Kaylee Stinebiser

Regex Exercise 3

Steps

Step 1

I did a search for special characters in the document. I only found an &, so I changed it to work with XML.

Find box: &

Replace box: &amp;

Step 2

I removed the extra spaces.

Find box: \n{3}

Replace box: \n\n

Step 2

I manually added the root element to the document.

Step 3

I added the speech element tags to the individual speaking parts.

Find box: \n\n

Replace: </speech>\0<speech>

I removed the extra speech tag from the top and added it to the bottom.

Step 4

I added the speaker tags in different steps.

Find box: (<speech>)([A-Z]+.+?\.)

Replace box: \1<speaker>\2</speaker>

Find box: (<speech>)([A-Z]+ [A-Z]+ )(\[)

Replace box: \1<speaker>\2</speaker>\3

After these two replaces, I searched <speech>[A-Z]+ [A-Z]+ to see how many of my speeches with names did not have the <speaker> tags added. Some of these were because the first find search did not bring them up, and others were because of the stage directions following the names. I did individual find searches for these names to add the <speaker> tags. All of the following replace boxes are the same; only the name changes.

Find box: (<speech>)(THE FLOWER GIRL) (\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(THE NOTE TAKER )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(THE SARCASTIC BYSTANDER )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(THE SARCASTIC ONE )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(MISS EYNSFORD HILL )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(THE PARLOR-MAID )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(THE BYSTANDERS GENERALLY )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(PICKERING )

Replace box: \1<speaker>\2</speaker>

Find box: (<speech>)(PICKERING )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(FREDDY )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(MRS. PEARCE )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(LIZA )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(HIGGINS )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(HIGGINS ) \*\*\*also done for Mrs. Higgins

Replace box: \1<speaker>\2</speaker>

Find box: (<speech>)(DOOLITTLE )(\[)

Replace box: \1<speaker>\2</speaker>\3

Find box: (<speech>)(CLARA )

Replace box: \1<speaker>\2</speaker>

After this, I did a Find of <speaker>[A-Z]+ and scanned to see if there were any other names that I missed. For the most part, it only selected lines that started with a capital letter. There were a few names, though, and I entered those in manually.

Step 5

I marked the Acts.

Find box: ^<speech>(ACT [IV]+)</speech>

Replace box: <act>\1</act>

Step 6

I marked the stage directions.

Find box: \[.+\]

Replace box: <stageDirection>\0</stageDirection>

Step 7

The clean-up step: I manually added the title, author, date, note, preface, and subTitle tags to their respective texts. I changed the <speech> tag on the paragraphs of the preface to be a <p> tag instead.