ASC Custom Tool Explanation

Mark Soulier

April 2024

1 Introduction

In this document, we describe the functionalities and the workflow of the ASC Custom Tool designed for managing and processing data and media files. This tool is specifically tailored to optimize the preprocessing steps necessary for efficient data analysis and video content creation.

2 Data Cleaning

The data cleaning module of the ASC Custom Tool is designed to handle and organize a variety of file types automatically. The process is delineated as follows:

1. The tool starts by accepting input in the form of a zip folder.
2. It then analyzes each item within the folder to categorize them based on file types:
   - CSV files are identified and tagged as data files.
   - Image files (PNG, JPG) and video files (MP4) are recognized as content files.
   - All other file types are grouped under a miscellaneous category named 'Other'.

   The ASC Custom Tool uses an advanced fuzzy logic algorithm to identify and match column names from the input files with a predefined set of keywords. This functionality is crucial for ensuring that data from different sources are standardized and harmonized for further processing. Below is an explanation of how fuzzy logic is applied:

   (a) **Keyword Set Definition:** The tool defines a comprehensive dictionary, *sensor_categories*, which maps each data attribute to a list of possible column names. For example, the attribute "Time" is associated with potential column names like "AlignedTimeSec", "Time", "Timestamp", and "Time_Stamp".

```
sensor_categories = {
    "Participant": ["Participant_ID", "Participant", "ID", "Respondent"],
    "Medium": ["Medium", "StimulusLabel", "SourceStimuliName", "SlideEvent"],
    "Time": ["AlignedTimeSec", "Time", "Timestamp", "Time_Stamp"],
    "DilatedPupilRight": [
        "DilatedPupil", "Pupil Size", "Pupil", "ET_PupiLeft"
    ],
    "DilatedPupilLeft": ["DilatiedPupilLeft", "Pupil Size Left", "ET_PupiLeft"]
    "GSR": ["GSR", "Conductance", "Skin Conductance"],
    "HR": ["HR", "Heart Rate"],
    "Blink": ["Blink", "Blinking"],
    "Eye_X": ["Eye_X", "X_Coordinate", "Gaze X"],
    "Eye_Y": ["Eye_Y", "Y_Coordinate", "Gaze Y"],
    "Joy": ["Joy"],
    "Anger": ["Anger"],
    "Surprise": ["Surprise"],
    "Fear": ["Fear"],
    "Disgust": ["Disgust"],
    "Sadness": ["Sadness"],
    "Contempt": ["Contempt"],
    "Frustration": ["Frustration"],
    "Confusion": ["Confusion"],
    "Neutral": ["Neutral"]
}
```

(b) **Fuzzy Matching Process:** When a file is read, each column name from the file is compared against the potential names listed in *sensor_categories* using a fuzzy matching algorithm. This algorithm calculates the similarity between the file's column name and each name in the list based on a set of linguistic rules that account for common variations and typographical errors.

(c) **Matching Outcome:** The highest scoring match above a certain threshold is selected as the standardized name for that column. This process ensures that even if a column is labeled slightly differently across files, it will still be recognized and correctly categorized by the tool.

(d) **Integration with Data Cleaning:** Once the columns are matched and standardized, they are incorporated into the cleaned data file as described in the Data Cleaning section. This ensures consistency and accuracy in subsequent data processing and analysis steps.

This implementation of fuzzy logic not only enhances the flexibility of the ASC Custom Tool but also significantly reduces the manual effort required to prepare and standardize data from diverse sources.

3. Subsequently, all identified data files are aggregated into a single cleaned data file. This consolidated DataFrame is structured with the following columns:

```
['Participant', 'Medium', 'Time', 'Dilated', 'GSR', 'HR', 'Blink', 'Eye_X',
'Eye_Y', 'Joy', 'Anger', 'Surprise', 'Fear', 'Disgust', 'Sadness', 'Contempt',
'Frustration', 'Confusion', 'Neutral']
```

4. Each data file undergoes a thorough scan whereby the tool detects and aligns the columns based on a predefined set of keywords. This ensures that the data across different files is standardized and ready for analysis

5. The tool will add in a participant column if none is detected with the participant name found at the top after #Participant Name

3 Video Creation

The video creation module leverages the content files sorted during the data cleaning process. This section is under development and will detail the steps involved in generating compelling video content using the identified media files.

1. Participants select an individual file through the drop down menu on which the tool reads in for video creation.

2. The selected file undergoes peak processing where Heart Rate and pupil dilation are standardized. The tool then fits the Galvanic Skin Response (GSR) data to a linear regression model. The residuals of this model are standardized according to the peak processing equations. These data parts are saved to peak processing file in results with those appended columns.

3. The Peak Processed data will then be used to construct a matplotlib graph that graphs the peak data and saves that off in results. Which will be used to display the graph at the bottom of the video.

4. The Video Creator module loads internal images such as emojis, GSR graphs, and heart rate icons. It sequentially processes participant data, line by line.

5. For visualization purposes, as the video creator goes line by line the tool saves the most recent valid data in the file, ignoring 'None' values. It displays the latest available data for each frame.

6. The tool processes the video frame by frame, calculating the actual time for each frame and linking it to the closest data entry to retrieve relevant data.

7. Each frame is created by synthesizing the processed data with the corresponding images, ensuring that all frames are accurately representative of the data timeline.