One of the challenges of HCI research when analyzing social data is coming up with computable definitions of what you want to measure. In this assignment, you will construct a computational definition for the nature of a social relationship. What questions and measures work best? Only with our efforts combined, will we discover the answers at the end of the assignment.

Read Gilbert et al. (2009) before starting this assignment. Their research explores social networking variables and their correlation to relationship strength. Today, messaging apps are overtaking social networking sites as a platform for online communication. We are going to look at messaging patterns' correlation with relationships, while also exploring different operational definitions of a social relationship. In other words, what can measures of messaging metadata tell us about your relationship with that person?

For January 29, please take a look at your own messaging records (just on your phone or on the web) and come up with a question that can capture an aspect of your relationship with a person. That aspect can still be relationship strength, but your question should be different from the 5 in Gilbert et al. Other aspects may be trustworthiness, dependency, or something else you come up with that is related to relationship strength. The answer should be scorable on a scale like those in Gilbert et al., i.e., (a number between 0–1).

After this, we will have a shared document with about 25 questions about the nature of a relationship (1 from each person in class, plus 5 from Gilbert et al.). Answer these questions for a random sample of 4 people in your messaging data (without looking at the data itself). To extract your messaging data, the easiest way may be Sochiatrist, a tool we have developed in our research group (http://sochiatrist.cs.brown.edu/ for an overview), but you are welcome to use your own techniques. You may talk to Talie or Gabby for support with Sochiatrist.

Please output this exact format, including the ISO 8601 style date and time but without the ‘T’, and get 4 random contacts by selecting random lines in the file (so you get a stratified sample of people rather than randomly sampling among the people you message) to make it more likely that frequent contacts will be selected. Keep all the messages (ignoring group messages if the platform has them) from the past 30 days for all 4 contacts.

Note that when reading or writing in this format, you should use the csv library that handles escaping quotes for you, not ','.split(). Next, hide the content of the messages by replacing the actual text content (not including the CSV escapes) with the ‘#’ character, and changing your contact’s name to an anonymous identifier that no one else would select. For example, the first two lines above would become:

2020-01-05 16:39:02, sent, #hello I am your friend how are you?, #Angelina Jolie, #facebookmessage
2020-01-05 16:40:07, received, #I am good, friend. I am writing you a message., #Angelina Jolie, #facebookmessage
2020-01-06 12:12:12, received, #hello I am your frenemy. do you want to come to the mall with me?, #Keanu Reeves, #texting
2020-01-06 12:19:11, received, #hello again, you did not answer my message. please come#, #Keanu Reeves, #texting

¹Messaging includes texting (SMS, iMessage), Facebook Messenger, WhatsApp, Twitter/Instagram DMs, etc.
Your next challenge will be to come up with a measure (like a “feature” in machine learning terms) for estimating the answer to your relationship question using an anonymous conversation as input. Your measure can take into account the time of day of the messages, speed of the responses, length of the typical conversations, frequency of communication, etc. Give a descriptive name for your function, which should take in one parameter: the filepath of the CSV file, and return a dict where each contact is assigned a number between 0 and 1, e.g., {"tomato3483": 0.346, "tomato5678": 0.882}. Consider what we discuss in class on January 29 about behavior data (weighting for intent, normalizing per person); I’d encourage you to choose a measure that is not a simple count. Test your function on your own CSV file (but ignore your own contacts when we test the correlation between your question and your measure). Write this function to be compatible with a recent version of python3, ideally something self-contained without external libraries (talk to me if you absolutely need an external library).

For the midpoint on February 5, do the following:

1. Extract your messaging data and hide the content like above.
2. Answer the 25 questions for each of the 4 contacts that you sample using random messages.
3. Come up with a measure for estimating the answer to your relationship question.

Then everyone will have a set of answers to the 25 questions, each question with about 80 answers (assuming 20 people in class). We will also cross-share the answers and output of the functions, creating a large table of measures for each relationship.

For the due date on February 12, hand in a short report (up to 4,000 words) at the beginning of class of what question and measure you chose, and why you believe the measure is a good estimation of the question’s answers. Discuss in your report whether your measure was correlated with the relationship questions you did not write, and whether you would ask your question differently, or how you might adjust your measure to be more accurate. Finally, compute a simple linear regression of the measures (treating them like features) so report how accurate your measure is and when did it worked well or poorly, what questions it best explains ($R^2$), and the relative importance of your measure for questions it explains well (standardized beta). Note the stats program called SPSS (Brown has a license) can do these calculations.

At the end as a class, we will have an understanding of what questions about a relationship’s strength best map to computable measures on an $N = 80$ participant dataset.

**Grading**

Your grade will be based on your function for calculating your measure (3 points), its appropriateness for the messaging metadata as behavior data (2 points), how well your measure estimates the score from your question (2 points), your multiple linear regression analysis (4 points), the descriptiveness of your report (3 points), completing your midpoint check-in (2 point), and whether your measure also correlates with other measures of relationship strength (2 points).