Folder with datafiles for the analysis

This folder contains the datafiles which are used in the Python-based analyses of our paper.

Contents of this folder

```
|   cleaned_database_data.csv
|   preprocessed_database_data.csv
|   preprocessed_demographic_data.csv
|   preprocessed_post_questionnaire_data.csv
|   preprocessed_pre_screening_data.csv
|   README.md/README.pdf
```

Filename	Content
cleaned_database_data.csv	Cleaned data from the database of the conversational sessions from the study.
preprocessed_database_data.csv	Preprocessed data from the database of the conversational sessions from the study.
preprocessed_demographic_data.csv	Preprocessed data from Prolific on demographics from the participants of the study.
preprocessed_post_questionnaire_data.csv	Preprocessed data from the post-questionnaire of the study.
preprocessed_pre_screening_data.csv	Preprocessed data from the pre-screening of the study.
README.md/README.pdf	The file you are currently reading - provides an overview of our data files.

Contents of the data files

cleaded_database_data.csv

This file contains the data gathered during the conversation sessions between the participants and the virtual coach. The data is already anonimized and cleaned using the script in scripts/Pre-processing.py. It contains the following fields:

- ID, the anonimized ID of the participant.
  - Question: NA
  - Possible values: p1 to p117
  - Note: p1 in this file is the same participant as p1 in the other data files.
- Session number, the session number corresponding to the data.
  - Question: NA

- Possible values: 1 to 5
  - Note: -

- Response type, the type of the data.
  - Question: NA
  - Possible values: prev_activity, mood, rest, available_time, self_motivation, self_efficacy, rl_action, initial_proposal, number_of_rejected_proposals, goal_achievability, self_efficacy_feedback, goal, rejection_reason
  - Note: -

- Response values, the actual data.
  - Question: NA
  - Possible values: NA
  - Note: The question and possible values depend on the type of data. These are listed in the fields of the preprocessed_database_data.csv.

preprocessed_database_data.csv

This file contains the data gathered during the conversation sessions between the participants and the virtual coach. The data is already preprocessed using the script in scripts/Pre-processing.py. It contains the following fields:

- ID, the anonimized id of the participant.
  - Question: NA
  - Possible values: p1 to p117
  - Note: p1 in this file is the same participant as p1 in the other data files.

- Self-motivation, the indicated self-motivation of the participant of the current session.
  - Question: How motivated or demotivated are you to go for a walk today on a scale from 0 to 10?
  - Possible values: 0 (I am very demotivated to go for a walk today) to 10 (I am very motivated to go for a walk today)
  - Note: -

- Self-efficacy, the indicated self-efficacy of the participant of the current session.
  - Question: If you were given a step goal of your average number of steps from the last few days, how confident or not confident are you that you can achieve that step goal today on a scale from 0 to 10?
  - Possible values: 0 (I'm not confident at all that I can achieve that step goal today) to 10 (I'm very confident that I can achieve that step goal today)
  - Note: -

- Rest, the indicated rest of the participant of the current session.
  - Question: How did you sleep last night on a scale from 0 to 10?
  - Possible values: 0 (I slept terribly last night) to 10 (I slept amazing last night)
  - Note: -

- Available time, the indicated available time of the particitpant of the current session.
  - Question: How much time do you have available today to go for a walk on a scale from 0 to 10?

- Possible values: 0 (I have no time at all to go for a walk today) to 10 (I have a lot of time to go for a walk today)
    - Note: -
- Mood, the indicated mood of the participant of the current session.
    - Question: How would you describe your current mood?
    - Possible values: glad, happy, pleased, delighted, serene, content, satisfied, relaxed, calm, excited, astonished, aroused, sleepy, neutral, tired, tense, alarmed, afraid, droopy, bored, angry, annoyed, frustrated, distressed, depressed, sad, gloomy, miserable
    - Note: -
- RL action, the action taken by the virtual coach to adjust the recommended step goal.
    - Question: NA
    - Possible values: dec, sdec, nothing, sinc, inc
    - Note: decrease (-400 steps), slightly decrease (-200 steps), do nothing, slightly increase (+200 steps), increase (+400 steps)
- Next self-motivation, the indicated self-motivation of the participant of the next session.
    - Question: How motivated or demotivated are you to go for a walk today on a scale from 0 to 10?
    - Possible values: 0 (I am very demotivated to go for a walk today) to 10 (I am very motivated to go for a walk today)
    - Note: -
- Next self-efficacy, the indicated self-efficacy of the participant of the next session.
    - Question: If you were given a step goal of your average number of steps from the last few days, how confident or not confident are you that you can achieve that step goal today on a scale from 0 to 10?
    - Possible values: 0 (I'm not confident at all that I can achieve that step goal today) to 10 (I'm very confident that I can achieve that step goal today)
    - Note: -
- Next rest, the indicated rest of the participant of the next session.
    - Question: How did you sleep last night on a scale from 0 to 10?
    - Possible values: 0 (I slept terribly last night) to 10 (I slept amazing last night)
    - Note: -
- Next available time, the indicated available time of the participant of the next session.
    - Question: How much time do you have available today to go for a walk on a scale from 0 to 10?
    - Possible values: 0 (I have no time at all to go for a walk today) to 10 (I have a lot of time to go for a walk today)
    - Note: -
- Next mood, the indicated mood of the participant of the next session.
    - Question: How would you describe your current mood?
    - Possible values: glad, happy, pleased, delighted, serene, content, satisfied, relaxed, calm, excited, astonished, aroused, sleepy, neutral, tired, tense, alarmed, afraid, droopy, bored, angry, annoyed, frustrated, distressed, depressed, sad, gloomy, miserable

