Commuter Preferences Data Collection Software
Requirements Document

Overview:

The goal of this project is to facilitate the collection of anonymized data for research purposes related to how individuals' preferences change over time. The produced software should include a locally installed app to collect data with limited user-interaction relating to an individual's daily transit, and servers to collect and group anonymized data and to verify participation for payment/incentivization purposes. Primary concerns are the validity/usefulness of the anonymized data, participants' data security and privacy, and convenience for participants.

Background:

The client, Dr. Judy Goldsmith, runs a research lab at UK that deals with computational models of human preferences. Her lab is interested in understanding and predicting how human preferences change over time, and they are interested in a data source that is nonproprietary and can be released on a repository for preferences-research data (preflib.org). To this end, the client desires software that can reliably capture data on a research participant's daily transit activities (e.g. which route they take from work, whether they stop to run errands on the way home, etc.) over a period of several months by leveraging hardware in mobile phones, such as the GPS.

Dr. Goldsmith works with a computational model of preferences called a CP-net that allows the expression of preferences over features of an outcome based on the value of other features. A person's preference over roads home might depend on features such as day of the week and time of day, but may also depend of factors such as what restaurant the person wants to stop at, whether the roads are icy or clear, or whether that route has been taken recently. In order for the client's lab to do analysis of the data, they will need time-stamped data indicating which route was taken and what the values of the other relevant features were. It may also be necessary for the software to ask the research participant to suggest or supply a relevant feature that affected their choice of route.

User Types:

Research Participants ("Participants"): A participant begins interacting with the software after downloading the app to a mobile device. The first N participants are enrolled in a periodic rewards program where they receive gift cards by mail for continued involvement in the research study. They should not be prompted for personally identifying information such as email and mailing address unless they are one of the first N people to download the app and opt to enroll in the research study. If they provide their personal information, they will receive a small gift card to start and the participant's data will become associated with an anonymous user key for the purpose of grouping data from the same user. The participant should remain anonymous with

respect to the content of uploaded data and the user key their data is tagged with. At the same time, it should be possible for an authorized researcher to determine who has continued to participate in the study for the purposes of sending out additional gift cards, but they should not be able to identify the participant a data set is associated with, even based on timestamps, etc.

Participants should be able to see summaries of the data collected so far, including a list of tracked locations (relevant sources and destinations for a route), features that are most predictive of which route they take, and their most common routes.

Researchers: Researchers should be able to access the uploaded data after the study. An authorized researcher should be able to determine periodically which participants have continued to participate in the study. Researchers should also be able to gather up-to-date information about the number of active participants, volume of data collected, and summary statistics including average number of observations per source-destination pair, per route, and per total outcome (include route taken and all other features).

Collected Data:

The app should keep a list of locations with numerical keys, and a total set of features which use a different set of numerical keys. Associated with each feature should be a set of possibilities, or "alternatives". E.g. the route taken feature might be feature 0 and have three options reflecting a short route (alternative 0:0), a fast route (alternative 0:1), and a low-traffic route (alternative 0:2). Data uploaded to the server should consist of a list of trips taken with each trip in the form <starting point, starting time, list of the alternatives observed for each tracked feature, ending point, ending time>.

The app should use statistical rules to distinguish between large locations (like a grocery store), and distinct locations near each other like the participant's apartment and the convenience store down the block. When the app uploads data to the server, it should also record time-stamped confidence values for the identification of locations and features.

Security:

Data for the researchers should be available online through secure connections with password-enabled access. Only anonymized data should be uploaded from a participant's device. A personal identification number or some other security measure should prevent nonparticpants from accessing the app on a participant's device, without significantly obstructing a legitimate participant. Web-based API's used by the software should not leave user-accessible data on a participant's mobile device.

Interfaces:

When participants first open the app, they should be asked for a short numerical PIN, and they should have the option to opt into the study. If applicable, they should have the option to

enter personally identifiable information in order to receive payment. Following that, every time the participants opens/views the app, they should be able to go to either a settings page where they can adjust active hours (to reduce power consumption at night, for example), edit mailing address or other information if applicable, and see (perhaps with a map) the locations that are being tracked by the app, or they should be able to go to a data summary page where they can see summary statistics about the routes they have taken while the app was active. The setttings page should also have a setting to disable the app.

The app should send the user a notification (on the device) if the app is unable to collect data (for more than a day) because the GPS is turned off, or if user input is needed to identify or indicate which are the relevant features for choosing a particular route, or asking if a new location has been properly identified. Responding to a notification about features, or accessing the settings or data summary page should require a participant to enter their PIN. If a PIN is forgotten, it should be resettable from the email address entered by the user when they first opened the app.

When a participant is queried if a newly identified location is relevant, the app should present them options of "Yes, this location affects how and when I go places," or, "This location is not relevant," or, "Please to not track routes which visit this location." When a participant is queried as to why they chose to take a particular route (which should occur every time they take a new route between locations previously visited), checkboxes should be presented with a list of possibly relevant features, as well as an option to add arbitrarily many new features. There should also be an checkable option that says those are always the set of features that correspond to taking that route.

When a researcher wants to view summary statistics about collected data, they should be presented with a screen where they can enter their password, and then should be able to view a dashboard presenting the statistics about uploaded data. A specially authorized researcher should be able to see a summary of participation, and should have the option to access individual email addresses or to send out a group notifications to those continuing to participate.

Summary:

The result of this project will be an app and server software to facilitate the collection of data related to individuals' preferences in daily transit. The software will allow researchers to track participation and volume of collected data, and the software will allow participants to manage what data is collected by their app and see summaries of their behavior as observed by the app.