How to Upload Data to Reverse Indexing

What you need:

- Access to the FHTL image server (if you don't have this, email the FHTL lab lead or Dr. Clement)
- To have already uploaded your ssh public key to the byu linux account manager
- A collection of segmented snippets of single words
- Transcriptions for each of the snippets
- Only upload snippets and transcriptions which:
    - Are single words (not full sentences)
    - Contain only alphanumeric characters
    - Appear at least 12 times in the dataset

The first thing you'll need to do is upload the images to the image server. Here's how:

- SSH into the server with the appropriate login credentials
    - "ssh adm.<username>@fhtlstore1.byu.edu -p 22222"
- Once logged in, you will be in a local directory. This is not where the files need to be. Navigate to the "/data" folder in the root directory
    - "cd /data"
- Make a new folder where the data will be uploaded. This can also be nested within another folder if you want. Think ahead as to how you want to organize your data so that you don't have to be changing filepaths frequently. REMEMBER THE FILEPATH!
    - "mkdir <folder name>"
- Log out of the image server
    - "exit"
- Use SCP to copy your snippets to the folder you created within the image server. This may take some time.
    - "scp -P 22222 <local path to snippets> adm.<username>@fhtlstore1.byu.edu:<path to project folder>
- If the file you copied was compressed, log back into the server and extract.

Now, you need to properly format a .tsv file that matches the snippets to the transcriptions. Here's how:

- You should already have a transcription associated with each of the snippets you uploaded, likely in the form of a .csv or similar. If not, do so. The file should have only two columns, the first will contain the path to the snippet on the image server, and the second will be the transcription of that snippet.
- The filepath column should contain the full filepath starting from the "/data" directory. So if I've uploaded a snippet called "my_snippet.png" to a folder called "my_project", then the proper path would be "my_project/my_snippet.png".
- The transcription column should contain the proper transcription for each snippet. While your output may be mixed case or contain non-alphanumeric characters, for reverse indexing purposes, these transcriptions must be lowercase and alphanumeric.

- Keep in mind that reverse indexing shows a user 12 snippets at a time, so if you have words that only show up once or twice in your data set, uploading these will make the user experience a bit strange. Personally, I just drop such cases from my dataset.
- The header for the path column must be “image_relative_path”, and the header for the transcription column should be “predicted_transcription”
- Ensure that the file is tab delimited and saved as a .tsv
- Email this file to the FHTL lab.