Leeds Status Report Feb 15 to Feb 21 2021

Weekly status report for Feb 15-Feb 21 2021

Scope	Schedule	Cost	Risks	Quality

Key Performance Indicators

- **Schedule:** **ON SCHEDULE**
  - Schedule Variance: 0 Days
  - Percent complete : **70%**
- **Labor:** **ON SCHEDULE**
  - Labor hours: 11 hours/person
- **Administration:** **COMPLETE**
  - Sponsor meeting, group meeting, attendance, hour log

Summary

This week seemed to be a pretty successful week. We were unable to meet with our sponsor but had two team meetings and it seemed like a lot of progress has been made. We have finished a python script that given the .zay file, it can construct the 99 text blobs and choose the optimum one. The script that uploads all the words to the database is being improved to incorporate better information. Lastly, all of the page images were put onto AWS so that it is easier to interact with and we no longer have to access them via the cluster.

Work planned for this week

- Choose optimal text blob, specifically picking 1 of the 99 outputs to work with CAP-76

Work completed this week

- Choose optimal text blob, specifically picking 1 of the 99 outputs to work with CAP-76
- Improve script to load word data to database
- Uploaded all images to AWS for easier access

Choose optimal text blob CAP-76

✓ This task has been completed. Process to solve:

1. Create a word dictionary
2. Iterate over all 99 text blobs, keeping count of how many times each bigram appears
3. Also keep track of each text blobs individual bag of words
4. Iterate over each individual text blob's bag of words summing up how many times each bigram in the individual BOW appeared in all 99 iterations
5. Pick blob that has the bigrams that appeared most in all 99 (optimal)

⚠ Right now, the text assembling gets jumbled up sometimes with the balance sheets and other tables. Need to improve this in the future.

Upload data from cluster to DB

✓ Improved script so that each word entry includes what page number they are on and also the variables that correlate to our sponsor's file naming

Images to AWS

All of the images were uploaded to AWS. This allows us more freedom to interact with them as it was cumbersome to work with on the cluster and the file sizes were very large to download on our local machines.

Plans for next week

- Improve text arrangement of blobs, CAP-74
- Extract information from optimal blob, CAP-77

Open Issues

Deliverables and Milestones for sprint (2/15 to 3/1)

Deliverable or Milestone	WBS	Planned	Forecasted	Actual	Status
Text blobs > 1 column	CAP-75	2/14/21	2/14/21		Complete
Choose optimal text blob	CAP-76	2/28/21	2/28/21	2/20/21	Complete
Extract info from optimal blob	CAP-77	3/1/21	3/1/21		In progress

---

Open Change Requests

Change Request Name	Change Request Number	Requested Date	Current Status