2017-10

Transcribe-a-thon for BU Recipe Boxes Event Materials

Build New Project

- See Zooniverse's tutorial on how to build a project: https://www.zooniverse.org/help
- Key settings:
    - Completeness statistic: Retirement count
    - Transcribe workflow
        - First, ask users to identify whether there is recipe content on the image
        - Then, ask them to transcribe text as written

Add Subject Set

- Batch resize images in Preview to 1000 pixels wide; images should be less than 1MB in size
- Create csv file with subject ID and file name
- Make sure all files are saved in the same folder, with the csv file
- Go to Subject sets/New subject set
- Drag and drop csv file and images into interface

Data Export

- Request new classification export
- Receive download link for csv file in email
- Cell contents in csv file will be in json

Data Cleaning

- Import csv into OpenRefine with ISO-8859-1 encoding (see exported json for OpenRefine)
- Reduce dataset to only those images with recipe content
    - Text filter on "annotations" column for "Yes"
- Identify filename for each image
    - Add new column **Filename** based on "subject_data" column, using GREL expression: value.split("image_name_1\":\"")[1].chomp("\"}}")
    - This expression splits the text after "image_name_1":" (so, the start of the filename) and creates a list of two items, takes the second item in this list (rbox12+0001.JPG"}}), removes the }} at the end of the filename, and puts this value in a new column.
- Identify classification count (or number of times the recipe was transcribed)
    - Add new column **Transcription count** based on "subject_data" column, using GREL expression: value.split("classifications_count\":")[1][0]
    - This expression splits the text after "classifications_count":" (right before the numerical value) and puts the first value of the second list item into another column.

- Extract transcription from json
    - Add new column **Recipe text** based on “annotations” column, using GREL expression: value.split("T0\",\"value\":\"")[1].chomp("\",\"task_label\":\"Transcribe text as written on the image.\"}]")
    - Each recipe text was preceded by “T0","value":" and followed by ","task_label":"Transcribe text as written on the image."}] so this expression uses these markers to extract the recipe text and add it to a new column.
- Export results to csv file
- Open file in Excel
- Find and Replace for:
    - \u0026 - plus sign
    - Degree symbol
- Use conditional formatting on Transcription count column to highlight cells with values greater than 1
- Compare transcriptions to find most accurate; delete other transcriptions
- Save as csv file

Generate Text Files

- Use catalog merge in Microsoft Word to generate text document (see template)
- Find and Replace for \n - ^p to insert paragraph breaks
- Save as txt file