Instructions for cleaning transcripts

The python script clean.py compiles doc file transcripts into two excel sheet formats, the convergence outputs and the collated output. Each doc file has one convergence output, which reflects the sequence of the conversation divided into individual speakers. The collated output is a single excel sheet that compiles all conversations into a single sheet, separated by speaker.

1. Place all transcripts (in .doc format) in samples folder.
2. Open Anaconda Command Shell and navigate to the current folder. To do so, type in

cd Desktop\cleantext

1. Type in the following command to run the script:

python clean.py

1. The convergence output sheets should now be in the “convergence\_outputs” folder. The collated output should be in the current folder, named “collated\_output”.

Compiled by Robyn Kwok, [jk355@duke.edu](mailto:jk355@duke.edu)