Feedback for BlueCookie:

Here are some of my thoughts and questions as I attempt to emulate what your classmates might have asked:

- Seeing Harry Potter text analysis was neat. Interesting to see a whole R package devoted to this.
- What was the list of words you had to remove? There were still names in some of your word clouds. Did you have a threshold for when to remove a name?
- Did an “academic” theme ever pop up?
- Seeing the bigram clouds was really cool.
- Another group had coherence scores/metrics to evaluate how many topics to have. Trying 7 is reasonable to see if they match books, but how did you choose 3 for within each book?
- What number of topics was suggested by the LDA metrics for the entire set of 7 books? You tried 7, but what did other measures actually suggest?
- Minor typos – you noted “tokeni” -> “tokenize”, one plot also had “occurrances” which should be “occurrences”
- From the sentiment plots with 100 word segments, it’s clear the books got longer, no? These segments, did they have the stop words removed? What fraction of words do you think (or can you compute) were lost?
- Similarly, for these 100 word segments, were words with 0 sentiment score included?
- Please describe how you constructed the confusion matrix. Some words were in multiple books, so were those dealt with? Also, since you said the topics didn’t really align with the books, how was this done?
- For the LDA description in the report, I would add more detail (not technical details), but your understanding of topic modeling. For example, it is finding patterns (like clustering) on text. Be sure to clearly describe the inputs, etc.
- What do you think might be different if you had analyzed sentence sentiment instead of word sentiment?