- Note: -
- Initial proposal, the first proposed step goal during the current session which could be changed during the session.
    - Question: NA
    - Possible values: 2,000 to 10,000
    - Note: this is the lowest of the three initial goal proposals.
- Previous_activity, the number of steps the participant took in the previous 9 days of the current session.
    - Question: How many steps did you take in the past 9 days?
    - Possible values: 9 numbers separated with a semicolon, ranging from 0 to 20000
    - Note: If the participant only filled in the past 5 to 8 days, the mean of those was used to fill until 9 numbers.
- Steps taken, the number of steps taken by the participant after the current session.
    - Question: How many steps did you take yesterday?
    - Possible values: 0 to 20,000
    - Note: this was asked during the next session.
- Goal achievability, the indicated goal achievability by the participant about the goal of the current session.
    - Question: Knowing what you know now, how achievable or unachievable do you think a step goal of X steps would have been yesterday on a scale from 0 to 10?
    - Possible values: 0 (I think that goal would have been very unachievable) to 10 (I think that goal would have been very achievable)
    - Note: this was asked during the next session, where X was replaced by the goal of the current session.
- Self-efficacy feedback, the indicated accuracy of the self-efficacy of the current session.
    - Question: Knowing what you know now, how would you evaluate your confidence for achieving a step goal yesterday on a scale from -5 to 5?
    - Possible values: -5 (I think my confidence yesterday should have been way lower) to 5 (I think my confidence yesterday should have been way higher)
    - Note: this was asked during the next session.
- Goal, the goal set during the current session.
    - Question: Which of the following step goals do you prefer: X, Y, or Z?
    - Possible values: 2,000 to 10,000
    - Note: this is the goal that was set during the session and can be different from the initial proposal. X, Y, and Z were replaced by three goal options, each 100 more than the previous one.
- Rejected proposals, the number of times that the goal proposals were rejected in the current session.
    - Question: NA
    - Possible values: 0 to 4
    - Note: The goal proposals were rejected if the participant chose the option 'other' instead of one of the three goal proposals.
- Rejection reason, the indicated reason by the participants why the goals were too high of the current session.
    - Question: Why do you think so?

- Possible values: weather, tired, time, other
    - Note: this was only asked once when the participant indicated that the goals were too high.
- Session number, the number of the current session
    - Question: NA
    - Possible values: 1 to 5
    - Note: -

preprocessed_demographic_data.csv

This file contains the data gathered from Prolific on the demographics of the participants. The data is from the Prolific profile of the participants. The data is already preprocessed using the script in scripts/Pre-processing.py. It contains the following fields:

- ID, the anonimized id of the participant.
    - Question: NA
    - Possible values: p1 to p117
    - Note: p1 in this file is the same participant as p1 in the other data files.
- Approvals, the number of approved submissions on Prolific of the participant.
    - Question: NA
    - Possible values: 1 to 100,000
    - Note: we were not able to select participants who had more than one approved submission, but we had to provide an upper limit. We chose 100,000 for that which should have given us all people with more than one approved submission.
- Gender, the gender of the participant.
    - Question: What gender are you currently?
    - Possible values: Man (including Trans Male/Trans Man), Woman (including Trans Female/Trans Woman), Non-binary (would like to give more detail), Rather not say
    - Note: -
- Education_level, the highest completed eduction of the participant.
    - Question: Which of these is the highest level of education you have completed?
    - Possible values: No formal qualifications, Secondary education (e.g. GED/GCSE), High school diploma/A-levels, Technical/community college, Undergraduate degree (BA/BSc/other), Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other), Don't know / not applicable
    - Note: -
- Exercise, the amount of time the participant engages in physical activity during a week.
    - Question: How often do you engage in physical exercise per week?
    - Possible values: Never (0 – 60 minutes per week), Sometimes (60 – 150 minutes per week), Often (more than 150 minutes), Rather not say
    - Note: the answers to this question and the same question in the pre-screening data might be different as this Prolific profile data can be outdated for some people. Also, we only selected people who answered

Never or Sometimes to participate, so only those answers will appear in the data.

