently, write a Python script that reads in the “Distinct\_Brands.txt” file, applies an approach to identify all the Ralph Laurent related strings and outputs those strings to another file. Please commit both your code and the output file containing the strings you identified to the GitHub repository.

# Task 2

This assignment deals with analyzing Hearful predictive model output and results.

Hearful uses rule-based models to predict the presence of specific themes in customer product reviews. These models are domain specific and look for a discrete set of themes depending on the product domain. For example, reviews on products from the fashion domain are analyzed for the presence of the following themes: fit, quality, style, use, and value. A given review text may contain zero or more themes.

Hearful also leverages sentiment models to identify the sentiment of a comment that mentions a specific theme. For example, if the product review mentions theme fit, it will also identify if theme fit was mentioned positively or negatively by the author. If a model identifies the presence of a theme within the review, it will also assign a sentiment to the given theme. Sentiment values are assigned to each theme independently. Therefore, a given review may have positive sentiment for one theme, but negative sentiment for another.

In this task, you will analyse the file APPAREL\_ids\_1\_2019.csv which represents sample output of one of Hearful’s models for the fashion domain. The contents of the file represent model predictions for 300 customer reviews. If a specific theme is identified to be present by the model, a 1 is placed in the associated theme column. Subsequently, the theme sentiment is assigned as pos or neg for positive or negative respectively in neighboring columns.

Your task is to study the provided file to understand the structure of the data and anything unusual that you may note about this data. Subsequently, you will write a Python script to read in the provided file and analyze the number of reviews for each theme and within each theme, how many reviews were positive and how many were negative. In your analysis, you may run across data points that seem unusual or improperly represented. You may drop some data points in your analysis provided you supply logic for your decision.

Please commit any code you write and the output of any analysis to the GitHub repository.

# Task 3

This assignment deals with evaluating Hearful predictive model performance against hand validated results.

Hearful relies upon a human validator to evaluate the performance of its models. The human validator is presented with the same set of 300 reviews that the model is run against to produce a similarly structured file containing predictions for the presence of specific themes as well as the sentiment associated with those themes.

For this task, you will be analyzing a human validated output for the same set of 300 reviews, contained in the file APPAREL\_ODOM\_1\_2019.csv. Your task is to write a Python script to evaluate the performance of the model on the following metrics: precision, recall, accuracy, and f-measure.

The metrics should be computed via standard formulas, using a 2x2 matrix, with true positive, true negative, false positive, and false negative values computed for each theme and each theme’s sentiment independently. The results should be presented similarly to the below table.

|  |  |  |
| --- | --- | --- |
|  | theme\_exists | theme\_sentiment |
| accuracies.fit | 0.123456 | 0.123456 |
| accuracies.quality | 0.123456 | 0.123456 |
| accuracies.style | 0.123456 | 0.123456 |
| accuracies.use | 0.123456 | 0.123456 |
| accuracies.value | 0.123456 | 0.123456 |
| fmeasures.fit | 0.123456 | 0.123456 |
| fmeasures.quality | 0.123456 | 0.123456 |
| fmeasures.style | 0.123456 | 0.123456 |
| fmeasures.use | 0.123456 | 0.123456 |
| fmeasures.value | 0.123456 | 0.123456 |
| precisions.fit | 0.123456 | 0.123456 |
| precisions.quality | 0.123456 | 0.123456 |
| precisions.style | 0.123456 | 0.123456 |
| precisions.use | 0.123456 | 0.123456 |
| precisions.value | 0.123456 | 0.123456 |
| recalls.fit | 0.123456 | 0.123456 |
| recalls.quality | 0.123456 | 0.123456 |
| recalls.style | 0.123456 | 0.123456 |
| recalls.use | 0.123456 | 0.123456 |
| recalls.value | 0.123456 | 0.123456 |

Please remember, that the human validator result is taken as the golden standard. Therefore, if a certain result is marked as present by the validator but is not present in the model prediction, that is counted as a false negative. Conversely, if a model predicted an outcome not marked by the validator, that is considered a false positive.

Please justify any assumptions you make in your computation and submit all code and output files to the GitHub repository.