Alexander Zurawski

azuraws2

CS 410

11/15/2021

CS410 Final Project Progress Report

- Which tasks have been completed?

The main thing I did was rescope the project. Instead of probing a lot of free Coursera classes, I will just be using the 3 I have access to. Instead of designing a crawler, I manually collected the data myself. I have also chosen to design a scoring function to be able to test the effectiveness of the different iterations of the ranking function. To do this, I will use the weekly quiz questions to design a set of queries that will be split into a testing and training group. After which I will rank the documents that are relevant as 1 and not relevant as 0. The rest of the project will continue as normal.

In terms of steps taken so far, the data has been collected.

- Which tasks are pending?
    - Conversion of data files into a bag of words representation
    - Create query set
    - Rank the data files in terms of relevance to query
    - Designing a ranking function
    - Designing a scoring function
    - Iterate on ranking function to optimize performance

- Are you facing any challenges?
    - Originally, I thought the data would be in the PPT file format, but it is in PDF format. I think this will be overall easier for the conversion.
    - Practical Statistical Learning is missing lectures from Week 5 onward. An entire week's worth of slides are all in one document rather than split into individual lecture videos. I may split them up or just keep them as is.
    - I think keep exclusively to the 3 classes I have could overfit the data. I hope splitting testing and training queries will avoid that.

- Response to specific feedback

Meta-Reviewer: I chose the three courses (CS410: Text Information Systems, CS598: Foundations of Data Curation, and CS598: Practical Statistical Learning) because of representation. The courses have some level of overlap, such as the EM algorithm, while also having a lot of unique information. I think it is an ideal test set.

Reviewer 1 & 4: I haven't thought of implementing a UI, I am focusing on the ranking and scoring. If I have time to build something pretty, I will. For now, it will run in the command line and output one or more data file. I can connect the data file to a URL that links to the lecture the data file is connected to.

Reviewer 3: I expected most people doing this project to focus on the transcript, so in a 'real-world' roll out, I would expect to mix multiple techniques to find the best clips. Great suggestion on using the quizzes to find queries, I am going to use that!