- Age, the age of the participant.
  - Question: What is your date of birth?
  - Possible values: 18 onwards
  - Note: We only selected people who answered between 18 and 65 to participate, so only those answers will appear in the data.

preprocessed_post_questionnaire_data.csv

This file contains the data gathered in the post-questionnaire of the study. The data is already preprocessed using the script in scripts/Pre-processing.py. It contains the following fields:

- ID, the anonimized id of the participant.
  - Question: NA
  - Possible values: p1 to p117
  - Note: p1 in this file is the same participant as p1 in the other data files.
- Steps, the number of steps the participant took on the previous day.
  - Question: How many steps did you take yesterday?
  - Possible values: 0 to 20,000
  - Note: -
- Goal_achievability, the indicated goal achievability of the goal of the previous day of the participant.
  - Question: Knowing what you know now, how achievable or unachievable do you think the step goal of yesterday was on a scale from 0 to 10?
  - Possible values: 0 (I think that goal would have been very unachievable) to 10 (I think that goal would have been very achievable)
  - Note: -
- Confidence_accuracy, the indicated accuracy of the self-efficacy of the session of the previous day.
  - Question: Knowing what you know now, how would you evaluate your confidence for achieving a step goal yesterday on a scale from -5 to 5?
  - Possible values: -5 (I think my confidence yesterday should have been way lower) to 5 (I think my confidence yesterday should have been way higher)
  - Note: -
- ASA1 to ASA24, the question from the short version of the ASAQ.
  - Question: the questions are given in the paper from the ASAQ by Fitrianie et al. [1]
  - Possible values: -3 (disagree) to 3 (agree)
  - Note: question 12, 17, 18, and 22 are reversed questions, so -3 would be a positive answer there. Also, we replaced [the agent] in the questions to Steph.
- Goal_personalization, the level of personalness that the goals felt to the participant.
  - Question: How personal (created especially for you) did the proposed goals in the conversation sessions feel to you on a scale of 0 to 10?
  - Possible values: 0 (The goals did not feel personal at all) to 10 (The goals felt very personal to me)

- Note: this was about all the goals of the study.
- Why, the reason for the goal personalization answer.
  - Question: Why do you think so?
  - Possible values: NA
  - Note: this was a free text response.
- Goal_difficulty, the indicated difficulty of the step goals of the participant.
  - Question: How easy or difficult did you find it to reach the daily step goals?
  - Possible values: -5 (It was very difficult to reach the daily goals) to 5 (It was very easy to reach the daily goals)
  - Note: this was about all the goals of the study.
- Why, the reason for the goal difficulty answer.
  - Question: Why do you think so?
  - Possible values: NA
  - Note: this was a free text response.
- Motivation, the indicated part of the conversation sessions that was the most motivating to the participant.
  - Question: What motivated you the most to go for a walk in the conversations/about the conversations?
  - Possible values: NA
  - Note: this was a free text response.

preprocessed_pre_screening_data.csv

This file contains the data gathered in the pre-screening of the study. The data is already preprocessed using the script in scripts/Pre-processing.py. It contains the following fields:

- ID, the anonimized id of the participant.
  - Question: NA
  - Possible values: p1 to p117
  - Note: p1 in this file is the same participant as p1 in the other data files.
- Exercise_per_week, the amount of time the participant engages in physical activity during a week.
  - Question: How often do you engage in physical exercise per week?
  - Possible values: Never (0 – 60 minutes per week), Sometimes (60 – 150 minutes per week), Often (more than 150 minutes per week)
  - Note: the answers to this question and the same question in the Prolific data might be different as this pre-screening data can be more recent for some people. Also, we only selected people who answered Never or Sometimes to participate, so only those answers will appear in the data.
- Steps_on_average, the average number of steps the participant took per day for the last week.
  - Question: How many steps did you take per day on average in the last week?
  - Possible_values: 0 to 20,000
  - Note: -
- Way_of_tracking_steps, the indicated way the participant was tracking the number of steps they did each day.
  - Question: How do you track the number of steps you take each day?

- Possible values: Using a Smartwatch, Using an iPhone's Health app, Using the Samsung Health app, and Other, namely
  - Note: -
- other, in case the participant answered 'Other, namely' to the way of tracking steps question, they can give their input.
  - Question: NA
  - Possible values: NA
  - Note: this was a free text response.
- TTM_stage, the stage of change the participant was in to become more physically active.
  - Question: If regular physical activity was defined as taking 10,000 steps per day, would you say you are doing regular physical activity according to this definition?
  - Possible values: values from the Exercise: Stages of Change (Short Form) by Nigg [2].
  - Note: We only invited people who answered No, but I intend to in the next 6 months, or No, but I intend to in the next 30 days, which is why only those values appear in the data.
- Godin1 to Godin3, the answers to the three questions of the Godin leisure time acitivity questions.
  - Question: the questions are given in the paper from Godin [3].
  - Possible values: 0 to 120
  - Note: -

References

[1]: Fitrianie et al., 2022, The Artificial-Social-Agent Questionnaire: Establishing the long and short questionnaire versions in IVA '22.
[2]: Nigg, 2023, Exercise: Stages of Change (Short Form) from https://web.uri.edu/cprc/measures/exercise/stages-of-change-short-form/.
[3]: Godin, 2011, The Godin-Shephard leisure-time physical activity questionnaire in Health & Fitness Journal of Canada.